SHARED-OBSERVATION CHAINING FOR HIERARCHICAL ADVERSARIAL NOISY BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study finite-time regret in adversarial noisy bandits with bounded conditional-mean feedback. Fixed-scale optimizer-cover scalars are not rate-complete: two finite-action classes agree on a nontrivial scale range yet have exponentially different regret. On a product routing model with 2^D templates, a known-horizon $O(D)$-time learner uses one route/reward observation to update all active levels and attains expected $O(\sqrt{T})$ regret against adaptive nonanticipating environments; an oblivious Bernoulli subclass matches this rate. Even the optimistic fixed-rate one-global-Shannon ExO expression, with arbitrary finite real-valued Borel scores, is $\Omega(\min\{T, \sqrt{DT}\})$. A simultaneous product packing matches the complete active-coordinate finite-horizon profile $\sum_{j\in J_+} \lambda_j \min\{\alpha_j T, \sqrt{T}\}$; in the information-limited regime this becomes $\sqrt{T}\sum_{j\in J_+}\lambda_j$, without a block-allocation depth loss. Under explicit margin, finite-range, matching, quadratic-regime, allocation, saturation, and terminal-control conditions, general optimizer hierarchies admit a conditional chain–packing comparison.

1 INTRODUCTION AND FLAGSHIP THEOREM

For $x, a \in \{-1, +1\}^D$, define the canonical product routing class by

$$q_x(a) = \prod_{j=1}^{D} \frac{1 + 2^{-j} x_j a_j}{2}, \qquad g_\theta(a) = \frac{1}{2} + \frac{1}{4}\sum_{j=1}^{D} \frac{2^{-j}}{1 - 2^{-D}} \theta_j a_j, \tag{1.1}$$

Write $(Q_D^{\mathrm{hier}}, \mathcal{G}_D^{\mathrm{hier}})$ and $\mathcal{M}_D^{\mathrm{mc}}$ for the resulting policy, reward, and observation classes.

Theorem 1.1 (Three logically separate flagship claims). **A. Actual minimax rate.**

$$\frac{1}{32}\sqrt{T} \le \mathfrak{M}_T^{\star,\mathrm{obl}} \le \mathfrak{M}_T^{\star,\mathrm{ad}} \le C_{\mathrm{sh}}\sqrt{T}, \qquad C_{\mathrm{sh}} = 2\log 2 + 1 + 2/\sqrt{3}. \tag{1.2}$$

The known-horizon upper bound is adaptive, uses bounded conditional-mean feedback, $O(D)$ time and memory, and one route/reward update for all levels; the lower bound is oblivious Bernoulli.

B. Optimistic expression.

$$\mathsf{OptFlatExO}_{D,T} := \min\left\{T, \inf_{\eta>0}\left[T\,\mathrm{ExO}_\eta(\mathcal{M}_D^{\mathrm{mc}}) + \frac{D\log 2}{\eta}\right]\right\} \ge \frac{1}{32}\min\{T, \sqrt{DT}\}. \tag{1.3}$$

This permits arbitrary finite real-valued Borel scores and is not asserted to be a regret certificate.

C. Validated affine certificate. The comparator-uniform restricted-affine certificate $\mathsf{StdFlatAExO}_{D,T}$ defined in Section 5 satisfies

$$\frac{1}{32}\min\{T, \sqrt{DT}\} \le \mathsf{StdFlatAExO}_{D,T} \le 4\min\{T, \sqrt{DT}\}. \tag{1.4}$$

For $D \le T$, the specified one-global-Shannon architecture is $\Omega(\sqrt{D})$ above the factorized guarantee. This does not lower-bound factorized or non-Shannon potentials, time-varying rates, all one-scale methods, unrestricted ExO algorithms as algorithms, or DEC itself.

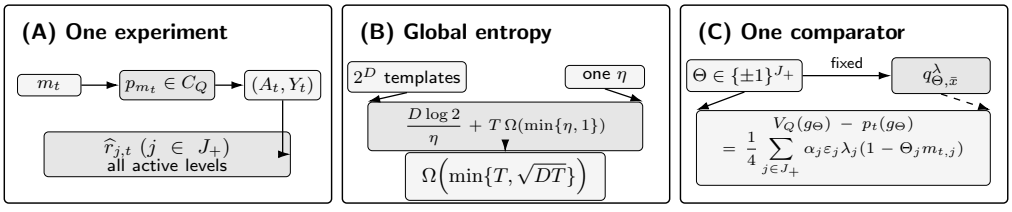


Figure 1. **(A)** One legal experiment updates every level. **(B)** One global Shannon expression pays complete-template entropy. **(C)** One fixed comparator certifies all active-coordinate gaps.

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

2 PROTOCOL, SCALAR INSUFFICIENCY, AND THE ROUTING CHANNEL

Contributions. **Optimizer-only insufficiency:** fixed-scale maximin-volume and distribution-cover scalars do not determine finite-time regret. **Sharp shared observation:** the product hierarchy has minimax $\Theta(\sqrt{T})$ regret although it contains 2^D templates. **Architecture and generalization:** a fixed-rate one-global-Shannon expression pays $\Theta(\min\{T, \sqrt{DT}\})$; simultaneous packing, primitive realizability, and fixed-design conic evaluation explain the mechanism beyond the canonical model.

At round t, the public history is $\mathscr{H}_{t-1} = (A_1, Y_1, \dots, A_{t-1}, Y_{t-1})$. The learner and environment choose $p_t \in \Delta(\mathcal{A})$ and (g_t, ν_t) from the public past, before the current private action draw, and then

$$\boxed{\mathscr{H}_{t-1} \longrightarrow (p_t, g_t, \nu_t) \longrightarrow A_t \sim p_t \longrightarrow Y_t \sim \nu_t(\cdot \mid \mathscr{H}_{t-1}, A_t), \quad \mathbb{E}_t Y_t = g_t(A_t).} \tag{2.1}$$

The environment knows the algorithm and public history but not A_t when choosing the current mean and kernel. We study expected regret against the best fixed action or mediator. An *oblivious-mean* environment fixes $g_{1:T}$ before play; upper bounds allow adaptive nonanticipating means. Bernoulli feedback is used only inside admissible lower-bound subclasses.

For $g \in \mathcal{G}$, let $\mathsf{A}_\alpha(g) = \{a : g(a) \geq \max_b g(b) - \alpha\}$, $\gamma_\alpha(\mathcal{G}) = \sup_p \inf_g p(\mathsf{A}_\alpha(g))$, and let $N_{\mathrm{dc}}(\mathcal{G}; \alpha, \beta)$ be the minimum number of distributions needed to put mass at least $1 - \beta$ on every such set.

Theorem 2.1 (Fixed-scale optimizer scalars are not rate-complete)**.** *For every $d \geq 45$, there are $K \geq e^{d/64}$ actions and compact convex classes $\mathcal{G}_{\mathrm{code}} \subseteq \mathcal{G}_{\mathrm{tab}} \subseteq [1/4, 3/4]^K$ such that, for all $\alpha \in [0, 1/8)$ and $\beta \in [0, 1/2)$,*

$$\gamma_\alpha(\mathcal{G}_{\mathrm{code}}) = \gamma_\alpha(\mathcal{G}_{\mathrm{tab}}) = K^{-1}, \qquad N_{\mathrm{dc}}(\mathcal{G}_{\mathrm{code}}; \alpha, \beta) = N_{\mathrm{dc}}(\mathcal{G}_{\mathrm{tab}}; \alpha, \beta) = K, \tag{2.2}$$

while, with $r = \dim \operatorname{span}(\mathcal{G}_{\mathrm{code}} - \mathcal{G}_{\mathrm{code}}) \leq d$,

$$\mathfrak{R}_T^{\star,\mathrm{obl}}(\mathcal{G}_{\mathrm{code}}) \leq \min\{T/2, 2\sqrt{rT \log K}\}, \qquad \mathfrak{R}_T^{\star,\mathrm{obl}}(\mathcal{G}_{\mathrm{tab}}) \geq \frac{1}{16\sqrt{2}} \min\{T, \sqrt{KT}\}. \tag{2.3}$$

At $T = K$, the ratio is exponential in d up to polynomial factors.

A constant-rate sign code gives rewards $g_v(x) = 1/2 + \langle v, x \rangle/(4d)$ with unique optimizer v and gap at least $1/8$, proving the scalar identities. The code class has reward rank at most d; the tabular lower bound uses fixed Bernoulli one-arm alternatives and a directed-KL change of measure. Appendix C gives all constants. The theorem equates only the displayed scalars on the stated rectangle, not complete near-optimality profiles.

Proposition 2.2 (The routing channel is not determined by its linear value feature)**.** *The template values satisfy*

$$q_x^\lambda(g_\theta) = \frac{1}{2} + \frac{1}{4} \sum_{j=1}^{D} \alpha_j \lambda_j x_j \theta_j, \tag{2.4}$$

so the value table is an anisotropic finite-action linear-bandit instance. For every $D \geq 2$, however, there are legal route laws $p, \widetilde{p}$ agreeing on every branch-additive reward while inducing different feedback laws.

Indeed, mix equally two templates with signs $(+,+)$ and $(-,-)$ in coordinates $1, 2$, and compare this correlated mixture with the product law having the same zero coordinate means. Both laws agree on every g_θ, but $\mathbb{E}_{\widetilde{p}}[A_1 A_2] = \lambda_1 \lambda_2$ whereas $\mathbb{E}_p[A_1 A_2] = 0$. The linear mean feature is therefore not information sufficient: the observed route channel carries the covariance structure used below.

3 ONE LEGAL ROUTE EXPERIMENT UPDATES EVERY LEVEL

Fix reliabilities $0 < \lambda_j \leq 1/2$ and amplitudes $\alpha_j \geq 0$ with $\sum_j \alpha_j \leq 1$. Templates and rewards are

$$q_x^\lambda(a) = \prod_{j=1}^{D} \frac{1 + \lambda_j x_j a_j}{2}, \qquad g_\theta(a) = \frac{1}{2} + \frac{1}{4} \sum_{j=1}^{D} \alpha_j \theta_j a_j. \tag{3.1}$$

Let $J_+ = \{j : \alpha_j > 0\}$. Inactive coordinates are fixed and omitted; if $J_+ = \varnothing$, regret is zero. For $m \in [-1,1]^D$, define

$$p_m(a) = \prod_{j=1}^{D} \frac{1 + \lambda_j m_j a_j}{2}, \qquad p_m = \sum_x \left[\prod_j \frac{1 + m_j x_j}{2}\right] q_x^\lambda. \tag{3.2}$$

Thus independently chosen level mixtures induce one legal physical experiment. Conversely every legal mixture has coordinate means $\mathbb{E}_p A_j = \lambda_j m_j$ for some $m_j \in [-1,1]$.

At level j, maintain exponential weights $\mu_{j,t}$ on $s \in \{\pm 1\}$ and $m_{j,t} = \sum_s s\mu_{j,t}(s)$. Draw one route $A_t \sim p_{m_t}$, observe Y_t, and use for every active level and both signs

$$\widehat{r}_{j,t}(s) = s\lambda_j \frac{A_{t,j} - \lambda_j m_{j,t}}{1 - \lambda_j^2 m_{j,t}^2} Y_t, \qquad W_{j,t+1}(s) = W_{j,t}(s) e^{\eta_j \widehat{r}_{j,t}(s)}, \quad \eta_j = \frac{1 - \lambda_j}{\lambda_j \sqrt{T}}. \tag{3.3}$$

This known-horizon algorithm requires $O(D)$ time and memory and never samples a latent tree node.

Lemma 3.1 (Simultaneous unbiasedness, energy, and range)**.** *For $r_{j,t}(s) = \alpha_j \lambda_j \theta_{t,j} s/4$,*

$$\mathbb{E}_t \widehat{r}_{j,t}(s) = r_{j,t}(s), \qquad \mathbb{E}_t \widehat{r}_{j,t}(s)^2 \le \frac{\lambda_j^2}{1 - \lambda_j^2}, \qquad |\widehat{r}_{j,t}(s)| \le \frac{\lambda_j}{1 - \lambda_j}. \tag{3.4}$$

Writing $b_{j,t} = \lambda_j m_{j,t}$, product independence and centering give

$$\mathbb{E}_t[(A_{t,j} - b_{j,t}) g_{\theta_t}(A_t)] = \frac{1}{4} \sum_k \alpha_k \theta_{t,k} \mathbb{E}_t[(A_{t,j} - b_{j,t}) A_{t,k}] = \frac{1}{4} \alpha_j \theta_{t,j} (1 - b_{j,t}^2),$$

which proves unbiasedness; the remaining claims follow from $Y_t \in [0,1]$ and $|b_{j,t}| \le \lambda_j$.

Theorem 3.2 (Adaptive shared-observation routing bound)**.** *For the known horizon T, under arbitrary adaptive bounded conditional-mean feedback,*

$$\mathbb{E}\,\mathrm{Reg}_T \le C_{\mathrm{sh}} \sqrt{T} \sum_{j \in J_+} \lambda_j. \tag{3.5}$$

4 Sharp minimax proof for the routing class

For a fixed sign s, the two-action potential gives, pathwise,

$$\sum_{t=1}^{T} \big[\widehat{r}_{j,t}(s) - \mu_{j,t}(\widehat{r}_{j,t})\big] \le \frac{\log 2}{\eta_j} + \eta_j \sum_{t=1}^{T} \mu_{j,t}(\widehat{r}_{j,t}^2). \tag{4.1}$$

The hindsight-optimal sign $s_j^\star$ is random under an adaptive environment. Put $\xi_{j,t}(s) = \widehat{r}_{j,t}(s) - r_{j,t}(s)$. Oddness gives $\xi_{j,t}(-1) = -\xi_{j,t}(+1)$ and hence

$$-\sum_t \xi_{j,t}(s_j^\star) \le \left|\sum_t \xi_{j,t}(+1)\right|.$$

Martingale orthogonality and Lemma 3.1 imply

$$\mathbb{E}\left|\sum_t \xi_{j,t}(+1)\right| \le \lambda_j \sqrt{\frac{T}{1 - \lambda_j^2}} \le \frac{2}{\sqrt{3}} \lambda_j \sqrt{T}. \tag{4.2}$$

The potential and quadratic terms are at most $2\lambda_j \sqrt{T} \log 2$ and $\lambda_j \sqrt{T}$. Summing over levels proves Theorem 3.2; for the canonical reliabilities, $\sum_j 2^{-j} < 1$.

For the lower bound, fix level k, draw a latent sign Σ, and use the constant reward $g_{\Sigma \varepsilon e_k}$ with Bernoulli observations. If the learner's conditional coordinate mean is $\lambda_k m_t$, the optimal template has one-round advantage

$$\frac{1}{4} \alpha_k \lambda_k \varepsilon (1 - \Sigma m_t).$$

At a common counterfactual history, the two Bernoulli means are $1/2 \pm \alpha_k \varepsilon A_{t,k}/4$, and

$$\mathrm{kl}\left(\frac{1}{2}+\frac{x}{4}\,\middle\|\,\frac{1}{2}-\frac{x}{4}\right) \le \frac{2}{3}x^2, \qquad x \in [0,1]. \tag{4.3}$$

The transcript chain rule and Pinsker, with $\varepsilon = \min\{1, \sqrt{3}/(2\alpha_k\sqrt{T})\}$, yield

$$\mathfrak{M}_T^{\star,\mathrm{obl}}(Q_{\lambda,D}, \mathcal{G}_{\alpha,D}) \ge \frac{\lambda_k}{8} \min\left\{\alpha_k T, \frac{\sqrt{3}}{2}\sqrt{T}\right\}. \tag{4.4}$$

For the canonical first coordinate, $\lambda_1 = 1/2$ and $\alpha_1 \ge 1/2$, giving $\sqrt{T}/32$ and completing Claim A of Theorem 1.1.

5 Fixed-rate one-global-Shannon obstruction and certificates

The standard comparator-uniform unrestricted expression is

$$\mathsf{StdFlatExO}_{D,T} := \min\left\{T, \inf_{\eta>0}\left[T\,\mathrm{ExO}_\eta(\mathcal{M}_D^{\mathrm{mc}}) + \frac{2D\log 2}{\eta}\right]\right\}. \tag{5.1}$$

A comparator-wise ExO+ supermartingale pays one $\log|Q|/\eta$ term through the Shannon potential and a second through aggregation over the random hindsight comparator. Consequently, with a predictable ξ_t-approximate one-step oracle,

$$\sup_{\mathsf{Adv}} \mathbb{E}\,\mathrm{Reg}_T^Q \le T\,\mathrm{ExO}_\eta(\mathcal{M}_{Q,\mathcal{G}}^{\mathrm{mc}}) + \sum_{t=1}^{T}\xi_t + \frac{2\log|Q|}{\eta}. \tag{5.2}$$

Thus $\mathsf{StdFlatExO}_{D,T}$ is a validated information-theoretic oracle certificate; $\mathsf{OptFlatExO}_{D,T}$ remains only the optimistic expression.

For the score-independent lower bound, let $S_\omega = \{q : \omega(q) > 0\}$. Scores are finite real-valued Borel functions, all ratios by $\omega(q)$ are evaluated only on S_ω, and an infinite exponential moment gives objective $+\infty$. Choose templates q^+, q^- differing only at coordinate one and reference law $(\delta_{q^+} + \delta_{q^-})/2$. After arbitrary ω and score G are chosen, use Bernoulli models with means $1/2 \pm tA_1/4$ and the matching comparators. On S_ω, define

$$D_{\omega,G}(q,a,y) = \frac{G(q^-;q,a,y) - G(q^+;q,a,y)}{\omega(q)}, \tag{5.3}$$

and set it arbitrarily off support. If P_+, P_- are the induced laws, $\zeta = P_+ + P_-$, and $p_\pm = dP_\pm/d\zeta$, then the opposite comparator produces $-D_{\omega,G}$ and pointwise AM–GM gives

$$p_+ e^{\eta D_{\omega,G}} + p_- e^{-\eta D_{\omega,G}} \ge 2\sqrt{p_+ p_-}. \tag{5.4}$$

The two models share the mediator/route marginal, and their conditional Bernoulli affinity is $\sqrt{1-t^2/4}$. Therefore

$$\mathrm{ExO}_\eta(\mathcal{M}_D^{\mathrm{mc}}) \ge \max_{0\le t\le\alpha_1}\left\{\frac{t}{8} - \frac{1-\sqrt{1-t^2/4}}{2\eta}\right\} \ge \frac{1}{32}\min\{\eta, 1\}. \tag{5.5}$$

Balancing with $D\log 2/\eta$ proves Claim B. For Claim C, the legal product design and affine score

$$\psi_x(a) = \sum_{j=1}^{D} \lambda_j (x_j - m_j)\frac{a_j - \lambda_j m_j}{1-\lambda_j^2 m_j^2} \tag{5.6}$$

represent every policy contrast exactly, have pairwise energy at most $16/9$ and range below 4, and imply $\mathsf{AExO}_\eta \le (16/9)\eta$ for $\eta \le 1/4$. Hence

$$\begin{aligned} \frac{1}{32}\min\{T, \sqrt{DT}\} \le \mathsf{OptFlatExO}_{D,T} \le \mathsf{StdFlatExO}_{D,T}, \\ \mathsf{StdFlatExO}_{D,T} \le \mathsf{StdFlatAExO}_{D,T} \le 4\min\{T, \sqrt{DT}\}. \end{aligned} \tag{5.7}$$

Appendices M and O give the complete comparator-uniform, support, envelope, and small-horizon proofs.

6 General shared-observation upper and compatible-packing lower principles

For a finite nonempty witness family $Q \subseteq \Delta(\mathcal{A})$, write $C_Q := \mathrm{co}(Q)$. A coherent chain represents each witness as $q_0 + d_1 + \cdots + d_J$, with finite increment classes $\mathcal{D}_j$, and requires that independently averaging one increment at every level gives one legal law in C_Q. The shared-observation certificate $\Gamma_T^{\mathrm{obs}}(Q, \mathcal{G})$ chooses a single physical law each round and updates every level from the same observation.

Theorem 6.1 (Shared-observation Chained-ExO)**.** *For every finite Q, compact convex $\mathcal{G}$, and adaptive nonanticipating bounded conditional-mean environment,*

$$\mathfrak{M}_T^{\star,\mathrm{ad}}(Q, \mathcal{G}) \le \Gamma_T^{\mathrm{obs}}(Q, \mathcal{G}). \tag{6.1}$$

At depth one, the certificate equals the restricted affine-ExO objective.

The proof combines a comparator-uniform exponential supermartingale at each level with one joint saddle problem for the common physical experiment; no inverse node-visitation factor appears.

A sequential optimizer packing is fixed before interaction. It uses deterministic time blocks, conditionally fresh child indices, and one leaf policy that remains approximately optimal on every ancestral block. Fano nodes use conditional list Fano; binary nodes use Le Cam affinity.

Theorem 6.2 (Sequential optimizer-packing lower game)**.** *For every $\rho \in (0, 1/2)$,*

$$\Pi_{T,\rho}^{\mathrm{seq}}(Q, \mathcal{G}) \le \mathfrak{M}_T^{\star,\mathrm{obl}}(Q, \mathcal{G}) \le \mathfrak{M}_T^{\star,\mathrm{ad}}(Q, \mathcal{G}). \tag{6.2}$$

The same leaf comparator and disjoint blocks prevent double counting. Universally,

$$\Pi_{T,\rho}^{\mathrm{seq}} \le \mathfrak{M}_T^{\star,\mathrm{obl}} \le \mathfrak{M}_T^{\star,\mathrm{ad}} \le \Gamma_T^{\mathrm{obs}}. \tag{6.3}$$

7 Primitive comparison and simultaneous packing

Local notation. Fix $\rho \in (0, 1/2)$. For finite nonempty $Q \subseteq \Delta(\mathcal{A})$ and compact convex $\mathcal{G}$, write

$$C_Q := \mathrm{co}(Q), \qquad V_Q(g) := \max_{q \in Q} q(g), \qquad \mathcal{G}_\rho := \mathcal{G} \cap [\rho, 1-\rho]^{\mathcal{A}}. \tag{7.1}$$

For an increment class $\mathcal{D}$ and a chain $\mathscr{C}$ with terminal map π_J, define

$$\mathrm{diam}_{\mathcal{G}}(\mathcal{D}) := \max_{d,r \in \mathcal{D}} \sup_{g \in \mathcal{G}} |(d-r)(g)|, \quad \mathrm{rad}_{\mathcal{G}}(\mathscr{C}) := \max_{q \in Q} \sup_{g \in \mathcal{G}_0} (q - \pi_J(q))(g), \quad \mathcal{G}_0 := \mathrm{co}(\mathcal{G} \cup \{0\}). \tag{7.2}$$

Theorem 7.1 (Primitive hierarchical comparison with explicit conditions)**.** *Let $\mathscr{C}$ be a coherent depth-J chain with increment classes $\mathcal{D}_j$, $H_j = \log|\mathcal{D}_j|$, and terminal radius $\mathrm{rad}_{\mathcal{G}}(\mathscr{C})$. Assume:*

1. ***Observable common design:*** *every full reference tuple induces one legal law serving all levels, and level j has an exactly representing affine score with energy V_j and finite pairwise range R_j. Put $B_j = \mathrm{diam}_{\mathcal{G}}(\mathcal{D}_j)$, $A_j = 4V_j + B_j^2$, $a_j = A_j H_j$, and $S = \sum_j a_j$.*
2. ***Strict gluing:*** *nested optimizer faces are exposed by $b_u + \lambda v_{u,i} \in \mathcal{G}_\rho$ for $\lambda \in [0, r_j]$, and $2r_j \max_i \|v_{u,i}\|_\infty < m_u$. Hence every path has one comparator optimal in every ancestral block.*
3. ***Switchability and information:*** *level j is a Fano level with list entropy $h_j \ge 4\log 2$, or a binary edge level; its decision scale is s_j and its quadratic direction energy is at most ζ_j. Put $c_j = \zeta_j/[\rho(1-\rho)]$, $\tau_j = h_j$ for Fano and 1 for edge levels, and $\chi_j = s_j^2 \tau_j / c_j$.*
4. ***Matching, quadratic regime, allocation, saturation:*** *for some $\kappa > 0$, $\chi_j \ge \kappa a_j$, $TA_j \ge 2R_j^2 H_j$, and positive integers n_j sum to T while*

$$n_j \ge \frac{(T-J)a_j}{S}, \qquad n_j r_j^2 c_j \ge \begin{cases} h_j/4, & \textit{Fano}, \\ 1, & \textit{edge}. \end{cases} \tag{7.3}$$

Assume $T \ge 2J$.

5. ***Terminal control:*** *retain the additive term $T\,\mathrm{rad}_{\mathcal{G}}(\mathscr{C})$; for the multiplicative sandwich below, assume the chain is exact so this radius is zero.*

Then

$$\Pi^{\mathrm{seq}}_{T,\rho}(Q,\mathcal{G}) \geq \frac{\sqrt{\kappa}}{2e}\sqrt{(T-J)S}, \tag{7.4}$$

and

$$\Gamma^{\mathrm{obs}}_{T}(Q,\mathcal{G}) \leq T\,\mathrm{rad}_{\mathcal{G}}(\mathscr{C}) + 8e\sqrt{J/\kappa}\,\Pi^{\mathrm{seq}}_{T,\rho}(Q,\mathcal{G}). \tag{7.5}$$

Observability supplies the range-certified upper profile; strict gluing identifies one leaf comparator; Bernoulli curvature converts quadratic direction energy into channel control; Fano/Le Cam at the information-critical amplitudes gives each block at least $(2e)^{-1}\sqrt{n_j\chi_j}$. Matching and proportional allocation yield the theorem. Matching, the quadratic regime, allocation, saturation, and terminal approximation are separate hypotheses.

Theorem 7.2 (Certified conditional cover–packing comparison)**.** *For an exact chain satisfying the preceding assumptions,*

$$\Pi^{\mathrm{seq}}_{T,\rho} \leq \mathfrak{M}^{\star,\mathrm{obl}}_{T} \leq \mathfrak{M}^{\star,\mathrm{ad}}_{T} \leq \Gamma^{\mathrm{obs}}_{T} \leq 8e\sqrt{J/\kappa}\,\Pi^{\mathrm{seq}}_{T,\rho}. \tag{7.6}$$

The disjoint-block factor is not intrinsic on product-realizable hierarchies. Define

$$\Psi_T(\lambda,\alpha) = \sum_{j\in J_+} \lambda_j \min\{\alpha_j T, \sqrt{T}\}. \tag{7.7}$$

Theorem 7.3 (Active-coordinate simultaneous product packing)**.** *If $J_+ = \varnothing$, both minimax regrets are zero. Otherwise, for all D, T,*

$$\frac{1}{8}\sqrt{\frac{3}{8}}\,\Psi_T(\lambda,\alpha) \leq \mathfrak{M}^{\star,\mathrm{obl}}_{T} \leq \mathfrak{M}^{\star,\mathrm{ad}}_{T} \leq C_{\mathrm{sh}}\,\Psi_T(\lambda,\alpha). \tag{7.8}$$

If $\alpha_{\min} = \min_{j\in J_+}\alpha_j$ and $T \geq 3/(8\alpha^2_{\min})$, then

$$\mathfrak{M}^{\star,\mathrm{obl}}_{T} \asymp \mathfrak{M}^{\star,\mathrm{ad}}_{T} \asymp \sqrt{T}\sum_{j\in J_+}\lambda_j. \tag{7.9}$$

The hard environment draws $\Theta \sim \mathrm{Unif}\{\pm 1\}^{J_+}$ once and activates every active coordinate on every round. Fix inactive template signs $\bar{x}$ deterministically and set $\varepsilon_j = 0$ for $j \notin J_+$. Every legal law has coordinate means $\lambda_j m_{t,j}$, and the one fixed comparator $q^{\lambda}_{\Theta,\bar{x}}$ gives the exact gap

$$V_Q(g^{\boldsymbol{\varepsilon}}_{\Theta}) - p_t(g^{\boldsymbol{\varepsilon}}_{\Theta}) = \frac{1}{4}\sum_{j\in J_+}\alpha_j\varepsilon_j\lambda_j(1-\Theta_j m_{t,j}). \tag{7.10}$$

For every $j \in J_+$, pairing active latent vectors that differ only at coordinate j, joint convexity of KL, and the transcript chain rule yield the lower bound. The upper bound chooses at each coordinate between the shared estimator and the zero-mean no-learning strategy; all choices still combine into the one legal product law (3.2).

8 Observation geometry, fixed-design computation, and closest work

Set

$$\mathsf{H}_0 := \{h \in \mathbb{R}^{\mathcal{A}} : \langle h, \mathbf{1}\rangle = 0\}, \qquad \mathcal{V} := \mathrm{span}(\mathcal{G}-\mathcal{G}), \qquad \mathcal{K} := \tfrac{1}{2}(\mathcal{G}-\mathcal{G}). \tag{8.1}$$

For finite nonempty $Q \subseteq \Delta(\mathcal{A})$, define

$$b_{\mathcal{G}}(Q) = \max_{g\in\mathcal{G}}\left[\max_{a} g(a) - \max_{q\in C_Q} q(g)\right]. \tag{8.2}$$

Proposition 8.1 (Exact optimizer-witness decomposition)**.** *For $\mathsf{s} \in \{\mathrm{ad}, \mathrm{obl}\}$,*

$$\mathfrak{R}^{\star,\mathsf{s}}_{T}(\mathcal{G}) = \min_{\substack{\varnothing\neq Q\subseteq\Delta(\mathcal{A})\\ |Q|<\infty}} \left\{T b_{\mathcal{G}}(Q) + \mathfrak{M}^{\star,\mathsf{s}}_{T}(Q,\mathcal{G})\right\}. \tag{8.3}$$

The useful direction is pathwise; equality is closed by the Dirac family. For a zero-mass contrast $h \in \mathsf{H}_0$, a physical design $p \in \Delta(\mathcal{A})$, and $\varepsilon \geq 0$, define

$$\begin{aligned} \mathcal{E}_\varepsilon(h;p) = \inf_{c,\psi} & \sum_a p(a)\psi(a)^2 \\ \text{s.t.} \quad & \sup_{g\in\mathcal{G}} |c + (p \odot \psi - h)(g)| \leq \varepsilon. \end{aligned} \tag{8.4}$$

The p-nullspace is

$$\mathcal{N}_p := \left\{ v \in \mathcal{V} : \sum_a p(a)v(a)^2 = 0 \right\}. \tag{8.5}$$

Theorem 8.2 (Riesz geometry of one-point affine estimation)**.** *With the closed perspective convention,*

$$\mathcal{E}_0(h;p) = \sup_{\substack{v\in\mathcal{V} \\ \sum_a p(a)v(a)^2 \leq 1}} |h(v)|^2, \tag{8.6}$$

$$\mathcal{E}_\varepsilon(h;p) = \sup_{v\in\mathcal{V}} \left\{ h(v) - \varepsilon \|v\|_{\mathcal{K}} - \tfrac{1}{4}\sum_a p(a)v(a)^2 \right\}. \tag{8.7}$$

Exact cost is finite precisely when $h(v) = 0$ for every $v \in \mathcal{N}_p$. For $\mathcal{G} = [0,1]^{\mathcal{A}}$ and $h = q - p$,

$$\mathcal{E}_0(q-p;p) = \begin{cases} \chi^2(q \,\|\, p), & q \ll p, \\ +\infty, & q \not\ll p. \end{cases} \tag{8.8}$$

For a linear reward span $v_\theta(a) = \langle \theta, \phi(a) \rangle$, write $m_h = \sum_a h(a)\phi(a)$ and $G_p = \sum_a p(a)\phi(a)\phi(a)^\top$. Then

$$\mathcal{E}_0(h;p) = \begin{cases} m_h^\top G_p^\dagger m_h, & m_h \in \operatorname{range}(G_p), \\ +\infty, & m_h \notin \operatorname{range}(G_p). \end{cases} \tag{8.9}$$

Theorem 8.3 (Fixed-design range-controlled quadratic score evaluation is an SOCP)**.** *Fix a finite coherent hierarchy, current reference tuple, induced common physical design, bias tolerances, finite pairwise score ranges, learning rates, and a polytopal reward class $\operatorname{co}\{g^1, \ldots, g^M\} \subseteq [0,1]^K$. If level j has n_j increments, the exact robust pairwise quadratic score problem is a polynomial-size SOCP using*

$$O\left(KM \sum_j n_j^2 \right) \tag{8.10}$$

three-dimensional rotated cones. For rational input, this is a polynomial-size rational SOCP representation. Standard conic methods compute numerically feasible, near-optimal solutions to prescribed tolerance on well-posed instances; predictable objective errors ξ_t add only $\sum_t \xi_t$ to regret.

The substitution $z_d(a) = p(a)\psi_d(a)$ yields the exact perspective cone $(p(a)/2, u_d(a), z_d(a)) \in \mathcal{Q}_{\mathrm{r}}^3$. The theorem does not optimize the hierarchy, witness family, reference laws, or physical design jointly with finite range, and it does not formulate the exact nonlinear exponential objective as an SOCP. The natural joint finite-range feasible set is nonconvex in the displayed variables.

Factorization and differentiated rates are not themselves new (Bubeck et al., 2019; Yang et al., 2022); low-rank experts and bandit linear optimization already exploit low-dimensional structure (Hazan et al., 2016; Bubeck et al., 2012). Our distinction is statistical: one legal route experiment supplies simultaneous estimators, while one legal route comparator supports the lower sum. General ExO, DEC, interactive Fano, mediator capacity, and sequential majorizing measures remain broader frameworks (Lattimore & Szepesvári, 2020; Foster et al., 2023; Chen et al., 2024; Eldowa et al., 2024; Block et al., 2021). Recent tree-similarity results show that hierarchy alone need not help under one-point feedback; the observed route is the additional structure here (Eldowa et al., 2026). Hierarchical configuration bandits and tree-MDP policy sharing use distinct models (Shabat et al., 2026; Shah et al., 2026).

Table 1: Closest frameworks and the additional statement proved here.

Framework	Already provides	Additional statement here
Multi-scale / hierarchical experts	Range-sensitive entropy and differentiated rates	One aggregate reward and the observed route yield all centered level estimators
Low-rank experts / linear bandits	Dimension- or rank-sensitive regret	Equal linear values need not imply equal routing feedback laws
DEC / general ExO	Broad decision–information theory; arbitrary scores	A scoped obstruction for one fixed-rate global Shannon architecture
Interactive Fano / mediator capacity	Adaptive testing; one-scale policy overlap	One optimizer tree with a legal global comparator and one common experiment
Tree similarity / configuration bandits	Hierarchical structure under other feedback models	Sharp one-point learning from a known stochastic route channel

9 Exact scope, impossibility, and proof map

Proposition 9.1 (An unconditional reverse comparison is impossible)**.** *For every $\rho \in (0, 1/2)$, there are finite Q and compact convex $\mathcal{G}$ such that, for every $T \geq 1$,*

$$\Pi^{\mathrm{seq}}_{T,\rho}(Q, \mathcal{G}) = 0, \qquad \mathfrak{M}^{\star,\mathrm{obl}}_{T}(Q, \mathcal{G}) \geq \min\left\{\frac{\rho T}{8}, \frac{\sqrt{T}}{12}\right\}. \tag{9.1}$$

Hence no universal finite constant reverses (6.3) *without an interior testing realization or an equivalent condition.*

The construction uses $\mathcal{A} = \{0, 1, 2, 3\}$, $Q = \{\delta_1, \delta_2\}$, and the reward segment $b + uv$, where $b = (1 - \rho, 1/2, 1/2, \rho)$ and $v = (\rho/4)(1, 1, -1, 1)$. The only ρ-interior reward is the neutral tie b, so no positive Fano or edge node exists, whereas the two endpoint Bernoulli alternatives switch the optimal witness and remain statistically close. Appendix O gives the directed-KL, transcript-TV, and all-horizon calculation.

All-horizon calculation. Let $s = \rho/4$ and compare the constant Bernoulli environments $g_\pm = b \pm \varepsilon v$. If $z_t = p_t(1) - p_t(2) \in [-1, 1]$, their one-round losses relative to δ_1 and δ_2 are

$$s\varepsilon(1 - z_t), \qquad s\varepsilon(1 + z_t). \tag{9.2}$$

At every common counterfactual history the action kernel is identical, while the two Bernoulli means are $1/2 \pm s\varepsilon$. Since $s\varepsilon \leq 1/8$,

$$D_{\mathrm{KL}}(P^{t-1}_+ \| P^{t-1}_-) \leq 18(t-1)s^2\varepsilon^2, \qquad \|P^{t-1}_+ - P^{t-1}_-\|_{\mathrm{TV}} \leq 3s\varepsilon\sqrt{t-1}. \tag{9.3}$$

With $\varepsilon = \min\{1, (6s\sqrt{T})^{-1}\}$, every prefix variation is at most $1/2$. Averaging (9.2) under the two models therefore gives

$$\frac{1}{2}\mathbb{E}_+ \operatorname{Reg}_T + \frac{1}{2}\mathbb{E}_- \operatorname{Reg}_T \geq \frac{s\varepsilon T}{2} = \min\left\{\frac{\rho T}{8}, \frac{\sqrt{T}}{12}\right\}.$$

This also shows why the counterexample is quantitative rather than a boundary-only pathology.

Why the primitive assumptions are logically distinct. *Observability does not imply switchability:* a direction may admit a bounded estimator while leaving a common optimizer optimal under every legal perturbation, so statistical difficulty creates no regret. *Switchability does not imply gluing:* for block rewards $g_1, \ldots, g_J$, the exact comparator identity is

$$\max_{q \in Q} \sum_{j,t \in I_j} [q(g_j) - p_t(g_j)] = \sum_{j,t \in I_j} [V_Q(g_j) - p_t(g_j)] - \min_{q \in Q} \sum_j n_j [V_Q(g_j) - q(g_j)]. \tag{9.4}$$

Thus local optimizer losses add exactly only when one comparator is optimal in all positive-length blocks; otherwise the final nonnegative term is a genuine double-counting correction. *Matching does not imply saturation:* even when $\chi_j \geq \kappa a_j$, a short block may require an information-critical amplitude larger than the legal radius r_j. Proportional allocation and the inequalities $n_j r_j^2 c_j \geq h_j/4$ or 1 must therefore be verified separately. Finally, terminal approximation is independent of all three primitive properties; it remains as $T \operatorname{rad}_{\mathcal{G}}(\mathscr{C})$ unless the chain is exact.

Why finite range or an equivalent tail condition is visible. Second moments alone do not control the exponential objective. For

$$X_M = \begin{cases} M, & \text{with probability } M^{-2}, \\ -M/(M^2-1), & \text{otherwise,} \end{cases}$$

one has $\mathbb{E}X_M = 0$ and $\mathbb{E}X_M^2 \le 4/3$, but $\mathbb{E}[e^{\eta X_M} - 1 - \eta X_M] \ge e^{\eta M}/M^2 - 1 \to \infty$. The paper uses the sufficient condition $|\eta X| \le 1$, under which $e^x - 1 \le x + x^2$. In the fixed-design SOCP this range restriction is linear after $z = p\psi$; if the physical probability and score offset are optimized jointly, constraints of the form $|pc + z| \le Lp$ are nonconvex in the displayed variables. This is the exact boundary between the proved conic subproblem and the unclaimed global design problem.

Table 2: Theorem scope. "Upper" and "lower" refer to what is actually proved.

Result	Environment	Observation	Conclusion
Routing upper	Adaptive, nonanticipating	Bounded conditional mean	$O(\sqrt{T})$
Routing / product lower	Oblivious mean	Bernoulli	Matching rate / additive packing
Global Shannon	Static coefficient game	Mean-compatible laws	Architectural obstruction
Primitive comparison	Adaptive upper; oblivious lower	Bounded mean; Bernoulli	Conditional comparison
Fixed-design evaluation	Deterministic finite program	Polytopal rewards	Polynomial-size SOCP

Main notation. $C_Q = \mathrm{co}(Q), V_Q(g) = \max_{q\in Q} q(g), \mathcal{G}_\rho = \mathcal{G} \cap [\rho, 1-\rho]^{\mathcal{A}}, \mathsf{H}_0 = \{h : \langle h, \mathbf{1}\rangle = 0\}, \mathcal{K} = (\mathcal{G} - \mathcal{G})/2$; the level diameter and terminal radius are those in (7.2). **Dependencies.** Common experiment ⇒ shared upper bound; optimizer switching plus one compatible comparator ⇒ packing lower bound; matching plus allocation and saturation ⇒ conditional comparison; fixed design plus finite score range ⇒ the SOCP subproblem.

Logical status of the global-potential quantities. The three numerical objects are deliberately distinct:

$\mathsf{OptFlatExO}_{D,T}$	Optimistic one-comparator expression; no regret guarantee is asserted.
$\mathsf{StdFlatExO}_{D,T}$	Valid adaptive unrestricted ExO+ oracle certificate, with the second entropy payment controlling the hindsight comparator.
$\mathsf{StdFlatAExO}_{D,T}$	Valid adaptive certificate after restricting the one-step score to the displayed affine class.

Limitations. The global-potential lower bound does not apply to factorized, product, non-Shannon, or time-varying regularizers, every one-scale algorithm, unrestricted ExO algorithms as algorithms, or DEC itself. The primitive comparison is sufficient, not necessary, and its reverse direction requires margin, finite range, matching, allocation, saturation, and terminal control. The SOCP theorem evaluates a fixed hierarchy and fixed common design; global hierarchy discovery is outside its scope. Upper results allow adaptive bounded conditional-mean noise, whereas lower bounds use admissible oblivious Bernoulli subclasses.

Proof map and conclusion. Appendices M and O prove the routing, certificate, arbitrary-score, active-coordinate, and impossibility results; Appendices G–N prove the general comparison; Appendix P gives the conic formulation. The three quantitative conclusions are

$$\begin{aligned} \mathfrak{M}_T^{\star,\mathrm{obl}}(Q_D^{\mathrm{hier}}, \mathcal{G}_D^{\mathrm{hier}}) &\asymp \mathfrak{M}_T^{\star,\mathrm{ad}}(Q_D^{\mathrm{hier}}, \mathcal{G}_D^{\mathrm{hier}}) \asymp \sqrt{T}, \\ \mathsf{OptFlatExO}_{D,T} &\asymp \min\{T, \sqrt{DT}\}, \\ \mathfrak{M}_T^{\star,\mathrm{obl}}(Q_{\lambda,D}, \mathcal{G}_{\alpha,D}) &\asymp \mathfrak{M}_T^{\star,\mathrm{ad}}(Q_{\lambda,D}, \mathcal{G}_{\alpha,D}) \asymp \Psi_T(\lambda, \alpha). \end{aligned} \tag{9.5}$$

They concern, respectively, canonical minimax regret, the specified fixed-rate global expression, and the complete active-coordinate product rate; none lower-bounds all flat algorithms. The paper isolates one principle: *one legal experiment updates the hierarchy, and one legal comparator certifies its lower bound.*

AI USE STATEMENT

Generative AI tools were used for literature discovery, research planning, refinement of definitions and theorem statements, proof-oriented drafting, consistency checks, code assistance, and LaTeX editing. Every AI-assisted mathematical claim and proof promoted in the submission was independently reconstructed and verified by the authors against the stated assumptions. All citations, theorem attributions, and external normalizations were checked against primary sources. Automated scripts provide additional compilation, transcription, constant, and scope checks, but do not replace the authors' mathematical verification. The authors take responsibility for the final text, claims, proofs, references, code, and supplementary material.

REPRODUCIBILITY STATEMENT

The probability model, quantifier order, theorem-specific assumptions, and adversary and noise scopes are stated in the main paper and Appendices A–P. Complete proofs of all promoted claims are included in the appendices. The supplementary source contains scripts checking finite-dimensional identities, displayed constants, hierarchical routing estimators, simultaneous-packing calculations, and SOCP formulations. All promoted theorem statements were independently reconstructed against their appendix proofs, and all external normalizations and citations were checked against primary sources.

REFERENCES

Farid Alizadeh and Donald Goldfarb. Second-order cone programming. *Mathematical Programming*, 95(1):3–51, 2003. doi: 10.1007/s10107-002-0339-5.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002. doi: 10.1137/S0097539701398375.

Adam Block, Yuval Dagan, and Alexander Rakhlin. Majorizing measures, sequential complexities, and online learning. In *Proceedings of the Thirty-Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 587–590. PMLR, 2021.

Sander Borst, Daniel Dadush, Neil Olver, and Makrand Sinha. Majorizing measures for the optimizer. In *12th Innovations in Theoretical Computer Science Conference (ITCS 2021)*, volume 185 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 73:1–73:20. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2021. doi: 10.4230/LIPIcs.ITCS.2021.73.

Sébastien Bubeck, Nicolò Cesa-Bianchi, and Sham M. Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pp. 41.1–41.14. PMLR, 2012.

Sébastien Bubeck, Nikhil R. Devanur, Zhiyi Huang, and Rad Niazadeh. Multi-scale online learning: Theory and applications to online auctions and pricing. *Journal of Machine Learning Research*, 20 (62):1–37, 2019. URL `https://www.jmlr.org/papers/v20/17-498.html`.

Clément L. Canonne. A short note on an inequality between KL and TV. *arXiv preprint arXiv:2202.07198*, 2022.

Nicolò Cesa-Bianchi, Pierre Gaillard, Claudio Gentile, and Sébastien Gerchinovitz. Algorithmic chaining and the role of partial feedback in online nonparametric learning. In Satyen Kale and Ohad Shamir (eds.), *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 465–481. PMLR, 2017.

Fan Chen, Dylan J. Foster, Yanjun Han, Jian Qian, Alexander Rakhlin, and Yunbei Xu. Assouad, Fano, and Le Cam with interaction: A unifying lower bound framework and characterization for bandit learnability. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Khaled Eldowa, Nicolò Cesa-Bianchi, Alberto Maria Metelli, and Marcello Restelli. Information capacity regret bounds for bandits with mediator feedback. *Journal of Machine Learning Research*, 25(353):1–36, 2024.

Khaled Eldowa, Thibaud Rahier, Augustin Cablant, Panayotis Mertikopoulos, and Pierre Gaillard. Leveraging similarities in multi-armed bandits. In Steve Hanneke and Tor Lattimore (eds.), *Proceedings of the Thirty-Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 2259–2306. PMLR, 2026.

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

Dylan J. Foster, Alexander Rakhlin, Ayush Sekhari, and Karthik Sridharan. On the complexity of adversarial decision making. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-2566.

Dylan J. Foster, Noah Golowich, and Yanjun Han. Tight guarantees for interactive decision making with the Decision-Estimation Coefficient. In *Proceedings of the Thirty-Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 3969–4043. PMLR, 2023.

Steve Hanneke and Kun Wang. A complete characterization of learnability for adversarial noisy bandits. *arXiv preprint arXiv:2605.09200*, 2026. URL https://arxiv.org/abs/2605.09200v2.

Elad Hazan, Tomer Koren, Roi Livni, and Yishay Mansour. Online learning with low rank experts. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir (eds.), *Proceedings of the 29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 1096–1114. PMLR, 2016.

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

J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960. doi: 10.4153/CJM-1960-030-4.

Tor Lattimore and Csaba Szepesvári. An information-theoretic approach to minimax regret in partial monitoring. In Alina Beygelzimer and Daniel Hsu (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 2111–2139. PMLR, 2019.

Tor Lattimore and Csaba Szepesvári. Exploration by optimisation in partial monitoring. In *Proceedings of the Thirty-Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2488–2515. PMLR, 2020.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Yury Polyanskiy and Yihong Wu. Dualizing Le Cam's method for functional estimation I: General theory. *The Annals of Statistics*, 54(1):1–24, 2026. doi: 10.1214/25-AOS2498.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018. doi: 10.1287/opre.2017.1663.

Gil Shabat, Chen Avin, Shie Mannor, Hanan Shteingart, Zvi Lotker, and Roey Yadgar. Hierarchical adversarial bandits for online configuration optimization. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 7899–7910. ACM, 2026. doi: 10.1145/3770855.3818431.

Anvay Shah, Ramsundar Anandanarayanan, Sharayu Moharir, and Shivaram Kalyanakrishnan. Online learning in tree MDPs by treating policies as bandit arms. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems*, pp. 1967–1976. IFAAMAS, 2026. doi: 10.65109/PYUD1139.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8:171–176, 1958. doi: 10.2140/pjm.1958.8.171.

Lin Yang, Yu-Zhen Janice Chen, Mohammad H. Hajiesmaili, Mark Herbster, and Don Towsley. Hierarchical learning algorithms for multi-scale expert problems. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 6(2):1–29, 2022. doi: 10.1145/3530900.

A Probability Model and Minimax Conventions

This appendix fixes the probability space, information pattern, strategy classes, and minimax conventions used throughout the paper. We use the probability-kernel formalism standard in structured adversarial decision making (Foster et al., 2022), while adopting the reward-function-before-action timing and arbitrary bounded-noise convention of adversarial noisy bandits (Hanneke & Wang, 2026). The formalization below is slightly stronger than the notation needed in the main text: the conditional observation law may depend arbitrarily on the public history, subject only to the prescribed conditional mean.

A.1 Canonical history space and probability kernels

Let the horizon be $T \in \mathbb{N}$, let $\mathcal{A}$ be a finite action set with $|\mathcal{A}| = K \geq 2$, and let

$$\mathcal{G} \subseteq [0,1]^{\mathcal{A}}$$

be a nonempty compact convex reward class. The observation space is $\mathcal{Y} := [0,1]$, equipped with its Borel σ-field. The action set is equipped with the discrete σ-field. For $t \in \{0, \dots, T\}$, define the canonical public-history space

$$\mathsf{Z}_0 := \{\varnothing\}, \qquad \mathsf{Z}_t := (\mathcal{A} \times \mathcal{Y})^t,$$

with the corresponding product σ-field. A generic element is written

$$h_t = (a_1, y_1, \dots, a_t, y_t) \in \mathsf{Z}_t.$$

The canonical sample space is $\Omega := \mathsf{Z}_T$. On this space, let

$$H_0 := \varnothing, \qquad H_t := (A_1, Y_1, \dots, A_t, Y_t), \qquad \mathscr{H}_t := \sigma(H_t), \qquad \mathscr{H}_0 := \{\varnothing, \Omega\}.$$

A probability kernel κ from a measurable space $(\mathsf{X}, \mathscr{X})$ to $(\mathsf{Y}, \mathscr{Y})$ assigns a probability measure $\kappa(\cdot \mid x)$ on $(\mathsf{Y}, \mathscr{Y})$ to each $x \in \mathsf{X}$, such that $x \mapsto \kappa(B \mid x)$ is $\mathscr{X}$-measurable for every $B \in \mathscr{Y}$. We write $\kappa : \mathsf{X} \rightsquigarrow \mathsf{Y}$.

A.2 Learner and environment strategies

Definition A.1 (Behavioral bandit algorithm)**.** A horizon-T algorithm is a sequence $\mathsf{Alg} = (\pi_t)_{t=1}^T$ of probability kernels

$$\pi_t : \mathsf{Z}_{t-1} \rightsquigarrow \mathcal{A}.$$

Equivalently, since $\mathcal{A}$ is finite, π_t is a Borel map from Z_{t-1} to $\Delta(\mathcal{A})$. The random mixed action used at round t is

$$p_t(\cdot) := \pi_t(\cdot \mid H_{t-1}).$$

The algorithm knows T, $\mathcal{A}$, and $\mathcal{G}$, and observes only the public history H_{t-1} before selecting p_t.

We take behavioral strategies as primitive. This is the standard kernel formalization for finite-horizon interactive decision problems; every algorithm used later in the paper is specified by such a sequence of conditional action laws.

Definition A.2 (Adaptive nonanticipating environment)**.** An admissible adaptive environment is a sequence

$$\mathsf{Adv} = \big(\mathbf{g}_t, \nu_t\big)_{t=1}^T$$

with the following properties.

1. The mean-reward selector $\mathbf{g}_t : \mathsf{Z}_{t-1} \to \mathcal{G}$ is Borel measurable.
2. The observation rule
$$\nu_t : \mathsf{Z}_{t-1} \times \mathcal{A} \rightsquigarrow \mathcal{Y}$$
is a probability kernel satisfying the pointwise mean constraint
$$\int_{\mathcal{Y}} y \, \nu_t(\mathrm{d}y \mid h_{t-1}, a) = \mathbf{g}_t(h_{t-1})(a) \tag{A.1}$$
for every $(h_{t-1}, a) \in \mathsf{Z}_{t-1} \times \mathcal{A}$.

The realized mean-reward vector is

$$g_t := \mathbf{g}_t(H_{t-1}) \in \mathcal{G}.$$

The class of all such environments is denoted by $\mathfrak{E}_T^{\mathrm{ad}}(\mathcal{G})$.

Definition A.2 is nonanticipating in the following precise sense. Conditional on $H_{t-1} = h_{t-1}$, the environment has fixed the entire menu

$$(\mathbf{g}_t(h_{t-1}), \{\nu_t(\cdot \mid h_{t-1}, a) : a \in \mathcal{A}\})$$

before the current action is sampled. The observation law may depend on the sampled action, but neither the reward vector nor the menu of observation laws may depend on the learner's current private draw. In the minimax criterion below the environment is selected after the algorithm, so it may know the algorithm and hence the map $h_{t-1} \mapsto p_t(\cdot \mid h_{t-1})$.

A.3 The induced trajectory law

Proposition A.3 (Existence and one-step factorization)**.** *For every algorithm* Alg *and every* $\mathsf{Adv} \in \mathfrak{E}_T^{\mathrm{ad}}(\mathcal{G})$*, there exists a unique probability measure* $\mathbb{P}_{\mathsf{Alg},\mathsf{Adv}}$ *on* Z_T *such that, for every* $t \in \{1, \dots, T\}$*,* $a \in \mathcal{A}$*, and Borel set* $B \subseteq \mathcal{Y}$*,*

$$\mathbb{P}_{\mathsf{Alg},\mathsf{Adv}}\big(A_t = a,\, Y_t \in B \mid \mathscr{H}_{t-1}\big) = \pi_t(a \mid H_{t-1})\nu_t(B \mid H_{t-1}, a) \quad \textit{a.s.} \tag{A.2}$$

Equivalently, for every bounded Borel function $\varphi : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$*,*

$$\mathbb{E}_{\mathsf{Alg},\mathsf{Adv}}\big[\varphi(A_t, Y_t) \mid \mathscr{H}_{t-1}\big] = \sum_{a \in \mathcal{A}} p_t(a) \int_{\mathcal{Y}} \varphi(a, y)\nu_t(\mathrm{d}y \mid H_{t-1}, a) \quad \textit{a.s.} \tag{A.3}$$

Proof. For $h_{t-1} \in \mathsf{Z}_{t-1}$, define the probability kernel $\kappa_t : \mathsf{Z}_{t-1} \rightsquigarrow (\mathcal{A} \times \mathcal{Y})$ by

$$\kappa_t(\{a\} \times B \mid h_{t-1}) := \pi_t(a \mid h_{t-1})\nu_t(B \mid h_{t-1}, a).$$

Measurability follows from finiteness of $\mathcal{A}$ and the kernel properties of π_t and ν_t. The finite-horizon Ionescu–Tulcea construction applied to $\kappa_1, \dots, \kappa_T$ yields a unique probability measure on Z_T with these conditional kernels. Equation (A.2) is the defining factorization, and (A.3) follows by integration. □

Unless ambiguity is possible, all expectations and conditional expectations below are taken under $\mathbb{P}_{\mathsf{Alg},\mathsf{Adv}}$. For $p \in \Delta(\mathcal{A})$ and $g \in \mathbb{R}^{\mathcal{A}}$, write

$$p(g) := \sum_{a \in \mathcal{A}} p(a)g(a).$$

Lemma A.4 (Conditional identities)**.** *For every admissible pair* $(\mathsf{Alg}, \mathsf{Adv})$ *and every* $t \in \{1, \dots, T\}$*, the following identities hold:*

$$\mathbb{E}[Y_t \mid \mathscr{H}_{t-1}, A_t] = g_t(A_t), \tag{A.4}$$

$$\mathbb{E}[Y_t \mid \mathscr{H}_{t-1}] = p_t(g_t), \tag{A.5}$$

$$\mathbb{E}[g_t(A_t) \mid \mathscr{H}_{t-1}] = p_t(g_t). \tag{A.6}$$

Define

$$\xi_t := Y_t - g_t(A_t), \qquad \zeta_t := g_t(A_t) - p_t(g_t).$$

Then $(\xi_t)_{t=1}^T$ *and* $(\zeta_t)_{t=1}^T$ *are martingale-difference sequences with respect to* $(\mathscr{H}_t)_{t=0}^T$*. Moreover,*

$$|\xi_t| \le 1, \qquad |\zeta_t| \le 1, \qquad \mathbb{E}[\xi_t^2 \mid \mathscr{H}_{t-1}, A_t] \le \frac{1}{4} \quad \textit{a.s.} \tag{A.7}$$

Proof. Equation (A.4) follows from (A.1) and Proposition A.3. Applying the tower property and summing over $A_t \sim p_t$ yields (A.5); the same calculation without Y_t gives (A.6). Hence $\mathbb{E}[\xi_t \mid \mathscr{H}_{t-1}] = 0$ and $\mathbb{E}[\zeta_t \mid \mathscr{H}_{t-1}] = 0$. The absolute bounds follow from $Y_t, g_t(A_t), p_t(g_t) \in [0, 1]$. Finally, a random variable supported on $[0, 1]$ has variance at most $1/4$, which proves the last inequality. □

A.4 Adaptive, oblivious-mean, and Bernoulli environments

Definition A.5 (Oblivious-mean environment)**.** An environment $\mathsf{Adv} = (\mathbf{g}_t, \nu_t)_{t=1}^T$ is *oblivious-mean* if there are vectors $g_1^\circ, \dots, g_T^\circ \in \mathcal{G}$ such that

$$\mathbf{g}_t(h_{t-1}) = g_t^\circ \qquad \text{for every } h_{t-1} \in \mathsf{Z}_{t-1}.$$

The observation kernels ν_t may remain history dependent, but must satisfy (A.1) with the fixed mean vector g_t°. The class of such environments is denoted by $\mathfrak{E}_T^{\mathrm{obl}}(\mathcal{G})$.

Thus "oblivious-mean" refers to the sequence of conditional mean-reward vectors: the sequence $(g_t^\circ)_{t=1}^T$ is fixed before interaction. This agrees with the standard reward-adversary distinction, under which an adaptive adversary may choose its next reward function using past actions and observed rewards (Hanneke & Wang, 2026). We deliberately retain robustness to history-dependent bounded noise even in the oblivious-mean case. In particular,

$$\mathfrak{E}_T^{\mathrm{obl}}(\mathcal{G}) \subseteq \mathfrak{E}_T^{\mathrm{ad}}(\mathcal{G}).$$

Definition A.6 (Bernoulli subclass)**.** Fix $\rho \in (0, 1/2)$. The adaptive Bernoulli subclass $\mathfrak{E}_{T,\rho}^{\mathrm{ad,Ber}}(\mathcal{G})$ consists of the environments in $\mathfrak{E}_T^{\mathrm{ad}}(\mathcal{G})$ satisfying

$$g_t(a) \in [\rho, 1-\rho] \quad \text{a.s. for every } t, a,$$

and

$$\nu_t(\cdot \mid h_{t-1}, a) = \mathrm{Ber}(\mathbf{g}_t(h_{t-1})(a)) \tag{A.8}$$

for every (h_{t-1}, a). The oblivious Bernoulli subclass is

$$\mathfrak{E}_{T,\rho}^{\mathrm{obl,Ber}}(\mathcal{G}) := \mathfrak{E}_{T,\rho}^{\mathrm{ad,Ber}}(\mathcal{G}) \cap \mathfrak{E}_T^{\mathrm{obl}}(\mathcal{G}).$$

When no interior parameter is required, the superscript Ber denotes the same definition with $\rho = 0$ and means allowed in $[0, 1]$.

We use the notation in Definition A.6 only when the corresponding subclass is nonempty; every explicit lower-bound class constructed in this paper lies in $[\rho, 1-\rho]^{\mathcal{A}}$ for a stated $\rho > 0$. The Bernoulli classes are used only for lower bounds. Since Bernoulli observations are a special case of bounded observations with the prescribed mean, every lower bound proved on a Bernoulli subclass is automatically a lower bound for the full bounded-noise model.

A.5 Regret and minimax values

Definition A.7 (Mean-reward regret)**.** For the interaction induced by an algorithm Alg and an environment Adv, define

$$\mathrm{Reg}_T := \max_{a \in \mathcal{A}} \sum_{t=1}^T g_t(a) - \sum_{t=1}^T g_t(A_t). \tag{A.9}$$

The comparator is selected after the realized sequence $(g_1, \dots, g_T)$ is determined. In particular, when the environment is adaptive, the comparator itself is generally random.

To prevent expectation identities from being mistaken for pathwise identities, we use three distinct symbols throughout this appendix:

$$\mathrm{Reg}_T^{\mathrm{real}} := \max_{a \in \mathcal{A}} \sum_{t=1}^T g_t(a) - \sum_{t=1}^T g_t(A_t), \tag{A.10}$$

$$\mathrm{Reg}_T^{\mathrm{mix}} := \max_{a \in \mathcal{A}} \sum_{t=1}^T g_t(a) - \sum_{t=1}^T p_t(g_t), \tag{A.11}$$

$$\mathrm{Reg}_T^{\mathrm{obs}} := \max_{a \in \mathcal{A}} \sum_{t=1}^T g_t(a) - \sum_{t=1}^T Y_t. \tag{A.12}$$

Thus (A.9) is $\mathrm{Reg}_T = \mathrm{Reg}_T^{\mathrm{real}}$; the other two quantities coincide with it only after taking expectation under an admissible trajectory law.

For a fixed algorithm, define its worst-case adaptive and oblivious expected regrets by

$$\mathfrak{R}_T^{\mathrm{ad}}(\mathsf{Alg}, \mathcal{G}) := \sup_{\mathsf{Adv} \in \mathfrak{E}_T^{\mathrm{ad}}(\mathcal{G})} \mathbb{E}_{\mathsf{Alg},\mathsf{Adv}}[\mathrm{Reg}_T], \tag{A.13}$$

$$\mathfrak{R}_T^{\mathrm{obl}}(\mathsf{Alg}, \mathcal{G}) := \sup_{\mathsf{Adv} \in \mathfrak{E}_T^{\mathrm{obl}}(\mathcal{G})} \mathbb{E}_{\mathsf{Alg},\mathsf{Adv}}[\mathrm{Reg}_T]. \tag{A.14}$$

The corresponding minimax values are

$$\mathfrak{R}_T^{\star,\mathrm{ad}}(\mathcal{G}) := \inf_{\mathsf{Alg}} \mathfrak{R}_T^{\mathrm{ad}}(\mathsf{Alg}, \mathcal{G}), \tag{A.15}$$

$$\mathfrak{R}_T^{\star,\mathrm{obl}}(\mathcal{G}) := \inf_{\mathsf{Alg}} \mathfrak{R}_T^{\mathrm{obl}}(\mathsf{Alg}, \mathcal{G}). \tag{A.16}$$

The infima range over all horizon-T behavioral algorithms from Definition A.1. No attainment of either infimum or supremum is assumed, and no interchange of $\inf$ and $\sup$ is used unless proved separately.

When the environment class in (A.13)–(A.16) is replaced by the Bernoulli subclasses from Definition A.6, we write $\mathfrak{R}_{T,\rho}^{\star,\mathrm{ad,Ber}}(\mathcal{G})$ and $\mathfrak{R}_{T,\rho}^{\star,\mathrm{obl,Ber}}(\mathcal{G})$.

Proposition A.8 (Equivalent expected-regret representations)**.** *For every admissible pair* $(\mathsf{Alg}, \mathsf{Adv})$*,*

$$\mathbb{E}_{\mathsf{Alg},\mathsf{Adv}}[\mathrm{Reg}_T^{\mathrm{real}}] = \mathbb{E}_{\mathsf{Alg},\mathsf{Adv}}[\mathrm{Reg}_T^{\mathrm{mix}}] \tag{A.17}$$

$$= \mathbb{E}_{\mathsf{Alg},\mathsf{Adv}}[\mathrm{Reg}_T^{\mathrm{obs}}]. \tag{A.18}$$

Moreover,

$$0 \le \mathfrak{R}_T^{\star,\mathrm{obl}}(\mathcal{G}) \le \mathfrak{R}_T^{\star,\mathrm{ad}}(\mathcal{G}) \le T. \tag{A.19}$$

For every $\rho \in (0, 1/2)$ *for which the corresponding Bernoulli subclass is nonempty,*

$$\mathfrak{R}_{T,\rho}^{\star,\mathrm{ad,Ber}}(\mathcal{G}) \le \mathfrak{R}_T^{\star,\mathrm{ad}}(\mathcal{G}), \tag{A.20}$$

$$\mathfrak{R}_{T,\rho}^{\star,\mathrm{obl,Ber}}(\mathcal{G}) \le \mathfrak{R}_T^{\star,\mathrm{obl}}(\mathcal{G}). \tag{A.21}$$

Proof. By Lemma A.4,

$$\mathbb{E}[g_t(A_t)] = \mathbb{E}[p_t(g_t)] = \mathbb{E}[Y_t]$$

for every t. Substituting these equalities into (A.9) proves (A.17) and (A.18).

The middle inequality in (A.19) follows from $\mathfrak{E}_T^{\mathrm{obl}}(\mathcal{G}) \subseteq \mathfrak{E}_T^{\mathrm{ad}}(\mathcal{G})$. Since all rewards lie in $[0, 1]$, $\mathrm{Reg}_T \le T$ pathwise, proving the upper bound. For the lower bound, fix any $g^\circ \in \mathcal{G}$ and consider the oblivious environment that repeats g° in every round. Its expected regret is

$$\mathbb{E}\left[\sum_{t=1}^{T} \big(\max_a g^\circ(a) - g^\circ(A_t)\big)\right] \ge 0$$

for every algorithm. Hence the supremum over oblivious environments is nonnegative. Finally, (A.20) and (A.21) follow by inclusion of environment classes. □

Remark A.9 (Expectation versus pathwise regret). Equations (A.17) and (A.18) are identities of expectations, not pathwise equalities. More precisely, if

$$\overline{\mathrm{Reg}}_T := \max_{a \in \mathcal{A}} \sum_{t=1}^{T} g_t(a) - \sum_{t=1}^{T} p_t(g_t),$$

then

$$\mathrm{Reg}_T - \overline{\mathrm{Reg}}_T = -\sum_{t=1}^{T} \zeta_t, \qquad \left(\max_a \sum_{t=1}^{T} g_t(a) - \sum_{t=1}^{T} Y_t\right) - \mathrm{Reg}_T = -\sum_{t=1}^{T} \xi_t.$$

Thus any high-probability transfer among these notions requires an explicit martingale argument; it does not follow from Proposition A.8 alone.

A.6 Randomized priors as a lower-bound device

Lower bounds will often introduce a latent index only for analysis. This does not change the minimax environment class.

Lemma A.10 (Bayes-to-worst-case reduction)**.** *Let Θ be a finite set, let $\mu \in \Delta(\Theta)$, and let $\{\mathsf{Adv}_\theta : \theta \in \Theta\}$ be any finite family contained in $\mathfrak{E}_T^{\mathrm{obl}}(\mathcal{G})$. For every behavioral algorithm Alg,*

$$\sup_{\mathsf{Adv} \in \mathfrak{E}_T^{\mathrm{obl}}(\mathcal{G})} \mathbb{E}_{\mathsf{Alg},\mathsf{Adv}}[\mathrm{Reg}_T] \geq \mathbb{E}_{\theta\sim\mu}\mathbb{E}_{\mathsf{Alg},\mathsf{Adv}_\theta}[\mathrm{Reg}_T]. \tag{A.22}$$

Consequently, if the right-hand side of (A.22) *is at least L for every Alg, then*

$$\mathfrak{R}_T^{\star,\mathrm{obl}}(\mathcal{G}) \geq L.$$

The same statement holds with the Bernoulli subclasses in place of the full oblivious class.

Proof. The right-hand side of (A.22) is a convex combination of expected regrets, each bounded above by the supremum on the left. Taking the infimum over Alg proves the second claim. □

Under the mixture law used in Lemma A.10, the latent θ is sampled once before interaction and is not included in the learner's filtration $(\mathscr{H}_t)$. Each conditional law $\mathbb{P}_{\mathsf{Alg},\mathsf{Adv}_\theta}$ corresponds to an individually admissible oblivious environment; the mixture over θ is a proof device and need not itself be identified with a deterministic adversary.

A.7 An oblivious Bernoulli convexification lemma

The core paper is stated directly for a convex class $\mathcal{G}$. The following auxiliary result records the exact convexification relation that is available without any trajectory-level subtlety. A related independent-mixture construction is used in the necessity argument of Hanneke & Wang (2026).

For a nonempty class $\mathcal{F} \subseteq [0,1]^{\mathcal{A}}$, let $\mathfrak{R}_T^{\star,\mathrm{obl,Ber}}(\mathcal{F})$ denote the oblivious minimax value when the adversary fixes a sequence in $\mathcal{F}^T$ and observations are conditionally Bernoulli with the corresponding means.

Proposition A.11 (Oblivious Bernoulli convexification)**.** *For every nonempty $\mathcal{F} \subseteq [0,1]^{\mathcal{A}}$ and every horizon T,*

$$\mathfrak{R}_T^{\star,\mathrm{obl,Ber}}(\mathcal{F}) = \mathfrak{R}_T^{\star,\mathrm{obl,Ber}}\big(\mathrm{co}(\mathcal{F})\big). \tag{A.23}$$

Proof. Since $\mathcal{F} \subseteq \mathrm{co}(\mathcal{F})$, the left-hand side of (A.23) is no larger than the right-hand side. We prove the reverse inequality.

Fix a behavioral algorithm Alg and a deterministic sequence

$$g_1, \dots, g_T \in \mathrm{co}(\mathcal{F}).$$

For each t, choose a finite convex representation

$$g_t = \sum_{i=1}^{m_t} \lambda_{t,i} f_{t,i}, \qquad f_{t,i} \in \mathcal{F}, \quad \lambda_{t,i} \geq 0, \quad \sum_{i=1}^{m_t} \lambda_{t,i} = 1.$$

Independently across t, and independently of the learner and observation randomness, draw an index I_t with $\mathbb{P}(I_t = i) = \lambda_{t,i}$, and set $F_t := f_{t,I_t}$. The entire sequence $(F_1, \dots, F_T)$ may be sampled before interaction. Conditional on this sequence, run the oblivious Bernoulli environment with mean vector F_t at round t.

We first compare the observable transcript with that generated by the deterministic convexified sequence $(g_t)_{t=1}^T$. Because F_t is independent of (H_{t-1}, A_t),

$$\mathbb{P}(Y_t = 1 \mid H_{t-1}, A_t) = \mathbb{E}[F_t(A_t) \mid H_{t-1}, A_t] = g_t(A_t).$$

A binary distribution is determined by its mean, so induction on t shows that the marginal law of H_T under the randomized $\mathcal{F}$-valued construction is identical to the transcript law under Bernoulli observations with deterministic mean sequence $(g_t)_{t=1}^T$.

Next, Jensen's inequality for the convex map $(x_1, \dots, x_T) \mapsto \max_a \sum_t x_t(a)$ gives

$$\mathbb{E}\left[\max_{a\in\mathcal{A}} \sum_{t=1}^{T} F_t(a)\right] \geq \max_{a\in\mathcal{A}} \sum_{t=1}^{T} g_t(a).$$

Moreover,

$$\mathbb{E}[F_t(A_t)] = \mathbb{E}\big[\mathbb{E}[F_t(A_t) \mid H_{t-1}, A_t]\big] = \mathbb{E}[g_t(A_t)].$$

Therefore the expected regret, averaged over $(I_1, \dots, I_T)$, is at least the expected regret of Alg against the deterministic convexified sequence $(g_t)_{t=1}^T$. Since the average of finitely many numbers cannot exceed their maximum, there is a deterministic realization $(f_1, \dots, f_T) \in \mathcal{F}^T$ whose expected regret is at least this value. Taking the supremum over $(g_t)_{t=1}^T$ and then the infimum over Alg proves the reverse inequality. □

Remark A.12 (Why no adaptive convexification equality is assumed). Proposition A.11 relies on the independence of the round-t mixture component F_t from the preceding transcript. Under an adaptive convexified environment, future reward vectors are functions of a transcript that is itself affected by the hidden mixture components. Although one can often match one-step observation marginals, the best-fixed-action comparator is a nonlinear functional of the entire latent reward sequence. We therefore do not invoke an adaptive analogue of (A.23). All adaptive upper bounds and the main cover–packing theorem are formulated directly for the compact convex class $\mathcal{G}$.

A.8 Standing conventions for subsequent appendices

The following conventions are in force throughout the remaining proofs.

1. All algorithms may depend on T, $\mathcal{A}$, and $\mathcal{G}$.
2. The symbol g_t always denotes the predictable mean-reward vector. More precisely, $g_t = \mathbf{g}_t(H_{t-1})$; it never denotes the noisy observation.
3. Unless a theorem explicitly states otherwise, upper bounds are with respect to $\mathfrak{E}_T^{\mathrm{ad}}(\mathcal{G})$ and use only the bounded conditional-mean property (A.1).
4. Lower bounds may restrict to a finite prior over the Bernoulli-oblivious class $\mathfrak{E}_{T,\rho}^{\mathrm{obl,Ber}}(\mathcal{G})$ and then invoke Lemma A.10.
5. The notation $\mathbb{E}_t[\cdot]$ abbreviates $\mathbb{E}[\cdot \mid \mathscr{H}_{t-1}]$; conditioning additionally on A_t is always written explicitly.
6. All optimization values are interpreted as infima or suprema. Existence of an optimizer is claimed only when proved. The core results are information-theoretic and may invoke exact finite-dimensional optimization oracles; no computational-efficiency claim is implicit.
7. Every occurrence of "adaptive," "oblivious," "Bernoulli," "expected," or "high probability" refers to the definitions in this appendix, unless the statement explicitly introduces a different convention.

B Optimizer Geometry and Hypergraph Duality

This appendix isolates the purely optimizer-theoretic content of generalized maximin volume and distribution covers. Both notions were introduced into the analysis of adversarial noisy bandits by Hanneke & Wang (2026). In the finite-action setting of this paper, they admit exact finite-dimensional hypergraph formulations. These formulations will be used in two ways. First, they yield algebraically exact identities and rounding bounds that remove any ambiguity about what the existing quantities measure. Second, they make precise what those quantities do *not* measure: they depend only on the family of near-optimal action sets and contain no information about how observations at one action constrain rewards at another action.

Throughout this appendix, fix a finite action set $\mathcal{A}$ with $|\mathcal{A}| = K \geq 2$, a nonempty reward class $\mathcal{G} \subseteq [0,1]^{\mathcal{A}}$, a scale $\alpha \in [0,1]$, and a cover slack $\beta \in [0,1)$. Convexity of $\mathcal{G}$ is not needed for the combinatorial identities below, although it remains a standing assumption of the main paper. Duplicate hyperedges are removed whenever a set system is formed; every near-optimal set is nonempty; empty coverage patterns may be discarded; and every reciprocal uses the extended-value convention $1/0 = +\infty$.

B.1 THE NEAR-OPTIMAL-SET HYPERGRAPH

For $g \in \mathcal{G}$, let

$$g^\star := \max_{a \in \mathcal{A}} g(a), \qquad \mathsf{A}_\alpha(g) := \{a \in \mathcal{A} : g^\star - g(a) \le \alpha\}.$$

Because $\mathcal{A}$ is finite, $\mathsf{A}_\alpha(g)$ is nonempty. Define the family of distinct α-optimal sets

$$\mathfrak{H}_\alpha(\mathcal{G}) := \{\mathsf{A}_\alpha(g) : g \in \mathcal{G}\} \subseteq 2^{\mathcal{A}} \setminus \{\varnothing\}. \tag{B.1}$$

Thus $\mathfrak{H}_\alpha(\mathcal{G})$ is a finite hypergraph on vertex set $\mathcal{A}$, even when $\mathcal{G}$ is infinite; in particular,

$$1 \le |\mathfrak{H}_\alpha(\mathcal{G})| \le 2^K - 1. \tag{B.2}$$

We recall the relevant covering and packing parameters for a nonempty finite hypergraph $\mathfrak{H} \subseteq 2^{\mathcal{A}} \setminus \{\varnothing\}$. Its integral transversal number is

$$\tau(\mathfrak{H}) := \min\{|S| : S \subseteq \mathcal{A},\ S \cap E \neq \varnothing \text{ for every } E \in \mathfrak{H}\}. \tag{B.3}$$

Its fractional transversal number is

$$\tau^\star(\mathfrak{H}) := \min_{w \in \mathbb{R}_+^{\mathcal{A}}} \left\{ \sum_{a \in \mathcal{A}} w(a) : \sum_{a \in E} w(a) \ge 1 \text{ for every } E \in \mathfrak{H} \right\}. \tag{B.4}$$

The dual fractional packing number is

$$\nu^\star(\mathfrak{H}) := \max_{\lambda \in \mathbb{R}_+^{\mathfrak{H}}} \left\{ \sum_{E \in \mathfrak{H}} \lambda(E) : \sum_{E \ni a} \lambda(E) \le 1 \text{ for every } a \in \mathcal{A} \right\}. \tag{B.5}$$

All three optimization problems attain their optima. The two fractional programs are a finite primal–dual linear-programming pair, and hence

$$\tau^\star(\mathfrak{H}) = \nu^\star(\mathfrak{H}). \tag{B.6}$$

Proposition B.1 (Maximin volume as reciprocal fractional transversality)**.** *Let $\mathfrak{H} = \mathfrak{H}_\alpha(\mathcal{G})$. Then*

$$\gamma_\alpha(\mathcal{G}) = \max_{p \in \Delta(\mathcal{A})} \min_{E \in \mathfrak{H}} p(E) \tag{B.7}$$

$$= \frac{1}{\tau^\star(\mathfrak{H})} = \frac{1}{\nu^\star(\mathfrak{H})} \tag{B.8}$$

$$= \min_{\lambda \in \Delta(\mathfrak{H})} \max_{a \in \mathcal{A}} \sum_{E \in \mathfrak{H}:\, a \in E} \lambda(E). \tag{B.9}$$

Here $p(E) := \sum_{a \in E} p(a)$, and $\Delta(\mathfrak{H})$ denotes the probability simplex over the distinct hyperedges of $\mathfrak{H}$.

Proof. The first equality follows directly from the definition of generalized maximin volume: the quantity $p(\mathsf{A}_\alpha(g))$ depends on g only through the hyperedge $\mathsf{A}_\alpha(g)$.

Set $\gamma := \max_{p \in \Delta(\mathcal{A})} \min_{E \in \mathfrak{H}} p(E)$. Since every edge is nonempty, the uniform distribution gives $\gamma \ge 1/K > 0$, and compactness of $\Delta(\mathcal{A})$ implies that the maximum is attained. If $p^\star$ attains γ, then $w(a) := p^\star(a)/\gamma$ is feasible for (B.4), with total mass $1/\gamma$. Hence

$$\tau^\star(\mathfrak{H}) \le \frac{1}{\gamma}.$$

Conversely, let w be feasible for (B.4) and write $W := \sum_a w(a)$. Necessarily $W > 0$. The normalized vector $p(a) := w(a)/W$ belongs to $\Delta(\mathcal{A})$ and satisfies $p(E) \ge 1/W$ for every $E \in \mathfrak{H}$. Therefore $\gamma \ge 1/W$. Minimizing over feasible w gives $\gamma \ge 1/\tau^\star(\mathfrak{H})$, proving the first equality in (B.8); the second follows from (B.6).

Finally, apply the finite minimax theorem to the matrix $M \in \{0,1\}^{\mathcal{A} \times \mathfrak{H}}$ with $M_{a,E} = \mathbf{1}\{a \in E\}$:

$$\max_{p \in \Delta(\mathcal{A})} \min_{\lambda \in \Delta(\mathfrak{H})} p^\top M \lambda = \min_{\lambda \in \Delta(\mathfrak{H})} \max_{p \in \Delta(\mathcal{A})} p^\top M \lambda.$$

The minimum over a simplex on the left is attained at a pure hyperedge, while the maximum over a simplex on the right is attained at a pure action. This is exactly (B.9). □

Equation (B.9) gives a useful congestion interpretation. An adversary randomizes over near-optimal sets according to λ; the load of action a is the probability that a random set contains a. Maximin volume is the smallest achievable maximum action load.

Corollary B.2 (Elementary bounds and extremal cases)**.** *For every finite-action reward class and every $\alpha \in [0,1]$,*

$$\frac{1}{K} \le \gamma_\alpha(\mathcal{G}) \le 1, \qquad 1 \le \tau^\star(\mathfrak{H}_\alpha(\mathcal{G})) \le \tau(\mathfrak{H}_\alpha(\mathcal{G})) \le K. \tag{B.10}$$

Moreover,

$$\gamma_\alpha(\mathcal{G}) = 1 \quad \Longleftrightarrow \quad \bigcap_{E \in \mathfrak{H}_\alpha(\mathcal{G})} E \neq \varnothing. \tag{B.11}$$

Proof. The uniform distribution gives the lower bound $1/K$, and all probabilities are at most one. The transversal inequalities are immediate from Proposition B.1 and the fact that every integral transversal induces a feasible fractional transversal. If the intersection in (B.11) contains a, then $p = \delta_a$ gives volume one. Conversely, volume one is attained by some p; $p(E) = 1$ for every edge E, so $\mathrm{supp}(p)$ is contained in every edge and is nonempty. □

In particular, qualitative positivity of maximin volume is automatic on a finite action set. The role of the quantity in this paper is therefore quantitative: its magnitude and the geometry behind its fractional transversal may or may not capture the finite-time regret rate.

Proposition B.3 (Finite randomized rounding)**.** *Let $\mathfrak{H} = \mathfrak{H}_\alpha(\mathcal{G})$ and $m_\alpha := |\mathfrak{H}|$. Then*

$$\tau(\mathfrak{H}) \le \min\left\{K, \left\lceil \frac{\log m_\alpha + 1}{\gamma_\alpha(\mathcal{G})} \right\rceil \right\}. \tag{B.12}$$

Consequently,

$$\tau(\mathfrak{H}) \le \big(\log m_\alpha + 2\big)\tau^\star(\mathfrak{H})\,. \tag{B.13}$$

Proof. Let p attain $\gamma_\alpha(\mathcal{G}) =: \gamma$, and draw n actions independently from p. For any $E \in \mathfrak{H}$, the probability that none of the sampled actions lies in E is

$$(1 - p(E))^n \le (1-\gamma)^n \le e^{-n\gamma}.$$

A union bound over the m_α distinct edges shows that the probability of missing at least one edge is at most $m_\alpha e^{-n\gamma}$. Taking $n = \lceil (\log m_\alpha + 1)/\gamma \rceil$ makes this probability strictly less than one. Hence some realization hits every edge; deleting repeated actions only reduces its cardinality. The bound K follows by taking all actions. Finally, use $1/\gamma = \tau^\star(\mathfrak{H}) \ge 1$ and $\lceil x \rceil \le x + 1$ to obtain (B.13). □

B.2 Hitting sets and zero-slack distribution covers

An α-hitting set for $\mathcal{G}$ is a set $S \subseteq \mathcal{A}$ satisfying $S \cap \mathsf{A}_\alpha(g) \neq \varnothing$ for every $g \in \mathcal{G}$. Its minimum size is therefore exactly

$$N_{\mathrm{hit}}(\mathcal{G}; \alpha) = \tau(\mathfrak{H}_\alpha(\mathcal{G}))\,. \tag{B.14}$$

Proposition B.4 (Zero-slack covers are exactly hitting sets)**.** *For every $\alpha \in [0,1]$,*

$$N_{\mathrm{dc}}(\mathcal{G}; \alpha, 0) = N_{\mathrm{hit}}(\mathcal{G}; \alpha) = \tau(\mathfrak{H}_\alpha(\mathcal{G}))\,. \tag{B.15}$$

In particular,

$$\frac{1}{\gamma_\alpha(\mathcal{G})} \le N_{\mathrm{dc}}(\mathcal{G}; \alpha, 0) \le \min\left\{K, \left\lceil \frac{\log m_\alpha + 1}{\gamma_\alpha(\mathcal{G})} \right\rceil \right\}. \tag{B.16}$$

Proof. If S is an α-hitting set, then $Q := \{\delta_a : a \in S\}$ is an $(\alpha, 0)$-distribution cover. Thus the left-hand side of (B.15) is no larger than the hitting number.

Conversely, let Q be an $(\alpha, 0)$-distribution cover. For every $q \in Q$, choose one action $s(q) \in \mathrm{supp}(q)$ and set $S := \{s(q) : q \in Q\}$. For each $g \in \mathcal{G}$, some $q \in Q$ satisfies $q(\mathsf{A}_\alpha(g)) = 1$. Since $\mathcal{A}$ is finite, this implies $\mathrm{supp}(q) \subseteq \mathsf{A}_\alpha(g)$, and hence $s(q) \in \mathsf{A}_\alpha(g)$. Thus S is a hitting set and $|S| \le |Q|$. This proves (B.15); the sandwich follows from Proposition B.1 and Proposition B.3. □

B.3 Distribution covers as a second-order set-cover problem

For a hypergraph $\mathfrak{H}$ on $\mathcal{A}$ and $q \in \Delta(\mathcal{A})$, define the β-coverage pattern of q by

$$\mathrm{pat}_\beta(q;\mathfrak{H}) := \{E \in \mathfrak{H} : q(E) \geq 1-\beta\}. \tag{B.17}$$

The family of realizable patterns is

$$\mathfrak{P}_\beta(\mathfrak{H}) := \big\{\mathrm{pat}_\beta(q;\mathfrak{H}) : q \in \Delta(\mathcal{A})\big\} \subseteq 2^{\mathfrak{H}}. \tag{B.18}$$

Although the simplex is uncountable, $\mathfrak{P}_\beta(\mathfrak{H})$ is finite, with cardinality at most $2^{|\mathfrak{H}|}$. A pattern records exactly which near-optimal sets are covered by one distribution; all further metric information about that distribution is discarded.

Define the integral and fractional second-order covering numbers by

$$\tau_\beta^{(2)}(\mathfrak{H}) := \min\left\{n : \exists P_1,\dots,P_n \in \mathfrak{P}_\beta(\mathfrak{H}),\ \mathfrak{H} = \bigcup_{i=1}^n P_i\right\}, \tag{B.19}$$

$$\tau_\beta^{(2),\star}(\mathfrak{H}) := \min_{w \in \mathbb{R}_+^{\mathfrak{P}_\beta(\mathfrak{H})}} \left\{\sum_{P \in \mathfrak{P}_\beta(\mathfrak{H})} w(P) : \sum_{P \ni E} w(P) \geq 1\ \forall E \in \mathfrak{H}\right\}. \tag{B.20}$$

The empty pattern, when realizable, is harmless and may be discarded from both programs.

Proposition B.5 (Exact second-order formulation of distribution covering)**.** *For $\mathfrak{H} = \mathfrak{H}_\alpha(\mathcal{G})$,*

$$N_{\mathrm{dc}}(\mathcal{G};\alpha,\beta) = \tau_\beta^{(2)}(\mathfrak{H}). \tag{B.21}$$

Equivalently, if

$$\mathcal{C}_{E,\beta} := \{q \in \Delta(\mathcal{A}) : q(E) \geq 1-\beta\}, \qquad E \in \mathfrak{H}, \tag{B.22}$$

then a finite $Q \subseteq \Delta(\mathcal{A})$ is a distribution cover if and only if it is a transversal of the second-order set system $\{\mathcal{C}_{E,\beta} : E \in \mathfrak{H}\}$ on vertex space $\Delta(\mathcal{A})$.

Proof. A finite family $Q = \{q_1,\dots,q_n\}$ is an (α,β)-distribution cover if and only if, for every $E \in \mathfrak{H}$, at least one q_i satisfies $q_i(E) \geq 1-\beta$. By (B.17), this is equivalent to

$$E \in \bigcup_{i=1}^n \mathrm{pat}_\beta(q_i;\mathfrak{H}) \qquad \forall E \in \mathfrak{H}.$$

Minimizing n gives (B.21). The transversal formulation is the same incidence relation with the roles of points and covered hyperedges written explicitly. □

The fractional relaxation in (B.20) has a direct maximin interpretation. Define the second-order fractional volume

$$\gamma_{\alpha,\beta}^{(2)}(\mathcal{G}) := \max_{\mu \in \Delta(\mathfrak{P}_\beta(\mathfrak{H}))} \min_{E \in \mathfrak{H}} \sum_{P \ni E} \mu(P), \qquad \mathfrak{H} = \mathfrak{H}_\alpha(\mathcal{G}). \tag{B.23}$$

Equivalently, one may maximize over probability measures on $\Delta(\mathcal{A})$ and write

$$\gamma_{\alpha,\beta}^{(2)}(\mathcal{G}) = \sup_{\mathsf{M} \in \mathcal{P}(\Delta(\mathcal{A}))} \inf_{g \in \mathcal{G}} \mathbb{P}_{q \sim \mathsf{M}}\left(q(\mathsf{A}_\alpha(g)) \geq 1-\beta\right). \tag{B.24}$$

The two forms agree because every distribution on q induces a distribution on its finite coverage pattern, and one may choose a representative q for every realizable pattern.

Proposition B.6 (Second-order fractional duality)**.** *Let $\mathfrak{H} = \mathfrak{H}_\alpha(\mathcal{G})$. Then*

$$\gamma_{\alpha,\beta}^{(2)}(\mathcal{G}) = \frac{1}{\tau_\beta^{(2),\star}(\mathfrak{H})} \tag{B.25}$$

$$= \min_{\lambda \in \Delta(\mathfrak{H})} \max_{P \in \mathfrak{P}_\beta(\mathfrak{H})} \sum_{E \in P} \lambda(E). \tag{B.26}$$

Moreover,

$$\gamma_\alpha(\mathcal{G}) \le \gamma^{(2)}_{\alpha,\beta}(\mathcal{G}), \qquad \frac{1}{\gamma^{(2)}_{\alpha,\beta}(\mathcal{G})} \le N_{\mathrm{dc}}(\mathcal{G};\alpha,\beta) \le \tau(\mathfrak{H})\,. \tag{B.27}$$

At zero slack, all first- and second-order fractional quantities coincide:

$$\gamma^{(2)}_{\alpha,0}(\mathcal{G}) = \gamma_\alpha(\mathcal{G})\,, \qquad \tau^{(2),\star}_0(\mathfrak{H}) = \tau^\star(\mathfrak{H})\,, \qquad \tau^{(2)}_0(\mathfrak{H}) = \tau(\mathfrak{H})\,. \tag{B.28}$$

Proof. The proof of (B.25) is the same rescaling argument as in Proposition B.1, now with patterns as covering objects and hyperedges as constraints. Explicitly, if μ guarantees coverage probability at least γ for every $E \in \mathfrak{H}$, then $w(P) = \mu(P)/\gamma$ is feasible for (B.20), with total mass $1/\gamma$. Conversely, normalizing any feasible w by its total mass gives a probability measure on patterns whose minimum edge-coverage probability is at least the reciprocal of that mass. Finite minimax duality for the incidence matrix $\mathbf{1}\{E \in P\}$ gives (B.26).

For every $a \in \mathcal{A}$ and every $\beta < 1$,

$$\mathrm{pat}_\beta(\delta_a;\mathfrak{H}) = \{E \in \mathfrak{H} : a \in E\}. \tag{B.29}$$

Thus the pattern family contains all patterns generated by pure actions. Allowing additional patterns can only decrease the integral and fractional cover numbers, which proves the first inequality in (B.27) after taking reciprocals, and the final inequality follows by using a hitting set as a family of Dirac distributions. The middle inequality is the standard relaxation inequality $\tau^{(2),\star}_\beta \le \tau^{(2)}_\beta$, together with Proposition B.5.

It remains to prove the zero-slack fractional equality. If $q(E) = 1$, then $\mathrm{supp}(q) \subseteq E$. Hence, for any $a \in \mathrm{supp}(q)$,

$$\mathrm{pat}_0(q;\mathfrak{H}) \subseteq \mathrm{pat}_0(\delta_a;\mathfrak{H}).$$

Every zero-slack pattern is therefore dominated by a Dirac pattern. Replacing a pattern by a superset cannot increase either an integral or fractional cover, so both second-order optima may be restricted to Dirac patterns. The resulting integral and fractional programs are exactly (B.3) and (B.4). The remaining identities in (B.28) follow from Proposition B.1 and Proposition B.4. □

Proposition B.7 (Second-order randomized rounding)**.** *Let* $m_\alpha = |\mathfrak{H}_\alpha(\mathcal{G})|$. *Then*

$$N_{\mathrm{dc}}(\mathcal{G};\alpha,\beta) \le \min\left\{K, \left\lceil \frac{\log m_\alpha + 1}{\gamma^{(2)}_{\alpha,\beta}(\mathcal{G})} \right\rceil\right\}. \tag{B.30}$$

Consequently,

$$N_{\mathrm{dc}}(\mathcal{G};\alpha,\beta) \le \big(\log m_\alpha + 2\big)\tau^{(2),\star}_\beta\left(\mathfrak{H}_\alpha(\mathcal{G})\right). \tag{B.31}$$

Proof. Let μ attain (B.23), and sample n patterns independently from μ. Each fixed edge E is absent from every sampled pattern with probability at most $(1-\gamma^{(2)}_{\alpha,\beta})^n \le \exp(-n\gamma^{(2)}_{\alpha,\beta})$. A union bound and the same choice of n as in Proposition B.3 produce a collection of patterns covering all edges. Choosing one realizing distribution for each sampled pattern gives a distribution cover. The K Dirac measures provide the other bound. Equation (B.31) follows from (B.25) and $\tau^{(2),\star}_\beta \ge 1$. □

Proposition B.8 (Monotonicity in approximation and cover slack)**.** *Let* $0 \le \alpha \le \alpha' \le 1$ *and* $0 \le \beta \le \beta' < 1$. *Then*

$$\gamma_\alpha(\mathcal{G}) \le \gamma_{\alpha'}(\mathcal{G})\,, \tag{B.32}$$

$$N_{\mathrm{hit}}(\mathcal{G};\alpha') \le N_{\mathrm{hit}}(\mathcal{G};\alpha), \tag{B.33}$$

$$N_{\mathrm{dc}}(\mathcal{G};\alpha',\beta') \le N_{\mathrm{dc}}(\mathcal{G};\alpha,\beta)\,, \tag{B.34}$$

$$\gamma^{(2)}_{\alpha,\beta}(\mathcal{G}) \le \gamma^{(2)}_{\alpha',\beta'}(\mathcal{G}). \tag{B.35}$$

In particular, $\gamma_1(\mathcal{G}) = 1$ *and* $N_{\mathrm{dc}}(\mathcal{G};1,\beta) = 1$ *for every* $\beta \in [0,1)$.

Proof. For every $g \in \mathcal{G}$, $\mathsf{A}_\alpha(g) \subseteq \mathsf{A}_{\alpha'}(g)$, while $1-\beta' \le 1-\beta$. Hence every distribution assigns at least as much mass to the enlarged near-optimal set, every α-hitting set is also an α'-hitting set, and every (α,β)-distribution cover is also an (α',β')-distribution cover. These observations prove (B.32)–(B.34). For (B.35), use the distributional form (B.24): for every fixed law M and reward vector g, the event

$$\{q : q(\mathsf{A}_\alpha(g)) \ge 1-\beta\}$$

is contained in its (α',β') analogue. Finally, $\mathsf{A}_1(g) = \mathcal{A}$ because rewards lie in $[0,1]$. □

B.4 A SINGLETON-FORCING CALCULATION

The next proposition is the exact optimizer-geometric calculation used in the quantitative separation developed in the next appendix. Its assumption is stable under adding arbitrary further hyperedges, which is crucial: convexifying a reward class may create new non-singleton near-optimal sets, but it cannot invalidate the conclusion once every singleton already occurs.

Proposition B.9 (Singleton forcing)**.** *Assume that every action occurs as a singleton edge:*

$$\big\{\{a\} : a \in \mathcal{A}\big\} \subseteq \mathfrak{H}_\alpha(\mathcal{G})\,. \tag{B.36}$$

For $\beta \in [0,1)$, define

$$m_\beta := \min\left\{K, \left\lfloor \frac{1}{1-\beta} \right\rfloor\right\}. \tag{B.37}$$

Then

$$\gamma_\alpha(\mathcal{G}) = \frac{1}{K}, \qquad \tau^\star(\mathfrak{H}_\alpha(\mathcal{G})) = K, \qquad \tau(\mathfrak{H}_\alpha(\mathcal{G})) = K, \tag{B.38}$$

$$\gamma^{(2)}_{\alpha,\beta}(\mathcal{G}) = \frac{m_\beta}{K}, \qquad \tau^{(2),\star}_\beta\left(\mathfrak{H}_\alpha(\mathcal{G})\right) = \frac{K}{m_\beta}, \qquad N_{\mathrm{dc}}(\mathcal{G};\alpha,\beta) = \left\lceil \frac{K}{m_\beta} \right\rceil. \tag{B.39}$$

In particular, if $\beta < 1/2$, then $m_\beta = 1$ and

$$\gamma^{(2)}_{\alpha,\beta}(\mathcal{G}) = \gamma_\alpha(\mathcal{G}) = \frac{1}{K}, \qquad N_{\mathrm{dc}}(\mathcal{G};\alpha,\beta) = K. \tag{B.40}$$

Proof. The singleton edges imply $\min_{E\in\mathfrak{H}} p(E) \le \min_a p(a) \le 1/K$ for every p, whereas the uniform distribution assigns at least $1/K$ to every nonempty edge. This proves the volume identity. A fractional transversal must satisfy $w(a) \ge 1$ for every singleton $\{a\}$, so its mass is at least K; the all-ones vector is feasible. Likewise, an integral transversal must contain every action. This proves (B.38).

Set $c := 1-\beta$. A distribution q covers the singleton $\{a\}$ exactly when $q(a) \ge c$. Since $\sum_a q(a) = 1$, the number of such actions is at most $\lfloor 1/c \rfloor$, and hence at most m_β. Therefore every distribution cover needs at least $\lceil K/m_\beta \rceil$ members.

For the matching upper bound, partition $\mathcal{A}$ into $\lceil K/m_\beta \rceil$ nonempty blocks, each of size at most m_β, and let q_B be uniform on block B. Since $m_\beta \le 1/(1-\beta)$, every $a \in B$ satisfies $q_B(a) = 1/|B| \ge 1/m_\beta \ge 1-\beta$. Thus these distributions cover every singleton. They also cover every nonempty edge E: choose $a \in E$ and use the block distribution that covers $\{a\}$, for which $q_B(E) \ge q_B(a) \ge 1-\beta$. This proves the integral identity in (B.39).

For the fractional volume, let M be any distribution over $q \in \Delta(\mathcal{A})$ and define

$$r_a := \mathbb{P}_{q\sim\mathsf{M}}\big(q(a) \ge 1-\beta\big).$$

Every q has at most m_β coordinates at least $1-\beta$, so

$$\sum_{a\in\mathcal{A}} r_a = \mathbb{E}_{q\sim\mathsf{M}}\,|\{a : q(a) \ge 1-\beta\}| \le m_\beta.$$

Hence some singleton edge has coverage probability at most m_β/K, proving $\gamma^{(2)}_{\alpha,\beta} \le m_\beta/K$.

For the reverse inequality, choose a subset $S \subseteq \mathcal{A}$ uniformly among all m_β-element subsets and let q_S be uniform on S. For every nonempty edge E, if $S \cap E \neq \varnothing$, then $q_S(E) \geq 1/m_\beta \geq 1-\beta$. Choosing any $a \in E$ gives

$$\mathbb{P}(S \cap E \neq \varnothing) \geq \mathbb{P}(a \in S) = \frac{m_\beta}{K}.$$

Thus every edge is covered with probability at least m_β/K, proving the matching lower bound on second-order volume. The fractional-cover identity follows from Proposition B.6. □

B.5 Expected-gap geometry and threshold conversion

Thresholded near-optimal sets discard the magnitude of suboptimality inside and outside the threshold. The following elementary conversion motivates the expected optimizer-witness bias used by the optimizer-witness reduction developed subsequently. For $g \in \mathcal{G}$, define the gap function and its expectation under q by

$$\Delta_g(a) := g^\star - g(a), \qquad \Delta_g(q) := q(\Delta_g) = g^\star - q(g). \tag{B.41}$$

Both lie in $[0,1]$.

Lemma B.10 (Sharp threshold-to-expectation inequalities)**.** *For every $g \in \mathcal{G}$, $q \in \Delta(\mathcal{A})$, and $\alpha \in [0,1]$,*

$$q(\mathsf{A}_\alpha(g)) \geq 1-\beta \quad \Longrightarrow \quad \Delta_g(q) \leq \alpha + (1-\alpha)\beta. \tag{B.42}$$

For every $\alpha > 0$,

$$q\big(\mathsf{A}_\alpha(g)^{\mathsf{c}}\big) \leq \frac{\Delta_g(q)}{\alpha}. \tag{B.43}$$

Both inequalities are best possible as functions of (α,β) and $\Delta_g(q)$, respectively.

Proof. Let $r := q(\mathsf{A}_\alpha(g)^{\mathsf{c}})$. On $\mathsf{A}_\alpha(g)$ the gap is at most α, while everywhere it is at most one. Therefore

$$\Delta_g(q) \leq \alpha(1-r) + r = \alpha + (1-\alpha)r,$$

which yields (B.42) when $r \leq \beta$. Moreover, $\mathsf{A}_\alpha(g)^{\mathsf{c}} = \{a : \Delta_g(a) > \alpha\}$, so Markov's inequality gives (B.43).

For sharpness of (B.42) when $\alpha < 1$, take a reward vector with available gaps 0, α, and 1, and let q assign masses $1-\beta$ and β to actions with the latter two gaps. When $\alpha = 1$, concentrating q on an action of gap one attains the bound. For sharpness of (B.43), place probability r on an action of gap $\alpha + \varepsilon$ and the remaining probability on an optimal action, and let $\varepsilon \downarrow 0$. □

For a finite nonempty $Q \subseteq \Delta(\mathcal{A})$, define its expected optimizer-witness bias by

$$b_{\mathcal{G}}(Q) := \sup_{g \in \mathcal{G}} \min_{q \in Q} \Delta_g(q) = \sup_{g \in \mathcal{G}} \left[g^\star - \max_{q \in Q} q(g) \right]. \tag{B.44}$$

Also define the minimum cardinality of an ε-witness family by

$$N_{\mathrm{wit}}(\mathcal{G};\varepsilon) := \min\left\{ |Q| : \varnothing \neq Q \subseteq \Delta(\mathcal{A}) \text{ finite}, \ b_{\mathcal{G}}(Q) \leq \varepsilon \right\}. \tag{B.45}$$

The minimum is well defined because the K Dirac measures have zero witness bias.

Corollary B.11 (Distribution covers versus expected-gap witnesses)**.** *For $\alpha \in (0,1]$ and $\beta \in [0,1)$,*

$$N_{\mathrm{wit}}\big(\mathcal{G};\alpha + (1-\alpha)\beta\big) \leq N_{\mathrm{dc}}(\mathcal{G};\alpha,\beta) \leq N_{\mathrm{wit}}(\mathcal{G};\alpha\beta). \tag{B.46}$$

Equivalently, every (α,β)-distribution cover Q satisfies

$$b_{\mathcal{G}}(Q) \leq \alpha + (1-\alpha)\beta, \tag{B.47}$$

whereas every finite Q satisfying $b_{\mathcal{G}}(Q) \leq \alpha\beta$ is an (α,β)-distribution cover.

Proof. For a distribution cover, apply (B.42) to the witnessing q for each g, then take the supremum over g. Conversely, for every g, choose $q \in Q$ minimizing $\Delta_g(q)$ and apply (B.43):

$$q\big(\mathsf{A}_\alpha(g)^{\mathsf{c}}\big) \leq \frac{\Delta_g(q)}{\alpha} \leq \beta.$$

Taking minimum cardinalities yields (B.46). □

The two tolerances in (B.46) are deliberately different. A threshold cover controls only whether the gap exceeds α, while expected-gap witnesses retain the magnitude of all gaps. The optimizer-witness appendix will show that the latter quantity enters an exact pathwise comparator reduction.

B.6 Optimizer-equivalence and the limit of set-system complexity

Definition B.12 (Optimizer-equivalence at scale α)**.** Two reward classes $\mathcal{G}_1, \mathcal{G}_2 \subseteq [0,1]^{\mathcal{A}}$ are *optimizer-equivalent at scale* α, written $\mathcal{G}_1 \equiv_\alpha \mathcal{G}_2$, if

$$\mathfrak{H}_\alpha(\mathcal{G}_1) = \mathfrak{H}_\alpha(\mathcal{G}_2)\,.$$

Proposition B.13 (Set-system invariance)**.** *If $\mathcal{G}_1 \equiv_\alpha \mathcal{G}_2$, then for every $\beta \in [0,1)$ the following quantities are identical for the two classes:*

$$\begin{gathered} \gamma_\alpha(\mathcal{G}_i)\,, \quad \tau^\star(\mathfrak{H}_\alpha(\mathcal{G}_i))\,, \quad \tau(\mathfrak{H}_\alpha(\mathcal{G}_i))\,, \\ \gamma^{(2)}_{\alpha,\beta}(\mathcal{G}_i), \quad \tau^{(2),\star}_\beta\left(\mathfrak{H}_\alpha(\mathcal{G}_i)\right), \quad N_{\mathrm{dc}}(\mathcal{G}_i;\alpha,\beta)\,, \qquad i \in \{1,2\}. \end{gathered} \tag{B.48}$$

Proof. Every quantity in (B.48) is defined solely from the common hypergraph and its realizable β-coverage patterns. For the latter, (B.17) depends on a reward class only through its family of hyperedges. Hence the pattern families, all associated integral and fractional covering programs, and their maximin values coincide. □

Proposition B.13 is the precise sense in which the existing optimizer-cover quantities are blind to observation geometry. They retain which actions can be approximately optimal, either individually or under a covering distribution, but they do not retain the linear or nonlinear relations among reward coordinates that determine what can be inferred from a one-point observation. The quantitative separation developed next exploits exactly this missing axis.

C A Quantitative Separation for Fixed-Scale Optimizer Geometry

This appendix proves the quantitative separation announced in the main text. The construction isolates a basic obstruction to any finite-time theory based only on maximin volume or distribution-cover cardinality: two reward classes can have the same values of all first- and second-order threshold-cover quantities from Appendix B at a nontrivial range of scales, while their oblivious minimax regrets differ exponentially in the ambient dimension. The structured upper bound is a noise-robust variant of the finite-action Exp2 argument with optimal experimental design (Bubeck et al., 2012); the tabular lower bound is proved from first principles by a Bernoulli change-of-measure argument, in the spirit of the classical adversarial multi-armed bandit lower bound (Auer et al., 2002).

We state the separation for oblivious mean-reward sequences. This is sufficient to show that the fixed-scale optimizer-cover scalars of Hanneke & Wang (2026) are insufficient to reconstruct finite-time regret even in the statistically simpler adversary model. The proof does not assert an adaptive-regret upper bound for the structured class; see Remark C.13.

C.1 A constant-rate separated code

For $x, v \in \{-1,+1\}^d$, let

$$d_{\mathrm{H}}(x,v) := \left|\{j \in [d] : x_j \neq v_j\}\right|$$

denote Hamming distance. The identity

$$\langle x, v\rangle = d - 2d_{\mathrm{H}}(x,v) \tag{C.1}$$

will be used repeatedly.

Lemma C.1 (Explicit constant-rate code)**.** *For every integer $d \geq 45$, there exists a set $\mathcal{C}_d \subseteq \{-1,+1\}^d$ such that*

$$K_d := |\mathcal{C}_d| \geq \exp(d/64), \qquad \min_{\substack{x,v\in\mathcal{C}_d \\ x\neq v}} d_{\mathrm{H}}(x,v) \geq d/4. \tag{C.2}$$

Proof. Let

$$M := \left\lfloor e^{d/32} \right\rfloor$$

and sample $X_1, \dots, X_M$ independently and uniformly from $\{-1, +1\}^d$. For each $i \neq j$, $d_{\mathrm{H}}(X_i, X_j)$ has the $\mathrm{Binomial}(d, 1/2)$ law. Hoeffding's inequality gives

$$\mathbb{P}(d_{\mathrm{H}}(X_i, X_j) < d/4) \leq e^{-d/8}.$$

Therefore, by a union bound,

$$\mathbb{P}\left(\min_{i \neq j} d_{\mathrm{H}}(X_i, X_j) < d/4 \right) \leq \binom{M}{2} e^{-d/8} \leq \frac{1}{2} e^{d/16} e^{-d/8} = \frac{1}{2} e^{-d/16} < 1.$$

Hence there is a realization with pairwise Hamming distance at least $d/4$. For $d \geq 45$, $e^{d/64} \geq 2$ and consequently $M \geq e^{d/32}/2 \geq e^{d/64}$. Taking the realized points as $\mathcal{C}_d$ proves the claim. Notice that distance zero is among the excluded bad events, so the selected points are distinct. □

Fix henceforth an integer $d \geq 45$ and a code $\mathcal{C}_d$ satisfying Lemma C.1. Set

$$\mathcal{A} := \mathcal{C}_d, \qquad K := K_d = |\mathcal{A}|, \qquad S := \mathrm{span}(\mathcal{A}) \subseteq \mathbb{R}^d, \qquad r := \dim(S) \leq d. \tag{C.3}$$

For $v \in \mathcal{C}_d$, define the reward vector $g_v \in [0,1]^{\mathcal{A}}$ by

$$g_v(x) := \frac{1}{2} + \frac{\langle v, x \rangle}{4d}, \qquad x \in \mathcal{C}_d. \tag{C.4}$$

We compare the convex structured class

$$\mathcal{G}_{\mathrm{code}} := \mathrm{co}\{g_v : v \in \mathcal{C}_d\} \tag{C.5}$$

with the tabular cube

$$\mathcal{G}_{\mathrm{tab}} := [1/4, 3/4]^{\mathcal{C}_d}. \tag{C.6}$$

Both are nonempty compact convex subsets of $[0,1]^{\mathcal{A}}$.

Lemma C.2 (Range, linear representation, and exposed optimizers)**.** *The following statements hold.*

1. *For every $v, x \in \mathcal{C}_d$,*
$$g_v(x) \in [1/4, 3/4]. \tag{C.7}$$
2. *Every $g \in \mathcal{G}_{\mathrm{code}}$ admits a representation*
$$g(x) = \frac{1}{2} + \langle u, x \rangle, \qquad u \in \frac{1}{4d} \mathrm{co}(\mathcal{C}_d) \subseteq S. \tag{C.8}$$
3. *For every $v \in \mathcal{C}_d$ and every $x \neq v$,*
$$g_v(v) - g_v(x) = \frac{d_{\mathrm{H}}(v, x)}{2d} \geq \frac{1}{8}. \tag{C.9}$$
Consequently,
$$\mathsf{A}_\alpha(g_v) = \{v\} \qquad \textit{for every } \alpha \in [0, 1/8). \tag{C.10}$$

Proof. Since $|\langle v, x \rangle| \leq d$, equation (C.4) implies (C.7). If $g = \sum_{i=1}^m \lambda_i g_{v_i}$, then

$$g(x) = \frac{1}{2} + \left\langle \frac{1}{4d} \sum_{i=1}^m \lambda_i v_i, x \right\rangle,$$

which proves (C.8). Finally, (C.1) gives

$$g_v(v) - g_v(x) = \frac{\langle v, v \rangle - \langle v, x \rangle}{4d} = \frac{2 d_{\mathrm{H}}(v, x)}{4d}.$$

Apply (C.2) to obtain (C.9) and (C.10). □

C.2 IDENTICAL FIXED-SCALE OPTIMIZER-COVER QUANTITIES

For each $v \in \mathcal{C}_d$, let $h_v \in \mathcal{G}_{\mathrm{tab}}$ be the spike reward vector

$$h_v(x) := \begin{cases} 3/4, & x = v, \\ 1/4, & x \neq v. \end{cases} \tag{C.11}$$

Then $\mathsf{A}_\alpha(h_v) = \{v\}$ for every $\alpha \in [0, 1/2)$. Together with (C.10), this allows us to invoke the singleton-forcing calculation from Proposition B.9.

Proposition C.3 (Coincidence of all threshold-cover scalars)**.** *For every $\alpha \in [0, 1/8)$ and $\beta \in [0, 1/2)$,*

$$\gamma_\alpha(\mathcal{G}_{\mathrm{code}}) = \gamma_\alpha(\mathcal{G}_{\mathrm{tab}}) = \frac{1}{K}, \tag{C.12}$$

$$N_{\mathrm{dc}}(\mathcal{G}_{\mathrm{code}}; \alpha, \beta) = N_{\mathrm{dc}}(\mathcal{G}_{\mathrm{tab}}; \alpha, \beta) = K. \tag{C.13}$$

More strongly, for either $\mathcal{G} \in \{\mathcal{G}_{\mathrm{code}}, \mathcal{G}_{\mathrm{tab}}\}$,

$$\begin{aligned} \tau^\star(\mathfrak{H}_\alpha(\mathcal{G})) = \tau(\mathfrak{H}_\alpha(\mathcal{G})) = K, \\ \gamma^{(2)}_{\alpha,\beta}(\mathcal{G}) = \frac{1}{K}, \qquad \tau^{(2),\star}_\beta(\mathfrak{H}_\alpha(\mathcal{G})) = K. \end{aligned} \tag{C.14}$$

Proof. By Lemma C.2, every singleton $\{v\}$ occurs in $\mathfrak{H}_\alpha(\mathcal{G}_{\mathrm{code}})$ when $\alpha < 1/8$. Equation (C.11) gives the same property for $\mathcal{G}_{\mathrm{tab}}$. Proposition B.9 therefore applies to both classes. Since $\beta < 1/2$, the integer m_β from (B.37) equals one, and (C.12)–(C.14) follow from (B.38)–(B.40). □

Remark C.4 (What Proposition C.3 does not say). The two near-optimal-set hypergraphs need not coincide, and the proposition makes no claim for $\alpha \geq 1/8$. In particular, this appendix proves fixed-scale insufficiency of the displayed scalar quantities, not equality of the complete profiles $\alpha \mapsto \gamma_\alpha(\mathcal{G})$ or $(\alpha, \beta) \mapsto N_{\mathrm{dc}}(\mathcal{G}; \alpha, \beta)$.

C.3 A NOISE-ROBUST FINITE-ACTION LINEAR UPPER BOUND

We next prove the structured regret upper bound under the bounded conditional-mean observation model of Appendix A. The argument is a self-contained adaptation of EXP2 with an optimal exploration design (Bubeck et al., 2012). The only place where obliviousness is used is that the best comparator below is fixed independently of the observation noise.

For a finite set $\mathcal{X} \subseteq \mathbb{R}^d$, let $S_\mathcal{X} = \mathrm{span}(\mathcal{X})$ and $r_\mathcal{X} = \dim(S_\mathcal{X})$. For $p \in \Delta(\mathcal{X})$ and $\varphi : \mathcal{X} \to \mathbb{R}$, write $p(\varphi) := \sum_{x \in \mathcal{X}} p(x)\varphi(x)$. Define the second-moment operator on $S_\mathcal{X}$ by

$$M(p) := \sum_{x \in \mathcal{X}} p(x) x x^\top. \tag{C.15}$$

All inverses below are understood as inverses on $S_\mathcal{X}$, or equivalently as Moore–Penrose pseudoinverses on $\mathbb{R}^d$.

Lemma C.5 (A G-optimal exploration design)**.** *For every finite nonempty $\mathcal{X} \subseteq \mathbb{R}^d$, there exists $\mu \in \Delta(\mathcal{X})$ such that $M(\mu)$ is positive definite on $S_\mathcal{X}$ and*

$$\max_{x \in \mathcal{X}} x^\top M(\mu)^\dagger x \leq r_\mathcal{X}. \tag{C.16}$$

Proof. Choose an orthonormal basis of $S_\mathcal{X}$ and identify that subspace with $\mathbb{R}^r$, where $r = r_\mathcal{X}$. Since $\mathcal{X}$ spans $\mathbb{R}^r$, some distribution has a positive-definite second-moment matrix. The continuous map $p \mapsto \det M(p)$ attains its maximum on the simplex; choose a maximizer μ with $\det M(\mu) > 0$. For any $x \in \mathcal{X}$ and $\varepsilon \in [0, 1]$, set $p_\varepsilon = (1 - \varepsilon)\mu + \varepsilon\delta_x$. Optimality of μ and differentiability of $\log\det$ at $M(\mu)$ imply

$$0 \geq \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \log\det M(p_\varepsilon)\Big|_{\varepsilon = 0^+} = \mathrm{tr}\big(M(\mu)^{-1}\big(xx^\top - M(\mu)\big)\big) = x^\top M(\mu)^{-1} x - r.$$

This proves (C.16) in the chosen coordinates, and hence the coordinate-free pseudoinverse statement. □

Theorem C.6 (Noisy EXP2 for an oblivious linear reward sequence)**.** *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be finite with $K = |\mathcal{X}| \geq 2$ and $r = r_{\mathcal{X}} \geq 1$. Let $u_1, \dots, u_T \in S_{\mathcal{X}}$ be fixed before interaction and satisfy*

$$|\langle u_t, x\rangle| \leq \frac{1}{4} \qquad \text{for every } t \in [T],\ x \in \mathcal{X}. \tag{C.17}$$

For each t, write $s_t(x) := \langle u_t, x\rangle$. Suppose that after selecting $X_t \in \mathcal{X}$ the learner observes $Z_t \in [-1/2, 1/2]$ satisfying

$$\mathbb{E}[Z_t \mid \mathscr{H}_{t-1}, X_t] = \langle u_t, X_t\rangle. \tag{C.18}$$

The conditional law of Z_t may otherwise depend arbitrarily on the public history. Then there exists a behavioral algorithm such that

$$\mathbb{E}\left[\max_{x\in\mathcal{X}} \sum_{t=1}^{T} \langle u_t, x\rangle - \sum_{t=1}^{T} \langle u_t, X_t\rangle\right] \leq \min\left\{\frac{T}{2}, 2\sqrt{rT\log K}\right\}. \tag{C.19}$$

Proof. The trivial strategy has regret at most $T/2$ because (C.17) implies that the one-round reward range is at most $1/2$. It therefore suffices to construct the second guarantee when

$$T > 16r\log K. \tag{C.20}$$

Let μ be the design from Lemma C.5, and set

$$\eta := \sqrt{\frac{\log K}{rT}}, \qquad \gamma := \frac{\eta r}{2}. \tag{C.21}$$

Condition (C.20) gives $\gamma < 1/8$.

Initialize $w_1(x) = 1$ for all $x \in \mathcal{X}$. At round t, let

$$q_t(x) := \frac{w_t(x)}{\sum_{z\in\mathcal{X}} w_t(z)}, \qquad p_t := (1-\gamma)q_t + \gamma\mu, \qquad M_t := M(p_t). \tag{C.22}$$

Sample $X_t \sim p_t$, observe Z_t, and define

$$\widehat{u}_t := M_t^{\dagger} X_t Z_t, \qquad \widehat{s}_t(x) := \langle \widehat{u}_t, x\rangle. \tag{C.23}$$

Update

$$w_{t+1}(x) := w_t(x)\exp(\eta\widehat{s}_t(x)). \tag{C.24}$$

We verify the three ingredients of the exponential-weights analysis. First, M_t is positive definite on $S_{\mathcal{X}}$ and, using (C.18),

$$\begin{aligned}\mathbb{E}[\widehat{u}_t \mid \mathscr{H}_{t-1}] &= M_t^{\dagger}\sum_{x\in\mathcal{X}} p_t(x)x\,\mathbb{E}[Z_t \mid \mathscr{H}_{t-1}, X_t = x] \\ &= M_t^{\dagger}M_t u_t = u_t.\end{aligned} \tag{C.25}$$

The final equality holds because $u_t \in S_{\mathcal{X}}$.

Second, $M_t \succeq \gamma M(\mu)$ on $S_{\mathcal{X}}$, so (C.16) gives

$$x^{\top}M_t^{\dagger}x \leq \frac{r}{\gamma} \qquad \text{for every } x \in \mathcal{X}. \tag{C.26}$$

By Cauchy–Schwarz in the inner product induced by $M_t^{\dagger}$,

$$|\widehat{s}_t(x)| \leq |Z_t|\sqrt{x^{\top}M_t^{\dagger}x}\sqrt{X_t^{\top}M_t^{\dagger}X_t} \leq \frac{r}{2\gamma}.$$

Thus (C.21) yields

$$|\eta\widehat{s}_t(x)| \leq 1 \qquad \text{for every } t, x. \tag{C.27}$$

Third, let $M(q_t) = \sum_x q_t(x)xx^\top$. From $M_t = (1-\gamma)M(q_t) + \gamma M(\mu)$, we have $M(q_t) \preceq M_t/(1-\gamma)$. Therefore, conditionally on $\mathscr{H}_{t-1}$,

$$\begin{aligned}\mathbb{E}\left[\sum_{x\in\mathcal{X}} q_t(x)\widehat{s}_t(x)^2 \middle| \mathscr{H}_{t-1}\right] &= \mathbb{E}\Big[Z_t^2 X_t^\top M_t^\dagger M(q_t) M_t^\dagger X_t \Big| \mathscr{H}_{t-1}\Big] \\ &\le \frac{1}{4(1-\gamma)}\mathbb{E}\Big[X_t^\top M_t^\dagger X_t \Big| \mathscr{H}_{t-1}\Big] \\ &= \frac{1}{4(1-\gamma)}\operatorname{tr}_{S_{\mathcal{X}}}(M_t^\dagger M_t) = \frac{r}{4(1-\gamma)}. \end{aligned} \tag{C.28}$$

For completeness, we now derive the deterministic potential inequality. Since $e^z \le 1 + z + z^2$ for $|z| \le 1$, equations (C.24) and (C.27) imply

$$\log \frac{\sum_x w_{t+1}(x)}{\sum_x w_t(x)} \le \eta \sum_x q_t(x)\widehat{s}_t(x) + \eta^2 \sum_x q_t(x)\widehat{s}_t(x)^2.$$

On the other hand, for every fixed $x^\star \in \mathcal{X}$,

$$\log \frac{\sum_x w_{T+1}(x)}{\sum_x w_1(x)} \ge \eta \sum_{t=1}^T \widehat{s}_t(x^\star) - \log K.$$

Combining and rearranging gives, pathwise,

$$\sum_{t=1}^T (\widehat{s}_t(x^\star) - q_t(\widehat{s}_t)) \le \frac{\log K}{\eta} + \eta \sum_{t=1}^T q_t(\widehat{s}_t^2). \tag{C.29}$$

Because $(u_t)_{t=1}^T$ is fixed before interaction, one may fix $x^\star \in \arg\max_x \sum_t \langle u_t, x\rangle$ before taking expectations. Using (C.25), (C.28), and (C.29),

$$\mathbb{E}\sum_{t=1}^T (\langle u_t, x^\star\rangle - q_t(s_t)) \le \frac{\log K}{\eta} + \frac{\eta r T}{4(1-\gamma)}. \tag{C.30}$$

Finally, since $p_t = (1-\gamma)q_t + \gamma\mu$ and all centered rewards lie in $[-1/4, 1/4]$,

$$q_t(s_t) - p_t(s_t) = \gamma\big(q_t(s_t) - \mu(s_t)\big) \le \frac{\gamma}{2}. \tag{C.31}$$

Combining (C.30) and (C.31), and using $\gamma < 1/8$, gives

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^T \langle u_t, x^\star\rangle - \sum_{t=1}^T \langle u_t, X_t\rangle\right] &\le \frac{\log K}{\eta} + \frac{\eta r T}{4(1-\gamma)} + \frac{\gamma T}{2} \\ &\le \sqrt{rT\log K} + \frac{2}{7}\sqrt{rT\log K} + \frac{1}{4}\sqrt{rT\log K} \\ &< 2\sqrt{rT\log K}.\end{aligned}$$

This proves the second term in (C.19); combining it with the trivial strategy proves the theorem for every T. □

Corollary C.7 (Structured-code upper bound)**.** *For every horizon $T \ge 1$,*

$$\mathfrak{R}_T^{\star,\mathrm{obl}}(\mathcal{G}_{\mathrm{code}}) \le \min\left\{\frac{T}{2}, 2\sqrt{rT\log K}\right\} \le \min\left\{\frac{T}{2}, 2\sqrt{dT\log K}\right\}. \tag{C.32}$$

The bound holds uniformly over all history-dependent observation kernels satisfying the bounded conditional-mean requirement of Definition A.5.

Proof. Fix an arbitrary oblivious sequence $g_1, \dots, g_T \in \mathcal{G}_{\mathrm{code}}$. By (C.8), there are fixed vectors $u_t \in (4d)^{-1}\operatorname{co}(\mathcal{C}_d) \subseteq S$ such that $g_t(x) = 1/2 + \langle u_t, x\rangle$. Set $Z_t = Y_t - 1/2$. Then $Z_t \in [-1/2, 1/2]$ and

$$\mathbb{E}[Z_t \mid \mathscr{H}_{t-1}, A_t] = \langle u_t, A_t\rangle.$$

Moreover, (C.7) and convexity imply $|\langle u_t, x\rangle| \le 1/4$. The additive constant $1/2$ cancels from regret, so Theorem C.6 applies. Taking the supremum over all oblivious environments and using $r \le d$ proves the result. □

C.4 A SELF-CONTAINED TABULAR BERNOULLI LOWER BOUND

We now prove the matching unstructured obstruction. The argument is stated for the oblivious Bernoulli subclass from Definition A.6; hence it also lower-bounds the full bounded-noise minimax value by Proposition A.8.

Fix an enumeration $\mathcal{C}_d = \{x_1, \dots, x_K\}$. For an algorithm Alg, let P_0 denote the transcript law when every arm has Bernoulli mean $1/2$. For $j \in [K]$ and $\varepsilon \in (0, 1/4]$, let P_j denote the transcript law when arm x_j has mean $1/2 + \varepsilon$ and every other arm has mean $1/2$, at every round. Let

$$N_j := \sum_{t=1}^{T} \mathbf{1}\{A_t = x_j\}. \tag{C.33}$$

All of these are legal oblivious Bernoulli environments with reward vectors in $\mathcal{G}_{\mathrm{tab}}$ and means in $[1/4, 3/4]$.

Lemma C.8 (Transcript divergence for a one-arm perturbation)**.** *For every behavioral algorithm and every $j \in [K]$,*

$$D_{\mathrm{KL}}(P_0 \| P_j) = \mathbb{E}_0[N_j]\, \mathrm{kl}\left(\frac{1}{2} \,\middle\|\, \frac{1}{2} + \varepsilon\right), \tag{C.34}$$

where kl *denotes binary relative entropy. Moreover, for $\varepsilon \in (0, 1/4]$,*

$$\mathrm{kl}\left(\frac{1}{2} \,\middle\|\, \frac{1}{2} + \varepsilon\right) = -\frac{1}{2}\log(1 - 4\varepsilon^2) \le 4\varepsilon^2. \tag{C.35}$$

Proof. The learner uses the same history-indexed action kernel under P_0 and P_j. Consequently the likelihood ratio of the complete transcript contains only the observation-kernel ratios. By the chain rule for relative entropy,

$$\begin{aligned} D_{\mathrm{KL}}(P_0 \| P_j) &= \sum_{t=1}^{T} \mathbb{E}_0[D_{\mathrm{KL}}(P_0(Y_t \in \cdot \mid \mathscr{H}_{t-1}, A_t) \| P_j(Y_t \in \cdot \mid \mathscr{H}_{t-1}, A_t))] \\ &= \sum_{t=1}^{T} \mathbb{E}_0[\mathbf{1}\{A_t = x_j\}]\, \mathrm{kl}\left(\frac{1}{2} \,\middle\|\, \frac{1}{2} + \varepsilon\right), \end{aligned}$$

which is (C.34). Direct calculation gives

$$\mathrm{kl}\left(\frac{1}{2} \,\middle\|\, \frac{1}{2} + \varepsilon\right) = \frac{1}{2}\log\frac{1/2}{1/2 + \varepsilon} + \frac{1}{2}\log\frac{1/2}{1/2 - \varepsilon} = -\frac{1}{2}\log(1 - 4\varepsilon^2).$$

For $x \in [0, 1/4]$, $-\log(1 - x) \le x/(1 - x) \le (4/3)x$. Substituting $x = 4\varepsilon^2$ yields the stronger bound $(8/3)\varepsilon^2$, and hence (C.35). □

Lemma C.9 (Average number of pulls of the favorable arm)**.** *For every behavioral algorithm and every $\varepsilon \in (0, 1/4]$,*

$$\frac{1}{K}\sum_{j=1}^{K} \mathbb{E}_j[N_j] \le \frac{T}{K} + \sqrt{2}\,\varepsilon T\sqrt{\frac{T}{K}}. \tag{C.36}$$

Proof. Since $0 \le N_j \le T$, the total-variation inequality and Pinsker's inequality give

$$\begin{aligned} \mathbb{E}_j[N_j] &\le \mathbb{E}_0[N_j] + T\,\|P_j - P_0\|_{\mathrm{TV}} \\ &\le \mathbb{E}_0[N_j] + T\sqrt{\frac{1}{2}D_{\mathrm{KL}}(P_0 \| P_j)} \\ &\le \mathbb{E}_0[N_j] + \sqrt{2}\,\varepsilon T\sqrt{\mathbb{E}_0[N_j]}, \end{aligned}$$

where the last step uses Lemma C.8. Average over j and apply Cauchy–Schwarz together with $\sum_j \mathbb{E}_0[N_j] = T$:

$$\frac{1}{K}\sum_{j=1}^{K}\sqrt{\mathbb{E}_0[N_j]} \le \sqrt{\frac{1}{K}\sum_{j=1}^{K}\mathbb{E}_0[N_j]} = \sqrt{\frac{T}{K}}.$$

This proves (C.36). □

Theorem C.10 (Tabular Bernoulli minimax lower bound)**.** *For every $K \geq 2$ and $T \geq 1$,*

$$\mathfrak{R}^{\star,\mathrm{obl},\mathrm{Ber}}_{T,1/4}(\mathcal{G}_{\mathrm{tab}}) \geq \frac{1}{16\sqrt{2}} \min\left\{T, \sqrt{KT}\right\}. \tag{C.37}$$

Consequently,

$$\mathfrak{R}^{\star,\mathrm{obl}}_{T}(\mathcal{G}_{\mathrm{tab}}) \geq \frac{1}{16\sqrt{2}} \min\left\{T, \sqrt{KT}\right\}. \tag{C.38}$$

Proof. Fix an arbitrary behavioral algorithm and set

$$\varepsilon := \min\left\{\frac{1}{4}, \frac{1}{4\sqrt{2}}\sqrt{\frac{K}{T}}\right\}. \tag{C.39}$$

Under P_j, arm x_j is uniquely optimal and the expected regret is

$$\mathbb{E}_j[\mathrm{Reg}_T] = \varepsilon\big(T - \mathbb{E}_j[N_j]\big). \tag{C.40}$$

Averaging (C.40) over a uniform j and invoking Lemma C.9 gives

$$\begin{aligned}\frac{1}{K}\sum_{j=1}^{K}\mathbb{E}_j[\mathrm{Reg}_T] &\geq \varepsilon T\left(1 - \frac{1}{K} - \sqrt{2}\,\varepsilon\sqrt{\frac{T}{K}}\right)\\ &\geq \frac{\varepsilon T}{4}.\end{aligned} \tag{C.41}$$

Indeed, $K \geq 2$ gives $1 - 1/K \geq 1/2$, and (C.39) gives $\sqrt{2}\varepsilon\sqrt{T/K} \leq 1/4$. Moreover,

$$\frac{\varepsilon T}{4} = \min\left\{\frac{T}{16}, \frac{\sqrt{KT}}{16\sqrt{2}}\right\} \geq \frac{1}{16\sqrt{2}}\min\left\{T, \sqrt{KT}\right\}.$$

The uniform prior over $j \in [K]$ is supported on legal oblivious Bernoulli environments. Lemma A.10, followed by the infimum over algorithms, proves (C.37). Finally, (C.38) follows from the inclusion of this Bernoulli subclass in the full oblivious bounded-noise environment class; see (A.21). □

C.5 The quantitative separation theorem

Theorem C.11 (Restatement of Theorem 2.1)**.** *For every integer $d \geq 45$, there are a finite action set $\mathcal{A} \subseteq \{-1, +1\}^d$ of cardinality $K \geq e^{d/64}$ and two compact convex reward classes $\mathcal{G}_{\mathrm{code}}, \mathcal{G}_{\mathrm{tab}} \subseteq [1/4, 3/4]^{\mathcal{A}}$ such that the following hold.*

1. *For every $\alpha \in [0, 1/8)$ and $\beta \in [0, 1/2)$,*

$$\gamma_\alpha(\mathcal{G}_{\mathrm{code}}) = \gamma_\alpha(\mathcal{G}_{\mathrm{tab}}) = \frac{1}{K}, \qquad N_{\mathrm{dc}}(\mathcal{G}_{\mathrm{code}}; \alpha, \beta) = N_{\mathrm{dc}}(\mathcal{G}_{\mathrm{tab}}; \alpha, \beta) = K. \tag{C.42}$$

   The same equality holds for the first- and second-order fractional and integral quantities listed in (C.14)*.*

2. *For every horizon $T \geq 1$,*

$$\mathfrak{R}^{\star,\mathrm{obl}}_{T}(\mathcal{G}_{\mathrm{code}}) \leq \min\left\{\frac{T}{2}, 2\sqrt{dT\log K}\right\}, \tag{C.43}$$

$$\mathfrak{R}^{\star,\mathrm{obl}}_{T}(\mathcal{G}_{\mathrm{tab}}) \geq \frac{1}{16\sqrt{2}}\min\left\{T, \sqrt{KT}\right\}. \tag{C.44}$$

In particular, at $T = K$,

$$\mathfrak{R}^{\star,\mathrm{obl}}_{K}(\mathcal{G}_{\mathrm{tab}}) \geq \frac{e^{d/128}}{32\sqrt{2\log 2}\,d}\mathfrak{R}^{\star,\mathrm{obl}}_{K}(\mathcal{G}_{\mathrm{code}}), \tag{C.45}$$

so the two values are separated by an exponential factor in d, up to a polynomial term.

Proof. Take the code from Lemma C.1 and the two classes from (C.5)–(C.6). Proposition C.3 proves (C.42) and its strengthened version. Corollary C.7 proves (C.43), and Theorem C.10 proves (C.44).

For $T = K$, equations (C.43) and (C.44) imply

$$\mathfrak{R}_K^{\star,\mathrm{obl}}(\mathcal{G}_{\mathrm{tab}}) \geq \frac{\sqrt{K}}{32\sqrt{2d\log K}}\mathfrak{R}_K^{\star,\mathrm{obl}}(\mathcal{G}_{\mathrm{code}}).$$

Indeed, the right-hand side is at most $\frac{\sqrt{K}}{32\sqrt{2d\log K}}\, 2\sqrt{dK\log K} = K/(16\sqrt{2})$, which is the tabular lower bound. Since $K \geq e^{d/64}$ and $K \leq 2^d$, we have $\sqrt{K} \geq e^{d/128}$ and $\log K \leq d\log 2$. Substitution yields (C.45). □

Corollary C.12 (No polylogarithmic-factor reconstruction from fixed-scale scalars)**.** *Fix any $\alpha \in [0, 1/8)$ and $\beta \in [0, 1/2)$. For every polynomial $P : \mathbb{R}_+ \to [1, \infty)$, there is no family of maps $\{\Psi_{T,K}\}$, depending on a K-action reward class only through*

$$\left(\gamma_\alpha(\mathcal{G})\,, N_{\mathrm{dc}}(\mathcal{G}; \alpha, \beta)\right),$$

that satisfies, for every finite-action compact convex class $\mathcal{G}$,

$$\frac{1}{P(\log K)}\mathfrak{R}_T^{\star,\mathrm{obl}}(\mathcal{G}) \leq \Psi_{T,K}(\mathcal{G}) \leq P(\log K)\mathfrak{R}_T^{\star,\mathrm{obl}}(\mathcal{G}). \tag{C.46}$$

The same impossibility holds when the input is augmented by all scalar quantities in (C.14).

Proof. For the classes in Theorem 2.1, $K \geq e^{d/64}$ implies $d \leq 64\log K$. Before the final substitution in the proof of that theorem we established the stronger inequality

$$\mathfrak{R}_K^{\star,\mathrm{obl}}(\mathcal{G}_{\mathrm{tab}}) \geq \frac{\sqrt{K}}{32\sqrt{2d\log K}}\mathfrak{R}_K^{\star,\mathrm{obl}}(\mathcal{G}_{\mathrm{code}}) \geq \frac{\sqrt{K}}{256\sqrt{2}\,\log K}\mathfrak{R}_K^{\star,\mathrm{obl}}(\mathcal{G}_{\mathrm{code}}).$$

The last factor dominates $P(\log K)^2$ for all sufficiently large K. The two classes have identical input scalars, so $\Psi_{K,K}$ takes the same value on both. Applying (C.46) to the two classes would imply

$$\mathfrak{R}_K^{\star,\mathrm{obl}}(\mathcal{G}_{\mathrm{tab}}) \leq P(\log K)^2\mathfrak{R}_K^{\star,\mathrm{obl}}(\mathcal{G}_{\mathrm{code}}),$$

a contradiction. The strengthened claim follows because the additional scalars also coincide by (C.14). □

Remark C.13 (Scope of the separation). The tabular lower bound is proved using oblivious environments and therefore also lower-bounds the corresponding adaptive minimax value. The structured upper bound in this appendix is deliberately stated for oblivious mean-reward sequences. The reason is conceptual rather than cosmetic: with adaptive rewards, the best fixed action in hindsight is itself a random function of the transcript, and an expected unbiased-estimator analysis against each deterministic comparator does not by itself control the expectation of this random maximum. The later chained upper theorem is formulated directly for the adaptive game and must address this issue through its own uniform martingale argument. No adaptive structured-rate claim is inferred from Theorem C.6.

Remark C.14 (Interpretation). For both classes, every action can be forced to be the unique optimizer at the scales in Theorem 2.1; threshold-based optimizer geometry therefore treats the K actions as maximally distinct. In $\mathcal{G}_{\mathrm{tab}}$, this is statistically accurate: observing one arm reveals nothing about any other arm. In $\mathcal{G}_{\mathrm{code}}$, however, every observation is a noisy linear measurement in an at most d-dimensional space. The exponential gap in (C.45) is thus an information-sharing gap, not an optimizer-covering gap. This is the precise reason the remainder of the paper equips optimizer witnesses with an observation-induced design geometry.

D Optimizer Witnesses and the Mediator Reduction

This appendix replaces thresholded optimizer coverage by an exact expected-gap notion and proves a variational reduction of the original bandit game to a family of structured mediator games. Distribution covers were introduced by Hanneke & Wang (2026) as a sufficient combinatorial device for adversarial

noisy-bandit learnability. Mediator feedback, in turn, models decisions as probability distributions over a common outcome space (Eldowa et al., 2024). Our reduction isolates the common structure behind these viewpoints: a finite family of action distributions approximates the best fixed action, while a second online problem accounts for the information required to learn among those distributions. All identities below hold under the adaptive nonanticipating protocol of Appendix A; no independence assumption on the bounded observation noise is needed.

D.1 Expected-gap witnesses

For $g \in \mathcal{G}$, write

$$g^\star := \max_{a \in \mathcal{A}} g(a), \qquad \Delta_g(a) := g^\star - g(a), \qquad \Delta_g(q) := q(\Delta_g) = g^\star - q(g).$$

Thus $\Delta_g(q) \in [0,1]$ is the expected suboptimality of the action distribution $q \in \Delta(\mathcal{A})$ under g.

Definition D.1 (Optimizer witness and witness bias)**.** Let $C \subseteq \Delta(\mathcal{A})$ be nonempty and compact. Its *optimizer-witness bias* for $\mathcal{G}$ is

$$b_{\mathcal{G}}(C) := \max_{g \in \mathcal{G}} \min_{q \in C} \Delta_g(q) = \max_{g \in \mathcal{G}} \left[g^\star - \max_{q \in C} q(g) \right]. \tag{D.1}$$

We call C an *ε-optimizer witness* if $b_{\mathcal{G}}(C) \le \varepsilon$. For a finite nonempty $Q \subseteq \Delta(\mathcal{A})$, write

$$C_Q := \mathrm{co}(Q)$$

and abbreviate $b_{\mathcal{G}}(Q) := b_{\mathcal{G}}(C_Q)$.

Compactness of $\mathcal{G}$ and C ensures that both extrema in (D.1) are attained. The use of C_Q is without loss because a linear functional has the same maximum over Q and over its convex hull.

Lemma D.2 (Elementary properties)**.** *For every nonempty compact $C \subseteq \Delta(\mathcal{A})$:*

1. $0 \le b_{\mathcal{G}}(C) \le 1$.
2. *If $C \subseteq C' \subseteq \Delta(\mathcal{A})$, then $b_{\mathcal{G}}(C') \le b_{\mathcal{G}}(C)$.*
3. *For $Q_{\mathcal{A}} := \{\delta_a : a \in \mathcal{A}\}$, $b_{\mathcal{G}}(Q_{\mathcal{A}}) = 0$.*
4. *$b_{\mathcal{G}}(C) = 0$ if and only if, for every $g \in \mathcal{G}$, there is $q \in C$ satisfying*

$$\mathrm{supp}(q) \subseteq \mathsf{A}_0(g)\,. \tag{D.2}$$

Proof. For every g and q, $\Delta_g(q)$ is an expectation of a random variable in $[0,1]$, proving the first claim. The second follows because enlarging the inner minimization domain can only decrease its value. For the third claim, choose a Dirac mass at any maximizer of g.

For the final claim, nonnegativity of Δ_g gives $\Delta_g(q) = 0$ if and only if q assigns probability one to actions of zero gap, which is exactly (D.2). Since the maxima and minima in (D.1) are attained, zero witness bias is equivalent to the existence of such a q for every g. □

The following layer-cake identity records exactly how expected-gap witnesses retain all threshold scales simultaneously.

Lemma D.3 (Layer-cake representation)**.** *For every $g \in \mathcal{G}$ and $q \in \Delta(\mathcal{A})$,*

$$\Delta_g(q) = \int_0^1 q\big(\mathsf{A}_\alpha(g)^{\mathsf{c}}\big)\, \mathrm{d}\alpha. \tag{D.3}$$

Consequently, for every finite nonempty Q,

$$b_{\mathcal{G}}(Q) = \max_{g \in \mathcal{G}} \min_{q \in C_Q} \int_0^1 q\big(\mathsf{A}_\alpha(g)^{\mathsf{c}}\big)\, \mathrm{d}\alpha. \tag{D.4}$$

Proof. For every action a, the gap $\Delta_g(a)$ lies in $[0,1]$ and

$$\Delta_g(a) = \int_0^1 \mathbf{1}\{\Delta_g(a) > \alpha\}\,\mathrm{d}\alpha = \int_0^1 \mathbf{1}\{a \notin \mathsf{A}_\alpha(g)\}\,\mathrm{d}\alpha.$$

Multiplying by $q(a)$, summing over the finite action set, and interchanging the finite sum and integral proves (D.3). Equation (D.4) follows from Definition D.1. □

Lemma D.3 also clarifies the distinction from a thresholded distribution cover: the distribution minimizing the integral for a given reward vector must control its complete gap-tail profile, not merely one chosen threshold. The sharp implications between the two notions are proved in Lemma B.10 and Corollary B.11.

D.2 A CONVEX-GEOMETRIC REPRESENTATION

For a nonempty compact convex set $D \subseteq \Delta(\mathcal{A})$, let

$$h_D(u) := \max_{q \in D} q(u), \qquad u \in \mathbb{R}^{\mathcal{A}},$$

be its support function under the canonical action–reward pairing. Let

$$\mathcal{G}^0 := \mathrm{co}\big(\mathcal{G} \cup \{0\}\big) = \{\lambda g : \lambda \in [0,1],\ g \in \mathcal{G}\}, \tag{D.5}$$

where the equality uses convexity of $\mathcal{G}$. Define the asymmetric reward-class discrepancy

$$d^+_{\mathcal{G}}(p,q) := \max_{u \in \mathcal{G}^0} (p-q)(u), \qquad p, q \in \Delta(\mathcal{A})\,. \tag{D.6}$$

The inclusion $0 \in \mathcal{G}^0$ makes $d^+_{\mathcal{G}}$ nonnegative. It need not be symmetric and need not separate distinct distributions.

Proposition D.4 (Directed Hausdorff representation)**.** *The map $d^+_{\mathcal{G}}$ is a hemimetric: for all p, q, r,*

$$d^+_{\mathcal{G}}(p,p) = 0, \qquad d^+_{\mathcal{G}}(p,r) \le d^+_{\mathcal{G}}(p,q) + d^+_{\mathcal{G}}(q,r).$$

Moreover, for every nonempty compact convex $C \subseteq \Delta(\mathcal{A})$,

$$b_{\mathcal{G}}(C) = \max_{u \in \mathcal{G}^0} \big[h_{\Delta(\mathcal{A})}(u) - h_C(u)\big] \tag{D.7}$$

$$= \max_{p \in \Delta(\mathcal{A})} \min_{q \in C} d^+_{\mathcal{G}}(p,q). \tag{D.8}$$

Hence C is an ε-optimizer witness if and only if it is a directed ε-net of the full action simplex under $d^+_{\mathcal{G}}$ in the sense of (D.8)*.*

Proof. Nonnegativity and $d^+_{\mathcal{G}}(p,p) = 0$ are immediate. For the triangle inequality,

$$\begin{aligned} d^+_{\mathcal{G}}(p,r) &= \max_{u \in \mathcal{G}^0} \big[(p-q)(u) + (q-r)(u)\big] \\ &\le \max_{u \in \mathcal{G}^0} (p-q)(u) + \max_{u \in \mathcal{G}^0} (q-r)(u). \end{aligned}$$

Since $C \subseteq \Delta(\mathcal{A})$, $h_{\Delta(\mathcal{A})}(g) - h_C(g) \ge 0$ for every g. Every $u \in \mathcal{G}^0$ has the form $u = \lambda g$ with $\lambda \in [0,1]$ and $g \in \mathcal{G}$, and support functions are positively homogeneous. Therefore

$$\max_{u \in \mathcal{G}^0} \big[h_{\Delta(\mathcal{A})}(u) - h_C(u)\big] = \max_{g \in \mathcal{G}} \big[g^\star - h_C(g)\big] = b_{\mathcal{G}}(C)\,,$$

which proves (D.7).

For the second representation, abbreviate $\Delta := \Delta(\mathcal{A})$. Using compactness and finite dimensionality,

$$\begin{aligned} b_{\mathcal{G}}(C) &= \max_{u \in \mathcal{G}^0} \left[\max_{p \in \Delta} p(u) - \max_{q \in C} q(u)\right] \\ &= \max_{p \in \Delta} \max_{u \in \mathcal{G}^0} \min_{q \in C} (p-q)(u). \end{aligned}$$

For each fixed p, the map $(u,q) \mapsto (p-q)(u)$ is continuous and bilinear on the compact convex sets $\mathcal{G}^0$ and C. Sion's minimax theorem (Sion, 1958) therefore gives

$$\max_{u \in \mathcal{G}^0} \min_{q \in C} (p-q)(u) = \min_{q \in C} \max_{u \in \mathcal{G}^0} (p-q)(u).$$

Substitution yields (D.8). □

The representation above is deliberately asymmetric. The full simplex contains all pure comparators, whereas C is an approximating witness polytope; only the directed error from the former to the latter enters regret.

Proposition D.5 (Exact worst-case approximation price)**.** *For every nonempty compact convex $C \subseteq \Delta(\mathcal{A})$ and every $T \geq 1$,*

$$\max_{g_1,\dots,g_T \in \mathcal{G}} \left\{ \max_{a \in \mathcal{A}} \sum_{t=1}^{T} g_t(a) - \max_{q \in C} \sum_{t=1}^{T} q(g_t) \right\} = T b_{\mathcal{G}}(C)\,. \tag{D.9}$$

In particular, the coefficient T multiplying witness bias in the reduction below is unimprovable in a uniform pathwise comparison.

Proof. For any $g_{1:T} \in \mathcal{G}^T$, convexity gives $\bar{g}_T := T^{-1} \sum_{t=1}^{T} g_t \in \mathcal{G}$, and hence

$$\max_{a} \sum_{t=1}^{T} g_t(a) - \max_{q \in C} \sum_{t=1}^{T} q(g_t) = T \left[\max_{a} \bar{g}_T(a) - \max_{q \in C} q(\bar{g}_T) \right] \leq T b_{\mathcal{G}}(C)\,.$$

Conversely, compactness yields a maximizer $g_C \in \mathcal{G}$ in (D.1). Taking $g_t = g_C$ for all t attains equality. □

D.3 The structured mediator game

Fix a finite nonempty $Q \subseteq \Delta(\mathcal{A})$ and its witness polytope $C_Q = \mathrm{co}(Q)$. A behavioral randomization over the policies in Q induces a physical action distribution in C_Q; conversely, every element of C_Q is such a mixture. We therefore formulate the game directly in terms of the induced physical distribution. This removes an inessential latent policy draw and makes the information pattern identical to Appendix A.

Definition D.6 (Q-mediated algorithm)**.** A Q-mediated algorithm is a behavioral bandit algorithm $\mathsf{Alg} = (\pi_t)_{t=1}^{T}$ in the sense of Definition A.1 satisfying

$$\pi_t(\cdot \mid h_{t-1}) \in C_Q \qquad \text{for every } t \text{ and } h_{t-1} \in \mathsf{Z}_{t-1}. \tag{D.10}$$

The class of all such algorithms is denoted by $\mathfrak{A}_T(Q)$.

For an interaction of a Q-mediated algorithm with an admissible environment, define its mediator regret by

$$\mathrm{Reg}_T^{Q} := \max_{q \in Q} \sum_{t=1}^{T} q(g_t) - \sum_{t=1}^{T} g_t(A_t). \tag{D.11}$$

The corresponding adaptive and oblivious minimax values are

$$\mathfrak{M}_T^{\star,\mathrm{ad}}(Q,\mathcal{G}) := \inf_{\mathsf{Alg} \in \mathfrak{A}_T(Q)} \sup_{\mathsf{Adv} \in \mathfrak{C}_T^{\mathrm{ad}}(\mathcal{G})} \mathbb{E}_{\mathsf{Alg},\mathsf{Adv}}[\mathrm{Reg}_T^{Q}], \tag{D.12}$$

$$\mathfrak{M}_T^{\star,\mathrm{obl}}(Q,\mathcal{G}) := \inf_{\mathsf{Alg} \in \mathfrak{A}_T(Q)} \sup_{\mathsf{Adv} \in \mathfrak{C}_T^{\mathrm{obl}}(\mathcal{G})} \mathbb{E}_{\mathsf{Alg},\mathsf{Adv}}[\mathrm{Reg}_T^{Q}]. \tag{D.13}$$

The environment observes exactly the same public history as in the original game; in particular, it is not given an auxiliary decomposition of the physical distribution $p_t \in C_Q$ into elements of Q.

This is a reward-class-restricted mediator-feedback problem: policies are action distributions, a policy mixture generates one sampled action, and only the reward at that action is observed. In contrast to an unrestricted mediator game, however, the adversary's reward vector is constrained to $\mathcal{G}$. This restriction is the source of the observation geometry developed in the next appendix. The broader adversarial decision-making framework of Foster et al. (2022) also contains (D.12) as a special case.

Proposition D.7 (Convex-hull invariance and the Dirac case)**.** *Let $Q, Q' \subseteq \Delta(\mathcal{A})$ be finite and nonempty. If $\mathrm{co}(Q) = \mathrm{co}(Q')$, then*

$$b_{\mathcal{G}}(Q) = b_{\mathcal{G}}(Q')\,, \qquad \mathfrak{M}_T^{\star,\mathsf{s}}(Q,\mathcal{G}) = \mathfrak{M}_T^{\star,\mathsf{s}}(Q',\mathcal{G}) \tag{D.14}$$

for $\mathsf{s} \in \{\mathrm{ad}, \mathrm{obl}\}$. Moreover, for $Q_{\mathcal{A}} = \{\delta_a : a \in \mathcal{A}\}$,

$$\mathfrak{M}_T^{\star,\mathrm{ad}}(Q_{\mathcal{A}},\mathcal{G}) = \mathfrak{R}_T^{\star,\mathrm{ad}}(\mathcal{G}), \qquad \mathfrak{M}_T^{\star,\mathrm{obl}}(Q_{\mathcal{A}},\mathcal{G}) = \mathfrak{R}_T^{\star,\mathrm{obl}}(\mathcal{G}). \tag{D.15}$$

Proof. The witness-bias equality follows because linear functionals have the same supremum over a set and its convex hull. The admissible algorithm class in (D.10) depends only on C_Q. Likewise, for every reward sequence,

$$\max_{q\in Q}\sum_{t=1}^{T} q(g_t) = \max_{q\in C_Q}\sum_{t=1}^{T} q(g_t),$$

so the comparator also depends only on C_Q. This proves (D.14).

For the Dirac family, $C_{Q_\mathcal{A}} = \Delta(\mathcal{A})$, so (D.10) imposes no restriction beyond Definition A.1. Furthermore,

$$\max_{q\in Q_\mathcal{A}}\sum_{t=1}^{T} q(g_t) = \max_{a\in\mathcal{A}}\sum_{t=1}^{T} g_t(a).$$

Thus both the strategy class and regret random variable coincide with those of the original problem, proving (D.15). □

D.4 Exact variational reduction

Lemma D.8 (Pathwise witness decomposition)**.** *For every finite nonempty $Q \subseteq \Delta(\mathcal{A})$, every sequence $g_{1:T} \in \mathcal{G}^T$, and every action sequence $A_{1:T} \in \mathcal{A}^T$,*

$$\mathrm{Reg}_T \le T b_\mathcal{G}(Q) + \mathrm{Reg}_T^Q\,. \tag{D.16}$$

The inequality is deterministic and therefore remains valid when g_t is selected adaptively from the public history.

Proof. Let $\bar g_T = T^{-1}\sum_{t=1}^{T} g_t$. Convexity of $\mathcal{G}$ gives $\bar g_T \in \mathcal{G}$. Adding and subtracting the best witness comparator,

$$\begin{aligned}\mathrm{Reg}_T &= \left[\max_a \sum_{t=1}^{T} g_t(a) - \max_{q\in Q}\sum_{t=1}^{T} q(g_t)\right] + \mathrm{Reg}_T^Q \\ &= T\left[\max_a \bar g_T(a) - \max_{q\in Q} q(\bar g_T)\right] + \mathrm{Reg}_T^Q \\ &\le T b_\mathcal{G}(Q) + \mathrm{Reg}_T^Q\,.\end{aligned}$$

□

Theorem D.9 (Restatement of Theorem 8.1)**.** *For every finite action set, nonempty compact convex reward class, and horizon $T \ge 1$,*

$$\mathfrak{R}_T^{\star,\mathrm{ad}}(\mathcal{G}) = \min_{\substack{Q\subseteq\Delta(\mathcal{A})\\ 1\le|Q|<\infty}} \left\{T b_\mathcal{G}(Q) + \mathfrak{M}_T^{\star,\mathrm{ad}}(Q,\mathcal{G})\right\}, \tag{D.17}$$

$$\mathfrak{R}_T^{\star,\mathrm{obl}}(\mathcal{G}) = \min_{\substack{Q\subseteq\Delta(\mathcal{A})\\ 1\le|Q|<\infty}} \left\{T b_\mathcal{G}(Q) + \mathfrak{M}_T^{\star,\mathrm{obl}}(Q,\mathcal{G})\right\}. \tag{D.18}$$

Both minima are attained by the Dirac family $Q_\mathcal{A}$.

Proof. Fix a finite nonempty Q and a Q-mediated algorithm. It is also an admissible algorithm for the original bandit game because $C_Q \subseteq \Delta(\mathcal{A})$. Applying Lemma D.8, taking expectation, then the supremum over adaptive environments, and finally the infimum over Q-mediated algorithms gives

$$\mathfrak{R}_T^{\star,\mathrm{ad}}(\mathcal{G}) \le T b_\mathcal{G}(Q) + \mathfrak{M}_T^{\star,\mathrm{ad}}(Q,\mathcal{G}).$$

Since this holds for every finite nonempty Q, it yields the "$\le$" direction in (D.17). For $Q = Q_\mathcal{A}$, Lemma D.2 and Proposition D.7 give

$$T b_\mathcal{G}(Q_\mathcal{A}) + \mathfrak{M}_T^{\star,\mathrm{ad}}(Q_\mathcal{A},\mathcal{G}) = \mathfrak{R}_T^{\star,\mathrm{ad}}(\mathcal{G}),$$

which proves the reverse inequality and attainment. The oblivious identity is proved identically after replacing $\mathfrak{E}_T^{\mathrm{ad}}(\mathcal{G})$ by $\mathfrak{E}_T^{\mathrm{obl}}(\mathcal{G})$ throughout. □

The variational identity separates the problem into an exact offline approximation term and an online estimation term. The first is entirely determined by how well the witness polytope approximates the action simplex in (D.8); the second depends on how one sampled action reveals the relative values of witnesses under the restricted reward class.

Corollary D.10 (Exact bias–estimation frontier)**.** *For* $\mathsf{s} \in \{\mathrm{ad}, \mathrm{obl}\}$ *and* $\varepsilon \in [0,1]$*, define*

$$\mathfrak{M}^{\star,\mathsf{s}}_{T,\varepsilon}(\mathcal{G}) := \inf_{\substack{Q \subseteq \Delta(\mathcal{A}),\ 1 \le |Q| < \infty \\ b_{\mathcal{G}}(Q) \le \varepsilon}} \mathfrak{M}^{\star,\mathsf{s}}_{T}(Q, \mathcal{G}). \tag{D.19}$$

Then

$$\mathfrak{R}^{\star,\mathsf{s}}_{T}(\mathcal{G}) = \inf_{\varepsilon \in [0,1]} \left\{ T\varepsilon + \mathfrak{M}^{\star,\mathsf{s}}_{T,\varepsilon}(\mathcal{G}) \right\}. \tag{D.20}$$

Proof. For any finite Q, choose $\varepsilon = b_{\mathcal{G}}(Q)$ in (D.19). This shows that the right-hand side of (D.20) is no larger than the corresponding minimum in Theorem D.9. Conversely, whenever $b_{\mathcal{G}}(Q) \le \varepsilon$,

$$T\varepsilon + \mathfrak{M}^{\star,\mathsf{s}}_{T}(Q, \mathcal{G}) \ge T b_{\mathcal{G}}(Q) + \mathfrak{M}^{\star,\mathsf{s}}_{T}(Q, \mathcal{G}) \ge \mathfrak{R}^{\star,\mathsf{s}}_{T}(\mathcal{G}),$$

where the last inequality is Theorem D.9. Taking both infima proves equality. □

Corollary D.11 (Distribution-cover reduction)**.** *Let* $\alpha \in (0,1]$*,* $\beta \in [0,1)$*, and let* $Q \subseteq \Delta(\mathcal{A})$ *be a finite* (α,β)*-distribution cover of* $\mathcal{G}$*. Then, for* $\mathsf{s} \in \{\mathrm{ad}, \mathrm{obl}\}$*,*

$$\mathfrak{R}^{\star,\mathsf{s}}_{T}(\mathcal{G}) \le T\big[\alpha + (1-\alpha)\beta\big] + \mathfrak{M}^{\star,\mathsf{s}}_{T}(Q, \mathcal{G}). \tag{D.21}$$

Conversely, if $b_{\mathcal{G}}(Q) \le \alpha\beta$*, then* Q *is an* (α,β)*-distribution cover.*

Proof. Corollary B.11 gives $b_{\mathcal{G}}(Q) \le \alpha + (1-\alpha)\beta$ for a distribution cover and gives the stated converse implication. Substituting the first inequality into Theorem D.9 proves (D.21). □

Replacing the mediator term in (D.21) by a generic finite-policy bandit guarantee recovers the flat distribution-cover strategy of Hanneke & Wang (2026). The identity itself is sharper: it leaves the mediator minimax value intact, allowing the subsequent observation-design theory to exploit relationships among rewards at different actions. In the unrestricted reward cube, those relationships disappear and the relevant one-step geometry reduces to the ordinary mediator-feedback χ^2 geometry of Eldowa et al. (2024); for structured reward classes it can be exponentially smaller, as Theorem 2.1 already anticipates.

D.5 Scope of the reduction

Theorem D.9 is an exact identity, but it is not by itself a finite-time complexity characterization: the Dirac witness family always attains the minimum and simply reproduces the original game. Its content is that any nontrivial rate theorem must control two mathematically distinct objects:

$$T b_{\mathcal{G}}(Q) \quad \text{(optimizer approximation)}, \qquad \mathfrak{M}^{\star,\mathsf{s}}_{T}(Q, \mathcal{G}) \quad \text{(online observation and estimation)}.$$

Proposition D.5 shows that the first term is the exact worst-case comparator loss associated with Q, while the second term is the only place where cross-action information can enter. The next appendix will develop a reward-class-restricted Riesz geometry for that mediator term. Until that appendix is established, no claim is made that witness bias or witness cardinality alone determines minimax regret.

E Riesz Geometry of One-Point Observation Designs

This appendix develops the local observation geometry underlying the mediator term in Theorem 8.1. A policy contrast is a signed action measure h with total mass zero. If one action $A \sim p$ and one noisy reward Y with conditional mean $g(A)$ are observed, then an affine statistic $c + \psi(A)Y$ can estimate the contrast $h(g)$ only through the restrictions that the reward class imposes across actions. We identify the minimum uniform stochastic second-moment budget of such an estimator with an exact Riesz dual norm, derive a finite-dimensional Fenchel dual for approximate estimation, and calculate the resulting geometry for tabular and linear reward classes.

For unrestricted rewards, the geometry reduces to the usual χ^2 divergence appearing in mediator-feedback capacity (Eldowa et al., 2024). For linear rewards, it becomes a pseudoinverse experimental-design norm of the type underlying classical optimal design (Kiefer & Wolfowitz, 1960) and adversarial linear bandits (Bubeck et al., 2012). The distinction between these two specializations is the information axis that the optimizer-cover quantities of Sections 2 and 2 cannot record.

E.1 Affine one-sample estimators

We identify functions on the finite action set with vectors in $\mathbb{R}^{\mathcal{A}}$ and use the canonical pairing $h(g) = \langle h, g\rangle = \sum_{a\in\mathcal{A}} h(a)g(a)$. Define the policy-contrast space

$$\mathsf{H}_0 := \left\{h \in \mathbb{R}^{\mathcal{A}} : \langle h, \mathbf{1}\rangle = 0\right\}. \tag{E.1}$$

Every difference $q - r$ of action distributions lies in H_0. The zero-mass condition makes policy contrasts invariant to a common reward offset.

The affine geometry of the reward class is encoded by

$$\mathcal{V} := \operatorname{span}(\mathcal{G} - \mathcal{G}), \qquad \mathcal{K} := \frac{1}{2}(\mathcal{G} - \mathcal{G}). \tag{E.2}$$

The set $\mathcal{K}$ is viewed as a compact convex subset of its linear span $\mathcal{V}$.

Lemma E.1 (Geometry of the centered reward body)**.** *The set $\mathcal{K}$ is compact, convex, centrally symmetric, and satisfies $\operatorname{span}(\mathcal{K}) = \mathcal{V}$ and $0 \in \operatorname{ri}_{\mathcal{V}}(\mathcal{K})$. Consequently, its Minkowski functional*

$$\|v\|_{\mathcal{K}} := \inf\{t > 0 : v \in t\mathcal{K}\}, \qquad v \in \mathcal{V}, \tag{E.3}$$

is a norm on $\mathcal{V}$. For every $r \in \mathbb{R}^{\mathcal{A}}$,

$$\sigma_{\mathcal{K}}(r) := \max_{v\in\mathcal{K}} \langle r, v\rangle = \frac{1}{2}\left[\max_{g\in\mathcal{G}} \langle r, g\rangle - \min_{g\in\mathcal{G}} \langle r, g\rangle\right]. \tag{E.4}$$

Proof. Compactness and convexity follow from compactness and convexity of $\mathcal{G}$. Central symmetry follows because $(g - g')/2 \in \mathcal{K}$ implies $(g' - g)/2 \in \mathcal{K}$. The span identity is immediate from (E.2). If 0 were not in the relative interior of $\mathcal{K}$, a nonzero linear functional on $\mathcal{V}$ would support $\mathcal{K}$ at zero. Central symmetry would then force that functional to vanish on all of $\mathcal{K}$, contradicting $\operatorname{span}(\mathcal{K}) = \mathcal{V}$. Hence $\mathcal{K}$ is absorbing in $\mathcal{V}$, and its gauge is finite. Convexity, symmetry, and the relative-interior property imply that the gauge is a norm. Finally,

$$\max_{v\in\mathcal{K}} \langle r, v\rangle = \frac{1}{2} \max_{g,g'\in\mathcal{G}} \langle r, g - g'\rangle,$$

and the maximum separates into the difference between the maximum and minimum of $\langle r, g\rangle$ over $\mathcal{G}$, proving (E.4). □

For $p \in \Delta(\mathcal{A})$ and $\psi \in \mathbb{R}^{\mathcal{A}}$, write

$$\|\psi\|_p^2 := \sum_{a\in\mathcal{A}} p(a)\psi(a)^2, \qquad p \odot \psi := \big(p(a)\psi(a)\big)_{a\in\mathcal{A}}.$$

Definition E.2 (Approximate one-point design cost)**.** For $h \in \mathsf{H}_0$, $p \in \Delta(\mathcal{A})$, and $\varepsilon \geq 0$, define

$$\mathcal{E}_\varepsilon(h; p) := \inf_{c\in\mathbb{R},\, \psi\in\mathbb{R}^{\mathcal{A}}} \left\{ \|\psi\|_p^2 : \sup_{g\in\mathcal{G}} |c + \langle p \odot \psi - h, g\rangle| \leq \varepsilon \right\}. \tag{E.5}$$

The infimum of the empty set is $+\infty$. Throughout this appendix, $\mathcal{E}_\varepsilon(h; p)$ is understood relative to the fixed reward class $\mathcal{G}$.

The scalar c is essential: the random observation need only estimate variation within the affine hull of $\mathcal{G}$; a deterministic offset can be supplied without sampling noise.

Proposition E.3 (Operational meaning). *Fix $g \in \mathcal{G}$, $p \in \Delta(\mathcal{A})$, and an observation pair (A, Y) satisfying $A \sim p$, $Y \in [0,1]$, and $\mathbb{E}[Y \mid A] = g(A)$. For any feasible (c, ψ) in* (E.5)*, the affine estimator*

$$\widehat{h}_{c,\psi}(A, Y) := c + \psi(A)Y \tag{E.6}$$

satisfies

$$\left|\mathbb{E}[\widehat{h}_{c,\psi}(A, Y)] - h(g)\right| \le \varepsilon, \tag{E.7}$$

$$\mathrm{Var}\Big(\widehat{h}_{c,\psi}(A, Y)\Big) \le \|\psi\|_p^2, \tag{E.8}$$

$$\mathbb{E}\Big[\big(\widehat{h}_{c,\psi}(A, Y) - h(g)\big)^2\Big] \le \|\psi\|_p^2 + \varepsilon^2. \tag{E.9}$$

The same statements hold conditionally on any past history under the protocol of Appendix A.

Proof. The conditional-mean property gives

$$\mathbb{E}[\widehat{h}_{c,\psi}(A, Y)] - h(g) = c + \sum_a p(a)\psi(a)g(a) - h(g),$$

which is bounded by ε by feasibility. Since c is deterministic and $0 \le Y \le 1$,

$$\mathrm{Var}(c + \psi(A)Y) \le \mathbb{E}[\psi(A)^2 Y^2] \le \sum_a p(a)\psi(a)^2.$$

The mean-squared-error bound is the variance–bias decomposition. Conditional versions follow by applying the same calculation under each regular conditional law given the public history. □

The next lemma eliminates the scalar c and exposes the intrinsic symmetric reward body.

Lemma E.4 (Optimal affine centering). *For every $r \in \mathbb{R}^{\mathcal{A}}$,*

$$\inf_{c \in \mathbb{R}} \sup_{g \in \mathcal{G}} |c + r(g)| = \sigma_{\mathcal{K}}(r). \tag{E.10}$$

If $M_r = \max_{g \in \mathcal{G}} r(g)$ and $m_r = \min_{g \in \mathcal{G}} r(g)$, then an optimal center is $c_r^\star = -(M_r + m_r)/2$. Consequently,

$$\mathcal{E}_\varepsilon(h; p) = \inf_{\psi \in \mathbb{R}^{\mathcal{A}}} \left\{ \|\psi\|_p^2 : \sigma_{\mathcal{K}}(p \odot \psi - h) \le \varepsilon \right\}. \tag{E.11}$$

Proof. The image $\{r(g) : g \in \mathcal{G}\}$ is the compact interval $[m_r, M_r]$, because $\mathcal{G}$ is compact and convex and r is linear. The smallest interval centered at $-c$ containing $[m_r, M_r]$ has radius $(M_r - m_r)/2$, attained at the midpoint. Equation (E.4) identifies this radius with $\sigma_{\mathcal{K}}(r)$. Substituting $r = p \odot \psi - h$ proves (E.11). □

It will be useful to replace the coefficient ψ by its signed observation measure $z = p \odot \psi$. Define the extended quadratic perspective

$$\jmath(z, s) := \begin{cases} z^2/s, & s > 0, \\ 0, & (z, s) = (0, 0), \\ +\infty, & s = 0 \text{ and } z \ne 0, \end{cases} \tag{E.12}$$

and

$$J_p(z) := \sum_{a \in \mathcal{A}} \jmath(z(a), p(a)). \tag{E.13}$$

The convention in (E.12) enforces $z(a) = 0$ whenever $p(a) = 0$.

Proposition E.5 (Signed-measure primal). *For every h, p, ε,*

$$\mathcal{E}_\varepsilon(h; p) = \inf_{z \in \mathbb{R}^{\mathcal{A}}} \{J_p(z) : \sigma_{\mathcal{K}}(z - h) \le \varepsilon\}. \tag{E.14}$$

Whenever the value is finite, the infimum is attained.

Proof. If $z = p \odot \psi$, then $J_p(z) = \|\psi\|_p^2$ after setting $\psi(a) = 0$ on $\{a : p(a) = 0\}$. Conversely, $J_p(z) < \infty$ implies $z(a) = 0$ whenever $p(a) = 0$, and defining $\psi(a) = z(a)/p(a)$ on the support of p recovers $z = p \odot \psi$ and the same objective. This proves the identity.

For attainment, let (z_n) be a minimizing sequence with uniformly bounded energy. Then $z_n(a) = 0$ on $\{p(a) = 0\}$ and $|z_n(a)|^2 \le Cp(a)$ on the support of p, so (z_n) is bounded. A convergent subsequence has a limit z satisfying the closed support-function constraint, and lower semicontinuity of J_p gives optimality. □

E.2 Exact observability and the Riesz representation

For $p \in \Delta(\mathcal{A})$, define the seminorm and its nullspace on the reward difference space by

$$\|v\|_p^2 := p(v^2) = \sum_a p(a)v(a)^2, \qquad \mathcal{N}_p := \{v \in \mathcal{V} : \|v\|_p = 0\}. \tag{E.15}$$

For $h \in \mathsf{H}_0$, define the extended dual seminorm

$$\|h\|_{\mathcal{V},p,\star} := \sup\left\{|h(v)| : v \in \mathcal{V},\ \|v\|_p \le 1\right\} \in [0, +\infty]. \tag{E.16}$$

The value is infinite precisely when h is nonzero on an observationally null reward direction.

Theorem E.6 (Exact Riesz-design identity)**.** *For every $h \in \mathsf{H}_0$ and $p \in \Delta(\mathcal{A})$,*

$$\mathcal{E}_0(h; p) = \|h\|_{\mathcal{V},p,\star}^2 . \tag{E.17}$$

The common value is finite if and only if

$$h(v) = 0 \qquad \textit{for every } v \in \mathcal{N}_p. \tag{E.18}$$

When (E.18) *holds, there exists a minimum-norm Riesz representer $\psi_{h,p}^\star \in \mathcal{V}$ satisfying*

$$p(\psi_{h,p}^\star v) = h(v) \quad \forall v \in \mathcal{V}, \qquad \left\|\psi_{h,p}^\star\right\|_p^2 = \mathcal{E}_0(h; p) . \tag{E.19}$$

Its equivalence class in the quotient Hilbert space $\mathcal{V}/\mathcal{N}_p$ is unique. For any $g_0 \in \mathcal{G}$, the choice

$$c_{h,p}^\star := h(g_0) - p(\psi_{h,p}^\star g_0) \tag{E.20}$$

makes $c_{h,p}^\star + \psi_{h,p}^\star(A)Y$ exactly unbiased for $h(g)$ for every $g \in \mathcal{G}$.

Proof. By Lemma E.4, exact feasibility is equivalent to

$$p(\psi v) = h(v) \qquad \forall v \in \mathcal{V},$$

because $\mathcal{K}$ spans $\mathcal{V}$. Any feasible ψ therefore satisfies, by Cauchy–Schwarz,

$$|h(v)| = |p(\psi v)| \le \|\psi\|_p \, \|v\|_p \qquad \forall v \in \mathcal{V}.$$

Taking the supremum over the unit ball in (E.16) proves $\|h\|_{\mathcal{V},p,\star}^2 \le \mathcal{E}_0(h; p)$ and also shows that finite exact cost requires (E.18).

Conversely, assume (E.18). Equip the quotient $\mathcal{V}/\mathcal{N}_p$ with the inner product

$$\langle [v], [w] \rangle_p := p(vw).$$

This is a finite-dimensional Hilbert space. Condition (E.18) makes $[v] \mapsto h(v)$ a well-defined linear functional, whose operator norm is exactly (E.16). The Riesz representation theorem gives a unique class $[\psi_{h,p}^\star]$ satisfying $h(v) = p(\psi_{h,p}^\star v)$ for every v and $\left\|\psi_{h,p}^\star\right\|_p = \|h\|_{\mathcal{V},p,\star}$. Choosing a representative in $\mathcal{V}$ proves (E.19) and the reverse inequality in (E.17).

Finally, $g - g_0 \in \mathcal{V}$ for every $g \in \mathcal{G}$, so

$$c_{h,p}^\star + p(\psi_{h,p}^\star g) - h(g) = p(\psi_{h,p}^\star(g - g_0)) - h(g - g_0) = 0.$$

Thus (E.20) gives exact unbiasedness. □

The identity admits a coordinate representation useful for calculations.

Corollary E.7 (Basis and pseudoinverse formula)**.** *Let $r = \dim(\mathcal{V})$ and let $B \in \mathbb{R}^{K\times r}$ have full column rank and column space $\mathcal{V}$. With $P = \mathrm{diag}(p)$, define*

$$G_{p,B} := B^\top P B, \qquad b_{h,B} := B^\top h. \tag{E.21}$$

Then

$$\mathcal{E}_0(h;p) = \begin{cases} b_{h,B}^\top G_{p,B}^\dagger b_{h,B}, & b_{h,B} \in \mathrm{range}(G_{p,B}), \\ +\infty, & b_{h,B} \notin \mathrm{range}(G_{p,B}). \end{cases} \tag{E.22}$$

In the finite case, one minimum-norm representer is

$$\psi_{h,p}^\star = B G_{p,B}^\dagger b_{h,B}. \tag{E.23}$$

The value in (E.22) *is independent of the chosen basis matrix B.*

Proof. Every $v \in \mathcal{V}$ has a unique representation $v = B\theta$. Hence $h(v) = b_{h,B}^\top \theta$ and $\|v\|_p^2 = \theta^\top G_{p,B}\theta$. The extended Rayleigh-quotient identity gives

$$\sup_{\theta:\, \theta^\top G_{p,B}\theta \le 1} |b_{h,B}^\top \theta|^2 = \begin{cases} b_{h,B}^\top G_{p,B}^\dagger b_{h,B}, & b_{h,B} \in \mathrm{range}(G_{p,B}), \\ +\infty, & \text{otherwise.} \end{cases}$$

Together with Theorem E.6, this proves the value formula. If the range condition holds, then $G_{p,B} G_{p,B}^\dagger b_{h,B} = b_{h,B}$, so

$$B^\top P \psi_{h,p}^\star = G_{p,B} G_{p,B}^\dagger b_{h,B} = b_{h,B}.$$

Thus (E.23) represents h on $\mathcal{V}$, and its p-norm squared is the value in (E.22). Basis independence follows because both sides equal the intrinsic quantity in (E.17). □

E.3 Approximate observability and Fenchel duality

The exact cost depends only on the span $\mathcal{V}$. At nonzero tolerance, the shape of the centered body $\mathcal{K}$ also matters. For a linear functional ℓ on $\mathcal{V}$, define

$$\|\ell\|_{\mathcal{K},\star} := \sup_{v\in\mathcal{K}} |\ell(v)|. \tag{E.24}$$

This is the dual norm of (E.3).

For $p \in \Delta(\mathcal{A})$, let

$$\mathscr{R}_p := \left\{ \ell_\psi : \mathcal{V} \to \mathbb{R} : \ell_\psi(v) = p(\psi v),\ \psi \in \mathbb{R}^{\mathcal{A}} \right\}. \tag{E.25}$$

Define the irreducible approximation error

$$\delta_p(h) := \inf_{\ell\in\mathscr{R}_p} \|h|_{\mathcal{V}} - \ell\|_{\mathcal{K},\star}. \tag{E.26}$$

Proposition E.8 (Null-direction characterization of approximate observability)**.** *The representable functionals are exactly the annihilator of $\mathcal{N}_p$:*

$$\mathscr{R}_p = \{\ell \in \mathcal{V}^\star : \ell(v) = 0\ \forall v \in \mathcal{N}_p\}. \tag{E.27}$$

Moreover,

$$\delta_p(h) = \sup\{|h(v)| : v \in \mathcal{N}_p,\ \|v\|_{\mathcal{K}} \le 1\}. \tag{E.28}$$

Consequently,

$$\mathcal{E}_\varepsilon(h;p) < +\infty \quad\Longleftrightarrow\quad \varepsilon \ge \delta_p(h). \tag{E.29}$$

Proof. Every $\ell_\psi \in \mathscr{R}_p$ vanishes on $\mathcal{N}_p$ by Cauchy–Schwarz. Conversely, any functional vanishing on $\mathcal{N}_p$ descends to the quotient Hilbert space $\mathcal{V}/\mathcal{N}_p$ and hence has a Riesz representer; this proves (E.27).

For any $\ell \in \mathscr{R}_p$ and any $v \in \mathcal{N}_p$, $(h-\ell)(v) = h(v)$, so the right-hand side of (E.28) lower bounds (E.26). Conversely, the restriction of h to $\mathcal{N}_p$ has norm equal to the right-hand side of (E.28). By

the finite-dimensional Hahn–Banach theorem, it extends to a functional r on $\mathcal{V}$ with the same norm. Then $h - r$ vanishes on $\mathcal{N}_p$, hence belongs to $\mathscr{R}_p$ by (E.27). Therefore $\delta_p(h) \le \|r\|_{\mathcal{K},\star}$, proving equality.

Finally, Lemma E.4 says that feasibility at tolerance ε is equivalent to approximating $h|_{\mathcal{V}}$ by an element of $\mathscr{R}_p$ within dual norm ε. The finite-dimensional subspace $\mathscr{R}_p$ is closed, so the distance is attained, and (E.29) follows. □

We next give a dual representation that will later turn observation-design obstructions into hard reward perturbations.

Theorem E.9 (Fenchel dual of approximate one-point design)**.** *For every $h \in \mathsf{H}_0$, $p \in \Delta(\mathcal{A})$, and $\varepsilon \ge 0$,*

$$\mathcal{E}_\varepsilon(h;p) = \sup_{v\in\mathcal{V}} \left\{ h(v) - \varepsilon \|v\|_{\mathcal{K}} - \frac{1}{4}\|v\|_p^2 \right\}, \tag{E.30}$$

where both sides may equal $+\infty$. If $\varepsilon < \delta_p(h)$, the supremum on the right is unbounded. If $\varepsilon \ge \delta_p(h)$, the primal infimum in (E.5) *is attained and the common value is finite.*

Proof. Let $T_p : \mathbb{R}^{\mathcal{A}} \to \mathcal{V}^\star$ be $(T_p\psi)(v) = p(\psi v)$. By Lemma E.4,

$$\mathcal{E}_\varepsilon(h;p) = \inf_{\psi} \left\{ \|\psi\|_p^2 : \|h|_{\mathcal{V}} - T_p\psi\|_{\mathcal{K},\star} \le \varepsilon \right\}. \tag{E.31}$$

The indicator of the closed dual-norm ball has the support representation

$$\iota_{\{\ell:\,\|\ell\|_{\mathcal{K},\star}\le\varepsilon\}}(\ell) = \sup_{v\in\mathcal{V}} \{\ell(v) - \varepsilon\|v\|_{\mathcal{K}}\}. \tag{E.32}$$

Indeed, the right-hand side is zero when the dual norm of ℓ is at most ε, and is $+\infty$ by positive scaling otherwise.

Assume first that $\varepsilon > \delta_p(h)$. Then (E.31) is strictly feasible. Applying finite-dimensional Fenchel duality to (E.31) and (E.32) yields

$$\mathcal{E}_\varepsilon(h;p) = \sup_{v\in\mathcal{V}} \inf_{\psi\in\mathbb{R}^{\mathcal{A}}} \left\{ \|\psi\|_p^2 + h(v) - p(\psi v) - \varepsilon\|v\|_{\mathcal{K}} \right\}.$$

For fixed v, completing the square coordinatewise gives

$$\inf_{\psi} \left\{ \|\psi\|_p^2 - p(\psi v) \right\} = -\frac{1}{4}\|v\|_p^2, \tag{E.33}$$

attained, on the support of p, at $\psi = v/2$. This proves (E.30) under strict feasibility.

If $\varepsilon = \delta_p(h)$, apply the preceding equality at $\varepsilon_n = \varepsilon + 1/n$. The primal values increase to the value at ε: choose each minimizing coefficient to vanish off $\mathrm{supp}(p)$. A feasible coefficient at the boundary uniformly bounds their p-norms, which on the finite support of p also bounds their Euclidean norms. Hence a convergent subsequence is feasible at the limit; the reverse inequality follows from nesting of the feasible sets. On the dual side, the objectives increase pointwise as $\varepsilon_n \downarrow \varepsilon$, and therefore their suprema increase to the supremum at ε. This proves equality at the boundary.

Finally, if $\varepsilon < \delta_p(h)$, compactness in (E.28) gives $v_0 \in \mathcal{N}_p$ such that, after changing its sign if necessary, $h(v_0) > \varepsilon\|v_0\|_{\mathcal{K}}$. Since $\|v_0\|_p = 0$, substituting tv_0 in the dual objective and sending $t \to \infty$ makes the right-hand side infinite. The primal is infeasible by (E.29). □

Theorem E.6 can alternatively be recovered from Theorem E.9 at $\varepsilon = 0$: optimizing the scale of each v gives

$$\sup_{v\in\mathcal{V}} \left\{ h(v) - \tfrac{1}{4}\|v\|_p^2 \right\} = \|h\|_{\mathcal{V},p,\star}^2.$$

The explicit quotient-space argument above additionally identifies the minimum-norm estimator.

Proposition E.10 (Regularity and invariances)**.** *For the fixed reward class $\mathcal{G}$, the following hold.*

1. *For each $\varepsilon \ge 0$, the map $(h,p) \mapsto \mathcal{E}_\varepsilon(h;p)$ is jointly convex and lower semicontinuous on $\mathsf{H}_0 \times \Delta(\mathcal{A})$.*

2. *The cost is even in h, nonincreasing and right-continuous in ε, and for every $\lambda > 0$,*

$$\mathcal{E}_{\lambda\varepsilon}(\lambda h; p) = \lambda^2 \mathcal{E}_\varepsilon(h; p)\,. \tag{E.34}$$

3. *One has*

$$\mathcal{E}_\varepsilon(h; p) = 0 \quad \Longleftrightarrow \quad \sigma_{\mathcal{K}}(h) \le \varepsilon. \tag{E.35}$$

4. *If $r \in \mathcal{V}^\perp \cap \mathsf{H}_0$, then*

$$\mathcal{E}_\varepsilon(h + r; p) = \mathcal{E}_\varepsilon(h; p)\,. \tag{E.36}$$

Thus the cost depends on h only through its restriction to $\mathcal{V}$.

5. *Translating the reward class does not change the cost: for every $u \in \mathbb{R}^{\mathcal{A}}$,*

$$\mathcal{E}_\varepsilon^{\mathcal{G}+u}(h; p) = \mathcal{E}_\varepsilon^{\mathcal{G}}(h; p), \tag{E.37}$$

where the superscript records the reward class used in Definition E.2.

Proof. Use the signed-measure formulation (E.14). The scalar perspective $(z, s) \mapsto \jmath(z, s)$ is closed and jointly convex, and the constraint $\sigma_{\mathcal{K}}(z - h) \le \varepsilon$ is convex in (z, h). Partial minimization over z therefore gives joint convexity. For lower semicontinuity, consider $(h_n, p_n) \to (h, p)$ with bounded liminf cost and choose nearly optimal z_n. Bounded energy implies $|z_n(a)|^2 \le C p_n(a)$, so a subsequence converges. Closedness of the constraint and lower semicontinuity of the perspective yield the claim.

Evenness follows by replacing z with $-z$; monotonicity follows from nesting of the constraint sets; right-continuity follows from the same compactness argument used at the boundary in Theorem E.9. Scaling (h, z, ε) by λ proves (E.34). Since $J_p(z) = 0$ only for $z = 0$, the signed-measure primal gives (E.35). If $r \in \mathcal{V}^\perp \cap \mathsf{H}_0$, then $\sigma_{\mathcal{K}}(z - h - r) = \sigma_{\mathcal{K}}(z - h)$, proving (E.36). Finally, translating $\mathcal{G}$ leaves $\mathcal{G} - \mathcal{G}$ and hence $\mathcal{K}$ unchanged, proving (E.37). □

E.4 Tabular and linear specializations

We now calculate the design geometry in the two classes used in the separation argument.

Proposition E.11 (Tabular reward cube)**.** *Suppose $\mathcal{G} = [0, 1]^{\mathcal{A}}$. Then $\mathcal{V} = \mathbb{R}^{\mathcal{A}}$, $\mathcal{K} = [-1/2, 1/2]^{\mathcal{A}}$, and for every $h \in \mathsf{H}_0$,*

$$\mathcal{E}_\varepsilon(h; p) = \min_{\substack{z(a)=0 \text{ if } p(a)=0 \\ \|z-h\|_1 \le 2\varepsilon}} \sum_{a:p(a)>0} \frac{z(a)^2}{p(a)}, \tag{E.38}$$

with value $+\infty$ if the constraint is empty. In particular,

$$\mathcal{E}_0(h; p) = \begin{cases} \displaystyle\sum_{a:p(a)>0} \frac{h(a)^2}{p(a)}, & h(a) = 0 \text{ whenever } p(a) = 0, \\ +\infty, & \text{otherwise}. \end{cases} \tag{E.39}$$

Consequently, for $q \in \Delta(\mathcal{A})$,

$$\mathcal{E}_0(q - p; p) = \chi^2(q \,\|\, p) := \begin{cases} \displaystyle\sum_{a:p(a)>0} p(a) \left(\frac{q(a)}{p(a)} - 1 \right)^2, & q \ll p, \\ +\infty, & q \not\ll p. \end{cases} \tag{E.40}$$

Proof. For the cube, $\sigma_{\mathcal{K}}(r) = \frac{1}{2} \|r\|_1$. Substituting this identity into Proposition E.5 gives (E.38). At $\varepsilon = 0$, the constraint forces $z = h$, giving (E.39). Taking $h = q - p$ yields (E.40) by direct algebra. □

The next proposition allows an affine baseline that need not be known through the sample; the deterministic centering constant absorbs it.

Proposition E.12 (Linear reward-difference classes)**.** *Suppose there is a feature map $\phi : \mathcal{A} \to \mathbb{R}^d$ such that*

$$\mathcal{V} = \left\{ v_\theta : a \mapsto \langle \theta, \phi(a) \rangle : \theta \in \mathbb{R}^d \right\}. \tag{E.41}$$

For $h \in \mathsf{H}_0$ and $p \in \Delta(\mathcal{A})$, define

$$m_h := \sum_a h(a)\phi(a), \qquad G_p := \sum_a p(a)\phi(a)\phi(a)^\top. \tag{E.42}$$

Then

$$\mathcal{E}_0(h; p) = \begin{cases} m_h^\top G_p^\dagger m_h, & m_h \in \mathrm{range}(G_p), \\ +\infty, & \textit{otherwise}. \end{cases} \tag{E.43}$$

Under the range condition, an optimal random coefficient is

$$\psi^\star_{h,p}(a) = \phi(a)^\top G_p^\dagger m_h. \tag{E.44}$$

Proof. For $v_\theta \in \mathcal{V}$,

$$h(v_\theta) = m_h^\top \theta, \qquad \|v_\theta\|_p^2 = \theta^\top G_p \theta.$$

The extended Rayleigh quotient in Theorem E.6 therefore equals the right-hand side of (E.43). Under the range condition,

$$p(\psi^\star_{h,p} v_\theta) = m_h^\top G_p^\dagger G_p \theta = m_h^\top \theta,$$

where the middle identity uses $G_p G_p^\dagger m_h = m_h$. Its squared p-norm is the finite value in (E.43), so it is optimal. □

The pseudoinverse quantity in (E.43) is the minimum coefficient energy of an unbiased linear estimator of the feature contrast m_h and hence a uniform upper bound on its stochastic second moment. Optimizing the sampling distribution p therefore gives a transductive optimal-design problem. Classical G-optimality identifies designs that control all pointwise prediction variances (Kiefer & Wolfowitz, 1960); the same information-matrix geometry is used in finite-action adversarial linear-bandit algorithms (Bubeck et al., 2012).

E.5 Restricted policy capacity and dimension collapse

The exact design cost can be aggregated over a distribution of witness policies. For a finite nonempty $Q \subseteq \Delta(\mathcal{A})$, define

$$\mathrm{Cap}_{\mathcal{G}}(Q) := \sup_{\mu \in \Delta(Q)} \sum_{q:\, \mu(q) > 0} \mu(q) \mathcal{E}_0(q - \bar{q}_\mu; \bar{q}_\mu), \qquad \bar{q}_\mu := \sum_{q \in Q} \mu(q) q. \tag{E.45}$$

If $\mu(q) > 0$, then $q \ll \bar{q}_\mu$, so every term in (E.45) is finite: the ordinary importance-weighted estimator is always feasible, even if the reward class has no structure.

Proposition E.13 (Tabular capacity and linear dimension bound)**.** *The following statements hold.*

1. *If $\mathcal{G} = [0,1]^{\mathcal{A}}$, then*

$$\mathrm{Cap}_{\mathcal{G}}(Q) = \sup_{\mu \in \Delta(Q)} \mathbb{E}_{q \sim \mu} \chi^2(q \,\|\, \bar{q}_\mu). \tag{E.46}$$

In particular, for the Dirac family $Q_{\mathcal{A}} = \{\delta_a : a \in \mathcal{A}\}$,

$$\mathrm{Cap}_{[0,1]^{\mathcal{A}}}(Q_{\mathcal{A}}) = K - 1. \tag{E.47}$$

2. *Under the linear-span assumption* (E.41)*,*

$$\mathrm{Cap}_{\mathcal{G}}(Q) \le \sup_{\mu \in \Delta(Q)} \mathrm{rank}(G_{\bar{q}_\mu}) \le d. \tag{E.48}$$

Proof. The first identity follows from (E.40). For the Dirac family, write the mixing law as $p \in \Delta(\mathcal{A})$. Then

$$\mathbb{E}_{A \sim p} \chi^2(\delta_A \,\|\, p) = \sum_{a: p(a) > 0} p(a) \left(\frac{1}{p(a)} - 1 \right) = |\,\mathrm{supp}(p)| - 1.$$

The supremum is $K-1$, attained by any full-support p.

For the linear class, fix μ and write $p=\bar{q}_\mu$, $m_q=\sum_a q(a)\phi(a)$, and $m_p=\sum_a p(a)\phi(a)$. Every q in the support of μ satisfies the range condition in Proposition E.12: if $u\in\ker(G_p)$, then $\langle u,\phi(a)\rangle=0$ for every a with $p(a)>0$, hence $\langle u, m_q-m_p\rangle=0$. Therefore

$$\mathbb{E}_{q\sim\mu}\mathcal{E}_0(q-p;p)=\operatorname{tr}\big(G_p^\dagger\operatorname{Cov}_{q\sim\mu}(m_q)\big)\,. \tag{E.49}$$

For each q, $m_q m_q^\top\preceq\mathbb{E}_{a\sim q}[\phi(a)\phi(a)^\top]$ by nonnegativity of covariance. Averaging and subtracting $m_p m_p^\top\succeq 0$ gives

$$0\preceq\operatorname{Cov}_{q\sim\mu}(m_q)\preceq G_p.$$

Conjugating by $G_p^{\dagger/2}$ shows that the trace in (E.49) is at most $\operatorname{rank}(G_p)$, proving (E.48). □

For the code class of Theorem 2.1, the Dirac witness family has $K=e^{\Omega(d)}$ elements. Proposition E.13 assigns it tabular capacity $K-1$ under unrestricted rewards but restricted capacity at most d under the code's linear reward span. This is the local geometric mechanism behind the exponential separation: identical optimizer witnesses can be statistically unrelated in one class and coupled through only d latent directions in the other.

E.6 A SINGULAR DESIGN WITH A SHARP APPROXIMATION THRESHOLD

The tolerance parameter in Definition E.2 is not cosmetic. Approximate estimation can remain possible when exact estimation is information-theoretically impossible.

Example E.14 (One unobserved tabular action)**.** Let $\mathcal{A}=\{1,2\}$, $\mathcal{G}=[0,1]^2$, $p=\delta_1$, and $h=\delta_2-\delta_1=(-1,1)$. Then

$$\mathcal{E}_\varepsilon(h;p)=\begin{cases}+\infty, & 0\le\varepsilon<1/2,\\ 4(1-\varepsilon)^2, & 1/2\le\varepsilon\le 1,\\ 0, & \varepsilon\ge 1.\end{cases} \tag{E.50}$$

Proof. By Proposition E.11, every finite-energy signed measure has the form $z=(z_1,0)$ and must satisfy

$$|z_1+1|+1\le 2\varepsilon.$$

The constraint is infeasible for $\varepsilon<1/2$. For $\varepsilon\in[1/2,1]$, the feasible interval is $[-2\varepsilon,2\varepsilon-2]$, whose point nearest zero is $z_1=2\varepsilon-2$; the objective is therefore $z_1^2=4(1-\varepsilon)^2$. For $\varepsilon\ge 1$, $z_1=0$ is feasible and has zero cost. □

At the critical tolerance $\varepsilon=1/2$, the estimator $1/2-Y$ has worst-case bias exactly $1/2$ for the unobserved contrast $g(2)-g(1)$ and stochastic cost one. No estimator based only on action 1 can achieve smaller uniform bias because the mean at action 2 can vary freely in $[0,1]$. The defect formula (E.28) recovers the same threshold.

E.7 WHAT THE RIESZ GEOMETRY DOES AND DOES NOT ESTABLISH

The results of this appendix are exact, static identities. They show that, for a fixed physical sampling distribution p and a fixed policy contrast h,

$$\boxed{\text{minimum one-sample stochastic budget} = \text{reward-class-restricted Riesz energy.}} \tag{E.51}$$

They also establish the exact dual obstruction (E.30), which will later be used to construct hard local reward alternatives.

These identities do not by themselves give an online regret bound. The Riesz coefficient can be signed and arbitrarily large on low-probability actions; a second-moment bound alone need not control the exponential potential used by a particular online algorithm. The next appendix therefore proves a one-scale structured-mediator theorem with an explicit regularizer and tail-control mechanism. Only after that theorem is established will the observation costs be composed across optimizer scales.

F ONE-SCALE RESTRICTED AFFINE EXO

This appendix converts the static one-observation geometry of Appendix E into an adaptive online-learning theorem for a fixed finite optimizer-witness family. A direct substitution of conditional second moments into the usual entropy-potential argument is not valid: the minimum-energy Riesz coefficients may be signed and arbitrarily large on actions of small sampling probability. We therefore begin with an exact exponential one-step certificate, specialized to affine reward estimators. The construction is a reward-only analogue of high-probability Exploration-by-Optimization (Foster et al., 2022); its robust moment-generating-function term is what allows the comparator to be selected in hindsight under an adaptive environment. We then prove that the Riesz design cost from Appendix E is the local quadratic geometry of this exact certificate. For unrestricted mediator feedback, the corresponding expected-regret geometry is the χ^2 policy-set capacity of Eldowa et al. (2024); the distinction between second-moment and exponential control is analogous to the role of implicit exploration in high-probability adversarial bandits (Neu, 2015).

Throughout this appendix, fix a finite nonempty witness family $Q \subseteq \Delta(\mathcal{A})$, write $N := |Q|$, and let $C_Q := \mathrm{co}(Q)$. The reward class $\mathcal{G}$ and the adaptive bounded conditional-mean protocol are those of Appendix A. We use the convention $\inf \varnothing = +\infty$.

F.1 MEAN-COMPATIBLE OBSERVATION LAWS AND AFFINE SCORE SYSTEMS

For $g \in \mathcal{G}$, let

$$\mathfrak{N}(g) := \prod_{a\in\mathcal{A}} \left\{ \nu_a \in \Delta([0,1]) : \int_0^1 y\, \nu_a(\mathrm{d}y) = g(a) \right\} \tag{F.1}$$

be the set of action-indexed observation laws with mean vector g. Since the action set is finite, an element $\nu = (\nu_a)_{a\in\mathcal{A}}$ specifies the conditional law $Y \mid A = a \sim \nu_a$.

Definition F.1 (Affine score system)**.** An affine score system is a pair

$$c \in \mathbb{R}^Q, \qquad \psi \in \mathbb{R}^{Q\times\mathcal{A}},$$

with score

$$S_{c,\psi}(q; a, y) := c(q) + \psi(q, a)y, \qquad q \in Q,\ a \in \mathcal{A},\ y \in [0,1]. \tag{F.2}$$

For $q, r \in Q$, define the pairwise score increment

$$D^{c,\psi}_{q,r}(a, y) := S_{c,\psi}(q; a, y) - S_{c,\psi}(r; a, y). \tag{F.3}$$

For $\eta > 0$, $\mu \in \Delta(Q)$, $p \in C_Q$, $q^\star \in Q$, and $g \in \mathcal{G}$, define the robust exponential increment

$$\Lambda_\eta(\mu, p, c, \psi; q^\star, g) := \sup_{\nu\in\mathfrak{N}(g)} \mathbb{E}_{\substack{A\sim p\\ Y\sim\nu_A}} \left[\sum_{q\in Q} \mu(q) \left(\exp\Big(\eta D^{c,\psi}_{q,q^\star}(A, Y) \Big) - 1 \right) \right]. \tag{F.4}$$

The supremum in (F.4) is not an additional stochastic assumption. It encodes exactly the worst conditional observation law allowed by the bounded conditional-mean protocol.

Proposition F.2 (Endpoint reduction for bounded conditional-mean noise)**.** *For every admissible tuple in* (F.4)*,*

$$\Lambda_\eta(\mu, p, c, \psi; q^\star, g) \tag{F.5}$$

$$= \sum_{a\in\mathcal{A}} p(a) \sum_{q\in Q} \mu(q) \Big[(1 - g(a)) \big(e^{\eta D^{c,\psi}_{q,q^\star}(a,0)} - 1 \big) + g(a) \big(e^{\eta D^{c,\psi}_{q,q^\star}(a,1)} - 1 \big) \Big]. \tag{F.6}$$

In particular, the supremum is attained by the endpoint law $Y \mid A = a \sim \mathrm{Ber}(g(a))$.

Proof. Fix $a \in \mathcal{A}$ and define

$$\Phi_a(y) := \sum_{q\in Q} \mu(q) \left(e^{\eta D^{c,\psi}_{q,q^\star}(a,y)} - 1 \right).$$

Each summand is the exponential of an affine function of y, hence Φ_a is convex on $[0,1]$. Therefore

$$\Phi_a(y) \le (1-y)\Phi_a(0) + y\Phi_a(1), \qquad y \in [0,1].$$

For any ν_a with mean $g(a)$, integration gives

$$\int \Phi_a(y)\,\nu_a(\mathrm{d}y) \le (1-g(a))\Phi_a(0) + g(a)\Phi_a(1).$$

Equality is achieved by the Bernoulli law supported on $\{0,1\}$. Summing with weights $p(a)$ proves (F.5). □

The endpoint reduction is exact for every affine score system, not merely an upper bound based on Hoeffding's lemma. It is the reason the one-scale theorem below remains valid under the full history-dependent observation class from Appendix A.

F.2 An exact affine Exploration-by-Optimization value

For $\mu \in \Delta(Q)$, define

$$\mathsf{aexo}_\eta(\mu; Q, \mathcal{G}) := \inf_{\substack{p \in C_Q,\ c \in \mathbb{R}^Q,\\ \psi \in \mathbb{R}^{Q\times\mathcal{A}}}} \sup_{\substack{q^\star \in Q\\ g \in \mathcal{G}}} \left\{ (q^\star - p)(g) + \frac{1}{\eta}\Lambda_\eta(\mu, p, c, \psi; q^\star, g) \right\}. \tag{F.7}$$

Its worst-reference value is

$$\mathsf{AExO}_\eta(Q, \mathcal{G}) := \sup_{\mu \in \Delta(Q)} \mathsf{aexo}_\eta(\mu; Q, \mathcal{G}). \tag{F.8}$$

The lower-case quantity is the optimization problem solved after the current exponential-weights distribution μ is known; the upper-case quantity is its uniform envelope.

Lemma F.3 (Finiteness and predictable approximate selectors)**.** *For every $\eta > 0$ and $\mu \in \Delta(Q)$, $\mathsf{aexo}_\eta(\mu; Q, \mathcal{G})$ is finite. Moreover, for every $\xi > 0$, there is a Borel map*

$$\mu \longmapsto (p_\xi(\mu), c_\xi(\mu), \psi_\xi(\mu))$$

with $p_\xi(\mu) \in C_Q$ whose objective in (F.7) *is at most $\mathsf{aexo}_\eta(\mu; Q, \mathcal{G}) + \xi$.*

Proof. Let $u_Q = N^{-1}\sum_{q\in Q} q \in C_Q$. If $u_Q(a) = 0$, then $q(a) = 0$ for every $q \in Q$. Otherwise $q(a)/u_Q(a) \le N$. Hence the choice

$$p = u_Q, \qquad c(q) = 0, \qquad \psi(q,a) = \begin{cases} q(a)/u_Q(a), & u_Q(a) > 0,\\ 0, & u_Q(a) = 0, \end{cases}$$

has finite scores on the compact observation interval. It gives a finite upper bound in (F.7). Conversely, $(q^\star - p)(g) \ge -1$ and $\Lambda_\eta > -1$, so the objective is bounded below by $-1 - \eta^{-1}$.

For the measurable-selection claim, enumerate a countable dense subset $\{z_m\}_{m\ge1}$ of $C_Q \times \mathbb{R}^Q \times \mathbb{R}^{Q\times\mathcal{A}}$, where $z_m = (p_m, c_m, \psi_m)$. By Proposition F.2, the objective after taking the supremum over $q^\star$ and g is continuous in (μ, z_m); compactness of $\mathcal{G}$ and finiteness of Q justify the maximum theorem. Continuity in z implies that its infimum over the dense sequence equals the infimum in (F.7). Select the first m whose value is at most the infimum plus ξ. Since a countable infimum of continuous functions is Borel, this first-index rule is Borel measurable. □

Remark F.4 (Relation to general ExO). The objective (F.7) is deliberately narrower than the general high-probability Exploration-by-Optimization objective of Foster et al. (2022): the decision space is the witness family, the physical experiment is an action distribution in C_Q, and the estimator is restricted to be affine in the one observed reward. The restriction yields the explicit endpoint formula (F.5) and permits a direct comparison with the Riesz geometry of Appendix E. We do not claim that (F.8) supersedes the general Decision–Estimation Coefficient.

F.3 RESTRICTED AFFINE EXO AND AN ADAPTIVE ONE-SCALE REGRET THEOREM

Fix $\eta > 0$ and a deterministic sequence $\xi_1, \dots, \xi_T > 0$. Initialize $W_1(q) = 1$ for every $q \in Q$. At round t, define

$$\mu_t(q) := \frac{W_t(q)}{\sum_{r \in Q} W_t(r)}. \tag{F.9}$$

Use Lemma F.3 to choose an $\mathscr{H}_{t-1}$-measurable ξ_t-optimal triple (p_t, c_t, ψ_t) for $\mathsf{aexo}_\eta(\mu_t; Q, \mathcal{G})$. Sample $A_t \sim p_t$, observe Y_t, form

$$\widehat{f}_t(q) := c_t(q) + \psi_t(q, A_t) Y_t, \tag{F.10}$$

and update

$$W_{t+1}(q) := W_t(q) e^{\eta \widehat{f}_t(q)}. \tag{F.11}$$

We call this procedure RESTRICTED AFFINE-EXO. It is information-theoretic: no polynomial-time claim is made for the saddle problem in (F.7).

Define the mixed mediator regret

$$\overline{\mathrm{Reg}}_T^Q := \max_{q \in Q} \sum_{t=1}^T q(g_t) - \sum_{t=1}^T p_t(g_t). \tag{F.12}$$

Its expectation equals that of Reg_T^Q, whereas their difference is the bounded action-sampling martingale

$$\mathrm{Reg}_T^Q - \overline{\mathrm{Reg}}_T^Q = \sum_{t=1}^T \big[p_t(g_t) - g_t(A_t)\big]. \tag{F.13}$$

Theorem F.5 (Adaptive one-scale restricted affine-ExO bound)**.** *For every finite nonempty Q, every $\eta > 0$, and every horizon T,*

$$\mathfrak{M}_T^{\star,\mathrm{ad}}(Q, \mathcal{G}) \le T\mathsf{AExO}_\eta(Q, \mathcal{G}) + \frac{2 \log N}{\eta}. \tag{F.14}$$

More precisely, RESTRICTED AFFINE-EXO *with oracle errors (ξ_t) satisfies*

$$\sup_{\mathsf{Adv} \in \mathfrak{C}_T^{\mathrm{ad}}(\mathcal{G})} \mathbb{E}_{\mathsf{Alg},\mathsf{Adv}}[\mathrm{Reg}_T^Q] \le T\mathsf{AExO}_\eta(Q, \mathcal{G}) + \sum_{t=1}^T \xi_t + \frac{2 \log N}{\eta}. \tag{F.15}$$

For every $\delta \in (0,1)$, under the same adaptive bounded-noise protocol, with probability at least $1 - \delta$,

$$\begin{aligned} \mathrm{Reg}_T^Q \le T\mathsf{AExO}_\eta(Q, \mathcal{G}) + \sum_{t=1}^T \xi_t + \frac{\log N + \log(2N/\delta)}{\eta} \\ + \sqrt{2T \log(2/\delta)}. \end{aligned} \tag{F.16}$$

When $N = 1$, the convention $\log N = 0$ applies.

The proof is self-contained and separates the deterministic exponential-weights potential from the stochastic exponential certificate.

Lemma F.6 (Exponential-weights potential)**.** *Let*

$$L_t := \frac{1}{\eta} \log \left(\sum_{q \in Q} \mu_t(q) e^{\eta \widehat{f}_t(q)} \right). \tag{F.17}$$

For every $q^\star \in Q$, pathwise,

$$\sum_{t=1}^T \big[\widehat{f}_t(q^\star) - L_t\big] \le \frac{\log N}{\eta}. \tag{F.18}$$

Proof. The update (F.11) gives

$$\sum_q W_{t+1}(q) = \left(\sum_q W_t(q)\right) \sum_q \mu_t(q) e^{\eta \widehat{f}_t(q)}.$$

Iterating and using $\sum_q W_1(q) = N$ yields

$$\sum_q W_{T+1}(q) = N \exp\left(\eta \sum_{t=1}^{T} L_t\right).$$

Since $W_{T+1}(q^\star) = \exp(\eta \sum_t \widehat{f}_t(q^\star))$ is one term in the left-hand sum, (F.18) follows after taking logarithms. □

For the selected triple, set

$$d_t := \mathsf{aexo}_\eta(\mu_t; Q, \mathcal{G}) + \xi_t \le \mathsf{AExO}_\eta(Q, \mathcal{G}) + \xi_t. \tag{F.19}$$

For $q^\star \in Q$, define

$$Z_t(q^\star) := (q^\star - p_t)(g_t) + L_t - \widehat{f}_t(q^\star) - d_t. \tag{F.20}$$

Lemma F.7 (Conditional exponential certificate)**.** *For every $q^\star \in Q$ and every round t,*

$$\mathbb{E}_t\left[e^{\eta Z_t(q^\star)}\right] \le 1. \tag{F.21}$$

Consequently,

$$\exp\left(\eta \sum_{s=1}^{t} Z_s(q^\star)\right), \qquad t = 0, \dots, T,$$

is a nonnegative supermartingale.

Proof. Condition on $\mathscr{H}_{t-1}$. The quantities $\mu_t, p_t, c_t, \psi_t, g_t$, and d_t are then deterministic. By (F.17),

$$\begin{aligned}\mathbb{E}_t\left[e^{\eta(L_t - \widehat{f}_t(q^\star))}\right] &= \mathbb{E}_t\left[\sum_{q\in Q} \mu_t(q) e^{\eta(\widehat{f}_t(q) - \widehat{f}_t(q^\star))}\right] \\ &\le 1 + \Lambda_\eta(\mu_t, p_t, c_t, \psi_t; q^\star, g_t),\end{aligned}$$

where the inequality holds because the realized conditional observation kernel is an element of $\mathfrak{N}(g_t)$. The weighted sum of exponentials defining $1 + \Lambda_\eta$ is strictly positive, hence $\Lambda_\eta > -1$. Therefore $1 + x \le e^x$ is applicable and

$$\begin{aligned}\mathbb{E}_t[e^{\eta Z_t(q^\star)}] &\le \exp(\eta[(q^\star - p_t)(g_t) - d_t] + \Lambda_\eta(\mu_t, p_t, c_t, \psi_t; q^\star, g_t)) \\ &\le 1,\end{aligned}$$

where the final inequality is exactly the ξ_t-optimality condition in (F.7). The supermartingale claim follows by iteration. □

Proof of Theorem F.5. For every $q^\star \in Q$, combining (F.20) and Lemma F.6 gives

$$\sum_{t=1}^{T} (q^\star - p_t)(g_t) \le \sum_{t=1}^{T} d_t + \sum_{t=1}^{T} Z_t(q^\star) + \frac{\log N}{\eta}. \tag{F.22}$$

By Lemma F.7, $\mathbb{E}[\exp(\eta \sum_t Z_t(q))] \le 1$ for each q. Hence

$$\begin{aligned}\mathbb{E}\left[\max_{q\in Q} \sum_{t=1}^{T} Z_t(q)\right] &\le \frac{1}{\eta} \mathbb{E}\left[\log \sum_{q\in Q} e^{\eta \sum_t Z_t(q)}\right] \\ &\le \frac{1}{\eta} \log \sum_{q\in Q} \mathbb{E} e^{\eta \sum_t Z_t(q)} \le \frac{\log N}{\eta}.\end{aligned} \tag{F.23}$$

Taking the maximum over $q^\star$ in (F.22), expectation, and using (F.19) proves (F.15) for $\overline{\mathrm{Reg}}_T^Q$. Equation (F.13) has conditional mean zero, so the same expected bound holds for Reg_T^Q. Since the oracle errors may be chosen with arbitrarily small total sum, taking the infimum over algorithms yields (F.14).

For the high-probability claim, the supermartingale bound and Markov's inequality imply

$$\max_{q\in Q}\sum_{t=1}^{T} Z_t(q) \le \frac{\log(2N/\delta)}{\eta} \tag{F.24}$$

with probability at least $1-\delta/2$. Indeed, if the maximum exceeded the right-hand side, then $\sum_q e^{\eta\sum_t Z_t(q)} > 2N/\delta$, whereas the expectation of this sum is at most N. Furthermore, the increments $p_t(g_t) - g_t(A_t)$ are conditionally mean zero and lie in $[-1,1]$, so Azuma's inequality gives

$$\sum_{t=1}^{T}[p_t(g_t) - g_t(A_t)] \le \sqrt{2T\log(2/\delta)} \tag{F.25}$$

with probability at least $1-\delta/2$. Combining (F.22), (F.19), (F.24), (F.25), and a union bound proves (F.16). □

Remark F.8 (Why a uniform exponential certificate is necessary). For an adaptive reward sequence, the best witness in hindsight is random. A conditional-unbiasedness argument for each deterministic q controls $\mathbb{E}[\sum_t (q-p_t)(g_t)]$ separately but does not justify exchanging expectation with $\max_{q\in Q}$. Lemma F.7 constructs one supermartingale for every comparator and controls their maximum uniformly. This is precisely the role played by high-probability ExO in the more general framework of Foster et al. (2022).

F.4 Quadratic envelopes from the Riesz design cost

The exact value (F.8) is nonlinear in the score tails. We next identify a regime in which it is upper bounded by the Riesz geometry of Appendix E.

Fix $\mu \in \Delta(Q)$, write

$$\bar{q}_\mu := \sum_{q\in Q}\mu(q)q, \qquad h_q^\mu := q - \bar{q}_\mu \in \mathsf{H}_0, \tag{F.26}$$

and for an affine score system define its conditional mean under (p,g) by

$$m_{p,c,\psi}(q;g) := c(q) + \sum_{a\in\mathcal{A}} p(a)\psi(q,a)g(a). \tag{F.27}$$

Its pairwise oscillation is

$$\begin{aligned}\mathrm{osc}(c,\psi) &:= \max_{q,r\in Q}\max_{a\in\mathcal{A}}\max_{y\in[0,1]} |D_{q,r}^{c,\psi}(a,y)| \\ &= \max_{q,r,a}\max\left\{|c(q)-c(r)|, |c(q)-c(r)+\psi(q,a)-\psi(r,a)|\right\}.\end{aligned} \tag{F.28}$$

The second equality holds because the absolute value of an affine function is convex and hence maximized at an endpoint.

Definition F.9 (Bias- and range-controlled score systems)**.** For $\varepsilon \ge 0$, $L \in [0,\infty]$, $\mu \in \Delta(Q)$, and $p \in C_Q$, let $\mathscr{S}_{\varepsilon,L}(\mu,p)$ be the set of affine score systems (c,ψ) satisfying

$$\max_{q\in Q}\sup_{g\in\mathcal{G}}\left|m_{p,c,\psi}(q;g) - h_q^\mu(g)\right| \le \varepsilon \tag{F.29}$$

and $\mathrm{osc}(c,\psi) \le L$. The oscillation constraint is omitted when $L=\infty$.

For a score system, define its robust pairwise second-moment cost

$$\mathsf{V}_\mu(p,c,\psi) := \sup_{\substack{q^\star\in Q\\ g\in\mathcal{G}}}\sum_{a\in\mathcal{A}} p(a)\sum_{q\in Q}\mu(q)\Big[(1-g(a))D_{q,q^\star}^{c,\psi}(a,0)^2 \tag{F.30}$$

$$+ g(a)D_{q,q^\star}^{c,\psi}(a,1)^2\Big]. \tag{F.31}$$

Exactly as in Proposition F.2, this is the supremum of $\mathbb{E}[\sum_q \mu(q) D_{q,q^\star}(A,Y)^2]$ over $\nu \in \mathfrak{N}(g)$. Define

$$\mathfrak{B}_{\varepsilon,L}(\mu,p) := \inf_{(c,\psi)\in\mathscr{S}_{\varepsilon,L}(\mu,p)} \mathsf{V}_\mu(p,c,\psi), \tag{F.32}$$

$$\mathfrak{U}_{\varepsilon,L}(\mu,p) := \inf_{(c,\psi)\in\mathscr{S}_{\varepsilon,L}(\mu,p)} \max_{q\in Q} \|\psi(q,\cdot)\|_p^2 . \tag{F.33}$$

Theorem F.10 (Quadratic envelope of restricted affine ExO)**.** *Let $\varepsilon \geq 0$, $L < \infty$, and $\eta L \leq 1$. Then, for every $\mu \in \Delta(Q)$,*

$$\mathsf{aexo}_\eta(\mu; Q, \mathcal{G}) \leq \inf_{p\in C_Q} \left\{ d_{\mathcal{G}}^+(\bar{q}_\mu, p) + 2\varepsilon + \eta \mathfrak{B}_{\varepsilon,L}(\mu,p) \right\}, \tag{F.34}$$

where $d_{\mathcal{G}}^+$ is the reward-class hemimetric from (D.6)*. Moreover,*

$$\mathfrak{B}_{\varepsilon,L}(\mu,p) \leq 4\mathfrak{U}_{\varepsilon,L}(\mu,p) + (1+2\varepsilon)^2, \tag{F.35}$$

and, without an oscillation constraint,

$$\mathfrak{U}_{\varepsilon,\infty}(\mu,p) = \max_{q\in Q} \mathcal{E}_\varepsilon(q - \bar{q}_\mu; p) . \tag{F.36}$$

If the right-hand side of (F.36) *is finite, then there is a finite $L_0 = L_0(\mu,p,\varepsilon)$ such that equality with the same value holds in* (F.36) *after replacing ∞ by every $L \geq L_0$.*

Proof. Fix $p \in C_Q$ and $(c,\psi) \in \mathscr{S}_{\varepsilon,L}(\mu,p)$. For $q^\star \in Q$ and $g \in \mathcal{G}$, let the expectation below be under any mean-compatible law. The elementary inequality $e^x \leq 1 + x + x^2$ for $x \leq 1$, together with $|\eta D_{q,q^\star}^{c,\psi}| \leq \eta L \leq 1$, gives

$$\frac{1}{\eta}\Lambda_\eta(\mu,p,c,\psi;q^\star,g) \leq \sum_q \mu(q)\mathbb{E}[D_{q,q^\star}^{c,\psi}(A,Y)] + \eta \sum_q \mu(q)\mathbb{E}[D_{q,q^\star}^{c,\psi}(A,Y)^2]. \tag{F.37}$$

Write

$$e_q(g) := m_{p,c,\psi}(q;g) - h_q^\mu(g), \qquad |e_q(g)| \leq \varepsilon.$$

Since $\sum_q \mu(q) h_q^\mu = 0$,

$$\begin{aligned}
&(q^\star - p)(g) + \sum_q \mu(q)\mathbb{E}[D_{q,q^\star}^{c,\psi}(A,Y)] \\
&\quad = (\bar{q}_\mu - p)(g) + \sum_q \mu(q) e_q(g) - e_{q^\star}(g) \\
&\quad \leq d_{\mathcal{G}}^+(\bar{q}_\mu, p) + 2\varepsilon.
\end{aligned}$$

Combining this with (F.37), taking the supremum over $q^\star, g$, the infimum over admissible score systems, and then the infimum over p proves (F.34).

To prove (F.35), fix an admissible score system and let $R := \max_q \|\psi(q,\cdot)\|_p^2$. For every $q, q^\star, g$ and every mean-compatible observation law,

$$\begin{aligned}
\mathrm{Var}\left(D_{q,q^\star}^{c,\psi}(A,Y)\right) &\leq \sum_a p(a)\big(\psi(q,a) - \psi(q^\star,a)\big)^2 \\
&\leq 4R.
\end{aligned}$$

Furthermore, by (F.29),

$$\left|\mathbb{E}[D_{q,q^\star}^{c,\psi}(A,Y)]\right| \leq |(q-q^\star)(g)| + 2\varepsilon \leq 1 + 2\varepsilon.$$

Thus every pairwise second moment is at most $4R + (1+2\varepsilon)^2$. Averaging over q, taking the supremum, and then taking the infimum over score systems proves (F.35).

When $L = \infty$, condition (F.29) separates over $q \in Q$, and the minimum squared coefficient norm for the qth constraint is exactly $\mathcal{E}_\varepsilon(q - \bar{q}_\mu; p)$ by Definition E.2. Taking the maximum proves (F.36). If all these design costs are finite, Proposition E.10 gives an attaining pair (c_q, ψ_q) for every q. Since Q and $\mathcal{A}$ are finite, the resulting family has finite pairwise oscillation; taking L_0 to be that oscillation completes the proof. □

Corollary F.11 (One-scale Riesz envelope)**.** *For every $\eta > 0$, $\varepsilon \geq 0$, and $L < \infty$ satisfying $\eta L \leq 1$,*

$$\mathfrak{M}_T^{\star,\mathrm{ad}}(Q,\mathcal{G}) \leq \frac{2\log N}{\eta} \tag{F.38}$$

$$+ T \sup_{\mu\in\Delta(Q)} \inf_{p\in C_Q} \Big\{ d_{\mathcal{G}}^{+}(\bar{q}_\mu, p) + 2\varepsilon \tag{F.39}$$

$$+ \eta\big[4\mathfrak{U}_{\varepsilon,L}(\mu,p) + (1+2\varepsilon)^2\big]\Big\}.$$

The same upper bound holds for the oblivious mediator value.

Proof. Combine Theorems F.5 and F.10 with (F.35). The oblivious value is no larger than the adaptive value. □

Corollary F.11 is deliberately stated with the range-controlled radius $\mathfrak{U}_{\varepsilon,L}$. Dropping the oscillation constraint while retaining only an average second moment is generally insufficient for a uniform exponential certificate.

F.5 The Riesz geometry as the local quadratic tangent

The preceding upper bound used a coarse exponential inequality. The next result identifies the exact second-order expansion.

Proposition F.12 (Local expansion of the exponential certificate)**.** *Fix $\mu, p, c, \psi, q^\star, g$ and a mean-compatible observation law. Suppose $\mathrm{osc}(c,\psi) \leq L$ and $\eta L \leq 1$. Then*

$$\left| \frac{1}{\eta}\mathbb{E}\left[\sum_{q\in Q}\mu(q)\left(e^{\eta D^{c,\psi}_{q,q^\star}(A,Y)} - 1\right)\right] - \sum_q \mu(q)\mathbb{E}[D^{c,\psi}_{q,q^\star}(A,Y)] \right.$$

$$\left. - \frac{\eta}{2}\sum_q \mu(q)\mathbb{E}[D^{c,\psi}_{q,q^\star}(A,Y)^2] \right| \leq \frac{e}{6}\eta^2 L^3. \tag{F.40}$$

Consequently, on every bounded score family, the first nonvanishing stochastic term in the affine ExO objective is the robust pairwise second moment from (F.30)*.*

Proof. For $|x| \leq 1$, Taylor's theorem with remainder gives

$$|e^x - 1 - x - x^2/2| \leq \frac{e}{6}|x|^3.$$

Apply this pointwise with $x = \eta D^{c,\psi}_{q,q^\star}(A,Y)$, average over q, A, Y, divide by η, and use $|D^{c,\psi}_{q,q^\star}| \leq L$. □

The proposition shows exactly how Appendix E enters the online problem. Bias-compatible affine scores have first-order expectation $(q - \bar{q}_\mu)(g)$; after cancellation with instantaneous decision regret, their second-order cost is governed by the minimum observation energy. The exact exponential objective retains the higher-order tail information needed outside the small-ηL regime.

F.6 Average capacity does not control comparator-uniform tails

The restricted capacity in (E.45) averages Riesz energies under the reference law μ. The adaptive theorem, by contrast, must control every hindsight comparator. The difference is already visible in the ordinary multi-armed bandit.

Proposition F.13 (Average–uniform gap for Dirac witnesses)**.** *Let $\mathcal{G} = [0,1]^{\mathcal{A}}$, $Q = Q_{\mathcal{A}} = \{\delta_a : a \in \mathcal{A}\}$, and let $\mu \in \Delta(Q)$ have full support. Identify μ with its induced action distribution p. Then, for every $a \in \mathcal{A}$,*

$$\mathcal{E}_0(\delta_a - p; p) = \frac{1}{p(a)} - 1, \tag{F.41}$$

while

$$\sum_a p(a)\mathcal{E}_0(\delta_a - p; p) = K - 1, \qquad \max_a \mathcal{E}_0(\delta_a - p; p) = \frac{1}{\min_a p(a)} - 1. \tag{F.42}$$

The average is independent of how close p is to the boundary of the simplex, but the comparator-uniform radius diverges as $\min_a p(a) \downarrow 0$.

Proof. The tabular specialization (E.39) gives

$$\mathcal{E}_0(\delta_a - p; p) = \frac{(1 - p(a))^2}{p(a)} + \sum_{b \neq a} \frac{p(b)^2}{p(b)} = \frac{1}{p(a)} - 1.$$

Averaging yields $\sum_a p(a)(1/p(a) - 1) = K - 1$, and the maximum is immediate. □

Proposition F.13 does not say that high-probability K-armed regret must depend on $1/\min_a p(a)$. Rather, it shows that a theorem based only on the naive maximum of unbiased Riesz energies would be too crude. Classical implicit-exploration estimators alter the bias–tail tradeoff precisely to avoid forced uniform exploration (Neu, 2015); general high-probability ExO instead optimizes an exponential certificate directly (Foster et al., 2022). The exact affine value (F.8) follows the latter route, while Theorem F.10 identifies the regime in which it can be controlled by the Riesz geometry.

F.7 Scope of the one-scale theorem

Theorem F.5 is an exact adaptive-adversary upper theorem for a fixed witness family. It requires neither unbiased scores nor bounded importance ratios: all tail behavior is accounted for by the robust exponential objective. Theorem F.10 then proves that the design cost of Appendix E is the local quadratic geometry of this objective under an explicit oscillation condition.

Three limitations are intentional. First, the theorem is information-theoretic and assumes access to approximate solutions of (F.7). Second, the quadratic envelope is a sufficient bound, not a claim that average Riesz capacity alone characterizes adaptive mediator regret. Third, this appendix contains no multiscale argument: Q is treated as one flat witness family. The next appendix will use the exact exponential certificate as the primitive object and couple several approximation levels through a single physical sampling distribution.

G Shared-Observation Chained Exploration by Optimization

This appendix couples the one-scale exponential certificates of Appendix F across a hierarchy of optimizer approximations. The central difficulty is that every scale must be learned from the *same* physical bandit observation. Running a separate bandit algorithm only when its node is visited would introduce inverse node-probability factors, while using conditional second moments without a comparator-uniform exponential certificate would fail for adaptive rewards. Our construction avoids both issues: every level maintains its own exponential potential, but all levels use one jointly optimized sampling law and are updated from the same observation. The resulting theorem is an observation-constrained analogue of algorithmic chaining under partial feedback (Cesa-Bianchi et al., 2017); unlike full-information sequential majorizing-measure bounds (Block et al., 2021), the local radii below are costs of implementable one-point experiments. The robust one-step primitive remains the affine Exploration-by-Optimization certificate from Appendix F, itself a specialization of the general adversarial ExO framework of Foster et al. (2022).

Throughout, fix a finite nonempty optimizer-witness family $Q \subseteq \Delta(\mathcal{A})$, let $C_Q := \mathrm{co}(Q)$, and write $N := |Q|$. All upper bounds use the adaptive bounded conditional-mean protocol of Appendix A. We use the conventions $\max \varnothing = -\infty$, $\inf \varnothing = +\infty$, and an empty sum equals zero.

G.1 Approximation chains and a deterministic telescoping inequality

Definition G.1 (Finite approximation chain)**.** A depth-J approximation chain for Q is a tuple

$$\mathscr{T} = \big((Q_j)_{j=0}^J, (\pi_j)_{j=0}^J, (\mathrm{par}_j)_{j=0}^{J-1}\big)$$

with the following properties.

1. $J \in \mathbb{N}_0$, every $Q_j \subseteq C_Q$ is finite, and $|Q_0| = 1$.
2. $\pi_j : Q \to Q_j$ and, writing $Q_0 = \{q_0\}$, $\pi_0(q) = q_0$ for every $q \in Q$.
3. For every $j < J$, there is a parent map $\mathrm{par}_j : Q_{j+1} \to Q_j$ such that $\pi_j = \mathrm{par}_j \circ \pi_{j+1}$.

The chain is *exact* when $\pi_J(q) = q$ for every $q \in Q$. For $j = 1, \dots, J$, define the level-j increment

$$d_j(q) := \pi_j(q) - \pi_{j-1}(q) \in \mathsf{H}_0, \tag{G.1}$$

and its finite increment class

$$\mathcal{D}_j(\mathscr{T}) := \{d_j(q) : q \in Q\}, \qquad N_j(\mathscr{T}) := |\mathcal{D}_j(\mathscr{T})|. \tag{G.2}$$

The terminal radius is

$$\mathrm{rad}_{\mathcal{G}}(\mathscr{T}) := \max_{q \in Q} d^+_{\mathcal{G}}\big(q, \pi_J(q)\big), \tag{G.3}$$

where $d^+_{\mathcal{G}}$ is the reward-class hemimetric from (D.6).

We call the chain *Talagrand-admissible* when $|Q_j| \le 2^{2^j}$ for every $j \ge 1$. This optional cardinality convention is standard in generic chaining and normalizes the amount of new combinatorial information introduced at each scale. None of the upper bounds below requires it: the theorems hold for arbitrary finite chains, so in particular every finite Q admits an exact depth-one chain. Later entropy estimates may restrict the infimum to Talagrand-admissible chains when doing so is useful.

Lemma G.2 (Pathwise chained comparator reduction)**.** *Let $\mathscr{T}$ be a depth-J approximation chain. For every deterministic sequence $g_1, \dots, g_T \in \mathcal{G}$ and every sequence $p_1, \dots, p_T \in C_Q$,*

$$\max_{q \in Q} \sum_{t=1}^{T} (q - p_t)(g_t) \le T \,\mathrm{rad}_{\mathcal{G}}(\mathscr{T}) + \sum_{t=1}^{T} (q_0 - p_t)(g_t) + \sum_{j=1}^{J} \max_{d \in \mathcal{D}_j(\mathscr{T})} \sum_{t=1}^{T} d(g_t). \tag{G.4}$$

Proof. For every $q \in Q$ and $g \in \mathcal{G}$, telescoping gives

$$(q - p)(g) = (q - \pi_J(q))(g) + (q_0 - p)(g) + \sum_{j=1}^{J} d_j(q)(g).$$

Since $g \in \mathcal{G} \subseteq \mathcal{G}^0$,

$$(q - \pi_J(q))(g) \le d^+_{\mathcal{G}}(q, \pi_J(q)) \le \mathrm{rad}_{\mathcal{G}}(\mathscr{T}).$$

Sum over t, maximize over q, and upper bound the maximum of the sum of level terms by the sum of the levelwise maxima. □

Lemma G.2 deliberately relaxes path consistency in the final step. This relaxation is what permits independent exponential potentials at the different scales. Consistency is not lost statistically: the physical sampling law below is optimized jointly for all levels, and a separate discrepancy term measures any incompatibility among their current mean increments.

G.2 Increment-level exponential certificates and the joint design value

Let $\mathcal{D} \subseteq \mathsf{H}_0$ be finite and nonempty. An affine score system on $\mathcal{D}$ is a pair $c \in \mathbb{R}^{\mathcal{D}}$ and $\psi \in \mathbb{R}^{\mathcal{D} \times \mathcal{A}}$ with

$$S_{c,\psi}(d; a, y) := c(d) + \psi(d, a) y.$$

For $d, r \in \mathcal{D}$, write $D^{c,\psi}_{d,r}(a, y) := S_{c,\psi}(d; a, y) - S_{c,\psi}(r; a, y)$. Given $\eta > 0$, $\mu \in \Delta(\mathcal{D})$, $p \in C_Q$, $d^\star \in \mathcal{D}$, and $g \in \mathcal{G}$, define

$$\Lambda^{\mathcal{D}}_\eta(\mu, p, c, \psi; d^\star, g) := \sup_{\nu \in \mathfrak{N}(g)} \mathbb{E}_{\substack{A \sim p \\ Y \sim \nu_A}} \left[\sum_{d \in \mathcal{D}} \mu(d) \left(e^{\eta D^{c,\psi}_{d,d^\star}(A,Y)} - 1 \right) \right]. \tag{G.5}$$

By Proposition F.2, applied with Q replaced by $\mathcal{D}$, this supremum is attained by conditionally Bernoulli observations and has the endpoint representation

$$\Lambda_\eta^{\mathcal{D}}(\mu,p,c,\psi;d^\star,g)=\sum_a p(a)\sum_d \mu(d)\Big[(1-g(a))\big(e^{\eta D^{c,\psi}_{d,d^\star}(a,0)}-1\big) + g(a)\big(e^{\eta D^{c,\psi}_{d,d^\star}(a,1)}-1\big)\Big]. \tag{G.6}$$

Fix a depth-J chain $\mathscr{T}$ and a vector of learning rates $\boldsymbol{\eta}=(\eta_1,\dots,\eta_J)\in(0,\infty)^J$. For a tuple of reference laws

$$\boldsymbol{\mu}=(\mu_1,\dots,\mu_J)\in\prod_{j=1}^J \Delta(\mathcal{D}_j(\mathscr{T}))\,,$$

For $p\in C_Q$, score systems $(c_j,\psi_j)_{j=1}^J$, a reward vector g, and comparators $d_j^\star\in\mathcal{D}_j$, set

$$\mathcal{L}_{\boldsymbol{\eta}}\big(p,(c_j,\psi_j)_j;g,(d_j^\star)_j\big):=(q_0-p)(g)+\sum_{j=1}^J\left[d_j^\star(g)+\frac{1}{\eta_j}\Lambda_{\eta_j}^{\mathcal{D}_j}(\mu_j,p,c_j,\psi_j;d_j^\star,g)\right]. \tag{G.7}$$

Define the local chained affine-ExO value by

$$\mathsf{caexo}_{\boldsymbol{\eta}}(\boldsymbol{\mu};\mathscr{T},\mathcal{G}):=\inf_{\substack{p\in C_Q,\\(c_j,\psi_j)_{j=1}^J}}\ \sup_{\substack{g\in\mathcal{G},\\ d_j^\star\in\mathcal{D}_j,\ 1\le j\le J}}\mathcal{L}_{\boldsymbol{\eta}}\big(p,(c_j,\psi_j)_j;g,(d_j^\star)_j\big), \tag{G.8}$$

where $c_j\in\mathbb{R}^{\mathcal{D}_j}$ and $\psi_j\in\mathbb{R}^{\mathcal{D}_j\times\mathcal{A}}$. Here and below, $\mathcal{D}_j=\mathcal{D}_j(\mathscr{T})$. Its uniform envelope is

$$\mathsf{CAExO}_{\boldsymbol{\eta}}(\mathscr{T},\mathcal{G}):=\sup_{\boldsymbol{\mu}\in\prod_{j=1}^J\Delta(\mathcal{D}_j)}\mathsf{caexo}_{\boldsymbol{\eta}}(\boldsymbol{\mu};\mathscr{T},\mathcal{G}). \tag{G.9}$$

When $J=0$, we set

$$\mathsf{CAExO}_{\varnothing}(\mathscr{T},\mathcal{G}):=\inf_{p\in C_Q}\sup_{g\in\mathcal{G}}(q_0-p)(g)=0.$$

The supremum in (G.8) uses one common reward vector g and one common physical law p across all levels. This is the implementable coupled quantity: the same observation may be informative about many increments simultaneously, while the shared sampling constraint prevents the fictitious use of a different physical experiment at every level.

Lemma G.3 (Finiteness and predictable joint selectors)**.** *For every chain $\mathscr{T}$, every $\boldsymbol{\eta}\in(0,\infty)^J$, and every reference tuple $\boldsymbol{\mu}$, the value in* (G.8) *is finite. Moreover, for every $\xi>0$ there is a Borel map from $\boldsymbol{\mu}$ to a ξ-optimal joint design*

$$\boldsymbol{\mu}\longmapsto\big(p_\xi(\boldsymbol{\mu}),(c_{j,\xi}(\boldsymbol{\mu}),\psi_{j,\xi}(\boldsymbol{\mu}))_{j=1}^J\big).$$

Proof. Let $u_Q=N^{-1}\sum_{q\in Q}q\in C_Q$. Every element of every Q_j belongs to C_Q and is therefore supported on $\operatorname{supp}(u_Q)$. Hence every $d\in\mathcal{D}_j$ is supported on $\operatorname{supp}(u_Q)$. The choice

$$p=u_Q,\qquad c_j(d)=0,\qquad \psi_j(d,a)=\begin{cases}d(a)/u_Q(a), & u_Q(a)>0,\\ 0, & u_Q(a)=0,\end{cases}$$

has finite scores and satisfies $\mathbb{E}[\psi_j(d,A)Y]=d(g)$ under every mean-compatible observation law. Thus the objective in (G.8) is finite. A finite lower bound follows from $(q_0-p)(g)\ge-1$, $d_j^\star(g)\ge-1$, and $\Lambda_{\eta_j}^{\mathcal{D}_j}>-1$.

For measurability, enumerate a countable dense subset of the finite-dimensional decision space

$$C_Q\times\prod_{j=1}^J\left(\mathbb{R}^{\mathcal{D}_j}\times\mathbb{R}^{\mathcal{D}_j\times\mathcal{A}}\right).$$

By the endpoint formula (G.6), the objective after taking the finite maximum over $(d_1^\star,\dots,d_J^\star)$ and the maximum over the compact set $\mathcal{G}$ is continuous in the reference tuple and every fixed decision vector. Its infimum over the dense sequence equals the infimum in (G.8). Selecting the first element within ξ of that infimum gives a Borel map, exactly as in Lemma F.3. □

G.3 THE CHAINED-EXO ALGORITHM AND ADAPTIVE REGRET THEOREM

Fix a chain $\mathscr{T}$, learning rates $\boldsymbol{\eta}$, and deterministic optimization errors $\xi_1, \dots, \xi_T > 0$. For every level j, initialize $W_{j,1}(d) = 1$ for $d \in \mathcal{D}_j$. At round t, form

$$\mu_{j,t}(d) := \frac{W_{j,t}(d)}{\sum_{r \in \mathcal{D}_j} W_{j,t}(r)}, \qquad j = 1, \dots, J. \tag{G.10}$$

Using Lemma G.3, choose an $\mathscr{H}_{t-1}$-measurable ξ_t-optimal joint design

$$\big(p_t, (c_{j,t}, \psi_{j,t})_{j=1}^J\big)$$

for $\mathsf{caexo}_{\boldsymbol{\eta}}((\mu_{j,t})_{j=1}^J; \mathscr{T}, \mathcal{G})$. Sample $A_t \sim p_t$, observe Y_t, and define, for every level and increment,

$$\widehat{f}_{j,t}(d) := c_{j,t}(d) + \psi_{j,t}(d, A_t) Y_t. \tag{G.11}$$

Update all levels from this same observation:

$$W_{j,t+1}(d) := W_{j,t}(d) e^{\eta_j \widehat{f}_{j,t}(d)}. \tag{G.12}$$

We call the resulting information-theoretic procedure CHAINED-EXO.

Define the observation-weighted chained complexity

$$\Gamma_T^{\mathrm{obs}}(Q, \mathcal{G}) := \inf_{\substack{J \in \mathbb{N}_0,\ \mathscr{T} \text{ a finite chain for } Q, \\ \boldsymbol{\eta} \in (0,\infty)^J}} \left\{ T \operatorname{rad}_{\mathcal{G}}(\mathscr{T}) + T\mathsf{CAExO}_{\boldsymbol{\eta}}(\mathscr{T}, \mathcal{G}) + 2 \sum_{j=1}^J \frac{\log N_j(\mathscr{T})}{\eta_j} \right\}. \tag{G.13}$$

Theorem G.4 (Chained-ExO upper bound)**.** *For every finite nonempty $Q \subseteq \Delta(\mathcal{A})$,*

$$\mathfrak{M}_T^{\star,\mathrm{ad}}(Q, \mathcal{G}) \le \Gamma_T^{\mathrm{obs}}(Q, \mathcal{G}). \tag{G.14}$$

Consequently,

$$\mathfrak{R}_T^{\star,\mathrm{ad}}(\mathcal{G}) \le \inf_{\substack{Q \subseteq \Delta(\mathcal{A}) \\ 1 \le |Q| < \infty}} \left\{ T b_{\mathcal{G}}(Q) + \Gamma_T^{\mathrm{obs}}(Q, \mathcal{G}) \right\}. \tag{G.15}$$

More precisely, for every fixed chain and learning-rate vector, CHAINED-EXO *satisfies*

$$\sup_{\mathsf{Adv} \in \mathfrak{C}_T^{\mathrm{ad}}(\mathcal{G})} \mathbb{E}_{\mathsf{Alg},\mathsf{Adv}}[\mathrm{Reg}_T^Q] \le T \operatorname{rad}_{\mathcal{G}}(\mathscr{T}) + T\mathsf{CAExO}_{\boldsymbol{\eta}}(\mathscr{T}, \mathcal{G}) + \sum_{t=1}^T \xi_t + 2 \sum_{j=1}^J \frac{\log N_j}{\eta_j}. \tag{G.16}$$

The theorem contains no inverse node-probability term: every level is updated on every round, and the sole physical experiment p_t is selected by the joint problem (G.8).

Theorem G.5 (High-probability chained bound)**.** *Let $N_\Sigma := \sum_{j=1}^J N_j$, with $N_\Sigma = 0$ when $J = 0$. For every $\delta \in (0,1)$, with probability at least $1 - \delta$,*

$$\mathrm{Reg}_T^Q \le T \operatorname{rad}_{\mathcal{G}}(\mathscr{T}) + T\mathsf{CAExO}_{\boldsymbol{\eta}}(\mathscr{T}, \mathcal{G}) + \sum_{t=1}^T \xi_t + \sum_{j=1}^J \frac{\log N_j + \log(2N_\Sigma/\delta)}{\eta_j} + \sqrt{2T \log(2/\delta)}. \tag{G.17}$$

When $J = 0$, the sum over levels is absent and the logarithm involving N_Σ is not used.

G.4 Proof of the chained upper bounds

For each level define the log-partition increment

$$L_{j,t} := \frac{1}{\eta_j} \log \left(\sum_{d \in \mathcal{D}_j} \mu_{j,t}(d) e^{\eta_j \widehat{f}_{j,t}(d)} \right). \tag{G.18}$$

The usual potential argument gives, pathwise, for every $d^\star \in \mathcal{D}_j$,

$$\sum_{t=1}^{T} \left[\widehat{f}_{j,t}(d^\star) - L_{j,t} \right] \le \frac{\log N_j}{\eta_j}. \tag{G.19}$$

For the selected joint design and the realized predictable reward vector g_t, set

$$r_{j,t} := \max_{d^\star \in \mathcal{D}_j} \left\{ d^\star(g_t) + \frac{1}{\eta_j} \Lambda_{\eta_j}^{\mathcal{D}_j}(\mu_{j,t}, p_t, c_{j,t}, \psi_{j,t}; d^\star, g_t) \right\}. \tag{G.20}$$

By ξ_t-optimality,

$$(q_0 - p_t)(g_t) + \sum_{j=1}^{J} r_{j,t} \le \mathsf{CAExO}_{\boldsymbol{\eta}}(\mathscr{T}, \mathcal{G}) + \xi_t. \tag{G.21}$$

For $d^\star \in \mathcal{D}_j$, define

$$Z_{j,t}(d^\star) := d^\star(g_t) + L_{j,t} - \widehat{f}_{j,t}(d^\star) - r_{j,t}. \tag{G.22}$$

Lemma G.6 (Levelwise conditional exponential certificate)**.** *For every j, $d^\star \in \mathcal{D}_j$, and t,*

$$\mathbb{E}_t \left[e^{\eta_j Z_{j,t}(d^\star)} \right] \le 1. \tag{G.23}$$

Consequently, $\exp(\eta_j \sum_{s=1}^{t} Z_{j,s}(d^\star))$ *is a nonnegative supermartingale.*

Proof. Condition on $\mathscr{H}_{t-1}$. From (G.18),

$$\begin{aligned} \mathbb{E}_t e^{\eta_j (L_{j,t} - \widehat{f}_{j,t}(d^\star))} &= \mathbb{E}_t \sum_{d \in \mathcal{D}_j} \mu_{j,t}(d) e^{\eta_j (\widehat{f}_{j,t}(d) - \widehat{f}_{j,t}(d^\star))} \\ &\le 1 + \Lambda_{\eta_j}^{\mathcal{D}_j}(\mu_{j,t}, p_t, c_{j,t}, \psi_{j,t}; d^\star, g_t), \end{aligned}$$

where the inequality uses that the realized observation kernel belongs to $\mathfrak{N}(g_t)$. The weighted exponential sum defining $1 + \Lambda_{\eta_j}^{\mathcal{D}_j}$ is strictly positive, hence $\Lambda_{\eta_j}^{\mathcal{D}_j} > -1$. Therefore $1 + x \le e^x$ applies, and $r_{j,t}$ dominates the defining certificate for $d^\star$,

$$\mathbb{E}_t e^{\eta_j Z_{j,t}(d^\star)} \le \exp\left(\eta_j [d^\star(g_t) - r_{j,t}] + \Lambda_{\eta_j}^{\mathcal{D}_j}(\cdots) \right) \le 1.$$

□

Proof of Theorem G.4. For every $d^\star \in \mathcal{D}_j$, rearranging (G.22) and using (G.19) gives

$$\sum_{t=1}^{T} d^\star(g_t) \le \sum_{t=1}^{T} r_{j,t} + \sum_{t=1}^{T} Z_{j,t}(d^\star) + \frac{\log N_j}{\eta_j}. \tag{G.24}$$

By Lemma G.6,

$$\mathbb{E} \max_{d \in \mathcal{D}_j} \sum_{t=1}^{T} Z_{j,t}(d) \le \frac{\log N_j}{\eta_j},$$

using the same log-sum-exp and Jensen argument as in (F.23). Combining Lemma G.2, (G.24), and (G.21), then taking expectation, yields

$$\mathbb{E}[\overline{\mathrm{Reg}}_T^{Q}] \le T \operatorname{rad}_{\mathcal{G}}(\mathscr{T}) + T \mathsf{CAExO}_{\boldsymbol{\eta}}(\mathscr{T}, \mathcal{G}) + \sum_t \xi_t + 2 \sum_j \frac{\log N_j}{\eta_j},$$

where $\overline{\mathrm{Reg}}_T^{Q}$ is the mixed mediator regret from (F.12). The action-sampling difference $\mathrm{Reg}_T^{Q} - \overline{\mathrm{Reg}}_T^{Q}$ is the mean-zero martingale in (F.13), proving (G.16). Letting the total oracle error tend to zero and optimizing over chains and learning rates proves (G.14). Equation (G.15) follows from the exact optimizer-witness identity in Theorem 8.1. □

Proof of Theorem G.5. For each pair (j, d), Lemma G.6 and Markov's inequality imply

$$\mathbb{P}\left(\sum_{t=1}^{T} Z_{j,t}(d) > \frac{\log(2N_\Sigma/\delta)}{\eta_j}\right) \le \frac{\delta}{2N_\Sigma}.$$

A union bound over all N_Σ level–increment pairs controls every maximum in (G.24) simultaneously with probability at least $1 - \delta/2$. Independently, Azuma's inequality applied to (F.13) gives

$$\mathrm{Reg}_T^Q - \overline{\mathrm{Reg}}_T^Q \le \sqrt{2T\log(2/\delta)}$$

with probability at least $1 - \delta/2$. Combine these events with Lemma G.2, (G.21), and a union bound. □

G.5 Recovery of the one-scale theorem

Proposition G.7 (One level recovers affine ExO exactly)**.** *Fix $q_0 \in C_Q$ and consider the exact depth-one chain $Q_0 = \{q_0\}$, $Q_1 = Q$, $\pi_1(q) = q$. Then*

$$\mathsf{CAExO}_{(\eta)}(\mathscr{T}, \mathcal{G}) = \mathsf{AExO}_\eta(Q, \mathcal{G}), \tag{G.25}$$

$N_1 = |Q|$, and $\mathrm{rad}_{\mathcal{G}}(\mathscr{T}) = 0$. Consequently, Theorem G.4 contains Theorem F.5 as the depth-one special case.

Proof. The map $q \mapsto d = q - q_0$ is a bijection from Q onto $\mathcal{D}_1$. Push every reference law on Q through this bijection and identify a score on d with the corresponding score on q. Pairwise score differences are unchanged. Furthermore,

$$(q_0 - p)(g) + d^\star(g) = (q^\star - p)(g).$$

Thus the local objectives in (G.8) and (F.7) agree pointwise, and so do their infima and outer suprema. □

G.6 A shared-design Riesz envelope

The exact functional (G.13) is nonlinear in score tails. We next derive a scale-sensitive quadratic envelope from the Riesz geometry of Appendix E. The crucial feature is that every level is evaluated under the same p.

For a finite nonempty $\mathcal{D} \subseteq \mathsf{H}_0$ and $\mu \in \Delta(\mathcal{D})$, write $\bar{d}_\mu := \sum_d \mu(d)d$. Define its reward diameter

$$\mathrm{diam}_{\mathcal{G}}(\mathcal{D}) := \max_{d,r\in\mathcal{D}} \sup_{g\in\mathcal{G}} |(d - r)(g)|. \tag{G.26}$$

For $\varepsilon \ge 0$, $L \in [0, \infty]$, and $p \in C_Q$, let $\mathscr{S}^{\mathcal{D}}_{\varepsilon,L}(\mu, p)$ be the affine score systems on $\mathcal{D}$ satisfying

$$\max_{d\in\mathcal{D}} \sup_{g\in\mathcal{G}} \left| c(d) + \sum_a p(a)\psi(d,a)g(a) - (d - \bar{d}_\mu)(g)\right| \le \varepsilon \tag{G.27}$$

and pairwise score oscillation at most L, with the latter condition omitted when $L = \infty$. Define

$$\mathsf{V}^{\mathcal{D}}_\mu(p, c, \psi) := \sup_{\substack{d^\star\in\mathcal{D}\\ g\in\mathcal{G}}} \sum_a p(a) \sum_d \mu(d)\Big[(1 - g(a))D^{c,\psi}_{d,d^\star}(a, 0)^2 + g(a)D^{c,\psi}_{d,d^\star}(a, 1)^2\Big], \tag{G.28}$$

and

$$\mathfrak{B}^{\mathcal{D}}_{\varepsilon,L}(\mu, p) := \inf_{(c,\psi)\in\mathscr{S}^{\mathcal{D}}_{\varepsilon,L}(\mu,p)} \mathsf{V}^{\mathcal{D}}_\mu(p, c, \psi), \tag{G.29}$$

$$\mathfrak{U}^{\mathcal{D}}_{\varepsilon,L}(\mu, p) := \inf_{(c,\psi)\in\mathscr{S}^{\mathcal{D}}_{\varepsilon,L}(\mu,p)} \max_{d\in\mathcal{D}} \|\psi(d, \cdot)\|_p^2. \tag{G.30}$$

Every $\bar{d}_\mu$ has total mass zero. For a vector $z \in \mathbb{R}^{\mathcal{A}}$ with $z(\mathbf{1}) = 1$ and $p \in \Delta(\mathcal{A})$, extend the reward discrepancy by

$$\mathfrak{d}^+_{\mathcal{G}}(z, p) := \max_{u \in \mathcal{G}^0} (z - p)(u). \tag{G.31}$$

The value is nonnegative because $0 \in \mathcal{G}^0$ and is invariant to common reward translations because $(z - p)(\mathbf{1}) = 0$.

For a reference tuple $\boldsymbol{\mu}$, set

$$\widetilde{q}_{\boldsymbol{\mu}} := q_0 + \sum_{j=1}^{J} \bar{d}_{\mu_j}. \tag{G.32}$$

This vector has total mass one, but need not be nonnegative unless the chain has the coherence property introduced below.

Theorem G.8 (Shared-design quadratic envelope)**.** *Fix tolerances $\varepsilon_1, \dots, \varepsilon_J \geq 0$ and finite ranges $L_1, \dots, L_J$ satisfying $\eta_j L_j \leq 1$. Then*

$$\begin{aligned} &\mathsf{CAExO}_{\boldsymbol{\eta}}(\mathscr{T}, \mathcal{G}) \\ &\leq \sup_{\boldsymbol{\mu}} \inf_{p \in C_Q} \left\{ \mathfrak{d}^+_{\mathcal{G}}(\widetilde{q}_{\boldsymbol{\mu}}, p) + 2 \sum_{j=1}^{J} \varepsilon_j + \sum_{j=1}^{J} \eta_j \mathfrak{B}^{\mathcal{D}_j}_{\varepsilon_j, L_j}(\mu_j, p) \right\}. \end{aligned} \tag{G.33}$$

Moreover, for every finite $\mathcal{D} \subseteq \mathsf{H}_0$,

$$\mathfrak{B}^{\mathcal{D}}_{\varepsilon, L}(\mu, p) \leq 4\mathfrak{U}^{\mathcal{D}}_{\varepsilon, L}(\mu, p) + \left[\mathrm{diam}_{\mathcal{G}}(\mathcal{D}) + 2\varepsilon\right]^2, \tag{G.34}$$

and

$$\mathfrak{U}^{\mathcal{D}}_{\varepsilon, \infty}(\mu, p) = \max_{d \in \mathcal{D}} \mathcal{E}_\varepsilon\left(d - \bar{d}_\mu; p\right). \tag{G.35}$$

If the right-hand side of (G.35) *is finite, the same value is attained with some finite oscillation bound L_0.*

Proof. Fix $\boldsymbol{\mu}$, p, and one score system from $\mathscr{S}^{\mathcal{D}_j}_{\varepsilon_j, L_j}(\mu_j, p)$ at every level. The inequality $e^x \leq 1 + x + x^2$ for $x \leq 1$ yields, for each j,

$$\frac{1}{\eta_j} \Lambda^{\mathcal{D}_j}_{\eta_j} \leq \sum_d \mu_j(d) \mathbb{E}[D_{d, d^\star_j}] + \eta_j \mathsf{V}^{\mathcal{D}_j}_{\mu_j}(p, c_j, \psi_j).$$

If $e_{j,d}(g)$ denotes the approximation error in (G.27), then

$$\begin{aligned} d^\star_j(g) + \sum_d \mu_j(d) \mathbb{E}[D_{d, d^\star_j}] &= \bar{d}_{\mu_j}(g) + \sum_d \mu_j(d) e_{j,d}(g) - e_{j, d^\star_j}(g) \\ &\leq \bar{d}_{\mu_j}(g) + 2\varepsilon_j. \end{aligned}$$

Summing over levels and adding $(q_0 - p)(g)$ gives $(\widetilde{q}_{\boldsymbol{\mu}} - p)(g) + 2\sum_j \varepsilon_j$. Taking the supremum over g, then the infimum over score systems and p, proves (G.33).

For (G.34), let $R = \max_d \|\psi(d, \cdot)\|_p^2$. Every pairwise score difference has variance at most $4R$, while its mean has absolute value at most $\mathrm{diam}_{\mathcal{G}}(\mathcal{D}) + 2\varepsilon$. Hence its second moment is at most the right-hand side of (G.34). Finally, when $L = \infty$, the constraints in (G.27) separate over d, and Definition E.2 gives (G.35). Attainment and finiteness of an oscillation bound follow from Proposition E.10 and finiteness of $\mathcal{D}$ and $\mathcal{A}$. □

G.7 Coherent chains and an entropy-sum corollary

Definition G.9 (Coherent approximation chain)**.** An approximation chain is *coherent* if

$$q_0 + \sum_{j=1}^{J} \mathrm{co}(\mathcal{D}_j) \subseteq C_Q. \tag{G.36}$$

Equivalently, every independent choice of mean increments from the levelwise convex hulls produces a valid physical policy in C_Q.

For a coherent chain and levelwise parameters (ε_j, L_j), define

$$V_j(\mathscr{T}; \varepsilon_j, L_j) := \sup_{\boldsymbol{\mu}} \mathfrak{U}^{\mathcal{D}_j}_{\varepsilon_j, L_j}(\mu_j, \widetilde{q}_{\boldsymbol{\mu}}). \tag{G.37}$$

Coherence ensures that the sampling law in (G.37) belongs to C_Q.

Corollary G.10 (Coherent-chain entropy sum)**.** *Let $\mathscr{T}$ be coherent. For every $\varepsilon_j \geq 0$, $L_j < \infty$, and $0 < \eta_j \leq 1/L_j$,*

$$\mathfrak{M}^{\star,\mathrm{ad}}_T(Q, \mathcal{G}) \leq T \operatorname{rad}_{\mathcal{G}}(\mathscr{T}) + 2T \sum_{j=1}^{J} \varepsilon_j + \sum_{j=1}^{J} \left\{ \frac{2 \log N_j}{\eta_j} + T\eta_j \left[4V_j(\mathscr{T}; \varepsilon_j, L_j) + \left(\operatorname{diam}_{\mathcal{G}}(\mathcal{D}_j) + 2\varepsilon_j \right)^2 \right] \right\}. \tag{G.38}$$

The same upper bound holds for the oblivious mediator value.

Proof. For every reference tuple, coherence permits the choice $p = \widetilde{q}_{\boldsymbol{\mu}}$, which makes the discrepancy in (G.33) zero. Apply Theorem G.8 and substitute the result into Theorem G.4. The oblivious value is no larger than the adaptive value. □

The sum in (G.38) is the operational analogue of a chaining entropy sum: each level pays its own logarithmic cardinality and its own observation-design radius. Unlike a flat bound, fine levels can use much larger learning rates when their policy increments are inexpensive to estimate.

G.8 A STRICT MULTISCALE GAIN ON A WEIGHTED PRODUCT CLASS

We conclude with a fully explicit class on which the chained Riesz bound removes an unbounded factor from its direct one-level specialization. The example also shows that the improvement is not an artifact of a loose lower bound: the chained rate is minimax-optimal up to constants.

Fix $m \geq 1$, let

$$Z_m := \sum_{j=1}^{m} 2^{-j} = 1 - 2^{-m}, \qquad w_j := \frac{2^{-j}}{Z_m},$$

and set

$$\mathcal{A}_m := [m] \times \{-1, +1\}.$$

For $x \in \{-1, +1\}^m$, define the policy

$$q_x(j, s) := w_j \mathbf{1}\{s = x_j\}, \qquad Q_m := \{q_x : x \in \{-1, +1\}^m\}. \tag{G.39}$$

Let

$$\mathcal{G}_m := \left\{ g_\theta : g_\theta(j, s) = \frac{1}{2} + \frac{1}{4} s\theta_j,\ \theta \in [-1, 1]^m \right\}. \tag{G.40}$$

Every q_x is exposed by g_x, so the witness family has 2^m genuinely relevant extreme policies.

Let $q_0(j, s) = w_j/2$. For $0 \leq \ell \leq m$, define

$$\pi_\ell(q_x)(j, s) := w_j \begin{cases} \mathbf{1}\{s = x_j\}, & j \leq \ell, \\ 1/2, & j > \ell. \end{cases} \tag{G.41}$$

The associated level increment class is $\mathcal{D}_j = \{r_j, -r_j\}$, where

$$r_j(k, s) := \frac{w_j}{2} \mathbf{1}\{k = j\} s. \tag{G.42}$$

For a truncation depth $J \leq m$, use (G.41) only through level J.

Proposition G.11 (Dimension-free chained rate and strict gain over the flat Riesz bound)**.** *For every* $m, T \geq 1$*,*

$$\frac{\sqrt{3}}{32}\sqrt{T} \leq \mathfrak{M}_T^{\star,\mathrm{obl}}(Q_m, \mathcal{G}_m) \leq \mathfrak{M}_T^{\star,\mathrm{ad}}(Q_m, \mathcal{G}_m) \leq 34\sqrt{T}. \tag{G.43}$$

In contrast, applying the one-level range-controlled Riesz envelope of Corollary F.11 directly to the 2^m *policies with the natural exact score system yields the bound*

$$\mathfrak{M}_T^{\star,\mathrm{ad}}(Q_m, \mathcal{G}_m) \leq 10\sqrt{mT}, \qquad 1 \leq m \leq T. \tag{G.44}$$

Thus the multiscale decomposition removes a factor of order $\sqrt{m}$ *from this flat one-level Riesz specialization.*

Proof. We first prove the upper bound. The convex hull C_{Q_m} is exactly the set of probability vectors p satisfying

$$p(j,+1) + p(j,-1) = w_j, \qquad j = 1, \dots, m. \tag{G.45}$$

Indeed, every mixture of the q_x has these block masses. Conversely, any p satisfying (G.45) is the expectation of q_X when the coordinates of X are independent and $\mathbb{P}(X_j = s) = p(j,s)/w_j$. It follows that every truncated chain above is coherent.

Fix reference laws μ_j on $\{r_j, -r_j\}$ and write $m_j := \mu_j(r_j) - \mu_j(-r_j)$. Then

$$p := \widetilde{q}_{\boldsymbol{\mu}} \quad \text{satisfies} \quad p(j,s) = \begin{cases} w_j(1 + s m_j)/2, & j \leq J, \\ w_j/2, & j > J. \end{cases} \tag{G.46}$$

For $\sigma \in \{-1, +1\}$, identify $\sigma r_j \in \mathcal{D}_j$ and define

$$\begin{aligned} \psi_j(\sigma r_j; (k,s)) &:= \mathbf{1}\{k = j\}(\sigma - m_j)s, \\ c_j(\sigma r_j) &:= -\frac{w_j m_j}{2}(\sigma - m_j). \end{aligned} \tag{G.47}$$

A direct calculation using (G.40) and (G.46) gives

$$c_j(\sigma r_j) + \sum_{k,s} p(k,s)\psi_j(\sigma r_j; (k,s)) g_\theta(k,s) = (\sigma r_j - m_j r_j)(g_\theta). \tag{G.48}$$

Thus the scores have zero bias. Moreover,

$$\|\psi_j(\sigma r_j; \cdot)\|_p^2 = w_j(\sigma - m_j)^2 \leq 4w_j, \qquad \mathrm{osc}(c_j, \psi_j) \leq 3, \tag{G.49}$$

and

$$\mathrm{diam}_{\mathcal{G}_m}(\mathcal{D}_j) = \frac{w_j}{2}. \tag{G.50}$$

Hence Corollary G.10 with $\varepsilon_j = 0$, $L_j = 3$, and $V_j \leq 4w_j$ gives, for every $J \leq m$,

$$\mathfrak{M}_T^{\star,\mathrm{ad}}(Q_m, \mathcal{G}_m) \leq \frac{T}{4}\sum_{j>J} w_j + \sum_{j=1}^{J} \inf_{0 < \eta_j \leq 1/3} \left\{ \frac{2\log 2}{\eta_j} + 17T\eta_j w_j \right\}. \tag{G.51}$$

The terminal-radius term is exact because

$$\sup_{\theta \in [-1,1]^m} (q_x - \pi_J(q_x))(g_\theta) = \frac{1}{4}\sum_{j>J} w_j.$$

Let J_T be the largest $j \leq m$ with $Tw_j \geq 1$, and set $J_T = 0$ if no such index exists. For $j \leq J_T$, choose

$$\eta_j = \sqrt{\frac{2\log 2}{17Tw_j}} \leq \sqrt{\frac{2\log 2}{17}} < \frac{1}{3}.$$

Geometric decay gives

$$\sum_{j>J_T} w_j \leq \frac{2}{T}, \qquad \sum_{j=1}^{J_T} \sqrt{w_j} \leq 2 + \sqrt{2}.$$

Substitution into (G.51) yields

$$\mathfrak{M}_T^{\star,\mathrm{ad}}(Q_m, \mathcal{G}_m) \le \frac{1}{2} + 2\sqrt{34 \log 2}\,(2+\sqrt{2})\sqrt{T} < 34\sqrt{T}.$$

For the flat comparison, fix an arbitrary reference law μ on Q_m, let $m_j = \mathbb{E}_{x\sim\mu}[x_j]$, and choose $p = \bar{q}_\mu$. For $q_x \in Q_m$, the score system

$$\psi_x(k,s) = (x_k - m_k)s,$$
$$c_x = -\frac{1}{2}\sum_{k=1}^{m} w_k m_k (x_k - m_k)$$

is exactly unbiased for $(q_x - p)(g_\theta)$, has coefficient energy at most 4, and pairwise oscillation at most 3. Furthermore, the reward diameter of $Q_m - p$ is at most $1/2$. Corollary F.11 therefore gives

$$\mathfrak{M}_T^{\star,\mathrm{ad}}(Q_m, \mathcal{G}_m) \le \inf_{0<\eta\le 1/3}\left\{\frac{2m\log 2}{\eta} + 17T\eta\right\}.$$

For $m \le T$, the unconstrained minimizer is admissible and the value is $2\sqrt{34mT\log 2} < 10\sqrt{mT}$, proving (G.44).

It remains to prove the lower bound in (G.43). Let $w_1 = 2^{-1}/Z_m \ge 1/2$ and consider the two oblivious Bernoulli environments $g_+ = g_{\varepsilon e_1}$ and $g_- = g_{-\varepsilon e_1}$. Every Q_m-mediated law has block-one mass w_1 by (G.45). Define the predictable sign bias

$$z_t := \frac{p_t(1,+1) - p_t(1,-1)}{w_1} \in [-1,1].$$

Under g_+ and g_-, the conditional mixed regrets are respectively $w_1\varepsilon(1-z_t)/4$ and $w_1\varepsilon(1+z_t)/4$. If P_+ and P_- denote the induced transcript laws, their average expected regret is at least

$$\frac{w_1\varepsilon T}{4}\left(1 - \mathrm{TV}(P_+, P_-)\right). \tag{G.52}$$

The adaptive-action KL chain rule gives

$$D_{\mathrm{KL}}(P_+\|P_-) = w_1 T\ \mathrm{kl}\left(\frac{1}{2} + \frac{\varepsilon}{4}\middle\|\frac{1}{2} - \frac{\varepsilon}{4}\right)$$
$$\le \frac{4}{3} w_1 T \varepsilon^2,$$

where the last inequality follows from $\mathrm{kl}(u\|v) \le (u-v)^2/[v(1-v)]$ and $v \in [1/4, 3/4]$. Pinsker's inequality therefore yields $\mathrm{TV}(P_+, P_-) \le \varepsilon\sqrt{2w_1T/3}$. Choosing $\varepsilon = \sqrt{3/(8w_1T)} \le 1$ makes this total variation at most $1/2$. Equation (G.52), the equality of expected mixed and realized mediator regret, and Lemma A.10 give

$$\mathfrak{M}_T^{\star,\mathrm{obl}}(Q_m, \mathcal{G}_m) \ge \frac{\sqrt{3w_1T}}{16\sqrt{2}} \ge \frac{\sqrt{3}}{32}\sqrt{T}.$$

□

Remark G.12 (What the strict-gain example establishes). Proposition G.11 compares two rigorous consequences of the same Riesz envelope: a single entropy penalty over 2^m witnesses versus levelwise penalties over m binary increments. It does not claim that the exact one-scale value $\mathsf{AExO}_\eta(Q_m, \mathcal{G}_m)$ is incapable of exploiting additional algebraic structure through some different score construction. The proposition establishes the narrower statement needed here: multiscale decomposition can be essential for turning the explicit observation geometry of Appendix E into a sharp regret guarantee.

G.9 Scope and limitations of the chained theorem

Theorem G.4 is a complete adaptive upper bound for every finite witness family and every finite approximation chain. It uses exact robust exponential certificates and is valid under arbitrary history-dependent bounded noise satisfying only the conditional-mean constraint. The theorem is

information-theoretic: it assumes predictable approximate solutions of the joint finite-dimensional saddle problem. Its pathwise comparator decomposition relaxes cross-level path consistency, and the discrepancy $\mathfrak{d}_{\mathcal{G}}^{+}(\widetilde{q}_{\boldsymbol{\mu}}, p)$ in Theorem G.8 records the price of that relaxation. Coherent chains eliminate the discrepancy exactly.

This appendix proves an upper theorem only. It does not assert that $\Gamma_T^{\mathrm{obs}}(Q, \mathcal{G})$ is necessary for arbitrary reward classes, nor that the admissible-cardinality convention alone yields a Talagrand majorizing measure. The subsequent lower-bound appendices must construct a fixed-environment sequential optimizer packing and compare it with the chained design under explicit local richness assumptions. Until that comparison is proved, the appropriate claim is "observation-weighted chained upper bound," not "exact minimax characterization."

H Interactive Information-Theoretic Preliminaries

This appendix collects the information-theoretic statements used by the sequential packing lower bounds. The principal issue is that the learner's action law is history dependent, so transcript divergences must be decomposed conditionally on the realized public history. Change-of-measure arguments of this form are standard in bandit lower bounds (Kaufmann et al., 2016); the interactive Fano framework of Chen et al. (2024) develops a substantially more general algorithm-dependent lower-bound theory. We prove below the exact finite-horizon forms required by our model, including conditional list-Fano and Assouad inequalities, so that no independence or fixed-design assumption is used implicitly. All logarithms are natural.

H.1 Divergences, testing, and transport

Let P and Q be probability measures on the same measurable space $(\mathsf{X}, \mathscr{X})$. Their Kullback–Leibler divergence and total-variation distance are

$$D_{\mathrm{KL}}(P\|Q) := \begin{cases} \displaystyle\int_{\mathsf{X}} \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P, & P \ll Q, \\ +\infty, & \text{otherwise}, \end{cases} \tag{H.1}$$

$$\|P - Q\|_{\mathrm{TV}} := \sup_{B \in \mathscr{X}} |P(B) - Q(B)|. \tag{H.2}$$

For $x, y \in [0,1]$, write

$$\mathrm{kl}(x\|y) := x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y}, \tag{H.3}$$

with the usual lower-semicontinuous conventions at the boundary. For a real number z, let $z_+ := \max\{z, 0\}$.

Lemma H.1 (Binary contraction, Pinsker, and Bretagnolle–Huber)**.** *For every event $B \in \mathscr{X}$,*

$$\mathrm{kl}\big(P(B)\|Q(B)\big) \le D_{\mathrm{KL}}(P\|Q). \tag{H.4}$$

Consequently,

$$\|P - Q\|_{\mathrm{TV}} \le \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P\|Q)}, \tag{H.5}$$

$$\|P - Q\|_{\mathrm{TV}} \le \sqrt{1 - e^{-D_{\mathrm{KL}}(P\|Q)}} \le 1 - \frac{1}{2} e^{-D_{\mathrm{KL}}(P\|Q)}, \tag{H.6}$$

and, for every event B,

$$P(B) + Q(B^{\mathsf{c}}) \ge \frac{1}{2} e^{-D_{\mathrm{KL}}(P\|Q)}. \tag{H.7}$$

If $Z : \mathsf{X} \to [a, b]$ is measurable, then

$$|\mathbb{E}_P Z - \mathbb{E}_Q Z| \le (b-a)\|P - Q\|_{\mathrm{TV}} \le (b-a)\sqrt{\frac{1}{2} D_{\mathrm{KL}}(P\|Q)}. \tag{H.8}$$

Proof. When $P \not\ll Q$, every displayed KL upper bound is trivial, so suppose $P \ll Q$. Applying the log-sum inequality to the measurable partition $\{B, B^{\mathsf{c}}\}$ gives (H.4). For $x, y \in [0,1]$, the function $f_y(x) = \mathrm{kl}(x\|y) - 2(x-y)^2$ satisfies $f_y(y) = f_y'(y) = 0$ and

$$f_y''(x) = \frac{1}{x(1-x)} - 4 \geq 0$$

on $(0,1)$; continuity handles the boundary. Hence $\mathrm{kl}(x\|y) \geq 2(x-y)^2$. Combining this with (H.4) and taking the supremum over B proves (H.5).

For (H.6), let $\lambda = P + Q$ and write $p = \mathrm{d}P/\mathrm{d}\lambda$, $q = \mathrm{d}Q/\mathrm{d}\lambda$. The Hellinger affinity $\mathsf{A}(P,Q) = \int \sqrt{pq}\,\mathrm{d}\lambda$ obeys

$$\|P - Q\|_{\mathrm{TV}}^2 \leq 1 - \mathsf{A}(P,Q)^2$$

by Cauchy–Schwarz. Jensen's inequality under P gives

$$\log \mathsf{A}(P,Q) = \log \mathbb{E}_P\left[\sqrt{\frac{\mathrm{d}Q}{\mathrm{d}P}}\right] \geq -\frac{1}{2} D_{\mathrm{KL}}(P\|Q),$$

where the claim is immediate if the affinity vanishes. Thus $\mathsf{A}(P,Q)^2 \geq e^{-D_{\mathrm{KL}}(P\|Q)}$. The second inequality in (H.6) follows from $\sqrt{1-u} \leq 1 - u/2$ for $u \in [0,1]$. Since $Q(B) - P(B) \leq \|P - Q\|_{\mathrm{TV}}$, (H.7) follows. This proof is the standard Bretagnolle–Huber argument; see also Canonne (2022).

The case $a = b$ is immediate. Otherwise, after replacing Z by $(Z-a)/(b-a)$, the layer-cake identity yields $|\mathbb{E}_P Z - \mathbb{E}_Q Z| \leq (b-a)\|P - Q\|_{\mathrm{TV}}$; Pinsker gives the last inequality in (H.8). □

Lemma H.2 (Two-model decision lower bound)**.** *Let P_0, P_1 be laws on X, let Z be a measurable decision space, and let $K : \mathsf{X} \rightsquigarrow \mathsf{Z}$ be any randomized decision rule. Suppose measurable losses $\ell_0, \ell_1 : \mathsf{Z} \to [0,\infty)$ satisfy*

$$\ell_0(z) + \ell_1(z) \geq 2\Delta \qquad \forall z \in \mathsf{Z} \tag{H.9}$$

for some $\Delta > 0$. If $Z \sim K(\cdot \mid X)$, then

$$\frac{1}{2}\left(\mathbb{E}_{P_0}\ell_0(Z) + \mathbb{E}_{P_1}\ell_1(Z)\right) \geq \Delta\big(1 - \|P_0 - P_1\|_{\mathrm{TV}}\big) \tag{H.10}$$

$$\geq \frac{\Delta}{2} e^{-D_{\mathrm{KL}}(P_0\|P_1)}. \tag{H.11}$$

It is also at least $\Delta\big(1 - \sqrt{D_{\mathrm{KL}}(P_0\|P_1)/2}\big)_+$.

Proof. Data processing for total variation gives $\|P_0K - P_1K\|_{\mathrm{TV}} \leq \|P_0 - P_1\|_{\mathrm{TV}}$. With respect to a measure dominating P_0K and P_1K, write the corresponding densities as p_0, p_1. Then

$$\begin{aligned}
\mathbb{E}_{P_0}\ell_0(Z) + \mathbb{E}_{P_1}\ell_1(Z) &\geq \int \min\{p_0, p_1\}\big(\ell_0 + \ell_1\big) \\
&\geq 2\Delta \int \min\{p_0, p_1\} \\
&= 2\Delta\big(1 - \|P_0K - P_1K\|_{\mathrm{TV}}\big).
\end{aligned}$$

Use Lemma H.1 to obtain the KL forms. □

H.2 Trajectory relative entropy under adaptive sampling

Fix a behavioral algorithm $\mathsf{Alg} = (\pi_t)_{t=1}^T$. Let Θ be a finite index set. For every $\theta \in \Theta$, let

$$\mathsf{Adv}_\theta = (\mathbf{g}_{\theta,t}, \nu_{\theta,t})_{t=1}^T \in \mathfrak{E}_T^{\mathrm{ad}}(\mathcal{G})$$

be an admissible environment in the sense of Definition A.2, and let P_θ denote the induced law of H_T under $(\mathsf{Alg}, \mathsf{Adv}_\theta)$. At a public history h_{t-1}, define the one-step joint kernel

$$K_{\theta,t}(\{a\} \times B \mid h_{t-1}) := \pi_t(a \mid h_{t-1})\nu_{\theta,t}(B \mid h_{t-1}, a). \tag{H.12}$$

The learner's action kernel is the same under every θ; only the environment's observation kernel changes.

Proposition H.3 (Interactive KL chain rule)**.** *For every $\theta, \theta' \in \Theta$,*

$$D_{\mathrm{KL}}(P_\theta \| P_{\theta'}) = \sum_{t=1}^{T} \mathbb{E}_\theta D_{\mathrm{KL}}(K_{\theta,t}(\cdot \mid H_{t-1}) \| K_{\theta',t}(\cdot \mid H_{t-1})) \tag{H.13}$$

$$= \sum_{t=1}^{T} \mathbb{E}_\theta \left[\sum_{a \in \mathcal{A}} p_t(a) D_{\mathrm{KL}}(\nu_{\theta,t}(\cdot \mid H_{t-1}, a) \| \nu_{\theta',t}(\cdot \mid H_{t-1}, a)) \right]. \tag{H.14}$$

Both equalities hold in $[0, \infty]$.

Proof. The relative-entropy chain rule on the finite product of standard Borel spaces gives (H.13). For completeness, it follows inductively from

$$D_{\mathrm{KL}}(P_{X,Y} \| Q_{X,Y}) = D_{\mathrm{KL}}(P_X \| Q_X) + \mathbb{E}_{X \sim P_X} D_{\mathrm{KL}}(P_{Y|X} \| Q_{Y|X}),$$

with the usual extended-value convention. At a fixed history h, the two kernels in (H.12) share the same action marginal $\pi_t(\cdot \mid h)$. Applying the same chain rule first to the action and then to the observation gives

$$D_{\mathrm{KL}}\big(K_{\theta,t}(\cdot \mid h) \| K_{\theta',t}(\cdot \mid h)\big) = \sum_a \pi_t(a \mid h) D_{\mathrm{KL}}\big(\nu_{\theta,t}(\cdot \mid h, a) \| \nu_{\theta',t}(\cdot \mid h, a)\big),$$

which proves (H.14). □

Remark H.4 (Counterfactual kernels are evaluated at the realized history). In (H.14), the kernel of environment θ' is evaluated at the same history H_{t-1} generated under environment θ. This is well defined because every admissible environment specifies its kernel at every public history, including histories having zero probability under that environment. Replacing this counterfactual evaluation by an expectation under $P_{\theta'}$ would be incorrect.

Corollary H.5 (Oblivious action-count decomposition)**.** *Suppose that under environment θ, action a always produces observations from a fixed law $\nu_{\theta,a}$, independently of time and history conditional on the action. With $N_a(T) = \sum_{t=1}^{T} \mathbf{1}\{A_t = a\}$,*

$$D_{\mathrm{KL}}(P_\theta \| P_{\theta'}) = \sum_{a \in \mathcal{A}} \mathbb{E}_\theta[N_a(T)] D_{\mathrm{KL}}(\nu_{\theta,a} \| \nu_{\theta',a}). \tag{H.15}$$

Proof. Substitute the history-independent kernels into (H.14) and exchange the finite sums. □

For $x, y \in [0,1]$, the Bernoulli divergence satisfies

$$2(x-y)^2 \le \mathrm{kl}(x \| y) \le \frac{(x-y)^2}{y(1-y)}. \tag{H.16}$$

The lower bound was proved in Lemma H.1. The upper bound follows from $\log u \le u - 1$:

$$\mathrm{kl}(x \| y) \le x \left(\frac{x}{y} - 1 \right) + (1-x) \left(\frac{1-x}{1-y} - 1 \right) = \frac{(x-y)^2}{y(1-y)}.$$

Corollary H.6 (Bernoulli information-energy comparison)**.** *Suppose both environments use Bernoulli observations, and write*

$$m_{\theta,t}(h, a) := \mathbf{g}_{\theta,t}(h)(a).$$

Then

$$D_{\mathrm{KL}}(P_\theta \| P_{\theta'}) = \mathbb{E}_\theta \sum_{t=1}^{T} \mathrm{kl}(m_{\theta,t}(H_{t-1}, A_t) \| m_{\theta',t}(H_{t-1}, A_t)). \tag{H.17}$$

If all means of environment θ' lie in $[\rho, 1-\rho]$, then

$$D_{\mathrm{KL}}(P_\theta \| P_{\theta'}) \le \frac{1}{\rho(1-\rho)} \mathbb{E}_\theta \sum_{t=1}^{T} [m_{\theta,t}(H_{t-1}, A_t) - m_{\theta',t}(H_{t-1}, A_t)]^2. \tag{H.18}$$

Without an interior assumption,

$$D_{\mathrm{KL}}(P_\theta \| P_{\theta'}) \ge 2\mathbb{E}_\theta \sum_{t=1}^{T} [m_{\theta,t}(H_{t-1}, A_t) - m_{\theta',t}(H_{t-1}, A_t)]^2. \tag{H.19}$$

Proof. Use Proposition H.3 and (H.16) pointwise. □

Continuation laws. For $1 \le s \le u \le T$ and a deterministic history $h_{s-1} \in \mathsf{Z}_{s-1}$, let $P^{s:u}_{\theta,h_{s-1}}$ denote the law of $(A_s, Y_s, \dots, A_u, Y_u)$ obtained by restarting the fixed algorithm and environment θ from history h_{s-1}. The Ionescu–Tulcea construction in Proposition A.3 makes this law well defined even when h_{s-1} has zero probability under P_θ.

Corollary H.7 (Blockwise counterfactual KL)**.** *For every h_{s-1} and every θ, θ', the continuation laws satisfy*

$$D_{\mathrm{KL}}\Big(P^{s:u}_{\theta,h_{s-1}}\Big\|P^{s:u}_{\theta',h_{s-1}}\Big) = \mathbb{E}_{\theta,h_{s-1}}\sum_{t=s}^{u} D_{\mathrm{KL}}(\nu_{\theta,t}(\cdot \mid H_{t-1}, A_t)\|\nu_{\theta',t}(\cdot \mid H_{t-1}, A_t))\,, \tag{H.20}$$

where $\mathbb{E}_{\theta,h_{s-1}}$ denotes expectation under the continuation law started from h_{s-1}.

Proof. Apply Proposition H.3 to the truncated interaction whose initial public history is fixed at h_{s-1}. □

H.3 Bayes mixtures and interactive mutual information

Let a latent index $\Theta \sim \pi \in \Delta(\Theta)$ be sampled before interaction, and conditional on $\Theta = \theta$ run environment Adv_θ against the fixed algorithm. We may and do remove zero-prior indices, so π has full support on Θ. Let $\mathbb{P}_\pi$ denote the resulting joint law of (Θ, H_T), and let

$$\overline{P}_\pi := \sum_{\theta\in\Theta} \pi(\theta) P_\theta$$

be the marginal transcript law. We write $\mathrm{I}_\pi(\Theta; X \mid Z)$ for conditional mutual information under $\mathbb{P}_\pi$.

Lemma H.8 (Information-radius identities)**.** *For every probability law Q on the transcript space,*

$$\sum_\theta \pi(\theta) D_{\mathrm{KL}}(P_\theta\|Q) = \mathrm{I}_\pi(\Theta; H_T) + D_{\mathrm{KL}}(\overline{P}_\pi\|Q). \tag{H.21}$$

Consequently,

$$\mathrm{I}_\pi(\Theta; H_T) \le \sum_\theta \pi(\theta) D_{\mathrm{KL}}(P_\theta\|Q), \tag{H.22}$$

$$\mathrm{I}_\pi(\Theta; H_T) \le \sum_{\theta,\theta'} \pi(\theta)\pi(\theta') D_{\mathrm{KL}}(P_\theta\|P_{\theta'}). \tag{H.23}$$

More generally, if U is finite, V and X take values in standard Borel spaces, and Q_v is any measurable family of reference laws for X, then

$$\mathrm{I}(U; X \mid V) \le \mathbb{E}_{U,V} D_{\mathrm{KL}}\big(P_{X|U,V}\big\|Q_V\big)\,. \tag{H.24}$$

Proof. Let λ dominate Q and all P_θ, with densities q, p_θ, and put $\bar p = \sum_\theta \pi(\theta) p_\theta$. Direct expansion gives

$$\sum_\theta \pi(\theta) \int p_\theta \log \frac{p_\theta}{q} = \sum_\theta \pi(\theta) \int p_\theta \log \frac{p_\theta}{\bar p} + \int \bar p \log \frac{\bar p}{q},$$

which is (H.21); standard approximation gives the same identity with extended values. Dropping the final nonnegative term gives (H.22). Apply it with $Q = P_{\theta'}$ and average over θ' to obtain (H.23). Applying the same argument under each regular conditional law given $V = v$ and integrating over v proves (H.24). □

Let

$$\pi_{t-1}(\theta) := \mathbb{P}_\pi(\Theta = \theta \mid \mathscr{H}_{t-1})$$

be a version of the posterior, and define the posterior-predictive observation kernel

$$\overline{\nu}_t(\cdot \mid H_{t-1}, a) := \sum_\theta \pi_{t-1}(\theta)\nu_{\theta,t}(\cdot \mid H_{t-1}, a). \tag{H.25}$$

Its value on histories of zero $\overline{P}_\pi$-probability is arbitrary.

Proposition H.9 (Interactive mutual-information chain rule)**.** *Under the Bayes mixture described above,*

$$\mathrm{I}_\pi(\Theta; H_T) = \sum_{t=1}^{T} \mathrm{I}_\pi(\Theta; Y_t \mid \mathscr{H}_{t-1}, A_t) \tag{H.26}$$

$$= \sum_{t=1}^{T} \mathbb{E}_\pi D_{\mathrm{KL}}(\nu_{\Theta,t}(\cdot \mid H_{t-1}, A_t) \| \overline{\nu}_t(\cdot \mid H_{t-1}, A_t)) \,. \tag{H.27}$$

In particular, the action draw contributes no new information conditionally on the public past:

$$\mathrm{I}_\pi(\Theta; A_t \mid \mathscr{H}_{t-1}) = 0. \tag{H.28}$$

Proof. The mutual-information chain rule gives

$$\mathrm{I}_\pi(\Theta; H_T) = \sum_{t=1}^{T} \mathrm{I}_\pi(\Theta; A_t, Y_t \mid \mathscr{H}_{t-1}).$$

Conditional on H_{t-1}, the algorithm samples A_t from the same kernel p_t under every value of Θ, proving (H.28). Applying the chain rule once more leaves $\mathrm{I}_\pi(\Theta; Y_t \mid \mathscr{H}_{t-1}, A_t)$. The conditional law of Y_t given (H_{t-1}, A_t) is exactly (H.25); the standard conditional-KL representation of mutual information yields (H.27). □

Corollary H.10 (Posterior-variance identity for Bernoulli feedback)**.** *Suppose every* Adv_θ *uses Bernoulli observations and all conditional means lie in* $[\rho, 1-\rho]$*. Put*

$$M_t := m_{\Theta,t}(H_{t-1}, A_t), \qquad \overline{M}_t := \mathbb{E}_\pi[M_t \mid \mathscr{H}_{t-1}, A_t]. \tag{H.29}$$

Then

$$\mathrm{I}_\pi(\Theta; Y_t \mid \mathscr{H}_{t-1}, A_t) = \mathbb{E}_\pi\left[\mathrm{kl}(M_t \| \overline{M}_t)\right], \tag{H.30}$$

and

$$2\mathbb{E}_\pi[\mathrm{Var}(M_t \mid \mathscr{H}_{t-1}, A_t)] \le \mathrm{I}_\pi(\Theta; Y_t \mid \mathscr{H}_{t-1}, A_t) \le \frac{1}{\rho(1-\rho)} \mathbb{E}_\pi[\mathrm{Var}(M_t \mid \mathscr{H}_{t-1}, A_t)] \,. \tag{H.31}$$

If Θ'_t *is a conditionally independent posterior copy of* Θ *given* $(\mathscr{H}_{t-1}, A_t)$ *on an enlarged probability space, then*

$$\mathrm{I}_\pi(\Theta; Y_t \mid \mathscr{H}_{t-1}, A_t) \le \frac{1}{2\rho(1-\rho)} \mathbb{E}_\pi\left[\left(m_{\Theta,t}(H_{t-1}, A_t) - m_{\Theta'_t,t}(H_{t-1}, A_t)\right)^2\right]. \tag{H.32}$$

Proof. A mixture of Bernoulli laws is Bernoulli, with mean equal to the mixture mean. Hence the posterior-predictive law in (H.25) is $\mathrm{Ber}(\overline{M}_t)$, which gives (H.30). Apply (H.16) pointwise and use $\mathbb{E}[(M_t - \overline{M}_t)^2 \mid \mathscr{H}_{t-1}, A_t] = \mathrm{Var}(M_t \mid \mathscr{H}_{t-1}, A_t)$ to prove (H.31). Finally, conditional on $(\mathscr{H}_{t-1}, A_t)$,

$$\mathbb{E}[(M_t - M'_t)^2] = 2\,\mathrm{Var}(M_t),$$

which gives (H.32). □

H.4 Conditional list-Fano inequalities

The list form below is useful when several optimizer cells may be simultaneously acceptable at one scale. The standard Fano inequality corresponds to list size one.

Theorem H.11 (Conditional list-Fano)**.** *Let* (U, V, X) *be random variables with* $U \in [M]$ *such that, conditional on* V*, the variable* U *is uniform on* $[M]$*, where* $M \ge 2$*. Let* $\widehat{\mathcal{L}} = \widehat{\mathcal{L}}(V, X)$ *be a random subset of* $[M]$ *satisfying* $|\widehat{\mathcal{L}}| \le L < M$ *almost surely. Then*

$$\mathbb{P}\big(U \notin \widehat{\mathcal{L}}\big) \ge \left[1 - \frac{\mathrm{I}(U; X \mid V) + \log 2}{\log(M/L)}\right]_+ . \tag{H.33}$$

For $L = 1$*, every estimator* $\widehat{U} = \widehat{U}(V, X)$ *satisfies*

$$\mathbb{P}(\widehat{U} \ne U) \ge \left[1 - \frac{\mathrm{I}(U; X \mid V) + \log 2}{\log M}\right]_+ . \tag{H.34}$$

Proof. Let $E = \mathbf{1}\{U \notin \widehat{\mathcal{L}}\}$. Since E is determined by (U, V, X),

$$\begin{aligned} H(U \mid V, X) &= H(E, U \mid V, X) \\ &\le H(E \mid V, X) + H(U \mid E, V, X) \\ &\le \log 2 + \mathbb{P}(E = 0) \log L + \mathbb{P}(E = 1) \log M. \end{aligned}$$

Because U is conditionally uniform, $H(U \mid V) = \log M$, and therefore

$$\log M - \mathrm{I}(U; X \mid V) \le \log 2 + \log L + \mathbb{P}(E = 1) \log(M/L).$$

Rearranging proves (H.33); the decoder case follows by taking $\widehat{\mathcal{L}} = \{\widehat{U}\}$. □

Corollary H.12 (Decision form of conditional Fano)**.** *Under the assumptions of Theorem H.11, let $Z = Z(V, X)$ take values in a measurable decision space Z. Suppose a nonnegative loss $\ell(u, v, z)$ satisfies, for every (v, z),*

$$|\{u : \ell(u, v, z) < \Delta\}| \le L \tag{H.35}$$

for some $\Delta > 0$. Then

$$\mathbb{E}[\ell(U, V, Z)] \ge \Delta \left[1 - \frac{\mathrm{I}(U; X \mid V) + \log 2}{\log(M/L)}\right]_+. \tag{H.36}$$

For any measurable family of reference laws Q_V for X, the mutual information in (H.36) *may be replaced by*

$$\mathbb{E}_{U,V} D_{\mathrm{KL}}(P_{X|U,V} \| Q_V). \tag{H.37}$$

Proof. Define the list $\widehat{\mathcal{L}}(v, z) = \{u : \ell(u, v, z) < \Delta\}$, padding it arbitrarily if necessary. On the event $U \notin \widehat{\mathcal{L}}(V, Z)$ one has $\ell(U, V, Z) \ge \Delta$. Apply Theorem H.11. The reference-law version follows from (H.24). □

H.5 Assouad mixtures for interactive experiments

Let $\Theta = \{-1, +1\}^m$ and let P_θ be arbitrary laws on a measurable space X. For $i \in [m]$, write $\theta^{(i)}$ for the vector obtained by flipping coordinate i, and define the coordinate mixtures

$$P_{i,+} := 2^{-(m-1)} \sum_{\theta : \theta_i = +1} P_\theta, \qquad P_{i,-} := 2^{-(m-1)} \sum_{\theta : \theta_i = -1} P_\theta. \tag{H.38}$$

Theorem H.13 (Assouad with interactive transcript laws)**.** *Let Θ be uniform on $\{-1, +1\}^m$, let $X \mid \Theta = \theta \sim P_\theta$, and let $\widehat{\Theta} = \widehat{\Theta}(X)$ be any possibly randomized estimator. Define*

$$\overline{D}_i := 2^{-(m-1)} \sum_{\theta : \theta_i = +1} D_{\mathrm{KL}}(P_\theta \| P_{\theta^{(i)}}). \tag{H.39}$$

Then

$$\mathbb{E}[d_{\mathrm{H}}(\widehat{\Theta}, \Theta)] \ge \frac{1}{2} \sum_{i=1}^m \big(1 - \|P_{i,+} - P_{i,-}\|_{\mathrm{TV}}\big) \tag{H.40}$$

$$\ge \frac{1}{2} \sum_{i=1}^m \left(1 - \sqrt{\frac{\overline{D}_i}{2}}\right)_+, \tag{H.41}$$

$$\ge \frac{1}{4} \sum_{i=1}^m e^{-\overline{D}_i}. \tag{H.42}$$

If the P_θ are transcript laws induced by a common behavioral algorithm and a family of environments, every edge divergence in (H.39) *admits the interactive decomposition of Proposition H.3.*

Proof. For coordinate i,

$$\mathbb{P}(\widehat{\Theta}_i \ne \Theta_i) = \frac{1}{2} P_{i,+}(\widehat{\Theta}_i = -1) + \frac{1}{2} P_{i,-}(\widehat{\Theta}_i = +1) \ge \frac{1}{2}\big(1 - \|P_{i,+} - P_{i,-}\|_{\mathrm{TV}}\big).$$

Summing over i gives (H.40). Joint convexity of KL and the pairing $\theta \leftrightarrow \theta^{(i)}$ imply

$$D_{\mathrm{KL}}(P_{i,+} \| P_{i,-}) \le \overline{D}_i. \tag{H.43}$$

Pinsker and Bretagnolle–Huber from Lemma H.1, applied to the coordinate mixtures, yield (H.41) and (H.42). □

Corollary H.14 (Bernoulli interactive Assouad)**.** *Suppose P_θ is the transcript law generated by an environment with Bernoulli mean map $m_{\theta,t}(h,a) \in [\rho, 1-\rho]$. Then*

$$\overline{D}_i \le \frac{2^{-(m-1)}}{\rho(1-\rho)} \sum_{\theta:\theta_i=+1} \mathbb{E}_\theta \sum_{t=1}^{T} \left[m_{\theta,t}(H_{t-1}, A_t) - m_{\theta^{(i)},t}(H_{t-1}, A_t) \right]^2. \tag{H.44}$$

Substitution into Theorem H.13 gives a lower bound expressed entirely through action-dependent squared mean separations accumulated by the algorithm.

Proof. Apply (H.18) to every hypercube edge and average. □

H.6 Blockwise conditional information and nodewise Fano

The subsequent sequential-packing construction reveals a branch index only during a designated block. The following statement isolates exactly the information accumulated in such a block. For each realization (u_0, v_0, h_{s-1}), write $P^{s:u}_{u_0,v_0,h_{s-1}}$ for the corresponding continuation law of $H_{s:u}$; thus $P^{s:u}_{U,V,H_{s-1}}$ denotes this probability kernel evaluated at the random triple (U, V, H_{s-1}).

Proposition H.15 (Blockwise interactive mutual information)**.** *Let V be arbitrary side information and let U be a finite latent variable. Fix $1 \le s \le u \le T$. Suppose*

$$U \perp\!\!\!\perp H_{s-1} \mid V, \tag{H.45}$$

and, conditional on (U, V), the interaction from round s onward uses a fixed behavioral algorithm and an environment indexed by (U, V). Then

$$\mathrm{I}(U; H_u \mid V) = \mathrm{I}(U; H_{s:u} \mid V, H_{s-1}) \tag{H.46}$$

$$= \sum_{t=s}^{u} \mathrm{I}(U; Y_t \mid V, H_{t-1}, A_t), \tag{H.47}$$

where $H_{s:u} = (A_s, Y_s, \ldots, A_u, Y_u)$.

Proof. By (H.45), $\mathrm{I}(U; H_{s-1} \mid V) = 0$. The chain rule therefore gives the first equality. Conditional on (V, H_{t-1}), the learner's action kernel does not depend on U, so $\mathrm{I}(U; A_t \mid V, H_{t-1}) = 0$. Applying the chain rule round by round proves the second equality. □

Theorem H.16 (Blockwise interactive list-Fano)**.** *Under the assumptions of Proposition H.15, suppose that $U \in [M]$ is uniform on $[M]$ conditional on V, and let $Z = Z(V, H_u)$ be any decision. We allow the decision rule to observe V; revealing this side information can only strengthen the learner and therefore preserves any lower bound. If a nonnegative block loss $\ell(U, V, Z)$ satisfies*

$$|\{u' : \ell(u', v, z) < \Delta\}| \le L < M \qquad \forall (v, z), \tag{H.48}$$

then

$$\mathbb{E}[\ell(U, V, Z)] \ge \Delta \left[1 - \frac{\sum_{t=s}^{u} \mathrm{I}(U; Y_t \mid V, H_{t-1}, A_t) + \log 2}{\log(M/L)} \right]_+. \tag{H.49}$$

For any family of reference continuation laws $Q^{s:u}_{v,h_{s-1}}$,

$$\mathbb{E}[\ell(U, V, Z)] \ge \Delta \left[1 - \frac{\mathbb{E} D_{\mathrm{KL}}\left(P^{s:u}_{U,V,H_{s-1}} \middle\| Q^{s:u}_{V,H_{s-1}} \right) + \log 2}{\log(M/L)} \right]_+, \tag{H.50}$$

where $P^{s:u}_{U,V,H_{s-1}}$ denotes the continuation law under the realized latent index and starting history.

Proof. Apply Corollary H.12 with side information (V, H_{s-1}) and observation $H_{s:u}$. Condition (H.45) preserves conditional uniformity of U. Then use Proposition H.15 to obtain (H.49), and Lemma H.8 conditionally on (V, H_{s-1}) to obtain (H.50). □

Corollary H.17 (Bernoulli block information)**.** *Under the conditions of Theorem H.16, suppose the block observations are Bernoulli with means in* $[\rho, 1-\rho]$*. Let*

$$M_t^U = m_{U,V,t}(H_{t-1}, A_t), \qquad \overline{M}_t = \mathbb{E}[M_t^U \mid V, H_{t-1}, A_t].$$

Then

$$\sum_{t=s}^{u} \mathrm{I}(U; Y_t \mid V, H_{t-1}, A_t) \le \frac{1}{\rho(1-\rho)} \sum_{t=s}^{u} \mathbb{E}\left[(M_t^U - \overline{M}_t)^2\right]. \tag{H.51}$$

Consequently, any upper bound on the posterior quadratic energy in (H.51)*, together with the decision separation* (H.48)*, yields an explicit block-regret lower bound through* (H.49)*.*

Proof. Apply Corollary H.10 conditionally on V and sum over the block. □

H.7 Standing lower-bound template

The subsequent lower-bound appendices use the results above in the following order. A latent branch variable is sampled before the block and is conditionally independent of the pre-block public history given its ancestors. A geometric separation lemma shows that any realized decision can be Δ-good for at most L of the M children. Proposition H.15 removes the learner's action draws from the information budget, and either Corollary H.10 or a reference-environment application of Proposition H.3 bounds the remaining observation information. Theorem H.16 then converts the unresolved branch entropy into block regret. Binary nodes may instead use Lemma H.2 or Theorem H.13. Finally, Lemma A.10 converts the Bayes lower bound into an oblivious minimax lower bound. The latent prior is only a proof device: every conditional environment indexed by a fixed leaf remains an individually admissible environment from Appendix A.

I Local Design–Testing Duality

This appendix turns the Riesz design geometry of Appendix E into explicit hard statistical experiments. The main point is that a large one-point design cost identifies a reward direction that changes a policy comparison substantially while producing little observation energy under the learner's sampling law. Such a direction becomes a valid lower-bound instance only if it can be realized inside the reward class while switching the relevant optimizer. We isolate this additional geometric requirement through *optimizer perturbation sets* and prove an exact local equivalence on regular exposed edges of the witness polytope.

The construction is deliberately local. It supplies the hard binary and multiway nodes used by the sequential packing in the subsequent sequential-packing appendix; it does not yet glue the nodes across time scales. Algorithm-dependent choices of hard alternatives and reference experiments are central to modern DEC lower bounds and to interactive Fano methods (Foster et al., 2023; Chen et al., 2024). Here the reward-only structure makes the local alternative explicit through the Fenchel–Riesz duality proved in Theorem E.9. The resulting estimator–testing correspondence is a sequential, action-designed analogue of the convex duality between affine estimators and Le Cam lower bounds for linear functional estimation (Polyanskiy & Wu, 2026); optimizer switching is the additional constraint specific to decision making.

Throughout, fix a finite nonempty witness family $Q \subseteq \Delta(\mathcal{A})$, write $C_Q = \mathrm{co}(Q)$, and fix $\rho \in (0, 1/2)$. Let

$$\mathcal{G}_\rho := \mathcal{G} \cap [\rho, 1-\rho]^{\mathcal{A}}. \tag{I.1}$$

All Bernoulli experiments below have mean vectors in $\mathcal{G}_\rho$.

$$\boxed{\text{large Riesz energy } \not\Rightarrow \text{ large regret}} \tag{I.2}$$

The missing implication is optimizer switching: a difficult contrast becomes a decision lower bound only when legal perturbations of the reward class change the relevant optimizer.

I.1 A separation–energy form of the Riesz dual

For $h \in \mathsf{H}_0$, $p \in \Delta(\mathcal{A})$, and $\varepsilon \geq 0$, define the excess contrast of a reward direction $v \in \mathcal{V}$ by

$$\mathfrak{s}_\varepsilon(h, v) := (|h(v)| - \varepsilon \|v\|_{\mathcal{K}})_+ . \tag{I.3}$$

For $x \geq 0$ and $y \geq 0$, use the extended-ratio convention

$$\frac{x^2}{y} := \begin{cases} x^2/y, & y > 0, \\ 0, & x = 0 = y, \\ +\infty, & x > 0 = y. \end{cases} \tag{I.4}$$

Proposition I.1 (Separation–energy representation)**.** *For every $h \in \mathsf{H}_0$, $p \in \Delta(\mathcal{A})$, and $\varepsilon \geq 0$,*

$$\mathcal{E}_\varepsilon(h; p) = \sup_{v \in \mathcal{V}} \frac{\mathfrak{s}_\varepsilon(h, v)^2}{\|v\|_p^2}. \tag{I.5}$$

The equality is in $[0, +\infty]$. In particular, the approximate design cost is the largest squared policy separation, after paying the geometric tolerance $\varepsilon \|v\|_{\mathcal{K}}$, per unit of observation energy.

Proof. By Theorem E.9 and central symmetry of $\mathcal{K}$,

$$\begin{aligned} \mathcal{E}_\varepsilon(h; p) &= \sup_{v \in \mathcal{V}} \left\{ h(v) - \varepsilon \|v\|_{\mathcal{K}} - \frac{1}{4} \|v\|_p^2 \right\} \\ &= \sup_{v \in \mathcal{V}} \sup_{t \geq 0} \left\{ t \left(|h(v)| - \varepsilon \|v\|_{\mathcal{K}} \right) - \frac{t^2}{4} \|v\|_p^2 \right\}. \end{aligned}$$

For a fixed direction v, if $\|v\|_p^2 > 0$, the inner supremum equals $\mathfrak{s}_\varepsilon(h, v)^2 / \|v\|_p^2$ and is attained at

$$t = \frac{2\mathfrak{s}_\varepsilon(h, v)}{\|v\|_p^2}.$$

If $\|v\|_p = 0$, the inner supremum is zero when $\mathfrak{s}_\varepsilon(h, v) = 0$ and is $+\infty$ otherwise. This is exactly the convention in (I.4). Taking the supremum over v proves (I.5). □

Remark I.2 (Exact observability as an infinite testing ratio). Equation (I.5) recovers the sharp feasibility threshold in Proposition E.8: if a p-null reward direction satisfies $|h(v)| > \varepsilon \|v\|_{\mathcal{K}}$, then it changes the target contrast without changing any observation available under p, and the ratio is infinite.

I.2 Optimizer perturbations and local testing coefficients

For $g^0 \in \mathcal{G}$, let

$$F_Q(g^0) := \arg\max_{q \in C_Q} q(g^0) \tag{I.6}$$

be the exposed optimizer face of the witness polytope. Fix distinct $q_+, q_- \in F_Q(g^0)$ and put

$$h := q_+ - q_- \in \mathsf{H}_0. \tag{I.7}$$

Because both policies are optimal at g^0, one has $h(g^0) = 0$.

Definition I.3 (Symmetric optimizer perturbation set)**.** The ρ-interior symmetric optimizer perturbation set associated with (g^0, q_+, q_-) is

$$\mathcal{W}_\rho(g^0; q_+, q_-) := \Big\{ w \in \mathcal{V} : g^0 + w, g^0 - w \in \mathcal{G}_\rho, \tag{I.8}$$

$$q_+ \in F_Q(g^0 + w), \quad q_- \in F_Q(g^0 - w) \Big\}. \tag{I.9}$$

For $p \in \Delta(\mathcal{A})$, define the local geometric testing modulus

$$\mathfrak{T}_{\varepsilon,\rho}(g^0, q_+, q_-; p) := \sup_{w \in \mathcal{W}_\rho(g^0; q_+, q_-)} \frac{(h(w) - \varepsilon \|w\|_{\mathcal{K}})_+^2}{\|w\|_p^2}, \tag{I.10}$$

with the convention (I.4).

Lemma I.4 (Elementary geometry of optimizer perturbations)**.** *The set $\mathcal{W}_\rho(g^0; q_+, q_-)$ is compact, contains 0, and is star-shaped about 0. Every w in this set satisfies $h(w) \geq 0$. Moreover,*

$$\mathfrak{T}_{\varepsilon,\rho}(g^0, q_+, q_-; p) \leq \mathcal{E}_\varepsilon(h; p)\,. \tag{I.11}$$

Proof. The membership conditions in (I.8) are closed, and $\mathcal{G}_\rho$ is compact, so the perturbation set is compact. The zero vector belongs to it because $q_+, q_- \in F_Q(g^0)$. If w is feasible and $\lambda \in [0,1]$, then $g^0 \pm \lambda w$ are convex combinations of g^0 and $g^0 \pm w$, hence belong to $\mathcal{G}_\rho$. For every $q \in C_Q$, the two functions

$$\lambda \mapsto (q_+ - q)(g^0 + \lambda w), \qquad \lambda \mapsto (q_- - q)(g^0 - \lambda w)$$

are affine and nonnegative at both endpoints $\lambda = 0, 1$; they are therefore nonnegative throughout $[0,1]$. This proves star-shapedness.

Comparing q_+ with q_- under $g^0 + w$ gives $h(g^0 + w) \geq 0$. Since $h(g^0) = 0$, one has $h(w) \geq 0$. Finally, $\mathcal{W}_\rho(g^0; q_+, q_-) \subseteq \mathcal{V}$, so (I.11) follows from Proposition I.1. □

Example I.5 (Observation difficulty without optimizer switching)**.** Let $\mathcal{A} = \{1, 2\}$, $Q = \{\delta_1, \delta_2\}$, and

$$\mathcal{G} = \{(1/2, 1/4 + t) : t \in [-1/8, 1/8]\}\,.$$

Then action 1 is uniquely optimal for every $g \in \mathcal{G}$, so no reward vector exposes an optimizer edge between δ_1 and δ_2. Nevertheless, with $h = \delta_1 - \delta_2$ and any p satisfying $p(2) > 0$,

$$\mathcal{E}_0(h; p) = \frac{1}{p(2)}.$$

Indeed, $\mathcal{V} = \operatorname{span}\{(0, 1)\}$ and the separation–energy quotient equals $1/p(2)$. Thus a large Riesz cost need not yield a regret lower bound unless the corresponding reward direction can be realized while changing the optimizer. The perturbation set in Definition I.3 records exactly this additional decision geometry.

For a feasible perturbation w, consider the binary Bernoulli experiment

$$U \sim \mathrm{Unif}\{-1, +1\}, \qquad A \sim p \ \text{ independently of } U, \qquad Y \mid (U, A) \sim \mathrm{Ber}(g^0(A) + Uw(A)). \tag{I.12}$$

Let

$$\mathcal{I}_p(w) := \mathrm{I}(U; A, Y) = \mathrm{I}(U; Y \mid A). \tag{I.13}$$

Define the operational binary testing coefficient

$$\mathfrak{B}_{\varepsilon,\rho}(g^0, q_+, q_-; p) := \sup_{w \in \mathcal{W}_\rho(g^0; q_+, q_-)} \frac{(h(w) - \varepsilon \|w\|_{\mathcal{K}})_+^2}{\mathcal{I}_p(w)}, \tag{I.14}$$

where a zero numerator contributes zero and a positive numerator divided by zero is $+\infty$.

Theorem I.6 (Local design–testing equivalence)**.** *For every $p \in \Delta(\mathcal{A})$ and $\varepsilon \geq 0$,*

$$2\,\|w\|_p^2 \leq \mathcal{I}_p(w) \leq \frac{\|w\|_p^2}{\rho(1-\rho)} \qquad \forall w \in \mathcal{W}_\rho(g^0; q_+, q_-). \tag{I.15}$$

Consequently,

$$\rho(1-\rho)\mathfrak{T}_{\varepsilon,\rho}(g^0, q_+, q_-; p) \leq \mathfrak{B}_{\varepsilon,\rho}(g^0, q_+, q_-; p) \leq \frac{1}{2}\mathfrak{T}_{\varepsilon,\rho}(g^0, q_+, q_-; p). \tag{I.16}$$

Thus the local geometric modulus is, up to constants depending only on the Bernoulli interiority parameter, the squared decision separation per nat of information in the best realizable binary optimizer-switching experiment.

Proof. Conditioned on $A = a$, the posterior-predictive Bernoulli mean in (I.12) is $g^0(a)$, while the random conditional mean is $g^0(a) + Uw(a)$. Its conditional variance over U is $w(a)^2$. Applying Corollary H.10 and averaging over $A \sim p$ gives (I.15). Dividing the squared excess contrast by these upper and lower information bounds and taking the supremum over w proves (I.16). □

Lemma I.7 (Pairwise decision separation)**.** *For $w \in \mathcal{W}_\rho(g^0; q_+, q_-)$, set $g_\pm = g^0 \pm w$ and define the mediator suboptimality*

$$\ell_\pm(p') := \max_{q \in Q} q(g_\pm) - p'(g_\pm), \qquad p' \in C_Q. \tag{I.17}$$

Then $q_\pm$ are valid maximizers and

$$\ell_+(p') + \ell_-(p') \geq h(w) \qquad \forall p' \in C_Q. \tag{I.18}$$

Consequently, for $s_\varepsilon(w) = (h(w) - \varepsilon \|w\|_{\mathcal{K}})_+$,

$$\ell_+(p') + \ell_-(p') \geq s_\varepsilon(w) \qquad \forall p' \in C_Q. \tag{I.19}$$

Proof. By definition of the perturbation set, q_+ and q_- are optimal under g_+ and g_-, respectively. Let $V^0 = \max_{q \in C_Q} q(g^0)$. Since $q_+, q_- \in F_Q(g^0)$,

$$\begin{aligned} \ell_+(p') + \ell_-(p') &= q_+(g^0 + w) + q_-(g^0 - w) - 2p'(g^0) \\ &= 2\big(V^0 - p'(g^0)\big) + h(w) \geq h(w). \end{aligned}$$

The second assertion follows because $s_\varepsilon(w) \leq h(w)$. □

I.3 Regular exposed edges and exact Riesz realizability

The preceding local testing modulus can be strictly smaller than the global Riesz cost: a statistically favorable reward direction need not switch a realizable pair of optimizers. The next condition identifies a broad local geometry for which no such loss occurs.

Definition I.8 (Regular exposed witness edge)**.** A quadruple

$$\mathfrak{e} = (g^0, q_+, q_-, r) \tag{I.20}$$

with $r > 0$ is a ρ-regular exposed edge of C_Q in $\mathcal{G}$ if:

1. $g^0 \in \mathcal{G}_\rho$, $q_+ \neq q_-$, and
$$F_Q(g^0) = \mathrm{co}\{q_+, q_-\}; \tag{I.21}$$
2. the relative ℓ_∞ ball of radius r is contained in the interior reward class,
$$g^0 + \{v \in \mathcal{V} : \|v\|_\infty \leq r\} \subseteq \mathcal{G}_\rho; \tag{I.22}$$
3. writing $V^0 = \max_{q \in C_Q} q(g^0)$, every witness outside the exposed edge has baseline gap at least $2r$,
$$\min_{q \in Q \setminus F_Q(g^0)} \big[V^0 - q(g^0)\big] \geq 2r, \tag{I.23}$$
where the minimum over the empty set is $+\infty$.

Condition (I.22) is relative to $\mathcal{V}$; it does not require $\mathcal{G}$ to have nonempty interior in the ambient space $\mathbb{R}^{\mathcal{A}}$.

Proposition I.9 (Every reward direction is optimizer-switching)**.** *Let $\mathfrak{e} = (g^0, q_+, q_-, r)$ be a ρ-regular exposed edge and put $h = q_+ - q_-$. If $v \in \mathcal{V}$ satisfies $\|v\|_\infty \leq r$ and $h(v) \geq 0$, then*

$$v \in \mathcal{W}_\rho(g^0; q_+, q_-). \tag{I.24}$$

Consequently, for every p and ε,

$$\mathfrak{T}_{\varepsilon,\rho}(g^0, q_+, q_-; p) = \mathcal{E}_\varepsilon(h; p)\,. \tag{I.25}$$

Combining with Theorem I.6,

$$\rho(1-\rho)\mathcal{E}_\varepsilon(h; p) \leq \mathfrak{B}_{\varepsilon,\rho}(g^0, q_+, q_-; p) \leq \frac{1}{2}\mathcal{E}_\varepsilon(h; p)\,. \tag{I.26}$$

Proof. Let v satisfy the hypotheses. Membership of $g^0 \pm v$ in $\mathcal{G}_\rho$ follows from (I.22). Consider first $q \in Q \cap F_Q(g^0)$. By (I.21), $q = \lambda q_+ + (1-\lambda)q_-$ for some $\lambda \in [0,1]$. Therefore

$$(q_+ - q)(v) = (1-\lambda)h(v) \geq 0, \qquad (q_- - q)(-v) = \lambda h(v) \geq 0.$$

For $q \in Q \setminus F_Q(g^0)$,

$$(q_+ - q)(g^0 + v) \geq 2r - |(q_+ - q)(v)| \geq 2r - 2\|v\|_\infty \geq 0,$$
$$(q_- - q)(g^0 - v) \geq 2r - |(q_- - q)(v)| \geq 0.$$

Thus q_+ and q_- dominate every element of Q under $g^0 + v$ and $g^0 - v$, respectively, and hence dominate all of C_Q. This proves (I.24).

The local modulus is at most the global Riesz cost by Lemma I.4. For the reverse inequality, fix any $u \in \mathcal{V}$ with $\mathfrak{s}_\varepsilon(h,u) > 0$. After replacing u by $-u$, assume $h(u) > 0$. Choose $t > 0$ so that $\|tu\|_\infty \leq r$. Then $tu \in \mathcal{W}_\rho(g^0; q_+, q_-)$, and homogeneity gives

$$\frac{(h(tu) - \varepsilon\|tu\|_{\mathcal{K}})_+^2}{\|tu\|_p^2} = \frac{\mathfrak{s}_\varepsilon(h,u)^2}{\|u\|_p^2}.$$

Taking the supremum over u and invoking Proposition I.1 proves (I.25). The final display follows from Theorem I.6. □

Corollary I.10 (Relative-interior exposed edges admit a certificate)**.** *Suppose g^0 belongs to the relative interior of $\mathcal{G}_\rho$ within the affine space $g^0 + \mathcal{V}$ and $F_Q(g^0) = \mathrm{co}\{q_+, q_-\}$. Then there exists $r > 0$ such that (g^0, q_+, q_-, r) is a ρ-regular exposed edge.*

Proof. Relative interior gives $r_{\mathrm{int}} > 0$ such that the relative ℓ_∞ ball of radius r_{int} about g^0 lies in $\mathcal{G}_\rho$. Since Q is finite, the minimum baseline gap over $Q \setminus F_Q(g^0)$ is strictly positive whenever this set is nonempty. Choose $r > 0$ smaller than r_{int} and at most half this minimum gap. □

Example I.11 (Regular edges in the tabular cube)**.** Let $\mathcal{G} = [\rho, 1-\rho]^{\mathcal{A}}$ and let $Q = \{\delta_a : a \in \mathcal{A}\}$. Fix distinct a, b and choose g^0 with $g^0(a) = g^0(b)$ strictly larger than all other coordinates, while every coordinate remains strictly inside $[\rho, 1-\rho]$. For sufficiently small $r > 0$, the tuple $(g^0, \delta_a, \delta_b, r)$ is a ρ-regular exposed edge. Thus every tabular pairwise contrast is locally Riesz-test rich with constant one.

Definition I.12 (Pairwise local richness)**.** The exposed tuple (g^0, q_+, q_-) is $(\varepsilon, \rho, \kappa)$-*Riesz-test rich* over a design set $\mathcal{P} \subseteq \Delta(\mathcal{A})$ if

$$\mathfrak{T}_{\varepsilon,\rho}(g^0, q_+, q_-; p) \geq \kappa \mathcal{E}_\varepsilon(q_+ - q_-; p) \qquad \forall p \in \mathcal{P}. \tag{I.27}$$

Every ρ-regular exposed edge is $(\varepsilon, \rho, 1)$-Riesz-test rich for every $\varepsilon \geq 0$ and every $\mathcal{P}$. This pairwise condition is the local prototype of the multiscale richness assumption used in the subsequent cover–packing appendix.

I.4 A ONE-NODE REGRET LOWER BOUND

The preceding equivalence concerns one fixed experiment. We now show how an algorithm's expected sampling design under a neutral edge model determines an oblivious two-point regret lower bound.

Let $\mathfrak{e} = (g^0, q_+, q_-, r)$ be a ρ-regular exposed edge, let $h = q_+ - q_-$, and fix $n \in \mathbb{N}$. For $p \in \Delta(\mathcal{A})$ define the finite-horizon truncated edge complexity

$$\mathcal{E}_{\varepsilon,n}^{\mathrm{edge}}(h; p \mid r, \rho) := \sup_{\substack{v \in \mathcal{V} \\ \|v\|_\infty \leq 1}} \mathfrak{s}_\varepsilon(h,v)^2 \min\left\{ nr^2, \frac{\rho(1-\rho)}{8\|v\|_p^2} \right\}. \tag{I.28}$$

When $\|v\|_p = 0$, the second term in the minimum is $+\infty$; when $\mathfrak{s}_\varepsilon(h,v) = 0$, the corresponding contribution is defined to be zero.

Proposition I.13 (Asymptotic recovery of the Riesz cost)**.** *For every fixed $h, p, r, \rho, \varepsilon$ with $r > 0$,*

$$\lim_{n\to\infty} \mathcal{E}_{\varepsilon,n}^{\mathrm{edge}}(h; p \mid r, \rho) = \frac{\rho(1-\rho)}{8} \mathcal{E}_\varepsilon(h; p)\,, \tag{I.29}$$

where both sides may be $+\infty$.

Proof. The integrand in (I.28) increases pointwise in n. Hence the supremum and monotone limit commute: $\lim_n \sup_v f_n(v) = \sup_v \sup_n f_n(v)$. Every nonzero direction can be rescaled to have ℓ_∞ norm at most one, and the quotient in Proposition I.1 is invariant under positive rescaling. Therefore

$$\lim_{n\to\infty} \mathcal{E}^{\mathrm{edge}}_{\varepsilon,n}(h; p \mid r, \rho) = \sup_{\substack{v\in\mathcal{V} \\ \|v\|_\infty \le 1}} \frac{\rho(1-\rho)}{8} \frac{\mathfrak{s}_\varepsilon(h,v)^2}{\|v\|_p^2} = \frac{\rho(1-\rho)}{8} \mathcal{E}_\varepsilon(h; p)\,.$$

The first equality remains valid in the extended sense when a positive excess contrast occurs in a p-null direction. □

Theorem I.14 (Neutral-reference one-node lower bound)**.** *Let $\mathfrak{e} = (g^0, q_+, q_-, r)$ be a ρ-regular exposed edge and let $h = q_+ - q_-$. Fix $n \ge 1$, $\varepsilon \ge 0$, and any Q-mediated behavioral algorithm* Alg*. Run* Alg *for n rounds under the neutral oblivious Bernoulli environment with constant mean g^0, and define its expected average sampling law*

$$\overline{p}^0_{\mathsf{Alg}} := \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_0[p_t] \in C_Q. \tag{I.30}$$

Then there are a direction $v \in \mathcal{V}$, an amplitude $\lambda \in [0, r]$, and a sign $\sigma \in \{-1, +1\}$ such that the constant mean vector $g_\sigma = g^0 + \sigma\lambda v$ belongs to $\mathcal{G}_\rho$ and

$$\mathbb{E}_\sigma[\mathrm{Reg}^Q_n] \ge \frac{\sqrt{n}}{4}\left[\mathcal{E}^{\mathrm{edge}}_{\varepsilon,n}\left(h; \overline{p}^0_{\mathsf{Alg}} \mid r, \rho\right)\right]^{1/2}. \tag{I.31}$$

Consequently,

$$\mathfrak{M}^{\star,\mathrm{obl}}_n(Q, \mathcal{G}) \ge \frac{\sqrt{n}}{4} \inf_{p\in C_Q} \left[\mathcal{E}^{\mathrm{edge}}_{\varepsilon,n}(h; p \mid r, \rho)\right]^{1/2}. \tag{I.32}$$

The hard alternative may depend on the algorithm, as permitted by the order $\inf_{\mathsf{Alg}} \sup_{\mathsf{Adv}}$ in the minimax value.

Proof. For fixed $\overline{p}^0_{\mathsf{Alg}}$, the integrand in (I.28) is continuous on the compact set $\{v \in \mathcal{V} : \|v\|_\infty \le 1\}$ under the convention in (I.28); hence its supremum is attained. Fix a maximizing v. If $\mathfrak{s}_\varepsilon(h, v) = 0$, the desired bound for this direction is trivial. Otherwise replace v by $-v$ if needed so that $h(v) > 0$, and set

$$\lambda := \min\left\{ r, \left(\frac{\rho(1-\rho)}{8n\,\|v\|^2_{\overline{p}^0_{\mathsf{Alg}}}}\right)^{1/2} \right\}, \tag{I.33}$$

where the second term is $+\infty$ when $\|v\|_{\overline{p}^0_{\mathsf{Alg}}} = 0$. Since $\|\lambda v\|_\infty \le r$ and $h(\lambda v) > 0$, Proposition I.9 gives $\lambda v \in \mathcal{W}_\rho(g^0; q_+, q_-)$. Let P_0, P_+, P_- be the prefix transcript laws under the constant Bernoulli means $g^0, g^0 + \lambda v, g^0 - \lambda v$, respectively.

For every $t \le n$ and $\sigma \in \{-1, +1\}$, the interactive KL chain rule and the Bernoulli upper bound in Appendix H give

$$D_{\mathrm{KL}}(P_0^{t-1} \| P_\sigma^{t-1}) \le \frac{\lambda^2}{\rho(1-\rho)} \mathbb{E}_0 \sum_{s=1}^{t-1} p_s(v^2) \tag{I.34}$$

$$\le \frac{\lambda^2 n}{\rho(1-\rho)} \|v\|^2_{\overline{p}^0_{\mathsf{Alg}}} \le \frac{1}{8}. \tag{I.35}$$

Thus Pinsker's inequality implies

$$\|P_0^{t-1} - P_\sigma^{t-1}\|_{\mathrm{TV}} \le \frac{1}{4}. \tag{I.36}$$

Set

$$a := \lambda \mathfrak{s}_\varepsilon(h, v), \qquad \widetilde{\ell}_\sigma(p') := \min\{\ell_\sigma(p'), a\}.$$

By Lemma I.7, $\ell_+(p') + \ell_-(p') \geq \lambda h(v) \geq a$. For nonnegative numbers whose sum is at least a, clipping both at a preserves a sum of at least a; hence

$$\widetilde{\ell}_+(p') + \widetilde{\ell}_-(p') \geq a \qquad \forall p' \in C_Q. \tag{I.37}$$

The random variable $\widetilde{\ell}_\sigma(p_t)$ is a measurable function of H_{t-1} with range in $[0, a]$. By the bounded transport inequality and (I.36),

$$\mathbb{E}_\sigma \widetilde{\ell}_\sigma(p_t) \geq \mathbb{E}_0 \widetilde{\ell}_\sigma(p_t) - \frac{a}{4}.$$

Averaging over the two signs and using (I.37) under P_0 yields

$$\frac{1}{2} \sum_{\sigma \in \{-1,+1\}} \mathbb{E}_\sigma \ell_\sigma(p_t) \geq \frac{a}{4}. \tag{I.38}$$

Since q_σ is optimal under g_σ and the expected physical reward at round t is $p_t(g_\sigma)$,

$$\mathbb{E}_\sigma[\mathrm{Reg}_n^Q] = \mathbb{E}_\sigma \sum_{t=1}^n \ell_\sigma(p_t).$$

Summing (I.38) gives

$$\max_{\sigma \in \{-1,+1\}} \mathbb{E}_\sigma[\mathrm{Reg}_n^Q] \geq \frac{n\lambda}{4} \mathfrak{s}_\varepsilon(h, v) \tag{I.39}$$

$$= \frac{\sqrt{n}}{4} \mathfrak{s}_\varepsilon(h, v) \left[\min \left\{ nr^2, \frac{\rho(1-\rho)}{8 \|v\|^2_{\overline{p}^0_{\mathsf{Alg}}}} \right\} \right]^{1/2}. \tag{I.40}$$

Using the maximizing v proves (I.31). Finally, $\overline{p}^0_{\mathsf{Alg}} \in C_Q$ because every $p_t \in C_Q$ and C_Q is convex. Taking the infimum over algorithms proves (I.32). □

Remark I.15 (Why the neutral reference is essential). The hard direction in Theorem I.14 is selected from the expected design generated by the neutral tie model g^0. This avoids a circular definition in which the sampling design is evaluated under an alternative that itself depends on that design. The reference environment is a valid member of the convex reward class, while the two alternatives remain statistically close to its transcript law by (I.34).

I.5 Multiway optimizer stars

Binary edges are sufficient for Assouad-type trees. Multiway nodes can be more efficient when one observation design must distinguish many optimizer cells. We record the exact local interface needed by the list-Fano theorem in Appendix H.

Definition I.16 (Interior optimizer star)**.** An M-ary ρ-interior optimizer star for Q is a collection

$$\mathfrak{S} = \{(g_i, q_i) : i \in [M]\} \subseteq \mathcal{G}_\rho \times Q \tag{I.41}$$

with $q_i \in F_Q(g_i)$ for every i. Its loss functions are

$$\ell_i(p) := q_i(g_i) - p(g_i), \qquad p \in C_Q. \tag{I.42}$$

The star is (Δ, L)-separated if $\Delta > 0$, $1 \leq L < M$, and

$$|\{i \in [M] : \ell_i(p) < \Delta\}| \leq L \qquad \forall p \in C_Q. \tag{I.43}$$

Let

$$\overline{g} := \frac{1}{M} \sum_{i=1}^M g_i \in \mathcal{G}_\rho, \qquad w_i := g_i - \overline{g}. \tag{I.44}$$

The average belongs to $\mathcal{G}_\rho$ by convexity.

Theorem I.17 (One-node multiway testing bound)**.** *Let $\mathfrak{S}$ be a (Δ, L)-separated M-ary ρ-interior optimizer star. Fix a Q-mediated behavioral algorithm and, for each $i \in [M]$, let P_i be its n-round transcript law under the constant oblivious Bernoulli mean g_i. Then*

$$\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_i[\mathrm{Reg}_n^Q] \geq n\Delta\left[1 - \frac{\mathcal{I}_{\mathfrak{S}}(\mathsf{Alg}, n) + \log 2}{\log(M/L)}\right]_+, \tag{I.45}$$

where

$$\mathcal{I}_{\mathfrak{S}}(\mathsf{Alg}, n) := \mathrm{I}(U; H_n) \tag{I.46}$$

$$\leq \frac{1}{M\rho(1-\rho)}\sum_{i=1}^{M}\mathbb{E}_i\sum_{t=1}^{n}p_t(w_i^2), \tag{I.47}$$

for $U \sim \mathrm{Unif}[M]$ and $H_n \mid U = i \sim P_i$. Thus a star with few simultaneously good optimizer cells and small average observation energy forces linear-in-block regret.

Proof. Under model i, q_i is optimal and the conditional expected physical reward is $p_t(g_i)$. Hence, with the random average sampling law $\overline{p}_n = n^{-1}\sum_{t=1}^{n} p_t \in C_Q$,

$$\mathbb{E}_i[\mathrm{Reg}_n^Q] = n\mathbb{E}_i[\ell_i(\overline{p}_n)]. \tag{I.48}$$

Apply Corollary H.12 with latent variable U, observation H_n, decision $\overline{p}_n$, and loss $\ell(U, p) = \ell_U(p)$. The list-separation condition (I.43) gives (I.45) with mutual information $\mathrm{I}(U; H_n)$.

Let $\overline{P}$ be the transcript law of the same algorithm under the constant Bernoulli mean $\overline{g}$. The information-radius identity in Lemma H.8 gives

$$\mathrm{I}(U; H_n) \leq \frac{1}{M}\sum_{i=1}^{M} D_{\mathrm{KL}}(P_i \| \overline{P}).$$

Using the interactive KL chain rule and the Bernoulli quadratic upper bound,

$$D_{\mathrm{KL}}(P_i\|\overline{P}) \leq \frac{1}{\rho(1-\rho)}\mathbb{E}_i\sum_{t=1}^{n}p_t\big((g_i - \overline{g})^2\big) = \frac{1}{\rho(1-\rho)}\mathbb{E}_i\sum_{t=1}^{n}p_t(w_i^2).$$

Averaging over i proves (I.46). □

Remark I.18 (Binary and multiway nodes use different testing inequalities). For $M = 2$ and $L = 1$, the additive $\log 2$ term makes the elementary Fano bound vacuous. Theorem I.14 instead uses the two-point transport inequality at every prefix and gives a nontrivial binary edge bound. For large branching number, Theorem I.17 captures the logarithmic cost of identifying one among many optimizer cells.

I.6 Interface with sequential optimizer packings

The results above provide the local ingredients for the subsequent sequential-packing appendix. A valid packing node will specify either:

1. a regular exposed edge, to which Theorem I.14 applies; or
2. a (Δ, L)-separated optimizer star, to which Theorem I.17 applies conditionally on the ancestral branch information using Theorem H.16.

The local Riesz-test richness condition (I.27) is geometric and independent of the eventual packing value. On a regular exposed edge it holds with constant one, and Theorem I.6 then identifies the Riesz cost with the optimal squared decision separation per unit of Bernoulli information, up to constants depending only on ρ. What remains is to choose disjoint or properly chargeable time blocks, preserve ancestral optimizer decisions under finer-scale perturbations, and aggregate the nodewise lower bounds without allowing the hard family to depend on the realized learner trajectory. These are the tasks of the subsequent sequential-packing and cover–packing appendices.

J SEQUENTIAL OPTIMIZER-PACKING LOWER GAMES

This appendix composes the local testing constructions of Appendix I across disjoint time blocks. The resulting object is an optimizer-indexed analogue of a sequential packing: every root-to-leaf path specifies an oblivious Bernoulli environment and a single fixed witness comparator, while every internal node encodes a statistically difficult decision among child optimizer cells. The construction is tailored to the reward-only mediator game from Appendix D. It uses the conditional information tools in Appendix H, but the packing itself is chosen before interaction and is independent of the learner's realized trajectory.

Interactive Fano and constrained-DEC lower bounds permit algorithm-dependent information acquisition and reference experiments (Chen et al., 2024; Foster et al., 2023). Our contribution here is a concrete tree factorization for adversarial noisy bandits: ancestral branch variables determine earlier reward blocks, descendant variables are conditionally fresh when their blocks begin, and one leaf witness simultaneously certifies all blockwise comparator losses. This last compatibility condition is what allows nodewise testing errors to add without double counting.

Throughout this appendix, fix a finite nonempty witness family $Q \subseteq \Delta(\mathcal{A})$, write $C_Q = \mathrm{co}(Q)$, and fix $\rho \in (0, 1/2)$. All reward vectors used below belong to $\mathcal{G}_\rho = \mathcal{G} \cap [\rho, 1-\rho]^{\mathcal{A}}$, and all observations are Bernoulli with their displayed means.

J.1 ROOTED TREES, BLOCK SCHEDULES, AND LATENT PATHS

A finite rooted tree of depth $D \geq 1$ is represented as a prefix-closed set

$$\mathbb{T} \subseteq \bigcup_{j=0}^{D} \mathbb{N}^j$$

containing the empty word $\varnothing$, such that every leaf has depth D. If $u \in \mathbb{T}$ has depth $|u| < D$, its children are

$$\mathrm{ch}(u) = \{u1, \dots, uM_u\}, \qquad M_u \geq 2.$$

Let $\mathbb{T}^\circ$ and $\mathbb{T}^\partial$ denote the internal nodes and leaves, respectively. For a leaf $\theta = (\theta_1, \dots, \theta_D)$, write $\theta_{<j} = (\theta_1, \dots, \theta_{j-1})$ and $\theta_{\leq j} = (\theta_1, \dots, \theta_j)$.

The canonical product-uniform prior on $\mathbb{T}^\partial$ is obtained by selecting, conditionally on the current node u, the next child uniformly from $[M_u]$. Its node masses are defined recursively by

$$\omega(\varnothing) = 1, \qquad \omega(ui) = \frac{\omega(u)}{M_u}. \tag{J.1}$$

Thus $\omega(u)$ is the probability that a random latent path visits u, and

$$\sum_{u \in \mathbb{T}:\, |u| = j} \omega(u) = 1 \qquad (j = 0, \dots, D). \tag{J.2}$$

A depth-D block schedule is a deterministic partition

$$0 = \tau_0 < \tau_1 < \cdots < \tau_D = T, \qquad I_j := \{\tau_{j-1} + 1, \dots, \tau_j\}, \qquad n_j := |I_j|. \tag{J.3}$$

The common schedule across all latent paths is essential: it gives disjoint blocks whose positions do not reveal the hidden branch.

J.2 SEQUENTIAL OPTIMIZER PACKINGS

For $g \in \mathcal{G}$ and $q \in Q$, define

$$V_Q(g) := \max_{r \in Q} r(g) = \max_{r \in C_Q} r(g), \qquad \ell_{g,q}(p) := q(g) - p(g), \quad p \in C_Q.$$

At a packing node, the anchor policy will be optimal, so its local loss is nonnegative.

Definition J.1 (Sequential optimizer packing)**.** A horizon-T, ρ-interior sequential optimizer packing for $(Q, \mathcal{G})$ is a tuple

$$\mathscr{P} = \big(\mathbb{T}, (\tau_j)_{j=0}^{D}, (q_\theta)_{\theta \in \mathbb{T}^\partial}, \mathbb{T}_{\mathrm{F}}, \mathbb{T}_{\mathrm{E}}, \mathsf{D}\big) \tag{J.4}$$

with the following data and properties.

1. The sets $\mathbb{T}_{\mathrm{F}}$ and $\mathbb{T}_{\mathrm{E}}$ form a partition of $\mathbb{T}^\circ$. They are called the *Fano nodes* and *edge nodes*.
2. Every leaf $\theta \in \mathbb{T}^\partial$ is assigned a fixed comparator $q_\theta \in Q$.
3. For every internal node u and child $i \in [M_u]$, the node data D specify
$$(g_{u,i}, q_{u,i}) \in \mathcal{G}_\rho \times Q, \qquad q_{u,i}(g_{u,i}) = V_Q(g_{u,i}). \tag{J.5}$$
Write
$$\ell_{u,i}(p) := V_Q(g_{u,i}) - p(g_{u,i}), \qquad p \in C_Q. \tag{J.6}$$
4. Every internal node u has a comparator-compatibility error $\varepsilon_u \in [0, 1]$ such that, for every child i and every descendant leaf $\theta \succeq ui$,
$$0 \le V_Q(g_{u,i}) - q_\theta(g_{u,i}) \le \varepsilon_u. \tag{J.7}$$
5. Every Fano node $u \in \mathbb{T}_{\mathrm{F}}$ has parameters $\Delta_u \in (0, 1]$ and $L_u \in \{1, \dots, M_u - 1\}$ satisfying
$$|\{i \in [M_u] : \ell_{u,i}(p) < \Delta_u\}| \le L_u \qquad \forall p \in C_Q. \tag{J.8}$$
6. Every edge node $u \in \mathbb{T}_{\mathrm{E}}$ has $M_u = 2$ and a parameter $\Delta_u \in (0, 1]$ satisfying
$$\ell_{u,1}(p) + \ell_{u,2}(p) \ge 2\Delta_u \qquad \forall p \in C_Q. \tag{J.9}$$

The set of all such packings is denoted by $\mathfrak{P}_{T,\rho}(Q, \mathcal{G})$.

Condition (J.7) is the global gluing requirement. It does not demand that q_θ be optimal in every ancestral block; it only requires a controlled approximation error. When all $\varepsilon_u = 0$, we call the packing *exact*. The node separation conditions are stated over the entire physical decision set C_Q, so they remain valid for arbitrary randomized Q-mediated algorithms.

J.3 The induced oblivious Bernoulli family

A leaf $\theta \in \mathbb{T}^\partial$ determines the piecewise-constant reward sequence

$$g_t^\theta := g_{\theta_{<j}, \theta_j} \qquad \text{for } t \in I_j, \quad j = 1, \dots, D. \tag{J.10}$$

Let Adv_θ be the oblivious Bernoulli environment with means $g_{1:T}^\theta$. Every Adv_θ belongs to $\mathfrak{E}_{T,\rho}^{\mathrm{obl,Ber}}(\mathcal{G})$.

Fix a Q-mediated behavioral algorithm Alg. Sample a leaf $\Theta = (\Theta_1, \dots, \Theta_D)$ from the product-uniform prior and then run Alg against Adv_Θ. Write $\mathbf{P}_{\mathscr{P}}^{\mathsf{Alg}}$ for the resulting joint law of (Θ, H_T). For an internal node u of depth $j - 1$, let

$$E_u := \{\Theta_{<j} = u\}. \tag{J.11}$$

Since $\omega(u) > 0$, all conditional laws below are well defined.

Lemma J.2 (Ancestral independence)**.** *For every internal node u of depth $j - 1$, under $\mathbf{P}_{\mathscr{P}}^{\mathsf{Alg}}(\cdot \mid E_u)$,*

$$\Theta_j \sim \mathrm{Unif}[M_u], \qquad \Theta_j \perp\!\!\!\perp H_{\tau_{j-1}}. \tag{J.12}$$

Moreover, conditional on E_u and $\Theta_j = i$, the reward vectors before τ_{j-1} do not depend on i, and the reward vector throughout I_j is the constant vector $g_{u,i}$.

Proof. The product-uniform prior chooses Θ_j only after the prefix $\Theta_{<j} = u$ has been fixed, and the choice is uniform on $[M_u]$. By (J.10), all reward vectors in blocks $I_1, \dots, I_{j-1}$ depend only on $\Theta_{<j}$. Conditional on E_u, the learner's action kernels and all pre-block observation laws therefore have the same conditional distribution for every value of Θ_j. Induction over rounds up to τ_{j-1} gives the conditional independence. The final assertion follows directly from (J.10). □

For u at depth $j-1$, define the random average physical decision in its block by

$$\overline{p}_u := \frac{1}{n_j} \sum_{t \in I_j} p_t \in C_Q. \tag{J.13}$$

This is an H_{τ_j}-measurable C_Q-valued decision.

For a Fano node, define the conditional block information

$$\begin{aligned} \mathcal{I}_u(\mathsf{Alg}; \mathscr{P}) &:= \mathrm{I}_{\mathbf{P}^{\mathsf{Alg}}_{\mathscr{P}}(\cdot \mid E_u)} \left(\Theta_j; H_{I_j} \mid H_{\tau_{j-1}}\right) \\ &= \sum_{t \in I_j} \mathrm{I}\left(\Theta_j; Y_t \mid E_u, H_{t-1}, A_t\right), \end{aligned} \tag{J.14}$$

where the second equality follows from Proposition H.15 and Lemma J.2.

For an edge node u, let $P^{\mathsf{Alg}}_{u,i}$ be the law of H_{τ_j} under $\mathbf{P}^{\mathsf{Alg}}_{\mathscr{P}}(\cdot \mid E_u, \Theta_j = i)$, and define its conditional testing affinity

$$\mathcal{A}_u(\mathsf{Alg}; \mathscr{P}) := 1 - \left\| P^{\mathsf{Alg}}_{u,1} - P^{\mathsf{Alg}}_{u,2} \right\|_{\mathrm{TV}}. \tag{J.15}$$

The pre-block marginal of $H_{\tau_{j-1}}$ is identical under the two child laws by Lemma J.2; hence (J.15) measures only the information created during I_j, although it is notationally convenient to use the full prefix law.

J.4 Nodewise decision lower bounds

Lemma J.3 (Conditional node risk)**.** *Fix a packing $\mathscr{P}$, a Q-mediated algorithm* Alg*, and an internal node u at depth $j-1$. Then*

$$\mathbb{E}\left[\ell_{u,\Theta_j}(\overline{p}_u) \mid E_u\right] \geq \mathfrak{r}_u(\mathsf{Alg}; \mathscr{P}), \tag{J.16}$$

where

$$\mathfrak{r}_u(\mathsf{Alg}; \mathscr{P}) := \begin{cases} \Delta_u \left[1 - \dfrac{\mathcal{I}_u(\mathsf{Alg}; \mathscr{P}) + \log 2}{\log(M_u/L_u)}\right]_+, & u \in \mathbb{T}_{\mathrm{F}}, \\ \Delta_u \, \mathcal{A}_u(\mathsf{Alg}; \mathscr{P}), & u \in \mathbb{T}_{\mathrm{E}}. \end{cases} \tag{J.17}$$

Proof. Suppose first that u is a Fano node. Conditional on E_u, the child Θ_j is uniform and independent of $H_{\tau_{j-1}}$ by Lemma J.2. Under the conditional law given E_u, apply Theorem H.16 with trivial side information, latent variable Θ_j, decision $\overline{p}_u$, and loss $\ell(i,p) = \ell_{u,i}(p)$. The list-separation hypothesis is exactly (J.8); this gives the first case of (J.17).

Now suppose u is an edge node. Apply the decision-theoretic Le Cam inequality from Lemma H.2 to the two conditional transcript laws $P^{\mathsf{Alg}}_{u,1}, P^{\mathsf{Alg}}_{u,2}$, the decision $\overline{p}_u$, and losses $\ell_{u,1}, \ell_{u,2}$. Condition (J.9) gives

$$\ell_{u,1}(p) + \ell_{u,2}(p) \geq 2\Delta_u \qquad \forall p \in C_Q,$$

so the conditional Bayes risk is at least $\Delta_u(1 - \|P^{\mathsf{Alg}}_{u,1} - P^{\mathsf{Alg}}_{u,2}\|_{\mathrm{TV}})$, which is the second case. □

J.5 The no-double-counting comparator decomposition

Proposition J.4 (Bayes comparator decomposition)**.** *For every packing $\mathscr{P}$ and every Q-mediated algorithm* Alg*,*

$$\mathbb{E}_\Theta \mathbb{E}_{\mathsf{Alg},\mathsf{Adv}_\Theta}[\mathrm{Reg}_T^Q] \geq \sum_{u \in \mathbb{T}^\circ} \omega(u) n_{|u|+1} \left[\mathfrak{r}_u(\mathsf{Alg}; \mathscr{P}) - \varepsilon_u\right]. \tag{J.18}$$

Proof. For each leaf θ, the maximum in Reg_T^Q dominates the fixed leaf comparator q_θ. Using $\mathbb{E}[g_t^\theta(A_t) \mid \mathscr{H}_{t-1}] = p_t(g_t^\theta)$ from Lemma A.4 therefore gives

$$\mathbb{E}_{\mathsf{Alg},\mathsf{Adv}_\theta}[\mathrm{Reg}_T^Q] \geq \mathbb{E}_{\mathsf{Alg},\mathsf{Adv}_\theta} \sum_{t=1}^{T} \left[q_\theta(g_t^\theta) - p_t(g_t^\theta)\right]. \tag{J.19}$$

For $t \in I_j$, put $u = \theta_{<j}$ and $i = \theta_j$. By (J.7),

$$q_\theta(g_{u,i}) - p_t(g_{u,i}) \geq \ell_{u,i}(p_t) - \varepsilon_u.$$

Averaging over the product-uniform prior, grouping leaves according to the node they visit, and using linearity of $p \mapsto \ell_{u,i}(p)$ gives

$$\mathbb{E}_\Theta \mathbb{E}_{\mathsf{Alg},\mathsf{Adv}_\Theta}[\mathrm{Reg}_T^Q] \geq \sum_{j=1}^{D} \sum_{\substack{u \in \mathbb{T}^\circ \\ |u|=j-1}} \omega(u) n_j \left\{ \mathbb{E}\left[\ell_{u,\Theta_j}(\overline{p}_u) \mid E_u\right] - \varepsilon_u \right\}.$$

Applying Lemma J.3 proves the claim. The blocks $I_1, \ldots, I_D$ are disjoint, and the same leaf comparator q_θ is used in all of them; hence no round or comparator contribution is counted more than once. □

J.6 Operational sequential packing value

For a packing and an algorithm, define

$$\Phi_T(\mathscr{P}, \mathsf{Alg}) := \sum_{u \in \mathbb{T}^\circ} \omega(u) n_{|u|+1} \left[\mathfrak{r}_u(\mathsf{Alg}; \mathscr{P}) - \varepsilon_u\right]. \tag{J.20}$$

The packing is fixed before the algorithm is run. Algorithm dependence enters only through the transcript information and testing affinities induced by that fixed family of environments.

$$\boxed{\sup_{\text{fixed packing } \mathscr{P}} \left[\inf_{\text{behavioral Alg}} \Phi_T(\mathscr{P}, \mathsf{Alg})\right]_+} \tag{J.21}$$

The environment family is algorithm-agnostic, whereas its operational value retains the learner-induced transcript information.

Definition J.5 (Operational sequential optimizer-packing value)**.** Define

$$\Pi^{\mathrm{seq}}_{T,\rho}(Q, \mathcal{G}) := \sup_{\mathscr{P} \in \mathfrak{P}_{T,\rho}(Q,\mathcal{G})} \left[\inf_{\mathsf{Alg} \in \mathfrak{A}_T(Q)} \Phi_T(\mathscr{P}, \mathsf{Alg})\right]_+. \tag{J.22}$$

Here $\mathfrak{A}_T(Q)$ is the class of Q-mediated behavioral algorithms from Appendix D.3; if the packing class is empty, the supremum in (J.22) is defined to be zero. The positive part is included because approximate comparator compatibility may make the displayed packing payoff negative, whereas the minimax regret is nonnegative.

Theorem J.6 (Restatement of Theorem 6.2)**.** *For every finite nonempty $Q \subseteq \Delta(\mathcal{A})$, $\rho \in (0, 1/2)$, and horizon T,*

$$\mathfrak{M}^{\star,\mathrm{obl}}_T(Q, \mathcal{G}) \geq \Pi^{\mathrm{seq}}_{T,\rho}(Q, \mathcal{G}). \tag{J.23}$$

Consequently, for the Dirac witness family $Q_\mathcal{A} = \{\delta_a : a \in \mathcal{A}\}$,

$$\mathfrak{R}^{\star,\mathrm{obl}}_T(\mathcal{G}) \geq \Pi^{\mathrm{seq}}_{T,\rho}(Q_\mathcal{A}, \mathcal{G}), \tag{J.24}$$

and the same lower bound holds for $\mathfrak{R}^{\star,\mathrm{ad}}_T(\mathcal{G})$.

Proof. Fix a packing $\mathscr{P}$ and a Q-mediated algorithm Alg. Every leaf θ defines an admissible oblivious Bernoulli environment. Hence

$$\begin{aligned}\sup_{\mathsf{Adv} \in \mathfrak{E}^{\mathrm{obl}}_T(\mathcal{G})} \mathbb{E}_{\mathsf{Alg},\mathsf{Adv}}[\mathrm{Reg}_T^Q] &\geq \max_{\theta \in \mathbb{T}^\partial} \mathbb{E}_{\mathsf{Alg},\mathsf{Adv}_\theta}[\mathrm{Reg}_T^Q] \\ &\geq \mathbb{E}_\Theta \mathbb{E}_{\mathsf{Alg},\mathsf{Adv}_\Theta}[\mathrm{Reg}_T^Q] \\ &\geq \Phi_T(\mathscr{P}, \mathsf{Alg}),\end{aligned}$$

where the last inequality is Proposition J.4. Taking the infimum over $\mathsf{Alg} \in \mathfrak{A}_T(Q)$ yields

$$\mathfrak{M}^{\star,\mathrm{obl}}_T(Q, \mathcal{G}) \geq \inf_{\mathsf{Alg} \in \mathfrak{A}_T(Q)} \Phi_T(\mathscr{P}, \mathsf{Alg}).$$

The left side is nonnegative: repeating any fixed $g \in \mathcal{G}$ yields nonnegative expected mediator regret because $p_t \in C_Q$ and $V_Q(g) - p_t(g) \geq 0$ conditionally in every round. It therefore also dominates the positive part of the right side. Taking the supremum over packings proves (J.23). For the Dirac family, the mediator and original games coincide by Proposition D.7, proving (J.24). Finally, $\mathfrak{R}_T^{\star,\mathrm{ad}} \geq \mathfrak{R}_T^{\star,\mathrm{obl}}$ by Proposition A.8. □

Remark J.7 (Non-tautological content). The complexity in (J.22) is a minimax game, but it is not defined through regret. A packing is specified only by reward vectors, optimizer witnesses, a tree prior, block lengths, compatibility errors, and local separation conditions. The learner enters through conditional mutual information and total-variation affinity, quantities that can be bounded independently by channel or Riesz geometry. The subsequent cover–packing appendix will compare this lower game with the chained upper design under an explicit multiscale richness condition.

J.7 Static channel relaxations

The operational value above retains the algorithm's actual information acquisition. For explicit constructions it is convenient to upper-bound that information by a static one-round channel capacity.

For a Fano node u, and $\lambda \in \Delta([M_u])$, let

$$\overline{g}_{u,\lambda} := \sum_{i=1}^{M_u} \lambda_i g_{u,i} \in \mathcal{G}_\rho. \tag{J.25}$$

Define its constrained Bernoulli channel capacity by

$$\mathsf{C}_u(Q) := \sup_{\substack{\lambda \in \Delta([M_u]) \\ p \in C_Q}} \sum_{a \in \mathcal{A}} p(a) \sum_{i=1}^{M_u} \lambda_i \,\mathrm{kl}\left(g_{u,i}(a) \,\middle\|\, \overline{g}_{u,\lambda}(a)\right). \tag{J.26}$$

For an edge node, define the directed one-round divergences

$$\mathsf{K}_u^{i \to k}(Q) := \sup_{p \in C_Q} \sum_{a \in \mathcal{A}} p(a) \,\mathrm{kl}\left(g_{u,i}(a) \,\|\, g_{u,k}(a)\right), \tag{J.27}$$

and

$$\mathsf{K}_u(Q) := \min\left\{\mathsf{K}_u^{1\to 2}(Q), \mathsf{K}_u^{2\to 1}(Q)\right\}. \tag{J.28}$$

Proposition J.8 (Uniform information bounds)**.** *For every packing, every Q-mediated algorithm, and every internal node u of depth $j-1$,*

$$\mathcal{I}_u(\mathsf{Alg}; \mathscr{P}) \leq n_j \mathsf{C}_u(Q), \qquad u \in \mathbb{T}_{\mathrm{F}}, \tag{J.29}$$

$$\mathcal{A}_u(\mathsf{Alg}; \mathscr{P}) \geq \frac{1}{2} \exp\left[-n_j \mathsf{K}_u(Q)\right], \qquad u \in \mathbb{T}_{\mathrm{E}}. \tag{J.30}$$

Moreover,

$$\mathsf{C}_u(Q) \leq \frac{1}{\rho(1-\rho)} \sup_{\substack{\lambda \in \Delta([M_u]) \\ p \in C_Q}} \sum_{i=1}^{M_u} \lambda_i p\left((g_{u,i} - \overline{g}_{u,\lambda})^2\right), \tag{J.31}$$

$$\mathsf{K}_u(Q) \leq \frac{1}{\rho(1-\rho)} \sup_{p \in C_Q} p\left((g_{u,1} - g_{u,2})^2\right). \tag{J.32}$$

Proof. Fix a Fano node and condition on E_u. At each $t \in I_j$, let λ_t be the posterior law of Θ_j given (E_u, H_{t-1}, A_t). Conditional on $A_t = a$, the observation is a mixture of Bernoulli laws with means $g_{u,i}(a)$ and posterior weights $\lambda_t(i)$; its mixture mean is $\overline{g}_{u,\lambda_t}(a)$. Therefore

$$\mathrm{I}(\Theta_j; Y_t \mid E_u, H_{t-1}, A_t) = \mathbb{E}\left[\sum_a p_t(a) \sum_i \lambda_t(i) \,\mathrm{kl}\left(g_{u,i}(a) \,\middle\|\, \overline{g}_{u,\lambda_t}(a)\right) \middle| E_u\right] \leq \mathsf{C}_u(Q),$$

where $p_t \in C_Q$. Summing over the block and using (J.14) proves (J.29).

For an edge node, apply the interactive KL chain rule conditionally on E_u. The child laws have the same pre-block history marginal, and throughout I_j their means are the fixed vectors $g_{u,1}$ and $g_{u,2}$. Thus

$$D_{\mathrm{KL}}(P^{\mathsf{Alg}}_{u,i} \| P^{\mathsf{Alg}}_{u,k}) \le n_j \mathsf{K}_u^{i\to k}(Q).$$

The Bretagnolle–Huber inequality in Lemma H.1 gives

$$1 - \|P^{\mathsf{Alg}}_{u,1} - P^{\mathsf{Alg}}_{u,2}\|_{\mathrm{TV}} \ge \frac{1}{2}\exp\left(-\min_{i\neq k} D_{\mathrm{KL}}(P^{\mathsf{Alg}}_{u,i} \| P^{\mathsf{Alg}}_{u,k})\right),$$

which proves (J.30). Finally, the Bernoulli inequality $\mathrm{kl}(x\|y) \le (x-y)^2/[y(1-y)]$ for $x, y \in [\rho, 1-\rho]$ yields (J.31) and (J.32). □

Define the static channel-relaxed payoff

$$\underline{\Phi}_T(\mathscr{P}) := \sum_{u\in\mathbb{T}_{\mathrm{F}}} \omega(u) n_{|u|+1} \left\{ \Delta_u \left[1 - \frac{n_{|u|+1}\mathsf{C}_u(Q) + \log 2}{\log(M_u/L_u)}\right]_+ - \varepsilon_u \right\} + \sum_{u\in\mathbb{T}_{\mathrm{E}}} \omega(u) n_{|u|+1} \left\{ \frac{\Delta_u}{2} e^{-n_{|u|+1}\mathsf{K}_u(Q)} - \varepsilon_u \right\}. \tag{J.33}$$

Corollary J.9 (Static channel packing lower bound)**.** *For every finite Q,*

$$\mathfrak{M}_T^{\star,\mathrm{obl}}(Q,\mathcal{G}) \ge \sup_{\mathscr{P}\in\mathfrak{P}_{T,\rho}(Q,\mathcal{G})} [\underline{\Phi}_T(\mathscr{P})]_+ . \tag{J.34}$$

Proof. Proposition J.8 implies $\Phi_T(\mathscr{P}, \mathsf{Alg}) \ge \underline{\Phi}_T(\mathscr{P})$ for every Alg. Substitute this inequality into Theorem J.6. □

J.8 Consistency checks and additive multiscale forms

Corollary J.10 (Depth-one recovery of local testing bounds)**.** *A depth-one Fano packing with exact leaf comparators recovers the optimizer-star lower bound of Theorem I.17. A depth-one edge packing with exact leaf comparators and pairwise loss separation $\ell_1 + \ell_2 \ge 2\Delta$ satisfies*

$$\mathfrak{M}_T^{\star,\mathrm{obl}}(Q,\mathcal{G}) \ge \frac{T\Delta}{2} e^{-T\mathsf{K}(Q)}, \tag{J.35}$$

where $\mathsf{K}(Q)$ is the edge capacity in (J.28)*.*

Proof. For a depth-one packing, the root has mass one and its block is the entire horizon. The two claims are the Fano and edge terms of Corollary J.9. □

In particular, let (g^0, q_+, q_-, r) be a ρ-regular exposed edge from Definition I.8, put $h = q_+ - q_-$, and choose $w \in \mathcal{W}_\rho(g^0; q_+, q_-)$. With $g_\pm = g^0 \pm w$, Lemma I.7 gives $\ell_+(p) + \ell_-(p) \ge h(w)$, so one may take $\Delta = h(w)/2$. Since $g_+ - g_- = 2w$,

$$\mathfrak{M}_T^{\star,\mathrm{obl}}(Q,\mathcal{G}) \ge \frac{Th(w)}{4}\exp\left\{-\frac{4T}{\rho(1-\rho)}\sup_{p\in C_Q} p(w^2)\right\}. \tag{J.36}$$

This is the fixed-direction, operational counterpart of the sharper neutral-reference theorem in Appendix I.4, where the hard Riesz direction may be selected after the algorithm is fixed.

Corollary J.11 (Additive depthwise lower bound)**.** *Suppose an exact packing has the following depth-homogeneous parameters. At every depth $j-1$, all Fano nodes have common $(M_j, L_j, \Delta_j, \mathsf{C}_j)$ and all edge nodes have common (Δ_j, K_j); let λ_j^{F} and $\lambda_j^{\mathrm{E}} = 1 - \lambda_j^{\mathrm{F}}$ be the total prior masses of the two node types at that depth. Then*

$$\mathfrak{M}_T^{\star,\mathrm{obl}}(Q,\mathcal{G}) \ge \sum_{j=1}^{D} n_j \left\{ \lambda_j^{\mathrm{F}} \Delta_j \left[1 - \frac{n_j\mathsf{C}_j + \log 2}{\log(M_j/L_j)}\right]_+ + \lambda_j^{\mathrm{E}} \frac{\Delta_j}{2} e^{-n_j\mathsf{K}_j} \right\}. \tag{J.37}$$

Proof. Apply Corollary J.9 and use (J.2) at every depth. Exact compatibility removes all ε_u terms. □

Corollary J.11 exhibits the multiscale mechanism explicitly. Each depth consumes a disjoint block and contributes a testing penalty at its own decision and information scales. The sum is valid because one leaf comparator is compatible with every ancestral block.

J.9 Scope and interface with the conditional comparison

Theorem J.6 is a complete lower theorem for every finite witness family and every finite tree satisfying Definition J.1. It uses only oblivious piecewise-constant Bernoulli environments, so the same lower bound applies a fortiori to adaptive environments and to the broader bounded-noise model. The theorem permits approximate leaf compatibility and mixes binary Le Cam nodes with high-branching list-Fano nodes.

The operational complexity in (J.22) retains the learner's actual information allocation and is therefore the sharp form established here. The static relaxation in Corollary J.9 is easier to evaluate but may be loose because it takes a worst-case one-round channel capacity before observing how the learner allocates samples. Appendix I supplies a sharper algorithm-dependent local construction through neutral reference designs.

What remains is not another information inequality. The subsequent cover–packing appendix must show that a chain used by the upper bound can be turned into a compatible family of optimizer-switching nodes without losing its local Riesz radii, and conversely that a large packing forces a large shared-design certificate. That comparison requires a tree-level richness condition, lacunary perturbation scales, and a convexification-stable gluing argument. Until that comparison is proved, Theorem J.6 should be described as a general sequential packing lower bound, not as an exact minimax characterization of Γ_T^{obs}.

K Certified Conditional Cover–Packing Comparison

Appendix G gives an adaptive upper bound by a hierarchy of optimizer-witness increments sharing one physical experiment, while Appendix J gives an oblivious lower bound by compatible optimizer-indexed testing trees. This appendix identifies an explicit geometric regime in which the two constructions match. The comparison has the same broad primal–dual form as classical chaining/packing theorems, where tree-valued upper and lower certificates are extracted from a common geometry (Block et al., 2021; Borst et al., 2021). The statistical ingredient is specific to the present problem: local Riesz costs must be realizable as optimizer-switching Bernoulli experiments, and the resulting local tests must admit one fixed witness comparator along every root-to-leaf path.

A universal comparison cannot follow from observation geometry alone. As Example I.5 shows, a large Riesz cost may occur in a direction that never changes the optimizer. Nor can arbitrary local optimizer switches simply be added across scales: without nested-face compatibility, there need not exist one policy that is simultaneously near-optimal in all ancestral blocks. We therefore state the strong theorem under a concrete *testing-realizable chain* condition. The condition is independent of minimax regret and of the packing value; it is formulated solely through (i) the chained Riesz certificates used by the upper bound, (ii) scalable interior reward families, (iii) optimizer margins, and (iv) one-round Bernoulli channel capacities.

Throughout, fix a finite nonempty witness family $Q \subseteq \Delta(\mathcal{A})$, write $C_Q = \mathrm{co}(Q)$, fix $\rho \in (0, 1/2)$, and let $T \geq 1$.

K.1 The universal weak sandwich

We first record the comparison that holds without any additional structure.

Proposition K.1 (Universal weak cover–packing sandwich)**.** *For every finite nonempty Q and every $\rho \in (0, 1/2)$,*

$$\Pi_{T,\rho}^{\mathrm{seq}}(Q, \mathcal{G}) \leq \mathfrak{M}_T^{\star,\mathrm{obl}}(Q, \mathcal{G}) \leq \mathfrak{M}_T^{\star,\mathrm{ad}}(Q, \mathcal{G}) \leq \Gamma_T^{\mathrm{obs}}(Q, \mathcal{G}). \tag{K.1}$$

For the Dirac family $Q_{\mathcal{A}} = \{\delta_a : a \in \mathcal{A}\}$, the two mediator values in (K.1) *equal the corresponding original-bandit minimax values.*

Proof. The first inequality is Theorem J.6, the middle inequality follows from inclusion of oblivious environments in adaptive environments, and the last inequality is Theorem G.4. The Dirac statement is Proposition D.7. □

The content below is the reverse comparison: under a verifiable local-to-global realizability condition, the chained upper certificate is no larger than the sequential packing value, up to an explicit depth factor and terminal approximation.

K.2 RANGE-CERTIFIED COHERENT CHAINS

Fix a depth-D coherent approximation chain $\mathscr{T}$ for Q in the sense of Definition G.9. We suppress singleton increment levels; hence $N_j := |\mathcal{D}_j| \geq 2$ for $j = 1, \ldots, D$. Let

$$H_j := \log N_j, \qquad B_j := \operatorname{diam}_{\mathcal{G}}(\mathcal{D}_j). \tag{K.2}$$

Definition K.2 (Range-certified upper profile)**.** A collection $(R_j, V_j, A_j)_{j=1}^{D}$ is a *range-certified upper profile* for $\mathscr{T}$ if $R_j \in (0, \infty)$, $V_j \in [0, \infty)$, $A_j \in (0, \infty)$, and

$$\sup_{\boldsymbol{\mu}} \mathfrak{U}_{0,R_j}^{\mathcal{D}_j}(\mu_j, \widetilde{q}_{\boldsymbol{\mu}}) \leq V_j, \tag{K.3}$$

$$A_j \geq 4V_j + B_j^2. \tag{K.4}$$

Here the supremum is over $\boldsymbol{\mu} \in \prod_{k=1}^{D} \Delta(\mathcal{D}_k)$, and $\widetilde{q}_{\boldsymbol{\mu}}$ is defined in (G.32). Define the level weight

$$a_j := A_j H_j, \qquad S := \sum_{j=1}^{D} a_j. \tag{K.5}$$

The profile is in the *quadratic regime at horizon* T if

$$TA_j \geq 2R_j^2 H_j \qquad \text{for every } j = 1, \ldots, D. \tag{K.6}$$

The range R_j controls the pairwise oscillation of the affine scores used at level j; V_j is their worst-policy coefficient energy under the common coherent sampling law. The number A_j therefore upper-bounds the complete quadratic term in Corollary G.10.

Proposition K.3 (Evaluation of a range-certified chain)**.** *If $\mathscr{T}$ has a range-certified upper profile in the quadratic regime, then*

$$\Gamma_T^{\mathrm{obs}}(Q, \mathcal{G}) \leq T \operatorname{rad}_{\mathcal{G}}(\mathscr{T}) + 2\sqrt{2T} \sum_{j=1}^{D} \sqrt{a_j}. \tag{K.7}$$

Consequently,

$$\Gamma_T^{\mathrm{obs}}(Q, \mathcal{G}) \leq T \operatorname{rad}_{\mathcal{G}}(\mathscr{T}) + 2\sqrt{2DTS}. \tag{K.8}$$

Proof. Apply Corollary G.10 with zero estimation tolerance and choose

$$\eta_j = \sqrt{\frac{2H_j}{TA_j}}.$$

Condition (K.6) gives $\eta_j \leq 1/R_j$. The jth entropy–energy term is therefore

$$\frac{2H_j}{\eta_j} + T\eta_j A_j = 2\sqrt{2Ta_j}.$$

Taking the infimum in the definition of Γ_T^{obs} proves (K.7). Cauchy–Schwarz gives $\sum_j \sqrt{a_j} \leq \sqrt{DS}$, proving (K.8). □

K.3 SCALABLE COMPATIBLE TESTING REALIZATIONS

We next define a lower certificate whose levels can be scaled to the information budget supplied by their assigned time blocks.

Definition K.4 (Scalable compatible testing realization)**.** A *ρ-interior scalable testing realization* of depth D consists of:

1. a finite rooted tree $\mathbb{T}$ of depth D, with every node at depth $j-1$ having M_j children, and a leaf map $\theta \mapsto q_\theta \in Q$;
2. a node type $\mathtt{t}_j \in \{\mathrm{F}, \mathrm{E}\}$ for every level;
3. positive scale parameters (r_j, s_j, c_j); and, when $\mathtt{t}_j = \mathrm{F}$, an integer $\ell_j \in \{1, \dots, M_j - 1\}$;
4. for every node u at depth $j-1$, every child $i \in [M_j]$, and every amplitude $\lambda \in [0, r_j]$, a reward vector $g^\lambda_{u,i} \in \mathcal{G}_\rho$.

These objects satisfy the following conditions.

Common neutral reward. For each node u, the vector $g^0_{u,i}$ does not depend on i.

Exact leaf compatibility. For every $\lambda \in [0, r_j]$, every child i, and every descendant leaf $\theta \succeq ui$,

$$q_\theta(g^\lambda_{u,i}) = V_Q(g^\lambda_{u,i}). \tag{K.9}$$

Fano levels. If $\mathtt{t}_j = \mathrm{F}$, then

$$h_j := \log(M_j/\ell_j) \geq 4 \log 2, \tag{K.10}$$

and, for every node u at depth $j-1$, $p \in C_Q$, and $\lambda \in [0, r_j]$,

$$\left|\left\{i \in [M_j] : V_Q(g^\lambda_{u,i}) - p(g^\lambda_{u,i}) < \lambda s_j\right\}\right| \leq \ell_j. \tag{K.11}$$

Moreover, the one-round constrained Bernoulli channel capacity satisfies

$$\sup_{\substack{\pi \in \Delta([M_j]) \\ p \in C_Q}} \sum_{a \in \mathcal{A}} p(a) \sum_{i=1}^{M_j} \pi_i \,\mathrm{kl}\left(g^\lambda_{u,i}(a) \,\middle\|\, \sum_{k=1}^{M_j} \pi_k g^\lambda_{u,k}(a) \right) \leq \lambda^2 c_j. \tag{K.12}$$

Edge levels. If $\mathtt{t}_j = \mathrm{E}$, then $M_j = 2$ and, for every node u, $p \in C_Q$, and $\lambda \in [0, r_j]$,

$$\sum_{i=1}^{2} \left[V_Q(g^\lambda_{u,i}) - p(g^\lambda_{u,i}) \right] \geq 2\lambda s_j, \tag{K.13}$$

and

$$\min_{i \neq k} \sup_{p \in C_Q} \sum_{a \in \mathcal{A}} p(a) \,\mathrm{kl}\big(g^\lambda_{u,i}(a) \,\big\|\, g^\lambda_{u,k}(a)\big) \leq \lambda^2 c_j. \tag{K.14}$$

For such a realization, define its levelwise testing mass by

$$\chi_j := \begin{cases} \dfrac{s_j^2 h_j}{c_j}, & \mathtt{t}_j = \mathrm{F}, \\[2ex] \dfrac{s_j^2}{c_j}, & \mathtt{t}_j = \mathrm{E}. \end{cases} \tag{K.15}$$

Definition K.5 (Matched testing-realizable chain)**.** A range-certified coherent chain and a scalable compatible testing realization of the same depth are *κ-matched*, for $\kappa \in (0, 1]$, if

$$\chi_j \geq \kappa a_j \qquad \text{for every } j = 1, \dots, D. \tag{K.16}$$

Condition (K.16) is a multiscale version of the local Riesz-test richness in Definition I.12. It compares only explicit local quantities: upper score energy and entropy on one side, and optimizer separation per unit Bernoulli information on the other. It does not refer to minimax regret, Γ^{obs}_T, or $\Pi^{\mathrm{seq}}_{T,\rho}$.

K.4 A CONVEXIFICATION-STABLE NESTED-FACE GLUING LEMMA

The exact compatibility in (K.9) can be verified through exposed faces of the witness polytope. The following lemma is useful for hierarchical constructions.

Lemma K.6 (Faces of a finite witness polytope)**.** *If F is a nonempty face of $C_Q = \mathrm{co}(Q)$, then*

$$F = \mathrm{co}(Q \cap F). \tag{K.17}$$

Proof. The inclusion $\mathrm{co}(Q \cap F) \subseteq F$ follows from convexity of the face. Conversely, write $x \in F$ as $x = \sum_i \lambda_i q_i$ with $q_i \in Q$, $\lambda_i > 0$, and $\sum_i \lambda_i = 1$. Since F is a face and $x \in F$, every q_i with positive coefficient belongs to F. Hence $x \in \mathrm{co}(Q \cap F)$. □

Lemma K.7 (Nested exposed-face gluing)**.** *Let $\mathbb{T}$ be a finite rooted tree and let $(F_u)_{u\in\mathbb{T}}$ be nonempty faces of C_Q such that $F_{ui} \subseteq F_u$ for every child ui, and every leaf face contains a designated $q_\theta \in Q$. Fix an internal node u. Suppose there are $b_u \in \mathcal{G}_\rho$, a margin $m_u > 0$, directions $v_{u,i} \in \mathcal{V}$, and a radius $r_u > 0$ such that:*

$$F_Q(b_u) = F_u, \tag{K.18}$$

$$V_Q(b_u) - q(b_u) \geq m_u \qquad \forall q \in Q \setminus F_u, \tag{K.19}$$

$$q(v_{u,i}) = \alpha_{u,i} \qquad \forall q \in Q \cap F_{ui}, \tag{K.20}$$

$$\alpha_{u,i} - q(v_{u,i}) > 0 \qquad \forall q \in (Q \cap F_u) \setminus F_{ui}, \tag{K.21}$$

$$b_u + \lambda v_{u,i} \in \mathcal{G}_\rho \qquad \forall \lambda \in [0, r_u], \tag{K.22}$$

and

$$2r_u \max_i \|v_{u,i}\|_\infty < m_u. \tag{K.23}$$

Then, for every i and every $\lambda \in (0, r_u]$,

$$F_Q(b_u + \lambda v_{u,i}) = F_{ui}. \tag{K.24}$$

Consequently, setting $g^\lambda_{u,i} = b_u + \lambda v_{u,i}$ and $g^0_{u,i} = b_u$ yields exact leaf compatibility at node u. The conclusion is stable under convexification of Q: every mixture in F_{ui} remains optimal.

Proof. Fix i, $\lambda \in (0, r_u]$, and $q_i \in Q \cap F_{ui}$. For $q \in (Q \cap F_u) \setminus F_{ui}$, the baseline values agree because $F_Q(b_u) = F_u$, while (K.20)–(K.21) give

$$(q_i - q)(b_u + \lambda v_{u,i}) = \lambda\big(\alpha_{u,i} - q(v_{u,i})\big) > 0.$$

For $q \in Q \setminus F_u$, one has

$$\begin{aligned}(q_i - q)(b_u + \lambda v_{u,i}) &\geq m_u - \lambda\,|(q_i - q)(v_{u,i})| \\ &\geq m_u - 2r_u\,\|v_{u,i}\|_\infty > 0.\end{aligned}$$

Thus every element of $Q \cap F_{ui}$ is optimal and every element of $Q \setminus F_{ui}$ is strictly suboptimal. Linear optimization over $C_Q = \mathrm{co}(Q)$ then gives $F_Q(b_u + \lambda v_{u,i}) = \mathrm{co}(Q \cap F_{ui}) = F_{ui}$, where the final equality is Lemma K.6. □

K.5 FROM MATCHED CHAINS TO SEQUENTIAL PACKINGS

Let $n_1, \ldots, n_D \in \mathbb{N}$ satisfy $\sum_j n_j = T$. We call this allocation *saturated* for a scalable testing realization if

$$n_j r_j^2 c_j \geq \begin{cases} h_j/4, & \mathtt{t}_j = \mathrm{F}, \\ 1, & \mathtt{t}_j = \mathrm{E}. \end{cases} \tag{K.25}$$

The condition says that the local reward class contains the information-critical amplitude selected for a block of length n_j.

Lemma K.8 (Block allocation lower bound)**.** *Suppose a scalable compatible testing realization is given. For every saturated allocation $(n_j)_{j=1}^D$,*

$$\Pi^{\mathrm{seq}}_{T,\rho}(Q, \mathcal{G}) \geq \frac{1}{2e} \sum_{j=1}^{D} \sqrt{n_j \chi_j}. \tag{K.26}$$

If the realization is κ-matched to an upper profile, then

$$\Pi^{\mathrm{seq}}_{T,\rho}(Q,\mathcal{G}) \geq \frac{\sqrt{\kappa}}{2e}\sum_{j=1}^{D}\sqrt{n_j a_j}. \tag{K.27}$$

Proof. Use the tree and leaf map from Definition K.4, the block lengths n_j, and zero compatibility errors. For every child ui, choose an arbitrary descendant leaf $\theta(ui) \succeq ui$ and use $q_{u,i} = q_{\theta(ui)}$ as the node optimizer; (K.9) guarantees that this policy is optimal for every selected amplitude. At a Fano level choose

$$\lambda_j = \sqrt{\frac{h_j}{4n_j c_j}} \leq r_j.$$

The node separation parameter is $\lambda_j s_j$ and the channel capacity is at most $\lambda_j^2 c_j$. Corollary J.9 gives the depth j contribution

$$\begin{aligned} n_j\lambda_j s_j\left[1-\frac{n_j\lambda_j^2 c_j+\log 2}{h_j}\right]_+ &\geq \frac{n_j\lambda_j s_j}{2} \\ &= \frac{1}{4}\sqrt{n_j\chi_j}, \end{aligned}$$

where $h_j \geq 4\log 2$ is used in the inequality.

At an edge level choose $\lambda_j = (n_j c_j)^{-1/2} \leq r_j$. Its separation parameter is $\lambda_j s_j$, and its directed channel capacity is at most $\lambda_j^2 c_j = 1/n_j$. The edge term of Corollary J.9 is therefore at least

$$\frac{n_j\lambda_j s_j}{2}e^{-1} = \frac{1}{2e}\sqrt{n_j\chi_j}.$$

At every depth the prior node masses sum to one, so the contributions add. Since $1/(2e) < 1/4$, this proves (K.26). The matched inequality (K.16) gives (K.27). □

Lemma K.9 (Proportional integer allocation)**.** *Let $a_1,\dots,a_D > 0$, $S = \sum_j a_j$, and $T \geq D$. There are positive integers $n_1,\dots,n_D$ with $\sum_j n_j = T$ such that*

$$n_j \geq \frac{(T-D)a_j}{S} \qquad \forall j, \tag{K.28}$$

and

$$\sum_{j=1}^{D}\sqrt{n_j a_j} \geq \sqrt{(T-D)S}. \tag{K.29}$$

Proof. Set $\widetilde{n}_j = 1 + \lfloor (T-D)a_j/S \rfloor$. Then $\sum_j \widetilde{n}_j \leq T$ and $\widetilde{n}_j \geq (T-D)a_j/S$. Add the remaining $T - \sum_j \widetilde{n}_j$ rounds to any coordinate. The resulting allocation satisfies (K.28), and hence

$$\sum_j \sqrt{n_j a_j} \geq \sum_j a_j\sqrt{\frac{T-D}{S}} = \sqrt{(T-D)S}.$$

□

Corollary K.10 (A sufficient horizon condition for saturation)**.** *Let $b_j = h_j/4$ at a Fano level and $b_j = 1$ at an edge level. If*

$$T \geq D + S\max_{1\leq j\leq D}\frac{b_j}{a_j r_j^2 c_j}, \tag{K.30}$$

then every proportional allocation satisfying $n_j \geq (T-D)a_j/S$ is saturated: $n_j r_j^2 c_j \geq b_j$ for every j.

Proof. Multiply the proportional lower bound by $r_j^2 c_j$ and use (K.30). □

K.6 CERTIFIED CONDITIONAL COMPARISON THEOREM

Theorem K.11 (Restatement of Theorem 7.2)**.** *Suppose a depth-D coherent chain $\mathscr{T}$ has a range-certified upper profile in the quadratic regime and a κ-matched scalable compatible testing realization. Suppose there are positive integers $n_1, \dots, n_D$ summing to T such that*

$$n_j \geq \frac{(T-D)a_j}{S} \quad \textit{and} \quad n_j r_j^2 c_j \geq \begin{cases} h_j/4, & \mathtt{t}_j = \mathrm{F}, \\ 1, & \mathtt{t}_j = \mathrm{E} \end{cases} \qquad (j = 1, \dots, D).$$

If $T \geq 2D$, then

$$\Pi^{\mathrm{seq}}_{T,\rho}(Q, \mathcal{G}) \geq \frac{\sqrt{\kappa}}{2e}\sqrt{(T-D)S}, \tag{K.31}$$

$$\Gamma^{\mathrm{obs}}_{T}(Q, \mathcal{G}) \leq T \operatorname{rad}_{\mathcal{G}}(\mathscr{T}) + 8e\sqrt{\frac{D}{\kappa}}\Pi^{\mathrm{seq}}_{T,\rho}(Q, \mathcal{G}). \tag{K.32}$$

In particular, if the chain is exact, then

$$\Pi^{\mathrm{seq}}_{T,\rho}(Q, \mathcal{G}) \leq \mathfrak{M}^{\star,\mathrm{obl}}_{T}(Q, \mathcal{G}) \leq \mathfrak{M}^{\star,\mathrm{ad}}_{T}(Q, \mathcal{G}) \leq \Gamma^{\mathrm{obs}}_{T}(Q, \mathcal{G}) \leq 8e\sqrt{\frac{D}{\kappa}}\Pi^{\mathrm{seq}}_{T,\rho}(Q, \mathcal{G}). \tag{K.33}$$

Proof. The first bound follows from Lemma K.8 and Lemma K.9. Proposition K.3 and Cauchy–Schwarz give

$$\Gamma^{\mathrm{obs}}_{T}(Q, \mathcal{G}) \leq T \operatorname{rad}_{\mathcal{G}}(\mathscr{T}) + 2\sqrt{2DTS}.$$

Since $T \geq 2D$, one has $T - D \geq T/2$, and (K.31) implies

$$8e\sqrt{\frac{D}{\kappa}}\Pi^{\mathrm{seq}}_{T,\rho}(Q, \mathcal{G}) \geq 2\sqrt{2DTS}.$$

This proves (K.32). If the chain is exact, its terminal radius is zero; combine with Proposition K.1 to obtain (K.33). □

The depth loss is the price of comparing a sum of levelwise upper costs with a lower construction that allocates disjoint blocks across levels. When the upper profile is geometrically summable, this loss disappears.

Corollary K.12 (Constant comparison for a lacunary profile)**.** *Under the assumptions of Theorem K.11, suppose there is a level $j_\star$ and a constant $\Lambda \geq 1$ such that*

$$\sum_{j=1}^{D} \sqrt{a_j} \leq \Lambda\sqrt{a_{j_\star}}. \tag{K.34}$$

Assume the single-level block of length T is saturated at $j_\star$. Then

$$\Gamma^{\mathrm{obs}}_{T}(Q, \mathcal{G}) \leq T \operatorname{rad}_{\mathcal{G}}(\mathscr{T}) + \frac{4\sqrt{2}e\Lambda}{\sqrt{\kappa}}\Pi^{\mathrm{seq}}_{T,\rho}(Q, \mathcal{G}). \tag{K.35}$$

For an exact chain, this is a constant-factor minimax comparison whenever Λ and κ^{-1} are bounded.

Proof. Choose any node at level $j_\star$ and retain only that node's child experiment and one descendant leaf comparator from each child. Exact leaf compatibility makes this a valid depth-one packing. The proof of Lemma K.8, now with block length T, gives

$$\Pi^{\mathrm{seq}}_{T,\rho}(Q, \mathcal{G}) \geq \frac{1}{2e}\sqrt{T\chi_{j_\star}} \geq \frac{\sqrt{\kappa}}{2e}\sqrt{Ta_{j_\star}}.$$

Combine this with (K.7) and (K.34). □

Corollary K.13 (Depth-efficient chains)**.** *If the exact chain in Theorem K.11 has*

$$D \leq c_0 + \lceil \log_2 \log_2 \max\{4, |Q|\} \rceil \tag{K.36}$$

for a universal constant c_0, then the chained upper and sequential packing lower values agree within a factor

$$O\left(\sqrt{\frac{\log\log|Q|}{\kappa}}\right).$$

K.7 VERIFICATION ON THE WEIGHTED-PRODUCT FAMILY

We verify that the comparison condition is nonvacuous on the weighted-product class from Appendix G.8. Retain its notation $w_j = 2^{-j}/(1-2^{-m})$, Q_m, and $\mathcal{G}_m$. Fix $T \geq 2$ and let J_T be the largest $j \leq m$ with $Tw_j \geq 1$.

Proposition K.14 (Constant cover–packing comparison for the weighted product)**.** *For the truncated coordinate chain through level J_T and $\rho = 1/4$,*

$$\Gamma_T^{\mathrm{obs}}(Q_m, \mathcal{G}_m) \leq \frac{1}{2} + \frac{4\sqrt{2}e(2+\sqrt{2})}{\sqrt{\kappa_0}} \Pi_{T,1/4}^{\mathrm{seq}}(Q_m, \mathcal{G}_m), \qquad \kappa_0 := \frac{3}{1088 \log 2}. \tag{K.37}$$

In particular,

$$\Gamma_T^{\mathrm{obs}}(Q_m, \mathcal{G}_m) \leq 900\, \Pi_{T,1/4}^{\mathrm{seq}}(Q_m, \mathcal{G}_m). \tag{K.38}$$

Proof. Appendix G.8 constructs a coherent coordinate chain with $\mathcal{D}_j = \{r_j, -r_j\}$, score range $R_j = 3$, coefficient energy at most $V_j = 4w_j$, and reward diameter $B_j = w_j/2$. We may therefore take $A_j = 17w_j$ and $H_j = \log 2$. Since $Tw_j \geq 1$ for $j \leq J_T$,

$$TA_j \geq 17 > 18 \log 2 = 2R_j^2 H_j,$$

so the truncated chain lies in the quadratic regime. Its terminal radius satisfies

$$T \operatorname{rad}_{\mathcal{G}_m}(\mathscr{T}) = \frac{T}{4} \sum_{j > J_T} w_j \leq \frac{1}{2}.$$

For the testing realization, take the binary tree $\{-1,+1\}^{\leq J_T}$ and map a leaf to any policy q_x extending its first J_T signs. At a level-j node, define

$$g_{u,\sigma}^{\lambda} = g_{\sigma\lambda e_j}, \qquad \sigma \in \{-1,+1\}, \quad \lambda \in [0,1].$$

Every descendant leaf with jth sign σ is optimal under this reward, so leaf compatibility is exact. For $p \in C_{Q_m}$ write $z_j = [p(j,+1) - p(j,-1)]/w_j \in [-1,1]$. The two local losses are

$$\frac{w_j \lambda}{4}(1 - z_j), \qquad \frac{w_j \lambda}{4}(1 + z_j),$$

whose sum is $w_j\lambda/2$. Thus $s_j = w_j/4$ in (K.13). Moreover, the two mean vectors differ by $\lambda s/2$ on block j and agree elsewhere. Since all means lie in $[1/4, 3/4]$, the Bernoulli quadratic bound gives

$$\min_{\sigma \neq \sigma'} \sup_{p \in C_{Q_m}} \sum_a p(a) \operatorname{kl}\big(g_{u,\sigma}^{\lambda}(a) \,\big\|\, g_{u,\sigma'}^{\lambda}(a)\big) \leq \frac{4}{3} w_j \lambda^2.$$

Hence $c_j = 4w_j/3$ and

$$\chi_j = \frac{s_j^2}{c_j} = \frac{3w_j}{64} \geq \kappa_0 A_j H_j.$$

The profile is lacunary because

$$\sum_{j=1}^{J_T} \sqrt{A_j H_j} \leq (2+\sqrt{2}) \sqrt{A_1 H_1}.$$

The level $j = 1$ is saturated for a block of length T because $Tc_1 \geq (4/3)T(1/2) > 1$. Corollary K.12 proves (K.37).

Finally, the depth-one edge construction also yields

$$\Pi_{T,1/4}^{\mathrm{seq}}(Q_m, \mathcal{G}_m) \geq \frac{1}{2e} \sqrt{\frac{3w_1 T}{64}} \geq \frac{1}{2e} \sqrt{\frac{3T}{128}}.$$

For $T \geq 2$, the additive $1/2$ in (K.37) is at most twenty times this lower bound. A direct numerical bound on the remaining displayed constant is less than 850, proving (K.38). □

Table 3: Domains and quantifiers in the frameworks compared below.

Framework	Decision	Experiment	Feedback	Divergence	Quantifiers	Relation proved here
Fractional cover	distribution on decisions	none	optimizer sets	none	$\sup_\mu \inf_M$	exact reciprocal maximin identity
DEC / ExO	model-dependent decisions	selected decision kernel	general observation	Hellinger / exponential	minimax saddle	general ExO $\leq$ restricted affine ExO
IDS / information ratio	posterior decisions	posterior experiment	model observation	mutual information	Bayesian average	exact χ^2 mediator specialization only
Mediator capacity	policies Q	action from chosen policy	sampled outcome	χ^2	$\sup_\mu$	exact for unrestricted rewards; reward-restricted upper refinement
Sequential chaining	function increments	sequential process	process feedback	process metric	entropy across scales	architectural analogy; not an implementability theorem
Present chain / packing	witnesses and tree branches	one physical law in C_Q	one bounded reward	exponential / KL	shared-design upper; fixed-tree lower	explicit common-experiment and fixed-leaf constraints

K.8 Scope of the comparison

Theorem K.11 is a conditional reverse comparison, but its scope is intentionally explicit.

First, it is not a universal majorizing-measure theorem for arbitrary reward classes. The local Riesz geometry can contain non-optimizer-switching directions, and local optimizer experiments can fail to admit one compatible leaf comparator. Both failures occur before any information inequality is applied.

Second, the factor $\sqrt{D}$ in (K.33) reflects the use of disjoint time blocks in the current sequential packing. A sharper universal theorem would require a simultaneous packing in which several scales contribute decision loss during the same round while their joint transcript information remains controlled. The present result removes the depth factor for lacunary profiles, which include the weighted product class, but does not claim that this is always possible.

Third, the comparison uses a range-certified quadratic envelope of the exact exponential Chained-ExO value. The exact nonlinear upper functional may be smaller. Thus the theorem proves that an explicit, implementable chained certificate is matched by an optimizer packing; it does not assert that every algebraically possible ExO certificate has a tree representation.

Within these limits, the result supplies the missing local-to-global theorem: Riesz-test richness converts upper design radii into statistical tests, nested exposed faces make those tests compatible under convexification, and proportional block allocation converts the entire multiscale profile into one sequential packing. This is the strongest comparison supported by the present proof architecture without silently assuming a universal bandit analogue of Talagrand's majorizing-measure theorem.

L Formal Comparisons with Existing Complexity Frameworks

This appendix locates the objects introduced in this paper relative to the main complexity frameworks for interactive decision making. The purpose is not to rename existing quantities. We prove exact reductions where they are valid, one-sided inequalities where a restriction is present, and explicit non-equivalences where two frameworks measure different geometric axes. The comparison involves five objects: convexified maximin volume and fractional covering (Hanneke & Wang, 2026; Chen et al., 2024); the Decision–Estimation Coefficient (DEC) and Exploration-by-Optimization (ExO) (Foster et al., 2021; Lattimore & Szepesvári, 2020; Foster et al., 2022; 2023); information ratios and mediator capacity (Russo & Van Roy, 2018; Lattimore & Szepesvári, 2019; Eldowa et al., 2024); and sequential majorizing measures (Block et al., 2021; Cesa-Bianchi et al., 2017). All normalizations are stated explicitly.

Throughout, fix a finite nonempty witness family $Q \subseteq \Delta(\mathcal{A})$, write $C_Q = \mathrm{co}(Q)$, and let $V_Q(g) = \max_{q \in Q} q(g)$. The reward class $\mathcal{G}$ is the nonempty compact convex class used in the rest of the paper. We first work with the full bounded-conditional-mean observation model and later restrict to interior Bernoulli models when comparing Hellinger information with the Riesz quadratic form.

L.1 Embedding the mediator game into structured decision making

Let the observation space be $\mathcal{O} = Q \times \mathcal{A} \times [0,1]$. For $g \in \mathcal{G}$ and $\nu = (\nu_a)_{a\in\mathcal{A}} \in \mathfrak{N}(g)$, define a model $M_{g,\nu}$ on decision space Q by

$$M_{g,\nu}(\mathrm{d}q', \mathrm{d}a, \mathrm{d}y \mid q) = \delta_q(\mathrm{d}q')q(a)\nu_a(\mathrm{d}y). \tag{L.1}$$

Thus the selected mediator policy q is recorded in the observation, an action is sampled from q, and only the corresponding reward is observed. Let

$$\mathcal{M}^{\mathrm{mc}}_{Q,\mathcal{G}} := \{M_{g,\nu} : g \in \mathcal{G},\ \nu \in \mathfrak{N}(g)\}. \tag{L.2}$$

The known reward map is $R(q', a, y) = y$. Hence the value and static suboptimality of $q \in Q$ under $M_{g,\nu}$ are

$$f^{M_{g,\nu}}(q) = q(g), \qquad \ell^{M_{g,\nu}}(q) = V_Q(g) - q(g). \tag{L.3}$$

For $\lambda \in \Delta(Q)$, write $p_\lambda = \sum_q \lambda(q)q \in C_Q$.

Proposition L.1 (Exact structured-observation embedding)**.** *The model class $\mathcal{M}^{\mathrm{mc}}_{Q,\mathcal{G}}$ is convex. Under a mixed decision $q \sim \lambda$, its expected reward and expected static loss are*

$$\mathbb{E}_{q\sim\lambda} f^{M_{g,\nu}}(q) = p_\lambda(g), \qquad \mathbb{E}_{q\sim\lambda}\ell^{M_{g,\nu}}(q) = V_Q(g) - p_\lambda(g). \tag{L.4}$$

Moreover, for every nonanticipating model sequence $(g_t, \nu_t)_{t=1}^T$ and every behavioral algorithm whose conditional mediator mixture is λ_t,

$$\mathbb{E}[\mathrm{Reg}^Q_T] = \mathbb{E}\max_{q^\star\in Q}\sum_{t=1}^T (q^\star - p_{\lambda_t})(g_t). \tag{L.5}$$

Consequently, the reward-class-restricted mediator game is a structured-observation problem with a fixed comparator, and its admissible physical experiments are exactly C_Q.

Proof. For $\theta \in [0,1]$, the pointwise mixture of $M_{g,\nu}$ and $M_{h,\xi}$ is $M_{\bar g,\bar\nu}$, where $\bar\nu_a = \theta\nu_a + (1-\theta)\xi_a$ and $\bar g = \theta g + (1-\theta)h \in \mathcal{G}$. Thus the model class is convex. Equation (L.4) follows by linearity. Conditional on $\mathscr{H}_{t-1}$, first drawing $q_t \sim \lambda_t$ and then $A_t \sim q_t$ gives marginal action law p_{λ_t}. Therefore $\mathbb{E}[g_t(A_t) \mid \mathscr{H}_{t-1}] = p_{\lambda_t}(g_t)$. Take conditional expectations in the definition of Reg^Q_T; the maximum over $q^\star$ is measurable because Q is finite. This proves (L.5). □

The selected policy component in (L.1) is available to a general structured-observation algorithm. The algorithms in this paper use only its physical marginal p_λ; this distinction becomes important in the ExO comparison below.

L.2 Fractional covering is witness-level maximin volume

For $\Delta \geq 0$, define the witness-level near-optimal set

$$\mathcal{Q}_\Delta(g) := \{q \in Q : V_Q(g) - q(g) \leq \Delta\}. \tag{L.6}$$

Its maximin mass and fractional covering number are

$$\gamma^Q_\Delta(\mathcal{G}) := \sup_{\mu\in\Delta(Q)}\inf_{g\in\mathcal{G}}\mu(\mathcal{Q}_\Delta(g)), \tag{L.7}$$

$$N^Q_{\mathrm{frac}}(\mathcal{G};\Delta) := \inf_{\mu\in\Delta(Q)}\sup_{g\in\mathcal{G}}\frac{1}{\mu(\mathcal{Q}_\Delta(g))}, \tag{L.8}$$

with $1/0 = +\infty$. Definition (L.8) is the fractional covering number of Chen et al. (2024), specialized to the model class (L.2) and decision set Q.

Let $\mathfrak{H}^Q_\Delta(\mathcal{G}) = \{\mathcal{Q}_\Delta(g) : g \in \mathcal{G}\}$ be the corresponding hypergraph on vertex set Q; duplicate edges are removed.

Proposition L.2 (Exact fractional-cover identities)**.** *For every $\Delta \geq 0$,*

$$N^{Q}_{\mathrm{frac}}(\mathcal{G};\Delta) = \frac{1}{\gamma^{Q}_{\Delta}(\mathcal{G})} = \tau^{*}\Big(\mathfrak{H}^{Q}_{\Delta}(\mathcal{G})\Big) = \nu^{*}\Big(\mathfrak{H}^{Q}_{\Delta}(\mathcal{G})\Big), \tag{L.9}$$

where τ^{*} and ν^{*} are the fractional transversal and packing numbers. For the Dirac family $Q_{\mathcal{A}} = \{\delta_a : a \in \mathcal{A}\}$,

$$\gamma^{Q_{\mathcal{A}}}_{\Delta}(\mathcal{G}) = \gamma_{\Delta}(\mathcal{G}), \qquad N^{Q_{\mathcal{A}}}_{\mathrm{frac}}(\mathcal{G};\Delta) = \frac{1}{\gamma_{\Delta}(\mathcal{G})}. \tag{L.10}$$

Proof. Since all near-optimal sets are nonempty and Q is finite, $\gamma^{Q}_{\Delta} > 0$. For every μ,

$$\sup_{g} \frac{1}{\mu(\mathcal{Q}_{\Delta}(g))} = \frac{1}{\inf_{g} \mu(\mathcal{Q}_{\Delta}(g))}.$$

Taking the infimum over μ proves the first equality in (L.9). The proof of Proposition B.1 applies verbatim after replacing the action vertex set by Q, proving the two linear-programming identities. For a Dirac policy,

$$V_{Q_{\mathcal{A}}}(g) - \delta_a(g) = \max_{b} g(b) - g(a),$$

so $\mathcal{Q}_{\Delta}(g)$ is identified with $\mathsf{A}_{\Delta}(g)$. Equation (L.10) follows. □

Thus maximin volume is not a competing information coefficient: for Dirac choices it is exactly the reciprocal fractional cover of the induced decision problem. Because $\mathcal{G}$ is convex throughout this appendix, this is already the convexified maximin-volume quantity relevant to adversarial learnability. The extension from actions to randomized optimizer witnesses replaces $\mathsf{A}_{\Delta}(g)$ by (L.6). Distribution covers are different: they ask whether one policy places most of its action mass inside an action-level near-optimal set. Corollary B.11 gives the precise bridge

$$N_{\mathrm{wit}}(\mathcal{G};\alpha + (1-\alpha)\beta) \leq N_{\mathrm{dc}}(\mathcal{G};\alpha,\beta) \leq N_{\mathrm{wit}}(\mathcal{G};\alpha\beta). \tag{L.11}$$

In particular, threshold covering and expected-gap approximation agree only after a change of tolerance; neither contains the observation law.

L.3 Affine ExO as a restricted general ExO objective

Let $G : Q \times Q \times \mathcal{A} \times [0,1] \to \mathbb{R}$ be a Borel score, let $\mu \in \Delta(Q)$ be the exponential-weights reference law, let $\lambda \in \Delta(Q)$ be the sampling law over mediator decisions, and put $p_{\lambda} = \sum_{q} \lambda(q) q$. Write

$$\Delta_G(r, q^{\star}; q, a, y) := G(r; q, a, y) - G(q^{\star}; q, a, y).$$

The high-probability ExO objective uses the importance-weighted score difference

$$\Gamma_{\eta}(\mu, \lambda, G; q^{\star}, g, \nu) := (q^{\star} - p_{\lambda})(g) + \frac{1}{\eta} \mathbb{E}_{\substack{q\sim\lambda,\ A\sim q \\ Y\sim\nu_A}} \left[\sum_{r\in Q} \mu(r) \exp\!\left(\frac{\eta\, \Delta_G(r, q^{\star}; q, A, Y)}{\lambda(q)} \right) - 1 \right]. \tag{L.12}$$

The ratio on the null set $\{q : \lambda(q) = 0\}$ may be defined arbitrarily. Let

$$\mathrm{ExO}_{\eta}(\mathcal{M}^{\mathrm{mc}}_{Q,\mathcal{G}}) := \sup_{\mu\in\Delta(Q)} \inf_{\substack{\lambda\in\Delta(Q) \\ G \text{ Borel}}} \sup_{\substack{q^{\star}\in Q,\ g\in\mathcal{G} \\ \nu\in\mathfrak{N}(g)}} \Gamma_{\eta}(\mu, \lambda, G; q^{\star}, g, \nu), \tag{L.13}$$

where scores for which the positive part of the exponential term is not integrable are assigned value $+\infty$. This is the high-probability ExO objective of Foster et al. (2022), specialized to (L.1); the factor $1/\lambda(q)$ is the importance weight for the selected mediator decision.

Proposition L.3 (Exact score-restriction comparison)**.** *For every $\eta > 0$,*

$$\mathrm{ExO}_{\eta}(\mathcal{M}^{\mathrm{mc}}_{Q,\mathcal{G}}) \leq \mathsf{AExO}_{\eta}(Q, \mathcal{G}). \tag{L.14}$$

More precisely, if the infimum in (L.13) *is restricted, jointly with λ, to scores of the form*

$$G_{\lambda,c,\psi}(r; q, a, y) = \lambda(q)[c(r) + \psi(r, a) y], \tag{L.15}$$

then the resulting value is exactly $\mathsf{AExO}_{\eta}(Q, \mathcal{G})$.

Proof. For every q with $\lambda(q) > 0$, substituting (L.15) cancels the importance weight and gives

$$\frac{G_{\lambda,c,\psi}(r;q,a,y) - G_{\lambda,c,\psi}(q^\star;q,a,y)}{\lambda(q)} = c(r) - c(q^\star) + [\psi(r,a) - \psi(q^\star,a)]y.$$

The right side is independent of the latent draw $q \sim \lambda$. Marginalizing that draw replaces $A \sim q$ by $A \sim p_\lambda$. Since $\{p_\lambda : \lambda \in \Delta(Q)\} = C_Q$, the restricted objective is exactly (F.7) after taking the outer supremum over μ. Removing a restriction from an infimum proves (L.14). □

To compare with DEC, use the half-Hellinger convention

$$D^2_{\mathrm{H}}(P,Q) := \frac{1}{2}\int \left(\sqrt{\frac{\mathrm{d}P}{\mathrm{d}\zeta}} - \sqrt{\frac{\mathrm{d}Q}{\mathrm{d}\zeta}}\right)^2 \mathrm{d}\zeta. \tag{L.16}$$

For a model class $\mathcal{M}$ and a reference model $\bar{M}$, define

$$\begin{aligned} \mathrm{rdec}^{\mathrm{o}}_\gamma(\mathcal{M},\bar{M}) &:= \inf_{\lambda\in\Delta(Q)} \sup_{M\in\mathcal{M}} \mathbb{E}_{q\sim\lambda}\left[\ell^M(q) - \gamma D^2_{\mathrm{H}}(M(q),\bar{M}(q))\right], \\ \mathrm{rdec}^{\mathrm{o}}_\gamma(\mathcal{M}) &:= \sup_{\bar{M}\in\mathcal{M}} \mathrm{rdec}^{\mathrm{o}}_\gamma(\mathcal{M},\bar{M}). \end{aligned} \tag{L.17}$$

The squared Hellinger distance in Foster et al. (2022) is twice (L.16). Translating their Corollary 3.1 into the present convention gives

$$\mathrm{rdec}^{\mathrm{o}}_{\gamma/2}(\mathrm{co}(\mathcal{M})) \le \mathrm{ExO}_{1/\gamma}(\mathcal{M}) \le \mathrm{rdec}^{\mathrm{o}}_{\gamma/4}(\mathrm{co}(\mathcal{M})). \tag{L.18}$$

The interactive-Fano framework of Chen et al. (2024) gives a complementary lower-bound interpretation of DEC and augments it with fractional covering to capture estimation complexity. Since $\mathcal{M}^{\mathrm{mc}}_{Q,\mathcal{G}}$ is convex, Proposition L.3 gives the formal chain

$$\mathrm{rdec}^{\mathrm{o}}_{1/(2\eta)}(\mathcal{M}^{\mathrm{mc}}_{Q,\mathcal{G}}) \le \mathrm{ExO}_\eta(\mathcal{M}^{\mathrm{mc}}_{Q,\mathcal{G}}) \le \mathsf{AExO}_\eta(Q,\mathcal{G}). \tag{L.19}$$

Combining this with Theorem F.10, for every $\varepsilon \ge 0$ and finite L satisfying $\eta L \le 1$, yields

$$\begin{aligned} \mathrm{rdec}^{\mathrm{o}}_{1/(2\eta)}(\mathcal{M}^{\mathrm{mc}}_{Q,\mathcal{G}}) &\le \mathsf{AExO}_\eta(Q,\mathcal{G}) \\ &\le \sup_{\mu\in\Delta(Q)} \inf_{p\in C_Q} \left\{ d^+_{\mathcal{G}}(\bar{q}_\mu, p) + 2\varepsilon + \eta\left[4\mathfrak{U}_{\varepsilon,L}(\mu,p) + (1+2\varepsilon)^2\right]\right\}. \end{aligned} \tag{L.20}$$

Equation (L.20) is the precise one-scale novelty boundary. Affine ExO is not a new DEC under a different name: it is a restricted ExO value and is therefore no smaller than the unrestricted objective. No strict separation is claimed here. Its role is that the restriction admits an exact endpoint reduction under arbitrary bounded conditional-mean noise and an explicit Riesz envelope. No converse upper comparison of affine ExO by DEC is claimed.

L.4 Hellinger DEC and the Riesz tangent geometry

Let $\mathcal{G}_\rho = \mathcal{G} \cap [\rho, 1-\rho]^{\mathcal{A}}$ and let M^{Ber}_g denote the Bernoulli model obtained from (L.1) by setting $\nu_a = \mathrm{Ber}(g(a))$.

Lemma L.4 (Bernoulli Hellinger–energy comparison)**.** *For every $x, y \in [\rho, 1-\rho]$,*

$$\frac{1}{2}(x-y)^2 \le D^2_{\mathrm{H}}(\mathrm{Ber}(x),\mathrm{Ber}(y)) \le \frac{(x-y)^2}{4\rho}. \tag{L.21}$$

Consequently, for every $p \in C_Q$ and $g, h \in \mathcal{G}_\rho$,

$$\frac{1}{2}p((g-h)^2) \le \mathbb{E}_{q\sim\lambda} D^2_{\mathrm{H}}(M^{\mathrm{Ber}}_g(q), M^{\mathrm{Ber}}_h(q)) \le \frac{1}{4\rho}p((g-h)^2), \tag{L.22}$$

whenever $p = p_\lambda$.

Proof. For Bernoulli laws,

$$D_{\mathrm{H}}^2 = \frac{1}{2}\left[(\sqrt{x}-\sqrt{y})^2 + (\sqrt{1-x}-\sqrt{1-y})^2\right].$$

The upper bound follows by rationalizing both differences and using $x, y, 1-x, 1-y \geq \rho$. The lower bound follows from $\|P-Q\|_{\mathrm{TV}} \leq \sqrt{2D_{\mathrm{H}}^2(P,Q)}$ and $\|\operatorname{Ber}(x) - \operatorname{Ber}(y)\|_{\mathrm{TV}} = |x-y|$. For (L.22), the selected policy component and the action marginal are identical under the two models, so Hellinger distance separates over (q,a); average (L.21) with weights $\lambda(q)q(a)$. □

Fix a reference $g^0 \in \mathcal{G}_\rho$. Define the quadratic offset coefficient

$$\mathrm{qdec}_\lambda(\mathcal{G}_\rho, g^0; Q) := \inf_{p \in C_Q} \sup_{g \in \mathcal{G}_\rho} \left\{V_Q(g) - p(g) - \lambda p((g-g^0)^2)\right\}. \tag{L.23}$$

Let $\mathrm{rdec}_\gamma^{\mathrm{o}}(\mathcal{M}_{Q,\mathcal{G}_\rho}^{\mathrm{Ber}}, M_{g^0}^{\mathrm{Ber}})$ be the offset DEC with this fixed reference model.

Proposition L.5 (Quadratic DEC sandwich)**.** *For every $\gamma > 0$,*

$$\mathrm{qdec}_{\gamma/(4\rho)}(\mathcal{G}_\rho, g^0; Q) \leq \mathrm{rdec}_\gamma^{\mathrm{o}}(\mathcal{M}_{Q,\mathcal{G}_\rho}^{\mathrm{Ber}}, M_{g^0}^{\mathrm{Ber}}) \leq \mathrm{qdec}_{\gamma/2}(\mathcal{G}_\rho, g^0; Q). \tag{L.24}$$

Proof. Under a mixed decision with physical marginal p, the DEC loss is $V_Q(g) - p(g)$ and the Hellinger information is the middle term in (L.22). Substitute the two bounds from that display into the offset objective and then take the infimum over p and supremum over g. □

Thus the squared $L_2(p)$ perturbation cost is the local tangent information of Hellinger DEC. The estimator-side Riesz quantity is its exact Fenchel dual: by Theorem E.9,

$$\mathcal{E}_\varepsilon(h;p) = \sup_{v \in \mathcal{V}} \left\{h(v) - \varepsilon\|v\|_{\mathcal{K}} - \frac{1}{4}\|v\|_p^2\right\}. \tag{L.25}$$

The next theorem makes this connection exact for the constrained DEC.

For $\kappa, r > 0$, define the constrained modulus

$$\omega_{\kappa,r}(h;p) := \sup\left\{|h(v)| : v \in \mathcal{V}, \|v\|_{\mathcal{K}} \leq \kappa, \|v\|_p \leq r\right\}. \tag{L.26}$$

Lemma L.6 (Exact infimal-convolution identity)**.** *For every $h \in \mathsf{H}_0$, $p \in \Delta(\mathcal{A})$, and $\kappa, r > 0$,*

$$\omega_{\kappa,r}(h;p) = \inf_{\varepsilon \geq 0}\left\{\kappa\varepsilon + r\sqrt{\mathcal{E}_\varepsilon(h;p)}\right\}. \tag{L.27}$$

Proof. The feasible set in (L.26) is centrally symmetric, so the absolute value can be replaced by $h(v)$. Slater's condition holds at $v = 0$. Finite-dimensional Lagrange duality therefore gives

$$\omega_{\kappa,r}(h;p) = \inf_{\varepsilon,\lambda \geq 0}\left\{\kappa\varepsilon + \lambda r^2 + \sup_{v \in \mathcal{V}}\left[h(v) - \varepsilon\|v\|_{\mathcal{K}} - \lambda\|v\|_p^2\right]\right\}.$$

For $\lambda > 0$, substitute $u = 4\lambda v$ and use (L.25); the inner supremum is $\mathcal{E}_\varepsilon(h;p)/(4\lambda)$. Finally,

$$\inf_{\lambda > 0}\left\{\lambda r^2 + \frac{E}{4\lambda}\right\} = r\sqrt{E}$$

for $E \in [0, +\infty]$, with the extended-value convention. This proves (L.27). □

Let $\mathfrak{e} = (g^0, q_+, q_-, r_0)$ be a ρ-regular exposed edge from Definition I.8, put $h = q_+ - q_-$, and let $\mathcal{K} = \mathcal{K}$. Set $R_\infty = \max_{v \in \mathcal{K}}\|v\|_\infty$ and choose $\kappa > 0$ with $\kappa R_\infty \leq r_0$. Define the local Bernoulli model family

$$\mathcal{M}_\kappa^{\mathrm{edge}} := \left\{M_{g^0+\sigma v}^{\mathrm{Ber}} : v \in \kappa\mathcal{K}, h(v) \geq 0, \sigma \in \{-1,+1\}\right\}. \tag{L.28}$$

For a model class $\mathcal{M}$ and reference $\bar{M}$, define the constrained DEC

$$\mathrm{rdec}_\epsilon^{\mathrm{c}}(\mathcal{M}, \bar{M}) := \inf_{\lambda \in \Delta(Q)} \sup_{M \in \mathcal{M}:\ \mathbb{E}_{q\sim\lambda} D_{\mathrm{H}}^2(M(q), \bar{M}(q)) \leq \epsilon^2} \mathbb{E}_{q\sim\lambda}\ell^M(q). \tag{L.29}$$

This is the constrained DEC of Foster et al. (2023) specialized to the present model.

Theorem L.7 (Local constrained-DEC–Riesz comparison)**.** *Let $C_\pm = \mathrm{co}\{q_+, q_-\}$. For every $\epsilon > 0$,*

$$\frac{1}{2} \inf_{p \in C_Q} \omega_{\kappa, 2\sqrt{\rho}\,\epsilon}(h; p) \le \mathrm{rdec}^{\mathrm{c}}_{\epsilon}(\mathcal{M}^{\mathrm{edge}}_{\kappa}, M^{\mathrm{Ber}}_{g^0}) \le \inf_{p \in C_\pm} \omega_{\kappa, \sqrt{2}\,\epsilon}(h; p). \tag{L.30}$$

Equivalently, by Lemma L.6,

$$\frac{1}{2} \inf_{p \in C_Q} \inf_{\delta \ge 0} \left\{ \kappa\delta + 2\sqrt{\rho}\,\epsilon \sqrt{\mathcal{E}_\delta(h; p)} \right\} \le \mathrm{rdec}^{\mathrm{c}}_{\epsilon}(\mathcal{M}^{\mathrm{edge}}_{\kappa}, M^{\mathrm{Ber}}_{g^0}) \le \inf_{p \in C_\pm} \inf_{\delta \ge 0} \left\{ \kappa\delta + \sqrt{2}\,\epsilon \sqrt{\mathcal{E}_\delta(h; p)} \right\}. \tag{L.31}$$

If $Q = \{q_+, q_-\}$, both design domains are $C_\pm$, so the constrained DEC is sandwiched, up to the displayed Bernoulli constants, by the same Riesz infimal-convolution.

Proof. Fix $\lambda \in \Delta(Q)$ and put $p = p_\lambda$. Let $v \in \kappa\mathcal{K}$ satisfy $h(v) \ge 0$ and $\|v\|_p \le 2\sqrt{\rho}\,\epsilon$. The choice of κ and Proposition I.9 ensure that $g^0 \pm v$ are legal and expose $q_\pm$. By the upper bound in (L.22), both models are feasible in (L.29). Lemma I.7 gives

$$\ell^{M_{g^0+v}}(p) + \ell^{M_{g^0-v}}(p) \ge h(v),$$

so at least one feasible model has loss at least $h(v)/2$. Maximize over v and then minimize over $p \in C_Q$ to prove the lower bound.

For the upper bound, restrict the learner to mixtures of q_+ and q_-. Write $p = \theta q_+ + (1-\theta) q_-$. For every model in (L.28), feasibility and the lower bound in (L.22) imply $\|v\|_p \le \sqrt{2}\,\epsilon$. Exact edge optimality and $h(g^0) = 0$ give

$$\ell^{M_{g^0+v}}(p) = (1-\theta) h(v) \le h(v), \qquad \ell^{M_{g^0-v}}(p) = \theta h(v) \le h(v).$$

Taking the supremum over feasible models and then the infimum over $p \in C_\pm$ proves the upper bound. Substitute (L.27) to obtain (L.31). □

Theorem L.7 is the closest formal analogue of the bias–variance duality for Le Cam's method (Polyanskiy & Wu, 2026). It also explains why the Riesz cost alone is not a universal lower complexity: the regular-edge assumption supplies the missing optimizer-switching geometry.

L.5 Information ratios and mediator capacity

Let $Q' \sim \mu \in \Delta(Q)$ and, conditionally on $Q' = q$, draw $A \sim q$. The marginal action law is $\bar{q}_\mu = \sum_q \mu(q) q$. Define the chi-square mutual information

$$I_{\chi^2}(Q'; A) := D_{\chi^2}(P_{Q',A} \| P_{Q'} P_A). \tag{L.32}$$

Proposition L.8 (Policy-set capacity is chi-square information)**.** *For every $\mu \in \Delta(Q)$,*

$$I_{\chi^2}(Q'; A) = \sum_{q \in Q} \mu(q) \chi^2(q \,\|\, \bar{q}_\mu), \tag{L.33}$$

$$I(Q'; A) \le \log\big(1 + I_{\chi^2}(Q'; A)\big) \le I_{\chi^2}(Q'; A). \tag{L.34}$$

Consequently,

$$\mathrm{Cap}_{[0,1]^{\mathcal{A}}}(Q) = \sup_{\mu \in \Delta(Q)} I_{\chi^2}(Q'; A), \qquad \mathrm{Cap}_{\mathcal{G}}(Q) \le \min\left\{ \sup_\mu I_{\chi^2}(Q'; A), \dim(\mathcal{V}) \right\}. \tag{L.35}$$

Proof. The likelihood ratio of $P_{Q',A}$ with respect to $P_{Q'} P_A$ at (q, a) is $q(a)/\bar{q}_\mu(a)$. Expanding the chi-square divergence proves (L.33). For every pair of measures, $D_{\mathrm{KL}}(P\|Q) \le \log(1 + D_{\chi^2}(P\|Q))$; applying this to the joint and product laws gives the first inequality in (L.34), and $\log(1+x) \le x$ gives the second. The unrestricted identity is (E.46). For a restricted reward class, every exact estimator feasible for the tabular cube remains feasible for $\mathcal{G}$, so its design cost can only decrease. The dimension bound is (E.48), after representing $\mathcal{V}$ in an arbitrary basis. □

Information-directed sampling chooses a decision distribution by minimizing squared posterior regret divided by Shannon information about the optimal decision (Russo & Van Roy, 2018); minimax partial-monitoring analyses use related information ratios (Lattimore & Szepesvári, 2019). The local coefficient in this paper has the reverse, lower-bound orientation. Under a regular exposed edge, Theorem I.6 and Proposition I.9 give

$$\rho(1-\rho)\mathcal{E}_\varepsilon(h;p) \le \sup_{w\in\mathcal{W}_\rho(g^0;q_+,q_-)} \frac{(h(w)-\varepsilon\,\|w\|_{\mathcal{K}})_+^2}{I(U;A,Y)} \le \frac{1}{2}\mathcal{E}_\varepsilon(h;p)\,. \tag{L.36}$$

Thus the Riesz cost is, up to Bernoulli constants, the largest realizable squared decision separation per nat of information for a fixed design. The quantifiers are different from Bayesian IDS: (L.36) takes a worst legal optimizer switch after the design is fixed, whereas IDS minimizes a posterior average regret–information ratio.

L.6 Sequential majorizing measures and the genuinely multiscale step

The upper chain in Appendix G resembles generic chaining in that each level pays an entropy term and an increment radius. The resemblance is exact at the level of the entropy sum, but not at the level of the metric. In sequential majorizing-measure theory, the increment geometry is generated by a sequential empirical process (Block et al., 2021); here it is generated by an implementable one-point observation design that must be shared across all levels.

Proposition L.9 (One-scale collapse and multiscale refinement)**.** *The following statements hold.*

1. *For the exact depth-one chain $Q_0 = \{q_0\}$, $Q_1 = Q$, and $\pi_1(q) = q$,*

$$\mathsf{CAExO}_{(\eta)}(\mathscr{T},\mathcal{G}) = \mathsf{AExO}_\eta(Q,\mathcal{G}). \tag{L.37}$$

2. *For every coherent chain and admissible level parameters, the shared-design Riesz envelope has the form*

$$\mathfrak{M}_T^{\star,\mathrm{ad}}(Q,\mathcal{G}) \le T\,\mathrm{rad}_{\mathcal{G}}(\mathscr{T}) + 2T\sum_{j=1}^{J}\varepsilon_j + \sum_{j=1}^{J}\left\{\frac{2\log N_j}{\eta_j} + T\eta_j\left[4V_j + (\mathrm{diam}_{\mathcal{G}}(\mathcal{D}_j) + 2\varepsilon_j)^2\right]\right\}. \tag{L.38}$$

   The same physical sampling law is used at every level.

3. *On the weighted-product family of Proposition G.11, the chained value is $O(\sqrt{T})$ and the minimax value is $\Omega(\sqrt{T})$, while the direct one-level range-controlled Riesz specialization gives only $O(\sqrt{mT})$ for $m \le T$.*

Proof. The first statement is Proposition G.7; the second is Corollary G.10; and the third is Proposition G.11. □

The formal novelty beyond one-scale ExO/DEC is therefore the joint experiment across levels, not a new name for a one-step saddle objective. The same observation (A_t, Y_t) updates every level, so no inverse node-probability penalty appears. On the lower side, Appendix J imposes optimizer-compatible leaf comparators before adding nodewise Fano and Le Cam risks. This is the analogue of a packing tree, but with decision compatibility and observation information in addition to metric separation.

The comparison in item 3 is intentionally narrow. It proves an unbounded advantage over the explicit flat *Riesz envelope*; it does not rule out a different nonlinear one-scale ExO score or prove that the unrestricted DEC is larger by the same factor.

L.7 Examples in which the three quantities vary independently

The preceding reductions show where the frameworks agree. The next result records examples in which the displayed optimizer, observation, and multiscale quantities can vary independently; it is not a universal independence theorem.

Theorem L.10 (No single scalar axis subsumes the framework)**.** *There are finite-action compact convex reward classes exhibiting each of the following separations.*

1. Optimizer geometry without observation geometry. *For every $d \geq 45$, Theorem C.11 constructs classes $\mathcal{G}_{\mathrm{code}}$ and $\mathcal{G}_{\mathrm{tab}}$ with identical maximin volume, distribution-cover number, and all first- and second-order cover/packing scalars at every $\alpha < 1/8$ and $\beta < 1/2$, yet at horizon $T = K$ their oblivious minimax regrets differ by an exponential factor in d up to a polynomial term. Hence no function of those fixed-scale optimizer scalars reconstructs regret within a polylogarithmic factor uniformly over reward classes.*
2. Observation geometry without optimizer switching. *In Example I.5, action 1 is uniquely optimal for every reward vector, while for $h = \delta_1 - \delta_2$ one has $\mathcal{E}_0(h; p) = 1/p(2)$ whenever $p(2) > 0$. Thus an arbitrarily large Riesz cost need not generate any local optimizer test or regret lower bound.*
3. A multiscale improvement of the explicit flat certificate. *On the weighted-product family, Proposition G.11 proves a $34\sqrt{T}$ chained certificate and a matching $\Omega(\sqrt{T})$ minimax lower bound, while the explicit direct one-level range-controlled Riesz specialization proved there evaluates to $10\sqrt{mT}$ for $m \leq T$. Thus the chained analysis removes a factor $\sqrt{m}$ from that particular flat Riesz certificate. This is a separation between two proved analyses, not a lower bound on every possible nonlinear one-scale ExO or DEC certificate.*

Proof. Item 1 is Theorem C.11 and Corollary C.12. Item 2 is the stated example together with Lemma I.4, which shows that optimizer-switching realizability is an additional condition. Item 3 is Proposition G.11. □

The theorem yields the following division of labor.

Framework	Quantity isolated	What it does not encode by itself
Maximin volume / fractional cover	Multiplicity of near-optimal decisions	Cost of distinguishing their reward models
DEC / information ratio	One-step decision loss versus information	A prescribed optimizer approximation hierarchy
Affine ExO / Riesz design	Robust one-sample estimation under a physical design	Arbitrary nonlinear scores; global optimizer compatibility
Sequential majorizing measures	Entropy accumulation across scales	Implementability of one shared bandit experiment
This paper's chain/packing pair	Witness bias, shared observation design, and compatible multiscale tests	Universal equivalence without the stated richness and gluing assumptions

L.8 Claim boundary

The formal comparisons support the following claims and no stronger ones.

1. At one scale, affine ExO is a restricted general ExO objective and is upper-bounded by an explicit Riesz envelope. It is not asserted to improve the unrestricted DEC.
2. On Bernoulli interior models, the Riesz quadratic form is the tangent information geometry of Hellinger DEC. On regular exposed edges, the constrained DEC and an infimal convolution of approximate Riesz costs agree up to explicit constants.
3. For Dirac decisions, convexified maximin volume is exactly reciprocal fractional covering. This is a decision-complexity identity, not a finite-time regret characterization.
4. The genuinely new multiscale object is the requirement that one physical experiment serve a hierarchy of optimizer increments, paired with a lower tree whose local tests share a single

leaf comparator. A universal reverse comparison still requires optimizer-switching richness and gluing; it does not follow from DEC, maximin volume, or generic chaining alone.

These distinctions are essential for interpreting the results: the framework factorizes approximation, observation, and scale interaction, while existing general complexities often combine some of these ingredients into one saddle objective.

M HIERARCHICAL ROUTING: MINIMAX LEARNING AND A GLOBAL-POTENTIAL OBSTRUCTION

This appendix proves the hierarchical routing theorem stated in the main text. It first defines the product routing model independently of the abstract comparison conditions, then proves the adaptive shared-observation upper bound and an oblivious Bernoulli lower bound. It also formalizes the fixed-rate one-global-Shannon architecture and proves the matching-order restricted-affine upper expression. The score-independent obstruction for unrestricted Borel ExO scores is proved in Appendix O.

The architectural lower bound does not apply to factorized potentials, non-Shannon or time-varying regularizers, every one-scale algorithm, or the abstract DEC framework. General ExO remains a broader partial-monitoring framework (Lattimore & Szepesvári, 2020).

M.1 MODEL AND MINIMAX VALUES

Fix an integer depth $D \geq 1$. A physical action is a complete binary route

$$\mathcal{A}_D := \{-1, +1\}^D. \tag{M.1}$$

A template $x = (x_1, \ldots, x_D) \in \{-1, +1\}^D$ specifies one preferred branch at each level. Let $\lambda_j \in (0, 1/2]$ be a known execution reliability and define

$$q_x^\lambda(a) := \prod_{j=1}^{D} \frac{1 + \lambda_j x_j a_j}{2}, \qquad Q_{\lambda,D} := \{q_x^\lambda : x \in \{-1, +1\}^D\}. \tag{M.2}$$

Under q_x^λ, the route coordinates are independent and $\mathbb{E}[A_j] = \lambda_j x_j$.

Let $\alpha_j \geq 0$ with $\sum_{j=1}^{D} \alpha_j \leq 1$. For $\theta \in [-1, 1]^D$, define

$$g_\theta(a) := \frac{1}{2} + \frac{1}{4} \sum_{j=1}^{D} \alpha_j \theta_j a_j, \qquad \mathcal{G}_{\alpha,D} := \{g_\theta : \theta \in [-1, 1]^D\}. \tag{M.3}$$

Then $\mathcal{G}_{\alpha,D}$ is compact, convex, and contained in $[1/4, 3/4]^{\mathcal{A}_D}$. Let $J_+ := \{j \in [D] : \alpha_j > 0\}$. If $J_+ = \varnothing$, all policy values are identical and minimax regret is zero; otherwise define $\alpha_{\min} = \min_{j \in J_+} \alpha_j$.

At round t, from the public history $\mathscr{H}_{t-1}$, the learner chooses a physical law $p_t \in C_{Q_{\lambda,D}}$ and a nonanticipating environment chooses $\theta_t \in [-1, 1]^D$ and an observation kernel. The learner then samples $A_t \sim p_t$ and observes $Y_t \in [0, 1]$ satisfying

$$\mathbb{E}[Y_t \mid \mathscr{H}_{t-1}, A_t] = g_{\theta_t}(A_t). \tag{M.4}$$

The mediator regret is

$$\mathrm{Reg}_T^Q := \max_{x \in \{-1,+1\}^D} \sum_{t=1}^{T} q_x^\lambda(g_{\theta_t}) - \sum_{t=1}^{T} g_{\theta_t}(A_t). \tag{M.5}$$

The values $\mathfrak{M}_T^{\star,\mathrm{ad}}(Q_{\lambda,D}, \mathcal{G}_{\alpha,D})$ and $\mathfrak{M}_T^{\star,\mathrm{obl}}(Q_{\lambda,D}, \mathcal{G}_{\alpha,D})$ are the infimum over behavioral learners of the worst expected regret over, respectively, adaptive nonanticipating and oblivious-mean environments.

The canonical hierarchy is

$$\lambda_j = 2^{-j}, \qquad \alpha_j = \frac{2^{-j}}{Z_D}, \qquad Z_D := 1 - 2^{-D}. \tag{M.6}$$

Write Q_D^{hier} and $\mathcal{G}_D^{\mathrm{hier}}$ for the corresponding classes.

Lemma M.1 (Policy values and common legal experiments)**.** *For every template x and reward parameter θ,*

$$q_x^\lambda(g_\theta) = \frac{1}{2} + \frac{1}{4} \sum_{j=1}^{D} \lambda_j \alpha_j x_j \theta_j. \tag{M.7}$$

Hence an optimizer is selected coordinatewise by $x_j \in \arg\max_{s\in\{-1,+1\}} s\theta_j$.

For every $m = (m_1, \dots, m_D) \in [-1,1]^D$*, the product law*

$$p_m(a) := \prod_{j=1}^{D} \frac{1 + \lambda_j m_j a_j}{2} \tag{M.8}$$

lies in $C_{Q_{\lambda,D}}$*. Conversely, every* $p \in C_{Q_{\lambda,D}}$ *satisfies*

$$\mathbb{E}_p[A_j] = \lambda_j m_j \quad \textit{for some } m_j \in [-1,1]. \tag{M.9}$$

Proof. The value identity follows from $\mathbb{E}_{q_x^\lambda}[A_j] = \lambda_j x_j$. For legality of p_m, put $\omega_m(x) = \prod_j (1 + m_j x_j)/2$. This is a probability law on templates and, for every route a,

$$p_m(a) = \sum_x \omega_m(x) q_x^\lambda(a) = \prod_j \sum_{x_j\in\{-1,+1\}} \frac{1 + x_j m_j}{2} \frac{1 + \lambda_j x_j a_j}{2}.$$

Thus $p_m \in C_{Q_{\lambda,D}}$. This product subclass need not exhaust the entire mediator hull. For the converse, write $p = \sum_x \omega(x) q_x^\lambda$ and set $m_j = \sum_x \omega(x) x_j$. Then $\mathbb{E}_p A_j = \lambda_j m_j$ and $m_j \in [-1,1]$. □

M.2 Fixed-rate one-global-Shannon expressions

For a finite witness family Q, an affine score is

$$S_{c,\psi}(q; a, y) = c(q) + \psi(q, a) y. \tag{M.10}$$

Write $D_{q,r}(a,y) = S_{c,\psi}(q;a,y) - S_{c,\psi}(r;a,y)$. For $\eta > 0$, reference law $\mu \in \Delta(Q)$, physical design $p \in C_Q$, comparator $q^\star \in Q$, and reward $g \in \mathcal{G}$, define

$$\Lambda_\eta(\mu, p, c, \psi; q^\star, g) := \sup_{\substack{\nu_a \in \Delta([0,1]) \\ \int y\, \nu_a(\mathrm{d}y) = g(a)}} \mathbb{E}_{A\sim p, Y\sim\nu_A} \sum_{q\in Q} \mu(q) \left(e^{\eta D_{q,q^\star}(A,Y)} - 1 \right). \tag{M.11}$$

Because the endpoint integrand is finite and convex in y, the supremum is attained by the Borel Bernoulli endpoint kernel with mean $g(a)$. The exact restricted coefficient is

$$\mathsf{AExO}_\eta(Q, \mathcal{G}) := \sup_{\mu\in\Delta(Q)} \inf_{p\in C_Q, c, \psi} \sup_{q^\star\in Q, g\in\mathcal{G}} \left\{ (q^\star - p)(g) + \frac{1}{\eta} \Lambda_\eta(\mu, p, c, \psi; q^\star, g) \right\}. \tag{M.12}$$

The unrestricted coefficient $\mathrm{ExO}_\eta(\mathcal{M}^{\mathrm{mc}}_{Q,\mathcal{G}})$ uses the larger Borel score class of the structured-observation model; hence

$$\mathrm{ExO}_\eta(\mathcal{M}^{\mathrm{mc}}_{Q,\mathcal{G}}) \le \mathsf{AExO}_\eta(Q, \mathcal{G}). \tag{M.13}$$

The architecture considered below uses one fixed learning rate and one global Shannon potential over all policies; it does not factor the potential across coordinates. For full-support initial and aggregation laws $\pi, \varrho \in \Delta(Q)$, define

$$\mathcal{H}(\pi, \varrho) := \log \left[\frac{1}{\varrho_{\min}} \sum_{q\in Q} \frac{\varrho(q)}{\pi(q)} \right], \qquad \varrho_{\min} = \min_q \varrho(q). \tag{M.14}$$

We reserve *optimistic expression* for the one-comparator entropy penalty and *standard comparator-uniform certificate* for the penalty justified by the uniform supermartingale aggregation.

For the canonical hierarchy, define

$$\mathsf{OptFlatExO}_{D,T} := \min \left\{ T, \inf_{\eta>0} \left[T\, \mathrm{ExO}_\eta(\mathcal{M}^{\mathrm{mc}}_D) + \frac{D \log 2}{\eta} \right] \right\}, \tag{M.15}$$

$$\mathsf{StdFlatExO}_{D,T} := \min \left\{ T, \inf_{\eta>0} \left[T\, \mathrm{ExO}_\eta(\mathcal{M}^{\mathrm{mc}}_D) + \frac{2D \log 2}{\eta} \right] \right\}, \tag{M.16}$$

$$\mathsf{StdFlatAExO}_{D,T} := \min \left\{ T, \inf_{\eta>0} \left[T\mathsf{AExO}_\eta(Q^{\mathrm{hier}}_D, \mathcal{G}^{\mathrm{hier}}_D) + \frac{2D \log 2}{\eta} \right] \right\}.$$

Lemma M.2 (Exact prior optimization)**.** *For every finite Q of cardinality N,*

$$\inf_{\pi} \max_{q} \log \frac{1}{\pi(q)} = \log N, \tag{M.17}$$

$$\inf_{\pi,\varrho} \mathcal{H}(\pi, \varrho) = 2 \log N. \tag{M.18}$$

Both infima are attained by uniform laws.

Proof. The first identity follows from $\min_q \pi(q) \le 1/N$. For the second, let $r = \varrho_{\min}$. Cauchy–Schwarz gives

$$\sum_{q} \frac{\varrho(q)}{\pi(q)} \ge \left(\sum_{q} \sqrt{\varrho(q)} \right)^2 \ge N^2 r.$$

Therefore $\mathcal{H}(\pi, \varrho) \ge 2 \log N$, with equality for uniform laws. □

M.2.1 Unrestricted comparator-uniform ExO+ certificate

Let Π be a finite decision set of size N. A model M assigns to every $\pi \in \Pi$ an observation law $M(\cdot \mid \pi)$ and a mean reward $f^M(\pi) \in [0, 1]$. For $\mu, p \in \Delta(\Pi)$ and a finite real-valued Borel score $G(r; \pi, z)$, define

$$\Gamma_{\mu,\eta}(p, G; \pi^\star, M) := f^M(\pi^\star) - \sum_{\pi} p(\pi) f^M(\pi) + \frac{1}{\eta} \mathbb{E}_{\substack{\pi \sim p,\ Z \sim M(\cdot|\pi) \\ R \sim \mu}} \left[\exp\left(\frac{\eta}{p(\pi)} \{ G(R; \pi, Z) - G(\pi^\star; \pi, Z) \} \right) - 1 \right]. \tag{M.19}$$

The quotient is evaluated only when $p(\pi) > 0$; zero-probability decisions do not enter the expectation. If an exponential expectation is infinite, the objective is $+\infty$. Put

$$\mathrm{ExO}_\eta(\mathcal{M}) := \sup_{\mu \in \Delta(\Pi)} \inf_{p, G} \sup_{M \in \mathcal{M},\, \pi^\star \in \Pi} \Gamma_{\mu,\eta}(p, G; \pi^\star, M). \tag{M.20}$$

A predictable ξ_t-approximate oracle, given μ_t, returns (p_t, G_t) whose inner supremum is at most $\mathrm{ExO}_\eta(\mathcal{M}) + \xi_t$.

Theorem M.3 (Comparator-uniform unrestricted ExO+ certificate)**.** *For every adaptive nonanticipating model sequence and every predictable approximate oracle,*

$$\mathbb{E} \max_{r \in \Pi} \sum_{t=1}^{T} \left[f^{M_t}(r) - \sum_{\pi} p_t(\pi) f^{M_t}(\pi) \right] \le T \, \mathrm{ExO}_\eta(\mathcal{M}) + \sum_{t=1}^{T} \xi_t + \frac{2 \log N}{\eta}. \tag{M.21}$$

Moreover, with probability at least $1 - \delta$, the same left side without the outer expectation is at most

$$T \, \mathrm{ExO}_\eta(\mathcal{M}) + \sum_{t=1}^{T} \xi_t + \frac{2 \log N + \log(1/\delta)}{\eta}. \tag{M.22}$$

Consequently, the standard-penalty unrestricted expression with $2 \log N / \eta$ is a valid adaptive expected-regret oracle certificate.

Proof. Initialize $w_1(r) = 1/N$ and let $\mu_t(r) = w_t(r) / \sum_s w_t(s)$. After the oracle returns (p_t, G_t), sample $\pi_t \sim p_t$, observe $Z_t \sim M_t(\cdot \mid \pi_t)$, form

$$\widehat{f}_t(r) = \frac{G_t(r; \pi_t, Z_t)}{p_t(\pi_t)}, \tag{M.23}$$

and update $w_{t+1}(r) = w_t(r) e^{\eta \widehat{f}_t(r)}$. For a fixed comparator r, define

$$L_t(r) = \frac{1}{\eta} \log \mathbb{E}_{R \sim \mu_t} e^{\eta (\widehat{f}_t(R) - \widehat{f}_t(r))}. \tag{M.24}$$

The exact exponential-weights potential identity gives, pathwise,

$$-\sum_{t=1}^{T} L_t(r) \le \frac{\log N}{\eta}. \tag{M.25}$$

Indeed, if $W_t = \sum_s w_t(s)$, then $\log W_{T+1} = \sum_t \log \mathbb{E}_{R\sim\mu_t} e^{\eta \widehat{f}_t(R)}$, whereas $W_{T+1} \ge N^{-1} e^{\eta \sum_t \widehat{f}_t(r)}$.

Let

$$\begin{aligned} d_t(r) &= f^{M_t}(r) - \sum_{\pi} p_t(\pi) f^{M_t}(\pi), \\ K_t(r) &= \mathbb{E}_{R\sim\mu_t} e^{\eta(\widehat{f}_t(R) - \widehat{f}_t(r))}, \\ z_t(r) &= \mathbb{E}_t K_t(r). \end{aligned}$$

The oracle inequality, applied to the predictable model M_t, is

$$d_t(r) + \frac{z_t(r) - 1}{\eta} \le \mathrm{ExO}_\eta(\mathcal{M}) + \xi_t. \tag{M.26}$$

Define

$$\mathcal{E}_t(r) = \exp\left(\eta \sum_{s=1}^{t} [d_s(r) + L_s(r) - \mathrm{ExO}_\eta(\mathcal{M}) - \xi_s]\right). \tag{M.27}$$

Since $e^{\eta L_t(r)} = K_t(r)$, inequality (M.26) and $\log z \le z - 1$ imply

$$\mathbb{E}_t \frac{\mathcal{E}_t(r)}{\mathcal{E}_{t-1}(r)} = e^{\eta(d_t(r) - \mathrm{ExO}_\eta - \xi_t)} z_t(r) \le e^{1 - z_t(r)} z_t(r) \le 1.$$

Thus $\mathcal{E}_t(r)$ is a nonnegative supermartingale and $\mathbb{E}\mathcal{E}_T(r) \le 1$. Rearranging (M.27) and using (M.25) gives, simultaneously for every fixed r,

$$\sum_t d_t(r) \le T\,\mathrm{ExO}_\eta + \sum_t \xi_t + \frac{\log N}{\eta} + \frac{\log \mathcal{E}_T(r)}{\eta}. \tag{M.28}$$

Finally,

$$\mathbb{E} \max_r \log \mathcal{E}_T(r) \le \mathbb{E} \log \sum_r \mathcal{E}_T(r) \le \log \sum_r \mathbb{E}\mathcal{E}_T(r) \le \log N,$$

which proves (M.21). For the high-probability statement, Markov's inequality applied to $N^{-1}\sum_r \mathcal{E}_T(r)$ gives $\mathcal{E}_T(r) \le N/\delta$ for every r on an event of probability at least $1 - \delta$; substitute this in (M.28). □

Corollary M.4 (Validity of $\mathsf{StdFlatExO}_{D,T}$)**.** *For the canonical routing model, $N = 2^D$. By Proposition A.8, the mixed mediator regret in Theorem M.3 has the same expectation as the realized routing mediator regret. Hence*

$$\sup_{\mathsf{Adv}} \mathbb{E}\,\mathrm{Reg}_T^Q \le T\,\mathrm{ExO}_\eta(\mathcal{M}_D^{\mathrm{mc}}) + \sum_{t=1}^{T} \xi_t + \frac{2D\log 2}{\eta}. \tag{M.29}$$

Therefore $\mathsf{StdFlatExO}_{D,T}$ *is a validated unrestricted ExO+ oracle certificate. The high-probability statement in Theorem M.3 controls mixed regret; a realized-regret version additionally includes the bounded action-sampling martingale deviation from Remark A.9. The adjective* oracle *records that no polynomial-time optimization over arbitrary Borel scores is asserted.*

Proof. Sampling a mediator q and then a route $A \sim q$ induces the legal physical law $\sum_q p_t(q) q$. Its mean reward is the corresponding mixture of $q(g_t)$, so (M.21), together with Proposition A.8, gives the expected routing regret bound. Optimize over η, let summable oracle errors tend to zero, and apply the trivial cap T. □

Proposition M.5 (Validity of the standard restricted-affine certificate)**.** *For every finite Q, compact reward class $\mathcal{G}$, and $\eta > 0$, the restricted affine-ExO algorithm with full-support initial law π and aggregation law ϱ satisfies*

$$\mathbb{E}\,\mathrm{Reg}_T^Q \le T\mathsf{AExO}_\eta(Q,\mathcal{G}) + \frac{\mathcal{H}(\pi,\varrho)}{\eta}. \tag{M.30}$$

Consequently, $\mathsf{StdFlatAExO}_{D,T}$ *is a valid adaptive expected-regret certificate.*

Proof. Initialize $W_1(q) = \pi(q)$ and $\mu_t(q) = W_t(q)/\sum_r W_t(r)$. Fix $\delta > 0$ and choose predictable ξ_t-minimizers of the finite endpoint objective, with $\xi_t = \delta/T$. Given $\mathscr{H}_{t-1}$, the design and score are measurable before (A_t, Y_t) is drawn. Update $W_{t+1}(q) = W_t(q)e^{\eta S_t(q;A_t,Y_t)}$.

For fixed $q^\star$, put $\Delta_t(q^\star) = (q^\star - p_t)(g_t)$. The local coefficient yields

$$\mathbb{E}_t \sum_q \mu_t(q) e^{\eta(S_t(q) - S_t(q^\star))} \le 1 + \eta(\mathsf{AExO}_\eta + \xi_t - \Delta_t(q^\star)).$$

The right side is positive because the left side is. Thus $1 + u \le e^u$ gives the one-step inequality

$$\mathbb{E}_t\left[e^{\eta(\Delta_t(q^\star) - \mathsf{AExO}_\eta - \xi_t)} \frac{\sum_q W_{t+1}(q)}{\sum_q W_t(q)} e^{-\eta S_t(q^\star;A_t,Y_t)}\right] \le 1. \tag{M.31}$$

The product of these adapted factors is a nonnegative supermartingale with initial value one. Telescoping and $\sum_q W_1(q) = 1$ imply

$$\mathbb{E}\exp\left(\eta\left[\sum_t \Delta_t(q^\star) - T\mathsf{AExO}_\eta - \sum_t \xi_t\right]\right) \le \frac{1}{\pi(q^\star)}.$$

For $z_q = \sum_t \Delta_t(q) - T\mathsf{AExO}_\eta - \sum_t \xi_t$,

$$e^{\eta \max_q z_q} \le \frac{1}{\varrho_{\min}} \sum_q \varrho(q) e^{\eta z_q}.$$

Taking expectations and applying Jensen yields

$$\mathbb{E}\max_q \sum_t \Delta_t(q) \le T\mathsf{AExO}_\eta + \delta + \frac{\mathcal{H}(\pi,\varrho)}{\eta}.$$

Let $\delta \downarrow 0$. The realized-minus-mixed reward is a bounded martingale-difference sum of mean zero, so the same expected bound holds for (M.5). □

M.3 Main theorem

Theorem M.6 (Hierarchical routing minimax rate and global-potential expressions)**.** *For the canonical hierarchy* (M.6) *and all $D, T \ge 1$:*

1. *There is a known-horizon $O(D)$-time and $O(D)$-memory algorithm satisfying*

$$\frac{1}{32}\sqrt{T} \le \mathfrak{M}_T^{\star,\mathrm{obl}}(Q_D^{\mathrm{hier}}, \mathcal{G}_D^{\mathrm{hier}}) \le \mathfrak{M}_T^{\star,\mathrm{ad}}(Q_D^{\mathrm{hier}}, \mathcal{G}_D^{\mathrm{hier}}) \le C_{\mathrm{sh}}\sqrt{T}, \tag{M.32}$$

   where $C_{\mathrm{sh}} = 2\log 2 + 1 + 2/\sqrt{3} < 3.55$. The upper bound permits arbitrary bounded conditional-mean observations; the lower bound uses two constant oblivious Bernoulli environments.

2. *The score-independent result of Theorem O.5, the score-class inclusion* (M.13)*, and the affine construction below imply*

$$\frac{1}{32}\min\{T, \sqrt{DT}\} \le \mathsf{OptFlatExO}_{D,T} \le \mathsf{StdFlatExO}_{D,T},$$
$$\mathsf{StdFlatExO}_{D,T} \le \mathsf{StdFlatAExO}_{D,T} \le 4\min\{T, \sqrt{DT}\}. \tag{M.33}$$

   Here $\mathsf{OptFlatExO}_{D,T}$ is only an optimistic expression, whereas Corollary M.4 and Proposition M.5 validate the standard unrestricted and restricted-affine oracle certificates.

3. *Consequently,*

$$\frac{\mathsf{OptFlatExO}_{D,T}}{C_{\mathrm{sh}}\sqrt{T}} \ge \frac{1}{114}\min\{\sqrt{D}, \sqrt{T}\}. \tag{M.34}$$

M.4 SHARED-OBSERVATION ALGORITHM AND ADAPTIVE UPPER BOUND

For $j \in J_+$ initialize $W_{j,1}(+1) = W_{j,1}(-1) = 1$; inactive coordinates are fixed arbitrarily and omitted. At round t, let

$$\mu_{j,t}(s) = \frac{W_{j,t}(s)}{W_{j,t}(+1) + W_{j,t}(-1)}, \qquad m_{j,t} = \sum_{s\in\{-1,+1\}} s\mu_{j,t}(s), \tag{M.35}$$

sample one route

$$A_t \sim p_{m_t}, \tag{M.36}$$

and after observing Y_t set

$$\widehat{r}_{j,t}(s) := s\lambda_j \frac{A_{t,j} - \lambda_j m_{j,t}}{1 - \lambda_j^2 m_{j,t}^2} Y_t. \tag{M.37}$$

Update

$$W_{j,t+1}(s) = W_{j,t}(s) e^{\eta_j \widehat{r}_{j,t}(s)}, \qquad \eta_j = \frac{1-\lambda_j}{\lambda_j\sqrt{T}}. \tag{M.38}$$

One physical pair (A_t, Y_t) updates every level.

Lemma M.7 (Simultaneous unbiasedness, energy, and range)**.** *Conditionally on the public past and predictable θ_t, define*

$$r_{j,t}(s) := \frac{1}{4}\lambda_j \alpha_j \theta_{t,j} s. \tag{M.39}$$

Then, for every j, t, s,

$$\mathbb{E}_t \widehat{r}_{j,t}(s) = r_{j,t}(s), \qquad \mathbb{E}_t \widehat{r}_{j,t}(s)^2 \le \frac{\lambda_j^2}{1-\lambda_j^2}, \qquad |\widehat{r}_{j,t}(s)| \le \frac{\lambda_j}{1-\lambda_j}. \tag{M.40}$$

Moreover, $\widehat{r}_{j,t}(-1) = -\widehat{r}_{j,t}(+1)$ and $r_{j,t}(-1) = -r_{j,t}(+1)$.

Proof. Under p_{m_t} the coordinates are independent and $b_{j,t} := \mathbb{E}_t A_{t,j} = \lambda_j m_{j,t}$. By (M.4) and (M.3),

$$\begin{aligned}\mathbb{E}_t[(A_{t,j} - b_{j,t})Y_t] &= \mathbb{E}_t[(A_{t,j} - b_{j,t}) g_{\theta_t}(A_t)] \\ &= \frac{1}{4}\alpha_j \theta_{t,j}(1 - b_{j,t}^2),\end{aligned}$$

because all cross-coordinate terms vanish by independence and centering. This gives unbiasedness. Since $Y_t^2 \le 1$,

$$\mathbb{E}_t \widehat{r}_{j,t}(s)^2 \le \lambda_j^2 \frac{\mathbb{E}_t(A_{t,j} - b_{j,t})^2}{(1-b_{j,t}^2)^2} = \frac{\lambda_j^2}{1-b_{j,t}^2} \le \frac{\lambda_j^2}{1-\lambda_j^2}.$$

For $A \in \{-1, +1\}$, $|A-b|/(1-b^2) \le (1-|b|)^{-1}$, and $|b_{j,t}| \le \lambda_j$, giving the range bound. Oddness in s is immediate. □

Proposition M.8 (Adaptive shared-observation upper bound)**.** *For all admissible λ, α, the algorithm* (M.35)–(M.38) *satisfies*

$$\mathbb{E}\,\mathrm{Reg}_T^Q \le C_{\mathrm{sh}}\sqrt{T} \sum_{j\in J_+} \lambda_j. \tag{M.41}$$

Proof. Fix a level j. Lemma M.7 implies $\eta_j|\widehat{r}_{j,t}(s)| \le T^{-1/2} \le 1$. The exponential-weights potential and $e^z \le 1 + z + z^2$ for $|z| \le 1$ give, simultaneously for $s = \pm 1$,

$$\sum_{t=1}^T [\widehat{r}_{j,t}(s) - \mu_{j,t}(\widehat{r}_{j,t})] \le \frac{\log 2}{\eta_j} + \eta_j \sum_{t=1}^T \mu_{j,t}(\widehat{r}_{j,t}^2). \tag{M.42}$$

Let $\xi_{j,t}(s) = \widehat{r}_{j,t}(s) - r_{j,t}(s)$. These are martingale differences in the public filtration, $\xi_{j,t}(-1) = -\xi_{j,t}(+1)$, and predictability of $\mu_{j,t}$ gives $\mathbb{E}\sum_t \mu_{j,t}(\xi_{j,t}) = 0$. Let

$$s_j^\star \in \operatorname*{arg\,max}_{s\in\{-1,+1\}} \sum_{t=1}^T r_{j,t}(s)$$

with deterministic tie-breaking. Although $s_j^\star$ is a random hindsight choice under an adaptive environment, oddness gives the pathwise inequality

$$-\sum_t \xi_{j,t}(s_j^\star) \le \left|\sum_t \xi_{j,t}(+1)\right|.$$

Martingale orthogonality and Lemma M.7 yield

$$\mathbb{E}\left|\sum_t \xi_{j,t}(+1)\right| \le \left(\sum_t \mathbb{E}\xi_{j,t}(+1)^2\right)^{1/2} \le \lambda_j\sqrt{\frac{T}{1-\lambda_j^2}}.$$

Taking expectations in (M.42) therefore gives

$$\begin{aligned}\mathbb{E}\sum_t [r_{j,t}(s_j^\star) - \mu_{j,t}(r_{j,t})] &\le \frac{\log 2}{\eta_j} + \eta_j T\frac{\lambda_j^2}{1-\lambda_j^2} + \lambda_j\sqrt{\frac{T}{1-\lambda_j^2}} \\ &\le \left(2\log 2 + 1 + \frac{2}{\sqrt{3}}\right)\lambda_j\sqrt{T},\end{aligned}$$

where we used $\lambda_j \le 1/2$.

By Lemma M.1,

$$\max_x \sum_t q_x^\lambda(g_{\theta_t}) = \frac{T}{2} + \sum_{j=1}^D \max_s \sum_t r_{j,t}(s),$$

while

$$\sum_t p_{m_t}(g_{\theta_t}) = \frac{T}{2} + \sum_{j=1}^D \sum_t \mu_{j,t}(r_{j,t}).$$

Summing the level bounds controls mixed mediator regret. Finally, $\sum_t [p_{m_t}(g_{\theta_t}) - g_{\theta_t}(A_t)]$ is a bounded martingale with mean zero, proving (M.41). □

For the canonical reliabilities, $\sum_{j\in J_+}\lambda_j < 1$, which proves the upper side of (M.32). The algorithm samples D independent bits and updates two weights per level, hence uses $O(D)$ time and memory per round.

M.5 Oblivious Bernoulli lower bound

Proposition M.9 (Two-environment lower bound)**.** *For every level k with $\alpha_k > 0$,*

$$\mathfrak{M}_T^{\star,\mathrm{obl}}(Q_{\lambda,D}, \mathcal{G}_{\alpha,D}) \ge \frac{\lambda_k}{8}\min\left\{\alpha_k T, \frac{\sqrt{3}}{2}\sqrt{T}\right\}. \tag{M.43}$$

Proof. Let Σ be uniform on $\{-1,+1\}$. Conditional on $\Sigma = \sigma$, use the constant reward $g_{\sigma\varepsilon e_k}$ and Bernoulli observations with that mean. Each realized environment is a deterministic oblivious-mean environment in $\mathcal{G}_{\alpha,D}$.

At a public history, any legal physical law satisfies $\mathbb{E}[A_k] = \lambda_k m_t$ for some predictable $m_t \in [-1,1]$ by Lemma M.1. Under sign σ, the conditional expected gap to an optimal template is

$$\frac{\lambda_k\alpha_k\varepsilon}{4}(1-\sigma m_t). \tag{M.44}$$

At the same counterfactual public history, the learner uses the same action kernel in the two environments. Conditional on A_k, the Bernoulli means are $1/2 \pm \alpha_k\varepsilon A_k/4$. For $x \in [0,1]$,

$$\mathrm{kl}\left(\frac{1}{2}+\frac{x}{4}\,\middle\|\,\frac{1}{2}-\frac{x}{4}\right) = \frac{x}{2}\log\frac{1+x/2}{1-x/2} \le \frac{2}{3}x^2. \tag{M.45}$$

The transcript KL chain rule and Pinsker's inequality imply

$$\mathrm{TV}(P_+^{t-1}, P_-^{t-1}) \le \alpha_k \varepsilon \sqrt{\frac{t-1}{3}}. \tag{M.46}$$

Averaging (M.44) over signs and using $|\mathbb{E}_+ m_t - \mathbb{E}_- m_t| \le 2\,\mathrm{TV}(P_+^{t-1}, P_-^{t-1})$ gives Bayes regret at least

$$\frac{\lambda_k \alpha_k \varepsilon}{4} \sum_{t=1}^{T} \left[1 - \mathrm{TV}(P_+^{t-1}, P_-^{t-1})\right].$$

Choose $\varepsilon = \min\{1, \sqrt{3}/(2\alpha_k\sqrt{T})\}$. Every total-variation term is at most $1/2$, so the Bayes regret is at least the right side of (M.43). Since both prior atoms are individually admissible, Bayes regret lower-bounds the worst-environment regret. □

For the canonical hierarchy, $k = 1$ has $\lambda_1 = 1/2$ and $\alpha_1 = (2Z_D)^{-1} \ge 1/2$. Since $T \ge \sqrt{T}$, the proposition gives $\sqrt{T}/32$, proving the lower side of (M.32).

M.6 A SCORE-INDEPENDENT AFFINE-EXO OBSTRUCTION

Lemma M.10 (Exact binary dual witness)**.** *Let $a_D := \alpha_1 = (2Z_D)^{-1}$. For every $\eta > 0$,*

$$\mathsf{AExO}_\eta(Q_D^{\mathrm{hier}}, \mathcal{G}_D^{\mathrm{hier}}) \ge \beta_D(\eta) := \max_{0 \le t \le a_D} \left\{ \frac{t}{8} - \frac{1 - \sqrt{1 - t^2/4}}{2\eta} \right\}. \tag{M.47}$$

The scalar maximum is

$$\beta_D(\eta) = \begin{cases} \dfrac{\sqrt{4+\eta^2} - 2}{4\eta}, & 0 < \eta \le \dfrac{2a_D}{\sqrt{4 - a_D^2}}, \\ \dfrac{a_D}{8} - \dfrac{1 - \sqrt{1 - a_D^2/4}}{2\eta}, & \eta \ge \dfrac{2a_D}{\sqrt{4 - a_D^2}}. \end{cases} \tag{M.48}$$

In particular,

$$\mathsf{AExO}_\eta(Q_D^{\mathrm{hier}}, \mathcal{G}_D^{\mathrm{hier}}) \ge \frac{1}{32} \min\{\eta, 1\}. \tag{M.49}$$

The bounds are uniform over all legal physical designs, all affine scores, all offsets, and all finite score ranges.

Proof. Choose templates $x^+ = (+1, +1, \ldots, +1)$ and $x^- = (-1, +1, \ldots, +1)$ and reference law $\mu = (\delta_{q_+} + \delta_{q_-})/2$. For $\varepsilon \in [0, 1]$, choose opposite rewards $g_\pm = g_{\pm\varepsilon e_1}$. Fix arbitrary p, c, ψ, put $D(a, y) = S_{c,\psi}(q_+; a, y) - S_{c,\psi}(q_-; a, y)$, and let $P_\pm$ be the endpoint laws of (A, Y) under p and $g_\pm$.

The two decision gaps have sum $\lambda_1\alpha_1\varepsilon/2$, so averaging the two local objectives gives

$$\frac{\lambda_1\alpha_1\varepsilon}{4} + \frac{1}{4\eta}\left[\mathbb{E}_{P_+} e^{-\eta D} + \mathbb{E}_{P_-} e^{\eta D} - 2\right]. \tag{M.50}$$

For densities $p_\pm$ relative to $P_+ + P_-$, pointwise AM–GM gives $p_+ e^{-\eta D} + p_- e^{\eta D} \ge 2\sqrt{p_+ p_-}$. Conditional on any route, the Bernoulli means are $1/2 \pm \alpha_1\varepsilon a_1/4$, whose Hellinger affinity is $\sqrt{1 - \alpha_1^2\varepsilon^2/4}$. Since $\lambda_1 = 1/2$, setting $t = \alpha_1\varepsilon$ proves (M.47) after taking the infimum over (p, c, ψ) and the supremum over μ.

For the closed form, the unconstrained maximizer of the scalar objective is $t_0(\eta) = 2\eta/\sqrt{4+\eta^2}$. If $t_0(\eta) \le a_D$, substitution gives $(\sqrt{4+\eta^2} - 2)/(4\eta)$; otherwise the maximum is attained at $t = a_D$. This proves (M.48).

Finally $a_D \ge 1/2$. Choose $t = \min\{\eta/2, 1/2\}$ and use $1 - \sqrt{1-u} \le u$. If $\eta \le 1$, the objective is at least $\eta/16 - \eta/32 = \eta/32$; if $\eta \ge 1$, it is at least $1/16 - 1/(32\eta) \ge 1/32$. This proves (M.49). □

M.7 A MATCHING-ORDER AFFINE-EXO UPPER CERTIFICATE

Lemma M.11 (Uniform affine-ExO upper certificate)**.** *For the canonical hierarchy and every* $0 < \eta \le 1/4$*,*

$$\mathsf{AExO}_\eta(Q_D^{\mathrm{hier}}, \mathcal{G}_D^{\mathrm{hier}}) \le \frac{16}{9}\eta. \tag{M.51}$$

Proof. Fix an arbitrary reference law μ over templates and put $m_j = \mathbb{E}_\mu X_j$, $b_j = \lambda_j m_j$. Use the legal product design p_m and zero offsets. For every template x, define

$$\psi_x(a) := \sum_{j=1}^{D} \lambda_j (x_j - m_j) \frac{a_j - b_j}{1 - b_j^2}. \tag{M.52}$$

Independence and centering give, for every $g \in \mathcal{G}_D^{\mathrm{hier}}$,

$$\mathbb{E}_{A \sim p_m, Y|A}[\psi_x(A) Y] = (q_x^\lambda - p_m)(g), \qquad \sum_x \mu(x)\psi_x(a) = 0 \quad \forall a. \tag{M.53}$$

Equivalently, p_m and the correlated mixture $\bar{q}_\mu = \sum_x \mu(x) q_x^\lambda$ have the same value on every reward in $\mathcal{G}_D^{\mathrm{hier}}$, although they need not be the same distribution on complete routes. For $x, x^\star$, let $\Delta\psi = \psi_x - \psi_{x^\star}$. Independence gives

$$\mathbb{E}_{p_m}[\Delta\psi(A)^2] = \sum_{j=1}^{D} \frac{\lambda_j^2 (x_j - x_j^\star)^2}{1 - b_j^2} \le 4 \sum_{j=1}^{D} \frac{\lambda_j^2}{1 - \lambda_j^2} \le \frac{16}{9}. \tag{M.54}$$

Also

$$|\Delta\psi(a)| \le 2 \sum_{j=1}^{D} \frac{\lambda_j}{1 - \lambda_j} \le \frac{10}{3} < 4. \tag{M.55}$$

Hence $|\eta \Delta\psi(a) y| \le 1$ for $\eta \le 1/4$. Applying $e^z \le 1 + z + z^2$ in the exact endpoint objective, the linear term equals $-(q^\star - p_m)(g)$ by (M.53) and cancels the decision gap. The quadratic term is at most

$$\eta \mathbb{E} \sum_x \mu(x) (\psi_x - \psi_{x^\star})^2 Y^2 \le \frac{16}{9}\eta.$$

Taking the supremum over comparator and reward, the infimum over legal certificates, and then the supremum over μ proves the claim. □

M.8 OPTIMIZATION AND COMPLETION OF THE MAIN THEOREM

Proof of Theorem M.6. The minimax sandwich follows from Propositions M.8 and M.9, using $\sum_j 2^{-j} < 1$, $\lambda_1 = 1/2$, and $\alpha_1 \ge 1/2$.

The score-independent obstruction in Theorem O.5 gives

$$\mathrm{ExO}_\eta(\mathcal{M}_D^{\mathrm{mc}}) \ge \frac{1}{32} \min\{\eta, 1\}.$$

For $0 < \eta \le 1$,

$$T \,\mathrm{ExO}_\eta(\mathcal{M}_D^{\mathrm{mc}}) + \frac{D \log 2}{\eta} \ge \frac{T\eta}{32} + \frac{D \log 2}{\eta} \ge 2\sqrt{\frac{DT \log 2}{32}} \ge \frac{1}{32}\sqrt{DT}.$$

For $\eta \ge 1$, the same expression is at least $T/32$. Taking the infimum and then the outer cap by T proves

$$\mathsf{OptFlatExO}_{D,T} \ge \frac{1}{32} \min\{T, \sqrt{DT}\}.$$

The inequality $\mathsf{OptFlatExO}_{D,T} \le \mathsf{StdFlatExO}_{D,T}$ follows from the larger entropy penalty in the latter expression. Since unrestricted ExO optimizes over a score class containing the restricted affine scores, (M.13) gives $\mathsf{StdFlatExO}_{D,T} \le \mathsf{StdFlatAExO}_{D,T}$.

For the final upper bound, combine Lemma M.11 with the definition of $\mathsf{StdFlatAExO}_{D,T}$. If $T \geq 18D\log 2$, choose

$$\eta^\star = \sqrt{\frac{9D\log 2}{8T}} \leq \frac{1}{4}.$$

Then

$$\frac{2D\log 2}{\eta^\star} + \frac{16}{9}T\eta^\star = 2\sqrt{\frac{32\log 2}{9}DT} < 4\sqrt{DT}.$$

If $T < D$, the outer cap gives $T = \min\{T, \sqrt{DT}\}$. If $D \leq T < 18D\log 2$, the cap gives $T < \sqrt{18\log 2}\sqrt{DT} < 4\sqrt{DT}$. This proves (M.33).

Finally, $32C_{\mathrm{sh}} < 114$ and $\min\{T, \sqrt{DT}\}/\sqrt{T} = \min\{\sqrt{T}, \sqrt{D}\}$, proving (M.34). □

M.9 Primitive geometry and exact scope

Proposition M.12 (Observable, switchable, and gluable)**.** *The routing hierarchy satisfies the following properties without invoking the abstract certificate assumptions of Appendix K.*

1. Observability. *Every level reward has the exact bounded affine estimator* (M.37)*, with the energy and range bounds in* (M.40)*.*
2. Switchability. *The rewards $g_{\pm\varepsilon e_j}$ expose the two optimizer faces $\{q_x : x_j = \pm 1\}$. For every legal physical design, the two one-round decision losses sum to $\lambda_j\alpha_j\varepsilon/2$, while the directed Bernoulli KL between the two observation laws is at most $(2/3)\alpha_j^2\varepsilon^2$.*
3. Gluability. *Every sequence of local signs is one legal global template, and that template is optimal for every reward whose nonzero coordinate signs agree with the sequence.*

Proof. Observability is Lemma M.7. For switchability, every legal design has $\mathbb{E}A_j = \lambda_j m_j$ for some $m_j \in [-1, 1]$. The losses under $g_{+\varepsilon e_j}$ and $g_{-\varepsilon e_j}$ are $\lambda_j\alpha_j\varepsilon(1 - m_j)/4$ and $\lambda_j\alpha_j\varepsilon(1 + m_j)/4$, respectively. Their sum is the claimed value, and the KL bound is (M.45). Gluability follows from the coordinatewise optimizer identity (M.7). □

The class is a noisy multistage routing/configuration model: an action is a complete route, the route coordinates are observed because the learner selected the route, and the scalar reward is additive across the realized branch decisions. Adjacent work studies hierarchical configuration bandits under metric Lipschitz structure and tree-MDP policy libraries with observation sharing, under different feedback and adversary models (Shabat et al., 2026; Shah et al., 2026). The present theorem uses neither cluster smoothness nor a stationary MDP transition model.

The precise architectural conclusion is:

> On the canonical routing class, a factorized shared-observation algorithm is minimax-sharp at order $\sqrt{T}$, while the specified fixed-rate one-global-Shannon expressions and validated oracle certificates have order $\min\{T, \sqrt{DT}\}$.

No claim is made against unrestricted ExO, DEC, nonlinear score families, non-Shannon regularizers, product or factorized potentials, time-varying learning-rate analyses, or all one-scale algorithms.

N HIERARCHICAL OPTIMIZER REALIZABILITY FROM PRIMITIVE GEOMETRY

The certified comparison theorem in Appendix K is stated through two operational certificates: a range-controlled upper profile and a scalable compatible testing realization. This appendix derives both certificates from three primitive properties of an optimizer hierarchy: *observability*, *switchability*, and *gluing*. The result is a sufficient-condition theorem, not an equivalence: the primitive conditions are strong enough to imply Definitions K.2– K.5, but the abstract certificates may also hold in models that do not admit the exposed-face representation used below.

Throughout, let $\mathcal{A}$ be finite, let $Q \subseteq \Delta(\mathcal{A})$ be finite and nonempty, let $C_Q = \mathrm{co}(Q)$, and let $\mathcal{G} \subseteq [0,1]^{\mathcal{A}}$ be nonempty, compact, and convex. Fix $\rho \in (0, 1/2)$ and write $\mathcal{G}_\rho = \mathcal{G} \cap [\rho, 1-\rho]^{\mathcal{A}}$. Let $\mathscr{T}$ be a depth-D coherent approximation chain with non-singleton increment classes $\mathcal{D}_1, \dots, \mathcal{D}_D$. Put

$$N_j := |\mathcal{D}_j|, \qquad H_j := \log N_j, \qquad B_j := \mathrm{diam}_{\mathcal{G}}(\mathcal{D}_j). \tag{N.1}$$

Null levels for which every increment is indistinguishable on $\mathcal{G}$ are removed; hence the quantities A_j defined below are positive.

N.1 PRIMITIVE UPPER OBSERVABILITY

For $\boldsymbol{\mu} \in \prod_{k=1}^{D} \Delta(\mathcal{D}_k)$, recall

$$\bar{d}_{\mu_j} = \sum_{d \in \mathcal{D}_j} \mu_j(d) d, \qquad \widetilde{q}_{\boldsymbol{\mu}} = q_0 + \sum_{k=1}^{D} \bar{d}_{\mu_k} \in C_Q,$$

where membership in C_Q follows from coherence.

Definition N.1 (Observable coherent hierarchy)**.** The chain $\mathscr{T}$ is $(R_j, \mathsf{V}_j)_{j=1}^{D}$-*observable* if, for every $\boldsymbol{\mu}$ and every level j, there is an affine score system

$$S_{j,\boldsymbol{\mu}}(d; a, y) = c_{j,\boldsymbol{\mu}}(d) + \psi_{j,\boldsymbol{\mu}}(d, a) y, \qquad d \in \mathcal{D}_j,$$

such that, with $p = \widetilde{q}_{\boldsymbol{\mu}}$,

$$c_{j,\boldsymbol{\mu}}(d) + \sum_{a \in \mathcal{A}} p(a) \psi_{j,\boldsymbol{\mu}}(d, a) g(a) = (d - \bar{d}_{\mu_j})(g) \qquad \forall d \in \mathcal{D}_j,\ g \in \mathcal{G}, \tag{N.2}$$

$$\max_{d \in \mathcal{D}_j} \sum_{a \in \mathcal{A}} p(a) \psi_{j,\boldsymbol{\mu}}(d, a)^2 \le \mathsf{V}_j, \tag{N.3}$$

$$\max_{d, r \in \mathcal{D}_j} \max_{a \in \mathcal{A},\ y \in [0,1]} |S_{j,\boldsymbol{\mu}}(d; a, y) - S_{j,\boldsymbol{\mu}}(r; a, y)| \le R_j. \tag{N.4}$$

Here $R_j \in (0, \infty)$ and $\mathsf{V}_j \in [0, \infty)$ are independent of $\boldsymbol{\mu}$.

This is a direct statistical condition: one legal physical experiment simultaneously supports exact affine estimators for every increment at level j, with controlled coefficient energy and score oscillation. It contains no optimizer-switching or lower-bound requirement.

Define

$$A_j := 4\mathsf{V}_j + B_j^2, \qquad a_j := A_j H_j, \qquad S := \sum_{j=1}^{D} a_j. \tag{N.5}$$

N.2 PRIMITIVE SWITCHABILITY AND GLUING

Let $\mathbb{T}$ be a finite rooted tree of depth D. Every node at depth $j-1$ has M_j children. To each node u associate a nonempty face F_u of C_Q. For a child ui, require $F_{ui} \subseteq F_u$. Each leaf θ has a designated witness

$$q_\theta \in Q \cap F_\theta. \tag{N.6}$$

At a node u of depth $j-1$, let $b_u \in \mathcal{G}_\rho$ be a neutral reward, let $m_u > 0$ be a parent-face margin, and let $v_{u,i} \in \mathcal{V}$ be a direction for child i. Let $\alpha_{u,i} \in \mathbb{R}$ satisfy

$$F_Q(b_u) = F_u, \tag{N.7}$$
$$V_Q(b_u) - q(b_u) \geq m_u \qquad \forall q \in Q \setminus F_u, \tag{N.8}$$
$$q(v_{u,i}) = \alpha_{u,i} \qquad \forall q \in Q \cap F_{ui}, \tag{N.9}$$
$$q(v_{u,i}) < \alpha_{u,i} \qquad \forall q \in (Q \cap F_u) \setminus F_{ui}. \tag{N.10}$$

For a level radius $r_j > 0$, define the normalized switching margin against a physical design $p \in C_Q$ by

$$\Delta_{u,i}(p) := \frac{V_Q(b_u) - p(b_u)}{r_j} + \alpha_{u,i} - p(v_{u,i}). \tag{N.11}$$

The first term charges policies outside the parent face; the second distinguishes children inside the parent face.

Definition N.2 (Switchable hierarchy)**.** The face hierarchy is *switchable* if every level j has a type $\mathtt{t}_j \in \{\mathrm{F}, \mathrm{E}\}$, constants $r_j, s_j, \zeta_j \in (0, \infty)$, and, at a Fano level, an integer $\ell_j \in \{1, \ldots, M_j - 1\}$, such that the following hold uniformly over all nodes u at depth $j-1$.

Reward feasibility. For every child i and every $\lambda \in [0, r_j]$,

$$b_u + \lambda v_{u,i} \in \mathcal{G}_\rho. \tag{N.12}$$

Fano switchability. If $\mathtt{t}_j = \mathrm{F}$, then

$$h_j := \log(M_j/\ell_j) \geq 4 \log 2, \tag{N.13}$$

and, for every $p \in C_Q$,

$$|\{i \in [M_j] : \Delta_{u,i}(p) < s_j\}| \leq \ell_j. \tag{N.14}$$

Moreover, with $\bar{v}_{u,\pi} = \sum_i \pi_i v_{u,i}$,

$$\sup_{\substack{p \in C_Q \\ \pi \in \Delta([M_j])}} \sum_{a \in \mathcal{A}} p(a) \sum_{i=1}^{M_j} \pi_i \big(v_{u,i}(a) - \bar{v}_{u,\pi}(a)\big)^2 \leq \zeta_j. \tag{N.15}$$

Edge switchability. If $\mathtt{t}_j = \mathrm{E}$, then $M_j = 2$ and, for every $p \in C_Q$,

$$\sum_{i=1}^{2} \Delta_{u,i}(p) \geq 2 s_j, \tag{N.16}$$

while

$$\sup_{p \in C_Q} \max_{i \neq k} \sum_{a \in \mathcal{A}} p(a) \big(v_{u,i}(a) - v_{u,k}(a)\big)^2 \leq \zeta_j. \tag{N.17}$$

Definition N.3 (Gluable hierarchy)**.** The switchable face hierarchy is *gluable* if its faces are nested as above, every leaf contains its designated witness, and, for every internal node u at level j,

$$2 r_j \max_{i \in [M_j]} \|v_{u,i}\|_\infty < m_u. \tag{N.18}$$

Switchability is a quantitative decision/information condition. Gluing is a global comparator condition: local child choices are nested inside one face tree, and the margin–radius inequality prevents a child perturbation from activating a policy outside the parent face.

N.3 Auxiliary lemmas

Lemma N.4 (Faces are generated by witness vertices)**.** *If F is a nonempty face of C_Q, then*

$$F = \mathrm{co}(Q \cap F). \tag{N.19}$$

Proof. The inclusion $\mathrm{co}(Q \cap F) \subseteq F$ follows from convexity. For the reverse inclusion, write $x \in F$ as $x = \sum_k \lambda_k q_k$ with $q_k \in Q$, $\lambda_k > 0$, and $\sum_k \lambda_k = 1$. Since F is a face and $x \in F$, every q_k with positive coefficient belongs to F. Thus $x \in \mathrm{co}(Q \cap F)$. □

Lemma N.5 (Primitive gluing gives exact child faces)**.** *Assume* (N.7)–(N.12) *and* (N.18)*. For every child i and every $\lambda \in (0, r_j]$,*

$$F_Q(b_u + \lambda v_{u,i}) = F_{ui}. \tag{N.20}$$

For $\lambda = 0$, the optimizer face is F_u.

Proof. The case $\lambda = 0$ is (N.7). Fix $\lambda \in (0, r_j]$, $q_i \in Q \cap F_{ui}$, and $q \in Q$. If $q \in (Q \cap F_u) \setminus F_{ui}$, then $q_i(b_u) = q(b_u) = V_Q(b_u)$ and (N.9)–(N.10) give

$$(q_i - q)(b_u + \lambda v_{u,i}) = \lambda\big(\alpha_{u,i} - q(v_{u,i})\big) > 0.$$

If $q \in Q \setminus F_u$, then

$$\begin{aligned}(q_i - q)(b_u + \lambda v_{u,i}) &\geq m_u - \lambda|(q_i - q)(v_{u,i})| \\ &\geq m_u - 2r_j\|v_{u,i}\|_\infty > 0.\end{aligned}$$

Thus every witness in $Q \cap F_{ui}$ is optimal and every witness outside this set is strictly suboptimal. Linear optimization over C_Q and Lemma N.4 yield $F_Q(b_u + \lambda v_{u,i}) = \mathrm{co}(Q \cap F_{ui}) = F_{ui}$. □

Lemma N.6 (Interior Bernoulli quadratic bound)**.** *For every $x, y \in [\rho, 1 - \rho]$,*

$$\mathrm{kl}(x\|y) \leq \frac{(x-y)^2}{y(1-y)} \leq \frac{(x-y)^2}{\rho(1-\rho)}. \tag{N.21}$$

Proof. Using $\log z \leq z - 1$ twice,

$$\begin{aligned}\mathrm{kl}(x\|y) &= x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y} \\ &\leq x\left(\frac{x}{y} - 1\right) + (1-x)\left(\frac{1-x}{1-y} - 1\right) \\ &= \frac{(x-y)^2}{y(1-y)}.\end{aligned}$$

The second inequality follows from $y(1-y) \geq \rho(1-\rho)$. □

Lemma N.7 (Proportional saturated allocation)**.** *Let $a_1, \ldots, a_D > 0$, $S = \sum_j a_j$, and $T \geq D$. There are positive integers $n_1, \ldots, n_D$ with $\sum_j n_j = T$ and*

$$n_j \geq \frac{(T-D)a_j}{S} \qquad (j = 1, \ldots, D). \tag{N.22}$$

If $r_j, c_j, \upsilon_j > 0$ and

$$T \geq D + S \max_j \frac{\upsilon_j}{a_j r_j^2 c_j}, \tag{N.23}$$

then every allocation satisfying (N.22) *obeys*

$$n_j r_j^2 c_j \geq \upsilon_j \qquad (j = 1, \ldots, D). \tag{N.24}$$

Proof. Set $x_j = (T-D)a_j/S$, so $\sum_j x_j = T - D$. Start from $n_j = 1 + \lfloor x_j \rfloor$. Their sum is at most T and each $n_j \geq x_j$. Distribute the remaining integer units arbitrarily. This proves the first claim. For the second, multiply (N.22) by $r_j^2 c_j$ and use (N.23). □

N.4 PRIMITIVE-TO-CERTIFICATE THEOREM

For every level define

$$\tau_j := \begin{cases} h_j, & \mathtt{t}_j = \mathrm{F}, \\ 1, & \mathtt{t}_j = \mathrm{E}, \end{cases} \qquad c_j := \frac{\zeta_j}{\rho(1-\rho)}, \tag{N.25}$$

and

$$\kappa_j := \frac{\rho(1-\rho)s_j^2\tau_j}{\zeta_j A_j H_j}, \qquad \kappa := \min\left\{1, \min_{1\le j\le D} \kappa_j\right\}. \tag{N.26}$$

All factors in (N.26) are positive, so $\kappa \in (0,1]$.

Let

$$\upsilon_j := \begin{cases} h_j/4, & \mathtt{t}_j = \mathrm{F}, \\ 1, & \mathtt{t}_j = \mathrm{E}. \end{cases} \tag{N.27}$$

Define the sufficient horizon

$$T_{\mathrm{HOR}} := \max\left\{2D, \max_j \frac{2R_j^2 H_j}{A_j}, D + S\max_j \frac{\upsilon_j}{a_j r_j^2 c_j}\right\}. \tag{N.28}$$

Since the horizon is integral, the condition below means $T \ge \lceil T_{\mathrm{HOR}} \rceil$.

Theorem N.8 (Hierarchical optimizer realizability)**.** *Suppose:*

1. *$\mathscr{T}$ is coherent and observable in the sense of Definition N.1;*
2. *the rooted face hierarchy is switchable in the sense of Definition N.2;*
3. *the hierarchy is gluable in the sense of Definition N.3.*

Then:

1. *$(R_j, \mathsf{V}_j, A_j)_{j=1}^D$ is a range-certified upper profile;*
2. *the rewards*
$$g_{u,i}^{\lambda} := b_u + \lambda v_{u,i}, \qquad 0 \le \lambda \le r_j, \tag{N.29}$$
form a ρ-interior scalable compatible testing realization with separation parameters s_j and channel parameters c_j from (N.25)*;*
3. *the upper and lower certificates are κ-matched, with the explicit κ in* (N.26)*.*

If $T \ge \lceil T_{\mathrm{HOR}} \rceil$, there is a saturated proportional block allocation, the upper profile is in the quadratic regime, and

$$\Pi_{T,\rho}^{\mathrm{seq}}(Q,\mathcal{G}) \ge \frac{\sqrt{\kappa}}{2e}\sqrt{(T-D)S}, \tag{N.30}$$

$$\Gamma_T^{\mathrm{obs}}(Q,\mathcal{G}) \le T\,\mathrm{rad}_{\mathcal{G}}(\mathscr{T}) + 8e\sqrt{\frac{D}{\kappa}}\Pi_{T,\rho}^{\mathrm{seq}}(Q,\mathcal{G}). \tag{N.31}$$

If the chain is exact, then

$$\Pi_{T,\rho}^{\mathrm{seq}} \le \mathfrak{M}_T^{\star,\mathrm{obl}} \le \mathfrak{M}_T^{\star,\mathrm{ad}} \le \Gamma_T^{\mathrm{obs}} \le 8e\sqrt{\frac{D}{\kappa}}\Pi_{T,\rho}^{\mathrm{seq}}, \tag{N.32}$$

where all quantities in (N.32) *are evaluated at $(Q,\mathcal{G})$.*

Proof. We prove the three certificate claims and then the finite-horizon conclusion.

Upper-profile implication. Fix j and $\boldsymbol{\mu}$. Coherence gives $p = \widetilde{q}_{\boldsymbol{\mu}} \in C_Q$. The score system in Definition N.1 belongs to $\mathscr{S}_{0,R_j}^{\mathcal{D}_j}(\mu_j, p)$ by (N.2) and (N.4). Equation (N.3) therefore gives

$$\mathfrak{U}_{0,R_j}^{\mathcal{D}_j}(\mu_j, \widetilde{q}_{\boldsymbol{\mu}}) \le \mathsf{V}_j.$$

Taking the supremum over $\boldsymbol{\mu}$ proves (K.3). By construction, $A_j = 4\mathsf{V}_j + B_j^2$, so (K.4) also holds.

Leaf-compatibility implication. Set $g^{\lambda}_{u,i}$ as in (N.29). Reward feasibility puts these vectors in $\mathcal{G}_\rho$. At $\lambda = 0$ the reward equals b_u, independently of i. Lemma N.5 gives $F_Q(g^{\lambda}_{u,i}) = F_{ui}$ for $\lambda > 0$ and $F_Q(g^{0}_{u,i}) = F_u$. If $\theta \succeq ui$, nesting gives $q_\theta \in F_\theta \subseteq F_{ui} \subseteq F_u$. Hence

$$q_\theta(g^{\lambda}_{u,i}) = V_Q(g^{\lambda}_{u,i}) \qquad \forall \lambda \in [0, r_j],$$

which is exact leaf compatibility.

Decision-separation implication. Fix a node u, a child i, a design $p \in C_Q$, and $\lambda \in [0, r_j]$. By Lemma N.5,

$$V_Q(g^{\lambda}_{u,i}) = V_Q(b_u) + \lambda \alpha_{u,i}.$$

Therefore

$$\begin{aligned} V_Q(g^{\lambda}_{u,i}) - p(g^{\lambda}_{u,i}) &= V_Q(b_u) - p(b_u) + \lambda(\alpha_{u,i} - p(v_{u,i})) && \text{(N.33)} \\ &= \lambda \Delta_{u,i}(p) + \left(1 - \frac{\lambda}{r_j}\right)[V_Q(b_u) - p(b_u)] && \text{(N.34)} \\ &\geq \lambda \Delta_{u,i}(p). && \text{(N.35)} \end{aligned}$$

At a Fano level, (N.14) and (N.33) imply that fewer than or equal to ℓ_j children have loss below λs_j. This is (K.11). At an edge level, summing (N.33) over the two children and using (N.16) gives (K.13).

Channel-control implication. At a Fano level, fix $p \in C_Q$ and $\pi \in \Delta([M_j])$. The mixture reward is

$$\bar{g}^{\lambda}_{u,\pi} = \sum_i \pi_i g^{\lambda}_{u,i} = b_u + \lambda \bar{v}_{u,\pi}.$$

It belongs to $\mathcal{G}_\rho$ by convexity. Applying Lemma N.6 pointwise and then (N.15) gives

$$\begin{aligned} &\sum_a p(a) \sum_i \pi_i \, \mathrm{kl}\big(g^{\lambda}_{u,i}(a) \,\big\|\, \bar{g}^{\lambda}_{u,\pi}(a)\big) \\ &\quad \leq \frac{\lambda^2}{\rho(1-\rho)} \sum_a p(a) \sum_i \pi_i \big(v_{u,i}(a) - \bar{v}_{u,\pi}(a)\big)^2 \\ &\quad \leq \lambda^2 c_j. \end{aligned}$$

This is (K.12). At an edge level, the same argument and (N.17) give, for both directions $i \neq k$,

$$\sup_{p \in C_Q} \sum_a p(a) \, \mathrm{kl}\big(g^{\lambda}_{u,i}(a) \,\big\|\, g^{\lambda}_{u,k}(a)\big) \leq \lambda^2 c_j,$$

which implies (K.14). Thus all clauses of Definition K.4 hold.

Explicit matching. By (K.15) and (N.25),

$$\chi_j = \frac{s_j^2 \tau_j}{c_j} = \frac{\rho(1-\rho) s_j^2 \tau_j}{\zeta_j} = \kappa_j A_j H_j = \kappa_j a_j \geq \kappa a_j.$$

Hence the chain and testing realization are κ-matched.

Quadratic regime and saturation. The second term in (N.28) gives $TA_j \geq 2R_j^2 H_j$ for every j, so the upper profile lies in the quadratic regime. Lemma N.7 supplies positive integers n_j summing to T and satisfying the proportional lower bound. The third term in (N.28) then gives $n_j r_j^2 c_j \geq \upsilon_j$, which is exactly the saturation requirement at Fano and edge levels. Finally $T \geq 2D$ by the first term.

Theorem K.11 now applies and yields (N.30)–(N.32). □

N.5 A CONVENIENT BINARY COROLLARY

The switchability inequalities are especially transparent for antipodal binary splits.

Corollary N.9 (Antipodal edge hierarchy)**.** *Suppose every internal node at level j has two children and there is one direction v_u such that*

$$v_{u,+} = v_u, \qquad v_{u,-} = -v_u,$$

while $F_{u,+}$ and $F_{u,-}$ are respectively the maximizer and minimizer faces of $q \mapsto q(v_u)$ over F_u. Let

$$w_u := \max_{q \in F_u} q(v_u) - \min_{q \in F_u} q(v_u), \qquad w_u \geq 2s_j > 0, \tag{N.36}$$

and assume

$$4 \sup_{p \in C_Q} p(v_u^2) \leq \zeta_j. \tag{N.37}$$

Then the edge-switchability conditions (N.16)–(N.17) *hold.*

Proof. Let $\alpha_{u,+} = \max_{q \in F_u} q(v_u)$ and $\alpha_{u,-} = \max_{q \in F_u} q(-v_u) = -\min_{q \in F_u} q(v_u)$. For every $p \in C_Q$,

$$\begin{aligned} \Delta_{u,+}(p) + \Delta_{u,-}(p) &= \frac{2[V_Q(b_u) - p(b_u)]}{r_j} + \alpha_{u,+} + \alpha_{u,-} \\ &\geq w_u \geq 2s_j. \end{aligned}$$

Moreover, $(v_{u,+} - v_{u,-})^2 = 4v_u^2$, so (N.37) implies (N.17). □

This corollary reduces binary switchability to an optimizer width and one quadratic observation-energy bound. The remaining gluing assumptions are the nesting, reward-feasibility, and margin–radius conditions.

N.6 WHY ALL THREE PRIMITIVE PROPERTIES ARE NEEDED

The next examples are not converses to Theorem N.8; they isolate the distinct roles of the three primitive properties and of saturation.

Example N.10 (Observable but not switchable)**.** Let $\mathcal{A} = \{1, 2\}$, $Q = \{\delta_1, \delta_2\}$, and

$$\mathcal{G} = \mathrm{co}\{(1/2, 1/2), (3/4, 1/2)\}.$$

Under the full-support design $(1/2, 1/2)$, every policy contrast has a bounded affine estimator. Nevertheless action 2 is never strictly optimal: the reward class can move only in a direction favoring action 1. Hence observation is possible but no two-child optimizer switch exists.

Example N.11 (Local switches without gluing)**.** Let C_Q be a triangle with vertices labelled q_{++}, q_{+-}, q_{-+}. The first local split distinguishes the edge $\mathrm{co}\{q_{++}, q_{+-}\}$ from the vertex q_{-+}; the second local split distinguishes the edge $\mathrm{co}\{q_{++}, q_{-+}\}$ from the vertex q_{+-}. Each split is exposed by a linear functional on the triangle. However, the formal path $(-, -)$ has no leaf witness: the corresponding child sets have empty intersection. Thus local optimizer switches do not by themselves provide one comparator across levels.

Example N.12 (Gluing without observability)**.** Let $Q = \{\delta_1, \delta_2\}$ and let $\mathcal{G} = [1/4, 3/4]^2$. The two vertex faces form a perfectly gluable depth-one optimizer tree. If the proposed common physical experiment is $p = \delta_1$, however, the contrast $\delta_2 - \delta_1$ does not annihilate the observation nullspace and has infinite exact Riesz cost. Geometric gluing cannot replace statistical observability.

Example N.13 (Matched geometry without saturation)**.** Consider one edge level with positive s, c but a reward radius $r > 0$ so small that $Tr^2c < 1$. The matching ratio $s^2/(cA \log N)$ can be positive, yet the information-critical edge amplitude $1/\sqrt{Tc}$ lies outside $[0, r]$. The local geometry is matched but no length-T block is saturated. This is why the horizon condition in (N.28) is logically separate from matching.

N.7 SCOPE

Theorem N.8 replaces the operational assumption stack in Appendix K by three primitive checks:

1. a common-design affine observability calculation;
2. optimizer-switching margins and quadratic reward-direction energies;
3. a nested exposed-face tree with a quantitative margin–radius condition.

It does not claim these conditions are necessary, nor does it remove the $\sqrt{D}$ factor caused by disjoint lower-bound blocks. Its contribution is a reusable implication: once observability, switchability, and gluing are verified, range certification, testing realizability, explicit κ-matching, and finite- horizon saturation follow with no additional minimax assumption.

O SIMULTANEOUS PRODUCT PACKINGS AND GLOBAL-POTENTIAL OBSTRUCTIONS

This appendix proves three complementary results: an obstruction to unconditional reverse chain–packing comparison, a simultaneous multiscale lower bound, and a score-independent obstruction for the fixed-rate one-global-Shannon architecture. All logarithms are natural, and every lower-bound prior is supported on individually admissible deterministic oblivious-mean Bernoulli environments.

O.1 NO UNCONDITIONAL REVERSE THEOREM

The interior testing requirement in the sequential packing is substantive.

Theorem O.1 (An unconditional reverse comparison is false)**.** *Fix $\rho \in (0, 1/2)$. There are a finite action set, a two-policy witness family Q, and a compact convex reward class $\mathcal{G} \subseteq [0,1]^{\mathcal{A}}$ such that, for every $T \geq 1$,*

$$\Pi^{\mathrm{seq}}_{T,\rho}(Q, \mathcal{G}) = 0, \qquad \mathfrak{M}^{\star,\mathrm{obl}}_{T}(Q, \mathcal{G}) \geq \min\left\{\frac{\rho T}{8}, \frac{\sqrt{T}}{12}\right\}. \tag{O.1}$$

In particular, the regret is at least $\sqrt{T}/12$ when $T \geq 4/(9\rho^2)$. Consequently, no finite universal constant C can satisfy $\Gamma^{\mathrm{obs}}_T \leq C\Pi^{\mathrm{seq}}_{T,\rho}$ over all compact convex reward classes.

Proof. Let $\mathcal{A} = \{0,1,2,3\}$ and $Q = \{\delta_1, \delta_2\}$. Put

$$b = (1-\rho, 1/2, 1/2, \rho), \qquad s = \rho/4, \qquad v = s(1,1,-1,1),$$

and define

$$\mathcal{G} = \{b + uv : u \in [-1,1]\}.$$

All coordinates lie in $[0,1]$. If $u > 0$, coordinate 0 exceeds $1-\rho$; if $u < 0$, coordinate 3 is below ρ. Hence

$$\mathcal{G}_\rho = \{b\}.$$

At b, the two witness policies tie. Every reward labeling a node of a ρ-interior optimizer packing must therefore equal b. The legal physical law $p = (\delta_1 + \delta_2)/2$ has $V_Q(b) - p(b) = 0$, so every prospective child has zero decision loss under this same experiment. Thus neither a positive Fano separation nor a positive edge separation can hold at any node. The only admissible packing is the trivial one, and (O.1) follows.

For the regret lower bound, let P_+ and P_- denote the trajectory laws generated by the constant mean rewards $g_+ = b + \varepsilon v$ and $g_- = b - \varepsilon v$, respectively, with Bernoulli observations. Each is an admissible deterministic oblivious-mean environment. At round t, write

$$z_t = p_t(1) - p_t(2) \in [-1,1].$$

The conditional expected losses relative to the optimal witnesses δ_1 and δ_2 are, respectively,

$$s\varepsilon(1 - z_t), \qquad s\varepsilon(1 + z_t). \tag{O.2}$$

At a common counterfactual public history, the learner uses the same action kernel in the two environments. Conditional on either sampled witness action, the Bernoulli means are $1/2 + \delta$ and $1/2 - \delta$, where $\delta = s\varepsilon \leq 1/8$. Therefore

$$\mathrm{kl}\left(\frac{1}{2} + \delta \,\middle\|\, \frac{1}{2} - \delta\right) \leq \frac{(2\delta)^2}{(\frac{1}{2} - \delta)(\frac{1}{2} + \delta)} \leq \frac{256}{15}\delta^2 \leq 18\delta^2. \tag{O.3}$$

The adaptive transcript chain rule gives

$$D_{\mathrm{KL}}(P_+^{t-1} \| P_-^{t-1}) \leq 18(t-1)s^2\varepsilon^2,$$

and Pinsker's inequality yields

$$\|P_+^{t-1} - P_-^{t-1}\|_{\mathrm{TV}} \leq 3s\varepsilon\sqrt{t-1}. \tag{O.4}$$

Choose

$$\varepsilon = \min\left\{1, \frac{1}{6s\sqrt{T}}\right\}.$$

For every $t \le T$, the right side of (O.4) is at most $1/2$. Indeed, this is immediate on the second branch, while $\varepsilon = 1$ implies $s\sqrt{T} \le 1/6$. Averaging the two losses in (O.2) gives

$$\begin{aligned}&\frac{s\varepsilon}{2}\left[\mathbb{E}_+(1-z_t) + \mathbb{E}_-(1+z_t)\right]\\ &\qquad = s\varepsilon\left[1 - \frac{1}{2}(\mathbb{E}_+ z_t - \mathbb{E}_- z_t)\right] \ge s\varepsilon\left[1 - \|P_+^{t-1} - P_-^{t-1}\|_{\mathrm{TV}}\right] \ge \frac{1}{2}s\varepsilon.\end{aligned}$$

Summing over t gives Bayes regret at least

$$\frac{s\varepsilon T}{2} = \min\left\{\frac{sT}{2}, \frac{\sqrt{T}}{12}\right\} = \min\left\{\frac{\rho T}{8}, \frac{\sqrt{T}}{12}\right\}.$$

One of the two deterministic prior atoms therefore has expected regret at least this value, proving (O.1). Finally, $\Gamma_T^{\mathrm{obs}} \ge \mathfrak{M}_T^{\star,\mathrm{ad}} \ge \mathfrak{M}_T^{\star,\mathrm{obl}}$, completing the proof. □

O.2 Simultaneous product packing removes the depth loss

Retain the hierarchical routing model of Appendix M with reliabilities $\lambda_j \in (0, 1/2]$, amplitudes $\alpha_j \ge 0$, and $\sum_j \alpha_j \le 1$. Define the active set

$$J_+ := \{j \in [D] : \alpha_j > 0\}.$$

If $J_+ = \varnothing$, all policies have the same value under every reward and both minimax regrets are zero. Hence assume $J_+ \ne \varnothing$. Let $\Theta = (\Theta_j)_{j \in J_+}$ be uniform on $\{-1, +1\}^{J_+}$, fix an inactive sign vector $\bar{x} \in \{-1, +1\}^{[D]\setminus J_+}$, and choose amplitudes $\varepsilon_j \in [0, 1]$ for $j \in J_+$ and $\varepsilon_j = 0$ otherwise. Write $q^{\lambda}_{\Theta,\bar{x}}$ for the template with active signs Θ and inactive signs $\bar{x}$. The constant mean reward is

$$g^{\boldsymbol{\varepsilon}}_{\Theta}(a) = \frac{1}{2} + \frac{1}{4}\sum_{j \in J_+} \alpha_j \varepsilon_j \Theta_j a_j. \tag{O.5}$$

Lemma O.2 (Global comparator and additive gap)**.** *For every $\Theta \in \{-1, +1\}^{J_+}$ and the fixed inactive extension $\bar{x}$,*

$$q^{\lambda}_{\Theta,\bar{x}} \in \operatorname*{arg\,max}_{q \in Q_{\lambda,D}} q(g^{\boldsymbol{\varepsilon}}_{\Theta}).$$

Moreover, if a legal physical law $p_t \in C_{Q_{\lambda,D}}$ has coordinate means $\mathbb{E}_t A_{t,j} = \lambda_j m_{t,j}$, then

$$V_Q(g^{\boldsymbol{\varepsilon}}_{\Theta}) - p_t(g^{\boldsymbol{\varepsilon}}_{\Theta}) = \frac{1}{4}\sum_{j \in J_+} \alpha_j \varepsilon_j \lambda_j (1 - \Theta_j m_{t,j}). \tag{O.6}$$

Proof. For every template x,

$$q^{\lambda}_x(g^{\boldsymbol{\varepsilon}}_{\Theta}) = \frac{1}{2} + \frac{1}{4}\sum_{j \in J_+} \alpha_j \varepsilon_j \lambda_j x_j \Theta_j.$$

The expression is maximized coordinatewise by $x_j = \Theta_j$ on J_+; set $x_j = \bar{x}_j$ off J_+. This proves the first assertion. Subtracting the value under p_t, whose jth coordinate mean is $\lambda_j m_{t,j}$, gives (O.6). □

$$\Psi_T(\lambda, \alpha) := \sum_{j \in J_+} \lambda_j \min\{\alpha_j T, \sqrt{T}\}, \qquad c_{\mathrm{pack}} := \frac{1}{8}\sqrt{\frac{3}{8}}. \tag{O.7}$$

Theorem O.3 (Active-coordinate simultaneous product packing). *If $J_+ = \varnothing$, both minimax regrets are zero. Otherwise, for every $D, T \geq 1$,*

$$\mathfrak{M}_T^{\star,\mathrm{obl}}(Q_{\lambda,D}, \mathcal{G}_{\alpha,D}) \geq \frac{1}{8} \sum_{j \in J_+} \lambda_j \min\left\{\alpha_j T, \sqrt{\frac{3T}{8}}\right\}, \tag{O.8}$$

$$\mathfrak{M}_T^{\star,\mathrm{ad}}(Q_{\lambda,D}, \mathcal{G}_{\alpha,D}) \leq \min\left\{\frac{T}{2}, \sum_{j \in J_+} \lambda_j \min\left\{\frac{\alpha_j T}{4}, C_{\mathrm{sh}}\sqrt{T}\right\}\right\}. \tag{O.9}$$

Consequently,

$$c_{\mathrm{pack}} \Psi_T(\lambda, \alpha) \leq \mathfrak{M}_T^{\star,\mathrm{obl}} \leq \mathfrak{M}_T^{\star,\mathrm{ad}} \leq C_{\mathrm{sh}} \Psi_T(\lambda, \alpha). \tag{O.10}$$

If

$$T \geq \frac{3}{8\alpha_{\min}^2}, \qquad \alpha_{\min} := \min_{j \in J_+} \alpha_j, \tag{O.11}$$

then

$$\frac{1}{8}\sqrt{\frac{3T}{8}} \sum_{j \in J_+} \lambda_j \leq \mathfrak{M}_T^{\star,\mathrm{obl}} \leq \mathfrak{M}_T^{\star,\mathrm{ad}} \leq C_{\mathrm{sh}}\sqrt{T} \sum_{j \in J_+} \lambda_j. \tag{O.12}$$

The lower bound uses one oblivious active-coordinate latent vector and one fixed legal comparator; the upper bound permits adaptive nonanticipating conditional-mean environments.

Proof. Fix a behavioral learner. For $j \in J_+$, define the prefix-mixture laws

$$P_{j,+}^{t-1} = 2^{-(|J_+|-1)} \sum_{\theta : \theta_j = +1} P_\theta^{t-1}, \qquad P_{j,-}^{t-1} = 2^{-(|J_+|-1)} \sum_{\theta : \theta_j = -1} P_\theta^{t-1}. \tag{O.13}$$

The predictable coordinate $m_{t,j}$ is an H_{t-1}-measurable function taking values in $[-1, 1]$. Therefore the variational characterization of total variation gives

$$\begin{aligned} \mathbb{E}[\Theta_j m_{t,j}] &= \frac{1}{2}\left(\mathbb{E}_{P_{j,+}^{t-1}} m_{t,j} - \mathbb{E}_{P_{j,-}^{t-1}} m_{t,j}\right) \\ &\leq \|P_{j,+}^{t-1} - P_{j,-}^{t-1}\|_{\mathrm{TV}}. \end{aligned} \tag{O.14}$$

For each $j \in J_+$, pair every active latent vector θ with $\theta^{\oplus j}$, obtained by flipping only coordinate j. At the same counterfactual public history the learner uses the same action kernel under the paired environments. Conditional on the same realized route, the two Bernoulli means differ by $\alpha_j \varepsilon_j a_j/2$ and both lie in $[1/4, 3/4]$. Hence

$$\mathrm{kl}(u\|v) \leq \frac{(u-v)^2}{v(1-v)} \leq \frac{4}{3}\alpha_j^2 \varepsilon_j^2. \tag{O.15}$$

The adaptive transcript chain rule gives the same bound per round for $D_{\mathrm{KL}}(P_\theta^{t-1} \| P_{\theta^{\oplus j}}^{t-1})$. Averaging over the paired vectors and applying joint convexity of relative entropy yields

$$D_{\mathrm{KL}}(P_{j,+}^{t-1} \| P_{j,-}^{t-1}) \leq \frac{4}{3}(t-1)\alpha_j^2 \varepsilon_j^2.$$

Pinsker's inequality therefore gives

$$\|P_{j,+}^{t-1} - P_{j,-}^{t-1}\|_{\mathrm{TV}} \leq \alpha_j \varepsilon_j \sqrt{\frac{2(t-1)}{3}}. \tag{O.16}$$

Choose

$$\varepsilon_j = \min\left\{1, \sqrt{\frac{3}{8}}\frac{1}{\alpha_j\sqrt{T}}\right\}, \qquad j \in J_+.$$

Then the right side of (O.16) is at most $1/2$ for every $t \leq T$. Lemma O.2, (O.14), and (O.16) yield

$$\mathbb{E}_\Theta \mathbb{E}_{\mathsf{Alg},\mathsf{Adv}_\Theta}[\mathrm{Reg}_T^Q] \geq \frac{T}{8} \sum_{j \in J_+} \alpha_j \varepsilon_j \lambda_j,$$

which is exactly (O.8). Each active latent vector, together with the fixed inactive extension, defines an individually admissible deterministic oblivious-mean Bernoulli environment, so one prior atom has expected regret at least the Bayes average. Under (O.11), every active coordinate uses the information-critical amplitude, proving (O.12). □

Proof of the upper and profile bounds in Theorem O.3. For an active coordinate j, the shared estimator of Appendix M gives adaptive expected coordinate regret at most $C_{\mathrm{sh}}\lambda_j\sqrt{T}$. There is also a direct no-learning option: fix $m_{t,j} = 0$ and perform no update on that coordinate. Its contribution relative to the best fixed sign is

$$\frac{1}{4}\alpha_j\lambda_j\left|\sum_{t=1}^{T}\theta_{t,j}\right| \le \frac{1}{4}\alpha_j\lambda_j T.$$

Choose the better option independently at each active coordinate and put $m_{t,j} = 0$ on inactive coordinates. Lemma M.1 shows that the resulting vector of coordinate means induces one legal product route law, so the same route/reward observation updates all coordinates that are learned. The coordinatewise value decomposition then yields

$$\mathfrak{M}_T^{\star,\mathrm{ad}} \le \sum_{j\in J_+}\lambda_j \min\left\{\frac{\alpha_j T}{4}, C_{\mathrm{sh}}\sqrt{T}\right\}.$$

Every policy value lies in $[1/4, 3/4]$, giving the additional cap $T/2$ and proving (O.9).

Let $c_0 = \sqrt{3/8}$. Since $\min\{a, c_0\sqrt{T}\} \ge c_0\min\{a, \sqrt{T}\}$ for $a \ge 0$, the lower bound implies the first inequality in (O.10). Since $C_{\mathrm{sh}} \ge 1/4$,

$$\min\left\{\frac{a}{4}, C_{\mathrm{sh}}\sqrt{T}\right\} \le C_{\mathrm{sh}}\min\{a, \sqrt{T}\},$$

which proves the last inequality. Under (O.11), every active coordinate is on the information-limited branch of the lower bound, while the shared estimator gives the displayed upper bound in (O.12). □

Corollary O.4 (Equal-depth additive hierarchy)**.** *Take $\lambda_j = 1/2$ and $\alpha_j = 1/D$. If $T \ge 3D^2/8$, then*

$$\frac{1}{16}\sqrt{\frac{3}{8}}\,D\sqrt{T} \le \mathfrak{M}_T^{\star,\mathrm{obl}} \le \mathfrak{M}_T^{\star,\mathrm{ad}} \le \frac{C_{\mathrm{sh}}}{2}D\sqrt{T}. \tag{O.17}$$

Hence the minimax regret is $\Theta(D\sqrt{T})$ in this regime. The horizon condition also implies $D\sqrt{T} \le \sqrt{8/3}\,T$, so the display is compatible with the trivial cap.

Proof. Substitute $\sum_j \lambda_j = D/2$ and $\alpha_{\min} = 1/D$ in (O.12). □

O.3 Arbitrary-score ExO and the corresponding offset-DEC expression

The affine restriction is not needed for the two-environment obstruction. We use the coefficient and quantifier order in (M.20). Scores are finite real-valued Borel functions. If an exponential expectation is $+\infty$, the desired lower bound is immediate. For a mediator law ω, write

$$S_\omega := \{q : \omega(q) > 0\}.$$

All importance-normalized differences are evaluated only on S_ω; their values on the null support are irrelevant.

Theorem O.5 (Score-independent global-potential obstruction)**.** *For the canonical hierarchical routing class and every $\eta > 0$,*

$$\mathrm{ExO}_\eta(\mathcal{M}_D^{\mathrm{mc}}) \ge \beta_D(\eta) \ge \frac{1}{32}\min\{\eta, 1\}, \tag{O.18}$$

where β_D is the binary envelope in Appendix M. Consequently,

$$\mathsf{OptFlatExO}_{D,T} \ge \frac{1}{32}\min\{T, \sqrt{DT}\}, \tag{O.19}$$

and the same lower bound holds for $\mathsf{StdFlatExO}_{D,T}$. Under the manuscript's half-Hellinger convention, it also holds for the corresponding fixed-rate one-global-Shannon coefficient expression obtained by replacing ExO_η by $\mathrm{rdec}^{\mathrm{o}}_{1/(4\eta)}(\mathrm{co}(\mathcal{M}_D^{\mathrm{mc}}))$.

Proof. Put

$$a_D := \alpha_1 = \frac{1}{2(1-2^{-D})} \in [1/2, 1].$$

Choose templates q^+, q^- that differ only at coordinate one, with first signs $+1$ and -1, and fix

$$\mu_0 = \frac{1}{2}\delta_{q^+} + \frac{1}{2}\delta_{q^-}.$$

For $t \in [0, a_D]$, define two legal branch-additive rewards and Bernoulli models by

$$g_{\sigma,t}(a) = \frac{1}{2} + \sigma\frac{t}{4}a_1, \qquad Y \mid A = a \sim \mathrm{Ber}(g_{\sigma,t}(a)), \qquad \sigma \in \{+1, -1\}. \tag{O.20}$$

Use comparator q^+ under the positive model and q^- under the negative model. After an arbitrary sampling law ω and score G are fixed, set

$$\mathsf{D}_{\omega,G}(q, a, y) = \frac{G(q^-; q, a, y) - G(q^+; q, a, y)}{\omega(q)} \quad (q \in S_\omega), \tag{O.21}$$

and set $\mathsf{D}_{\omega,G} = 0$ off S_ω. This function is finite and Borel. Under the negative model with comparator q^-, the nonzero reference term is exactly $-\mathsf{D}_{\omega,G}$.

Let

$$m_\omega = \sum_x \omega(q_x)x_1.$$

Because $\mathbb{E}_{q_x} A_1 = x_1/2$, the two decision losses are

$$q^+(g_{+,t}) - p_\omega(g_{+,t}) = \frac{t}{8}(1 - m_\omega), \qquad q^-(g_{-,t}) - p_\omega(g_{-,t}) = \frac{t}{8}(1 + m_\omega),$$

and their average is exactly $t/8$.

For $\sigma \in \{+1, -1\}$, define the law of the selected mediator, route, and observation by

$$P^\omega_{\sigma,t}(q, a, y) = \omega(q)q(a)[g_{\sigma,t}(a)]^y[1 - g_{\sigma,t}(a)]^{1-y}, \qquad y \in \{0, 1\}. \tag{O.22}$$

The two laws share the same (Q, A) marginal. The two ExO objectives for the chosen models and comparators are therefore

$$\Gamma_+ = \frac{t}{8}(1 - m_\omega) + \frac{1}{2\eta}\left(\mathbb{E}_{P^\omega_{+,t}} e^{\eta \mathsf{D}_{\omega,G}} - 1\right), \tag{O.23}$$

$$\Gamma_- = \frac{t}{8}(1 + m_\omega) + \frac{1}{2\eta}\left(\mathbb{E}_{P^\omega_{-,t}} e^{-\eta \mathsf{D}_{\omega,G}} - 1\right). \tag{O.24}$$

If either expectation is infinite, the result follows. Otherwise let $\zeta = P^\omega_{+,t} + P^\omega_{-,t}$ and $p_\pm = dP^\omega_{\pm,t}/d\zeta$. Pointwise AM–GM gives

$$p_+ e^{\eta \mathsf{D}_{\omega,G}} + p_- e^{-\eta \mathsf{D}_{\omega,G}} \geq 2\sqrt{p_+ p_-}. \tag{O.25}$$

Conditional on any mediator and route, the Bernoulli affinity equals

$$\sqrt{\left(\frac{1}{2} + \frac{t}{4}a_1\right)\left(\frac{1}{2} - \frac{t}{4}a_1\right)} + \sqrt{\left(\frac{1}{2} - \frac{t}{4}a_1\right)\left(\frac{1}{2} + \frac{t}{4}a_1\right)} = \sqrt{1 - \frac{t^2}{4}}, \tag{O.26}$$

independently of q, a, and ω. Averaging (O.23)–(O.24), integrating (O.25), and using (O.26) yield

$$\sup_{q^\star, M} \Gamma_{\mu_0,\eta}(\omega, G; q^\star, M) \geq \frac{t}{8} - \frac{1 - \sqrt{1 - t^2/4}}{2\eta}. \tag{O.27}$$

The right side is independent of ω and G, so the infimum over both and the outer supremum over μ give

$$\mathrm{ExO}_\eta(\mathcal{M}^{\mathrm{mc}}_D) \geq \beta_D(\eta) := \max_{0 \leq t \leq a_D} \left\{\frac{t}{8} - \frac{1 - \sqrt{1 - t^2/4}}{2\eta}\right\}.$$

For $0 < \eta \le 1$, choose $t = \eta/2$ and use $1 - \sqrt{1-u} \le u$ to obtain $\beta_D(\eta) \ge \eta/32$. For $\eta \ge 1$, choose $t = 1/2$ to obtain $\beta_D(\eta) \ge 1/32$. This proves (O.18). Balancing this lower bound with $D \log 2/\eta$ in the cases $D \le T$ and $D > T$ gives (O.19); the larger standard entropy penalty inherits the same lower bound.

Under the half-Hellinger convention, the comparison

$$\mathrm{ExO}_\eta(\mathcal{M}) \le \mathrm{rdec}^{\mathrm{o}}_{1/(4\eta)}(\mathrm{co}(\mathcal{M}))$$

shows that replacing ExO by the displayed offset-DEC coefficient can only increase the same fixed-rate global-potential expression. □

Remark O.6 (Scope). The theorem lower-bounds the optimistic and standard fixed-rate one-global-Shannon coefficient expressions, even with arbitrary Borel ExO scores. It does not lower-bound factorized potentials, product or non-Shannon regularizers, time-varying potentials, every algorithm derived from DEC, or the abstract minimax power of DEC. The shared routing algorithm itself achieves $O(\sqrt{T})$ on the canonical hierarchy.

P FIXED-DESIGN SOCP EVALUATION OF RANGE-CONTROLLED QUADRATIC SCORES

This appendix gives an exact finite-dimensional conic formulation of the score construction used by the range-controlled quadratic Chained-ExO envelope. The result addresses a deliberately local computational question. The approximation chain, the current levelwise reference laws, the legal physical experiment, the bias tolerances, and the score-range bounds are fixed. Under a polytopal reward class, the minimum affine-score energy and the exact robust pairwise second moment are then second-order cone programs (SOCPs) of polynomial size. For a coherent chain, the current reference laws determine a legal common physical design, so one block-diagonal SOCP computes all level scores used in the round.

The distinction between fixed-design score construction and global hierarchy selection is essential. The former is conic; optimizing a finite score-range constraint jointly over the physical law and the score offsets is nonconvex in general. The theorem therefore does not claim a polynomial-time procedure for selecting the witness family or the hierarchy. Accordingly, the computational result concerns score construction for a fixed coherent hierarchy and fixed common design, rather than global hierarchy selection or the exact nonlinear exponential saddle objective.

We use the rotated quadratic cone

$$\mathcal{Q}_{\mathrm{r}}^3 := \left\{(x,y,z) \in \mathbb{R}^3 : x \geq 0,\ y \geq 0,\ 2xy \geq z^2\right\}. \tag{P.1}$$

Rotated-cone representations of quadratic-over-linear perspectives are standard in SOCP modeling; see, for example, Alizadeh & Goldfarb (2003). In particular,

$$z^2 \leq pu, quadp, u \geq 0 \quad \Longleftrightarrow \quad (p/2, u, z) \in \mathcal{Q}_{\mathrm{r}}^3. \tag{P.2}$$

P.1 FINITE DATA AND ENDPOINT REDUCTIONS

Let

$$\mathcal{A} = \{1, \dots, K\}, \qquad \mathcal{G} = \mathrm{co}\{g^1, \dots, g^M\} \subseteq [0,1]^K.$$

Let

$$\mathcal{D} = \{d^1, \dots, d^n\} \subseteq \mathsf{H}_0, \qquad \mu \in \Delta([n]), \qquad \bar{d}_\mu = \sum_{i=1}^n \mu_i d^i,$$

and put

$$h_i := d^i - \bar{d}_\mu.$$

Fix a physical law $p \in \Delta(\mathcal{A})$, a bias tolerance $\varepsilon \geq 0$, and a finite score-range bound $L \geq 0$. Write

$$I_p := \{a : p_a > 0\}, \qquad s := |I_p|.$$

An affine score system is

$$S_i(a,y) = c_i + \psi_{i,a} y.$$

It is (ε, L)-feasible if

$$\max_i \sup_{g \in \mathcal{G}} \left| c_i + \sum_a p_a \psi_{i,a} g(a) - h_i(g) \right| \leq \varepsilon, \tag{P.3}$$

$$\max_{i,r,a,y \in [0,1]} |S_i(a,y) - S_r(a,y)| \leq L. \tag{P.4}$$

Define the uniform score-energy value

$$\mathfrak{U}_{\varepsilon,L}^{\mathcal{D}}(\mu, p) := \inf_{(c,\psi)\ \text{feasible}} \max_{1 \leq i \leq n} \sum_a p_a \psi_{i,a}^2, \tag{P.5}$$

and the exact robust pairwise second-moment value

$$\mathfrak{B}_{\varepsilon,L}^{\mathcal{D}}(\mu, p) := \inf_{(c,\psi)\ \text{feasible}} \max_{\substack{1 \leq r \leq n \\ g \in \mathcal{G}}} \sum_a p_a \sum_i \mu_i \Big[(1 - g(a))(c_i - c_r)^2 + g(a)(c_i - c_r + \psi_{i,a} - \psi_{r,a})^2 \Big]. \tag{P.6}$$

The value is $+\infty$ when the feasible set is empty.

Lemma P.1 (Support, vertex, and endpoint reduction)**.** *In* (P.5) *and* (P.6)*, one may impose* $\psi_{i,a} = 0$ *for* $a \notin I_p$ *without changing either value. Moreover,* (P.3) *is equivalent to its restriction to* $g^1, \dots, g^M$*,* (P.4) *is equivalent to checking* $y \in \{0, 1\}$*, and the maximum over* g *in* (P.6) *equals the maximum over the* M *listed reward vertices.*

Proof. If $p_a = 0$, replacing every $\psi_{i,a}$ by zero changes neither the bias nor the energy nor the pairwise second moment. At that action, every endpoint-1 difference becomes $c_i - c_r$, which is already controlled by the endpoint-0 range constraint, so the score range does not increase.

For fixed (c, ψ, i), the signed bias error is affine in g. Its upper and lower bounds are therefore verified on the vertices of the reward polytope. Equivalently, if $g = \sum_m \lambda_m g^m$, then the error equals the same convex combination of the vertex errors, and its absolute value is at most ε. For fixed (i, r, a), the score difference is affine in y, so the absolute value, being convex, attains its maximum over $[0, 1]$ at an endpoint. Finally, for fixed scores and comparator r, the expression in (P.6) is affine in g, and hence its maximum over the polytope is attained at a listed vertex. □

P.2 Exact SOCP for the uniform Riesz radius

For $a \in I_p$, introduce the weighted coefficient

$$z_{i,a} := p_a \psi_{i,a}.$$

Consider the following conic program, denoted U-SOCP:

$$\text{minimize} \quad U \tag{P.7}$$

subject to, for every i, m,

$$-\varepsilon \le c_i + \sum_{a \in I_p} z_{i,a} g_a^m - h_i(g^m) \le \varepsilon, \tag{P.8}$$

for every i, r,

$$-L \le c_i - c_r \le L, \tag{P.9}$$

for every i, r and $a \in I_p$,

$$-L p_a \le p_a(c_i - c_r) + z_{i,a} - z_{r,a} \le L p_a, \tag{P.10}$$

and, for every $i, a \in I_p$,

$$(p_a/2, u_{i,a}, z_{i,a}) \in \mathcal{Q}_{\mathrm{r}}^3, \tag{P.11}$$

$$\sum_{a \in I_p} u_{i,a} \le U. \tag{P.12}$$

Theorem P.2 (Exact uniform-energy SOCP)**.** *The optimal value of* U-SOCP *equals* $\mathfrak{U}_{\varepsilon,L}^{\mathcal{P}}(\mu, p)$ *in the extended-real sense. If the feasible set is nonempty, the optimum is attained. The formulation uses*

$$O(ns + n)$$

variables, ns *three-dimensional rotated-cone constraints, and*

$$O(nM + n^2 s)$$

linear inequalities.

Proof. Let (c, ψ) be feasible in (P.5), with $\psi_{i,a} = 0$ off I_p. Set

$$z_{i,a} = p_a \psi_{i,a}, \qquad u_{i,a} = p_a \psi_{i,a}^2, \qquad U = \max_i \sum_a p_a \psi_{i,a}^2.$$

The bias constraints become (P.8). The endpoint-0 range constraints are (P.9); multiplying each endpoint-1 constraint by the positive number p_a gives (P.10). Finally, $z_{i,a}^2 = p_a u_{i,a}$, so (P.11) holds with equality, and (P.12) holds. Thus the conic optimum is no larger than the original value.

Conversely, let (c, z, u, U) be conic feasible. Define

$$\psi_{i,a} = \begin{cases} z_{i,a}/p_a, & a \in I_p, \\ 0, & a \notin I_p. \end{cases}$$

The bias and range constraints follow from (P.8)–(P.10) and Lemma P.1. The rotated-cone constraint gives $z_{i,a}^2 \le p_a u_{i,a}$, hence

$$\sum_a p_a \psi_{i,a}^2 = \sum_{a \in I_p} \frac{z_{i,a}^2}{p_a} \le \sum_{a \in I_p} u_{i,a} \le U.$$

Therefore the original value is no larger than the conic optimum, proving equality.

For attainment, assume feasibility and choose a minimizing sequence with $U \le U_0$. Replacing each $u_{i,a}$ by $z_{i,a}^2/p_a$ cannot violate any constraint. The energy bound gives $|z_{i,a}| \le \sqrt{p_a U_0}$. Fix one reward vertex g^1. The bias constraint then gives

$$|c_i| \le \varepsilon + |h_i(g^1)| + \sum_{a \in I_p} |z_{i,a}|\,|g_a^1|,$$

so the offsets are bounded. The canonical u variables are also bounded. A convergent subsequence exists, the conic feasible set is closed, and the objective is continuous. Its limit attains the infimum. □

P.3 Exact SOCP for the robust pairwise second moment

The exact quadratic Chained-ExO envelope uses the pairwise value (P.6), rather than only its uniform-energy upper bound. It also has a rational-data SOCP representation.

Use the variables (c, z) and the constraints (P.8)–(P.10). For every score index i, comparator index r, reward vertex m, and action $a \in I_p$, introduce $t_{i,r,m,a}^0, t_{i,r,m,a}^1 \ge 0$ and impose

$$\left(\frac{(1-g_a^m)p_a}{2}, t_{i,r,m,a}^0, (1-g_a^m)p_a(c_i - c_r)\right) \in \mathcal{Q}_{\mathrm{r}}^3, \tag{P.13}$$

$$\left(\frac{g_a^m p_a}{2}, t_{i,r,m,a}^1, g_a^m\big[p_a(c_i - c_r) + z_{i,a} - z_{r,a}\big]\right) \in \mathcal{Q}_{\mathrm{r}}^3. \tag{P.14}$$

For every (r, m) impose

$$\sum_{i=1}^{n} \sum_{a \in I_p} \mu_i \big(t_{i,r,m,a}^0 + t_{i,r,m,a}^1\big) \le B, \tag{P.15}$$

and minimize B. Call this program B-SOCP.

Theorem P.3 (Exact pairwise-moment SOCP)**.** *The optimal value of* B-SOCP *equals* $\mathfrak{B}_{\varepsilon,L}^{\mathcal{D}}(\mu, p)$*. If the score-feasibility set is nonempty, the optimum is attained. A rational-data formulation has*

$$O(n^2 M s)$$

variables and three-dimensional rotated-cone constraints, and $O(nM + n^2 s)$ *linear inequalities.*

Proof. For a feasible score system, set $z_{i,a} = p_a \psi_{i,a}$ and choose

$$t_{i,r,m,a}^0 = p_a(1 - g_a^m)(c_i - c_r)^2,$$
$$t_{i,r,m,a}^1 = p_a g_a^m (c_i - c_r + \psi_{i,a} - \psi_{r,a})^2.$$

Both cone constraints hold with equality. By Lemma P.1, the robust pairwise second moment is the maximum of the left side of (P.15) over (r, m). Thus the SOCP value is no larger than (P.6).

Conversely, the rotated-cone constraints imply

$$t_{i,r,m,a}^0 \ge p_a(1 - g_a^m)(c_i - c_r)^2,$$
$$t_{i,r,m,a}^1 \ge \frac{g_a^m}{p_a}\big[p_a(c_i - c_r) + z_{i,a} - z_{r,a}\big]^2$$
$$= p_a g_a^m (c_i - c_r + \psi_{i,a} - \psi_{r,a})^2,$$

where $\psi = z/p$ on I_p. Hence every conic feasible point induces a feasible score system whose robust pairwise second moment is at most B. This proves exact value equivalence.

It remains to justify attainment despite the fact that pairwise differences are invariant under adding one common affine score to every index. Fix an anchor i_0. The range constraints give

$$|c_i - c_{i_0}| \le L, \qquad |\psi_{i,a} - \psi_{i_0,a}| \le 2L.$$

Thus all anchored differences are bounded. Let $x = (c_{i_0}, (\psi_{i_0,a})_{a\in I_p})$ and define the fixed linear map

$$(Ax)_m = c_{i_0} + \sum_{a\in I_p} p_a \psi_{i_0,a} g_a^m.$$

For fixed anchored differences, the bias constraints place every coordinate of Ax in a bounded interval, because the targets $h_i(g^m)$ and all anchored differences are bounded. Since Ax belongs to $\mathrm{range}(A)$, replace x by the minimum-Euclidean-norm preimage $A^\dagger Ax$. This preserves every vertex expectation, every bias constraint, every pairwise difference, and the objective, while giving a uniformly bounded common component. Hence every minimizing sequence has a bounded representative. The epigraph variables may be replaced by their minimal quadratic values and are then bounded by the finite range. Closedness and compactness yield an optimizer. □

P.4 One block-diagonal SOCP for a coherent chain

Fix a finite coherent depth-J chain $\mathcal{C}$ with increment classes $\mathcal{D}_j$, current reference laws $\boldsymbol{\mu} = (\mu_1, \dots, \mu_J)$, tolerances ε_j, finite ranges L_j, and learning rates $\eta_j > 0$ satisfying $\eta_j L_j \le 1$. Coherence gives the legal common design

$$p_{\boldsymbol{\mu}} = q_0 + \sum_{j=1}^{J} \bar{d}_{\mu_j} \in C_Q. \tag{P.16}$$

Let

$$B_j^* = \mathfrak{B}^{\mathcal{D}_j}_{\varepsilon_j, L_j}(\mu_j, p_{\boldsymbol{\mu}}).$$

Define the fixed-reference quadratic chain certificate

$$\mathsf{QChain}_T(\boldsymbol{\mu}; \mathcal{C}) := T \,\mathrm{rad}_{\mathcal{G}}(\mathcal{C}) + 2T \sum_{j=1}^{J} \varepsilon_j + \sum_{j=1}^{J} \left[\frac{2 \log |\mathcal{D}_j|}{\eta_j} + T \eta_j B_j^* \right]. \tag{P.17}$$

Singleton levels are omitted.

Theorem P.4 (Fixed-design range-controlled quadratic score evaluation)**.** *For a fixed finite coherent chain $\mathcal{C}$, current reference tuple $\boldsymbol{\mu}$, coherence-induced common physical design $p_{\boldsymbol{\mu}}$, bias tolerances $\boldsymbol{\varepsilon}$, finite score ranges $\boldsymbol{L}$, and learning rates $\boldsymbol{\eta}$, the nonconstant score part of* (P.17) *is the optimal value of one block-diagonal SOCP: place one copy of* B-SOCP *at every level, use the common fixed law $p_{\boldsymbol{\mu}}$, and minimize*

$$\sum_{j=1}^{J} T \eta_j B_j.$$

If $n_j = |\mathcal{D}_j|$, the formulation uses

$$O\left(KM \sum_{j=1}^{J} n_j^2 \right) \tag{P.18}$$

three-dimensional rotated cones and a polynomial number of variables and linear constraints. Consequently, the current-round range-controlled quadratic Chained-ExO score construction is computationally evaluable without enumerating reward vectors outside the M supplied vertices.

A smaller SOCP is obtained by replacing B_j^* by

$$4U_j^* + \left[\mathrm{diam}_{\mathcal{G}}(\mathcal{D}_j) + 2\varepsilon_j \right]^2,$$

where U_j^* is the value of U-SOCP. *This formulation uses only $O(K \sum_j n_j)$ rotated cones.*

Proof. The shared-design quadratic-envelope theorem bounds the exact chained exponential coefficient by

$$2\sum_j \varepsilon_j + \sum_j \eta_j \mathfrak{B}^{\mathcal{D}_j}_{\varepsilon_j, L_j}(\mu_j, p_{\boldsymbol{\mu}}),$$

because coherence makes the reward-discrepancy term zero at $p_{\boldsymbol{\mu}}$. For fixed $p_{\boldsymbol{\mu}}$, the score variables of different levels are disjoint. Therefore the infimum of the weighted sum is the sum of the levelwise infima, and Theorem P.3 supplies one SOCP block per level. Adding the entropy and terminal-radius constants yields (P.17).

The exact cone count follows from Theorem P.3; replacing the support size by K gives (P.18). The smaller formulation uses Theorem P.2 and the deterministic inequality

$$\mathfrak{B}^{\mathcal{D}}_{\varepsilon,L}(\mu, p) \le 4\mathfrak{U}^{\mathcal{D}}_{\varepsilon,L}(\mu, p) + [\mathrm{diam}_{\mathcal{G}}(\mathcal{D}) + 2\varepsilon]^2.$$

□

Corollary P.5 (Closed-form learning-rate optimization)**.** *After the level values B_j^* have been computed, the optimal fixed learning rate under $0 < \eta_j \le 1/L_j$ is*

$$\eta_j^* = \min\left\{\frac{1}{L_j}, \sqrt{\frac{2\log n_j}{TB_j^*}}\right\} \tag{P.19}$$

when $B_j^* > 0$ and $L_j > 0$. The corresponding contribution is

$$\begin{cases} 2\sqrt{2TB_j^* \log n_j}, & TB_j^* \ge 2L_j^2 \log n_j, \\ 2L_j \log n_j + TB_j^*/L_j, & TB_j^* < 2L_j^2 \log n_j. \end{cases} \tag{P.20}$$

If $B_j^* = 0$, the value is $2L_j \log n_j$ for $L_j > 0$ and zero for $L_j = 0$. The evident unconstrained formula applies when $L_j = 0$ and $B_j^* > 0$.

Proof. Minimize the convex scalar function

$$\eta \mapsto \frac{2\log n_j}{\eta} + TB_j^* \eta$$

over $(0, 1/L_j]$. Its unconstrained minimizer is the square-root term in (P.19); comparison with the boundary gives (P.20). The zero cases follow by monotonicity. □

P.5 Accuracy, online oracle errors, and joint-design boundary

Assume the action probabilities, reward vertices, increments, reference weights, tolerances, ranges, and learning rates are rational. The formulations above are polynomial-size rational SOCP representations because the quadratic perspectives are encoded by rotated cones without taking square roots. Standard conic optimization methods compute numerically feasible, near-optimal solutions to prescribed tolerance on well-posed instances; see Alizadeh & Goldfarb (2003). We do not assert a complete Turing-model bit-complexity theorem covering weak infeasibility, conditioning, or exact rational-time computation of irrational optima.

Corollary P.6 (Approximate online implementation)**.** *Suppose the fixed chain and reward polytope are given explicitly. At round t, let $\boldsymbol{\mu}_t$ be the predictable levelwise reference tuple and let coherence define p_t by* (P.16)*. Solve the unscaled local block SOCP with objective $\sum_j \eta_j B_j$. If its returned objective is within ξ_t of the local optimum, the adaptive Chained-ExO regret theorem holds with $\sum_{t=1}^T \xi_t$ added to its right-hand side. Choosing $\xi_t = \delta/T$ gives total optimization loss at most δ.*

Proof. The SOCP returns feasible scores whose quadratic local objective is at most ξ_t above the fixed-reference optimum. This is exactly a predictable local oracle error in the approximate-saddle version of the Chained-ExO theorem. Summing those errors over rounds proves the claim. □

Finite score range is compatible with SOCP computation at a fixed design, but not with unrestricted joint optimization over the design and offsets.

Proposition P.7 (The joint finite-range feasible set is nonconvex)**.** *When both the physical probability p_a and the score offset difference $c_i - c_r$ are variables, the endpoint-1 range constraint*

$$|p_a(c_i - c_r) + z_{i,a} - z_{r,a}| \le L p_a \tag{P.21}$$

need not define a convex set, even together with $|c_i - c_r| \le L$. Consequently, this joint feasible set cannot have an exact standard SOCP representation through affine lifting in these variables.

Proof. Take $L = 1$ and abbreviate $p = p_a$, $c = c_i - c_r$, and $z = z_{i,a} - z_{r,a}$. The two points

$$(p, c, z) = (0.1, 1, -0.2), \qquad (p, c, z) = (0.9, -1, 1.8)$$

satisfy $|c| \le 1$ and (P.21) with equality. Their midpoint is

$$(p, c, z) = (0.5, 0, 0.8),$$

which violates (P.21) because $0.8 > 0.5$. Since feasible sets of SOCPs and their affine projections are convex, there is no exact standard SOCP formulation through affine lifting in the displayed variables. This statement does not rule out a problem-specific nonlinear reparameterization. The fixed-design theorem avoids this obstruction, and a coherent chain supplies precisely such a legal fixed design at each round. □

For completeness, if the finite-range constraint is removed, the joint uniform-energy design is again conic.

Proposition P.8 (Joint design without score-range control)**.** *Let Q be finite and represent $p \in C_Q$ as $p = \sum_{q \in Q} \lambda_q q$, $\lambda \in \Delta(Q)$. The value*

$$\inf_{p \in C_Q} \mathfrak{U}^{\mathcal{D}}_{\varepsilon,\infty}(\mu, p) \tag{P.22}$$

is an SOCP: use variables (λ, p, c, z, u, U), the linear mixture equations for p, the bias constraints (P.8)*, the cones $(p_a/2, u_{i,a}, z_{i,a}) \in \mathcal{Q}^3_{\mathrm{r}}$, and $\sum_a u_{i,a} \le U$.*

Proof. The conic constraints enforce the closed quadratic-over-linear perspective with the usual convention that $p_a = 0$ forces $z_{i,a} = 0$. The two-direction construction in Theorem P.2 applies verbatim, now with p also a variable constrained linearly to lie in C_Q. No range constraint introduces a product between p_a and an unknown offset difference. □

P.6 Computational claim boundary

The exact conclusion is:

> For a fixed finite coherent chain, current reference tuple, coherence-induced common physical design, fixed bias tolerances, fixed finite score ranges, fixed learning rates, and a polytopal reward class, the exact range-controlled pairwise quadratic score-construction subproblem is one polynomial-size SOCP. A smaller SOCP computes the uniform-Riesz envelope, and the learning rates are optimized in closed form. Joint design is conic without range control, but finite-range joint design is nonconvex in general. (P.23)

This result makes the certificate operational at a fixed hierarchy. It does not solve the combinatorial problem of selecting an optimal witness family or hierarchy, and it does not claim that the exact exponential Chained-ExO saddle objective is an SOCP.