TASKS DO NOT TENSORIZE: LEARNING WHEN TO RESET OR CONTINUE

Anonymous authors
Paper under double-blind review

ABSTRACT

Meta-reinforcement learning usually treats the grouping of source trajectories into tasks as fixed. We study an acquisition model in which the learner decides whether the next controlled trajectory should retain the current hidden task or begin an independent one. This changes the statistical experiment. For every $r \geq 2$, we construct finite meta-MDPs in which every episodewise experiment has zero information about two target-incompatible models while one bounded persistent experiment is informative; under perturbation the persistent-to-episodewise rate ratio is arbitrarily large. For every fixed h, the zero-terminal-value h-step receding-horizon information optimizer can likewise be arbitrarily inefficient. For finite persistent meta-MDPs with bounded task depth, a finite observational quotient, and disjoint single-answer cells, we characterize the exact fixed-confidence source cost over all randomized history-dependent experiment trees. The characteristic information is a Bellman–Chernoff game: a dynamic program chooses the next episodic policy and whether to stop, while a least-favorable target alternative prices information. Certified column generation computes a sparse design, and vanishing certified optimization error preserves the characteristic constant. A cross-modal calibration model illustrates how continuation reveals correspondence while resetting supplies independent breadth. Thus task grouping is not bookkeeping: it is an information-control action.

1 INTRODUCTION

Meta-reinforcement learning usually treats the grouping of source trajectories into tasks as part of the access model. We make that grouping a decision. Before each source episode, the learner chooses whether to RESET to an independent latent task or CONTINUE the current latent task while restarting the physical state. For fixed episode policies, these choices induce different experiments:

$$\int \bigotimes_{j=1}^{m} P_{\theta,z}^{\pi_j} \, \mu_\theta(dz) \neq \bigotimes_{j=1}^{m} \int P_{\theta,z}^{\pi_j} \, \mu_\theta(dz). \tag{1.1}$$

Thus the learner controls not only what data are collected, but which data share an unobserved cause.

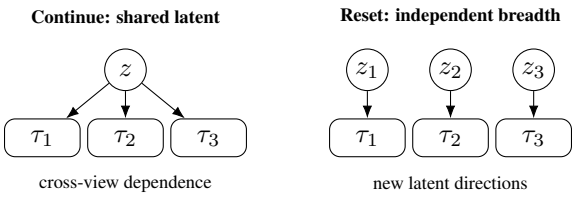


Figure 1: Same policies and the same number of trajectories; only latent-task coupling changes. The learner chooses this coupling.

Two obstructions follow. If every one-episode marginal is invariant, fresh tasks alone reveal nothing about cross-view correspondence. Conversely, unlimited episodes on one task reveal no latent directions absent from that task. The central question is therefore:

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.

Table 1: Closest access models. "Retain/resample" asks whether the learner controls which observations share a task-specific latent.

Family	shared latent	retain/resample	policy traj.	within-task stop	random cost	fresh target
Grouped samples	yes	no	no	no	no	no
Active multitask	indexed ID	choose ID	sometimes	no	possible	yes
Infinite-arm bandits	arm	yes	repeated pull	across pulls	standard	no
Active sequential testing	hypothesis	no	sensing action	global	usually fixed	label
Cost-aware testing	hypothesis	no	sensing action	cancellation	yes	no
RL data acquisition	one MDP	no	yes	global	interaction	usually same MDP
This work	hidden task	yes	yes	yes	yes	independent

Given everything observed from the current hidden task, is another controlled trajectory worth more than resetting?

Contributions. *Non-tensorization and persistence hierarchy.* For every $r \geq 2$, we construct a finite meta-MDP in which all adaptive source interactions using fewer than r same-task views have identical law under two target-incompatible models, while completing an r-view task creates positive information. Episodewise information can be zero while persistent information is positive, with an arbitrarily large positive-rate ratio under perturbation. Moreover, for every fixed planning depth h, the zero-terminal-value h-step receding-horizon information rule defined in Appendix F can be arbitrarily worse than the full Bellman policy (Theorems 3.1, F.1 and F.2).

Exact reset-or-continue complexity. For finite persistent meta-MDPs and maximum same-task depth L, we optimize over the complete deterministic experiment-tree class. A constructive random tape reduction represents every randomized history-dependent bounded experiment. After quotienting observationally equivalent models and imposing disjoint single-answer cells, the exact fixed-confidence constant is a Bellman–Chernoff value. Exact and certified approximate designs attain that constant under conditional sub-gamma likelihood increments (Theorems 4.1 to 4.3).

Breadth versus correspondence. In group-action models, continuation reveals cross-view correspondence and resetting supplies independent anchor breadth. We obtain quotient identifiability, near-invariance bounds, orthogonal paired-view rates, and fresh-target control consequences. A multi-step cross-modal calibration environment exhibits the Bellman stopping rule (Theorem 5.1 and proposition G.2).

Operational interpretation. The latent task may represent a replayable simulator seed, user/session, physical object, or system instance. CONTINUE preserves that latent configuration while physically reinitializing the episode; RESET requests an independent configuration. The theory requires precisely this retain-versus-resample capability, not that ordinary environments expose it.

Closest theory and novelty boundary. Active sequential testing dynamically chooses sensing actions under a fixed hypothesis; cost-aware testing optimizes expected information per expected random cost; RL data acquisition controls trajectories inside one MDP; and active multitask or infinite-arm models allocate among already indexed or candidate source objects (Naghshvar & Javidi, 2013; Garivier & Kaufmann, 2016; Vershinin et al., 2025; Chen et al., 2022). We do not claim novelty for Bellman recursion, Chernoff design, fractional programming, Track-and-Stop, column generation, compact-group tools, or Procrustes estimation. The contribution is the experiment created when the learner controls whether policy-generated trajectories share one sampled task latent, may stop that same-task interaction from its observations, and is evaluated on an independent fresh target.

Appendix R gives a theorem-level comparison.

2 MODEL: CONTROLLED TASK PERSISTENCE

A fixed unknown global parameter $\theta \in \Theta$ governs a task-latent law $Z \sim \mu_\theta$. Conditional on (θ, Z), one physical episode is a finite-horizon controlled process of horizon H. Before source episode t, the learner chooses $U_t^{\mathrm{pers}} \in \{\mathsf{reset}, \mathsf{continue}\}$ predictably from the observed source history. A reset

draws a fresh $Z_t \sim \mu_\theta$; continuation sets $Z_t = Z_{t-1}$. Every physical episode restarts its physical state. The source cost is

$$\mathsf{C}_t = t + \kappa \sum_{s=1}^{t} \mathbf{1}\{U_s^{\mathrm{pers}} = \mathsf{reset}\}.$$

Appendix A gives the full kernel and filtration construction, including distinct but related source, target-adaptation, and deployment interfaces.

After source stopping, an independent target task $Z_\star \sim \mu_\theta$ is drawn. A target strategy g receives B adaptation episodes and one deployment episode. Let $\mathcal{L}_B(\theta, g)$ be its expected target loss and

$$\mathcal{L}_B^\star(\theta) = \inf_g \mathcal{L}_B(\theta, g).$$

For $\varepsilon > 0$, define the acceptable strategies and pairwise incompatible alternatives

$$\mathcal{G}_{B,\varepsilon}(\theta) \coloneqq \{g : \mathcal{L}_B(\theta, g) \le \mathcal{L}_B^\star(\theta) + \varepsilon\}, \tag{2.1}$$

$$\mathrm{Alt}_{B,\varepsilon}(\theta) \coloneqq \{\theta' : \mathcal{G}_{B,\varepsilon}(\theta) \cap \mathcal{G}_{B,\varepsilon}(\theta') = \varnothing\}. \tag{2.2}$$

The fixed-confidence objective is to output an acceptable target strategy with probability at least $1 - \delta$ under every model while minimizing instance-dependent expected source cost.

Complete bounded experiments and the decision-relevant quotient. For the flagship theorem, all primitive alphabets and the physical horizon are finite, Θ is finite, and one source task may be continued for at most L episodes. A deterministic experiment tree specifies, at every reachable within-task history, the next deterministic episodic policy or the decision to stop, with mandatory stopping by depth L. The finite class of all such trees is $\mathfrak{E}_L$. Tree e induces a stopped transcript law Q_θ^e, expected cost $c_\theta(e)$, and divergence $d_e(\theta, \theta') = D_{\mathrm{KL}}(Q_\theta^e \| Q_{\theta'}^e)$. Crucially, Theorem D.2 proves that every randomized history-dependent depth-L experiment is, under every model, a model-independent mixture over $\mathfrak{E}_L$. Thus optimizing over $\mathfrak{E}_L$ does not hide the reset-or-continue problem inside a supplied experiment menu.

Define observational equivalence by

$$\theta \sim_L \theta' \quad \Longleftrightarrow \quad Q_\theta^e = Q_{\theta'}^e \quad \text{for every } e \in \mathfrak{E}_L, \tag{2.3}$$

and let $\overline{\Theta}_L \coloneqq \Theta/\!\sim_L$. Write $\xi = [\theta]$ for a quotient class. Since the realized tree cost is part of the stopped transcript, $Q_\theta^e = Q_{\theta'}^e$ also implies $c_\theta(e) = c_{\theta'}(e)$. Hence Q_ξ^e, $c_\xi(e)$, and $d_e(\xi, \xi') = D_{\mathrm{KL}}(Q_\xi^e \| Q_{\xi'}^e)$ are representative invariant.

Assumption 2.1 (Single-answer quotient partition)**.** There are a finite nonempty answer set $\mathcal{I}$, a surjective map $\bar{a} : \overline{\Theta}_L \to \mathcal{I}$, and canonical target strategies $(g_i)_{i \in \mathcal{I}}$ such that:

1. for every quotient class ξ, every representative $\theta \in \xi$, and $i = \bar{a}(\xi)$, $g_i \in \mathcal{G}_{B,\varepsilon}(\theta)$;
2. if $\bar{a}(\xi) \neq \bar{a}(\xi')$, then for every $\theta \in \xi$ and $\theta' \in \xi'$,

$$\mathcal{G}_{B,\varepsilon}(\theta) \cap \mathcal{G}_{B,\varepsilon}(\theta') = \varnothing.$$

Under Assumption 2.1, the target-incompatible quotient set is exactly

$$\mathcal{J}_\xi \coloneqq \{\xi' \in \overline{\Theta}_L : \bar{a}(\xi') \neq \bar{a}(\xi)\}. \tag{2.4}$$

Thus the flagship theorem identifies only distinctions that are source observable and capable of changing the target answer. Problems with overlapping valid answers require an answer-specific game and are outside the single-answer theorem proved here.

3 Persistence Hierarchies

Grouped observations can reveal latent structure that is absent from every proper marginal (Vandermeulen & Scott, 2019). We show that, when the grouping is selected through *controlled task persistence*, this effect induces an arbitrary hierarchy of target-relevant source-access requirements.

A parity-tilted meta-MDP. Fix an integer $r \geq 2$ and a signal level $\varrho \in (0, 1)$. The global parameter is $\sigma \in \{-1, +1\}$, and a fresh task has latent $X = (X_1, \ldots, X_r) \in \{-1, +1\}^r$ drawn from

$$\mu_\sigma^{(r,\varrho)}(x) \coloneqq 2^{-r} \left(1 + \varrho\sigma\chi(x)\right), \qquad \chi(x) \coloneqq \prod_{j=1}^{r} x_j. \tag{3.1}$$

The horizon-one MDP has a safe action 0 with reward $1/2$ and coordinate actions $j \in [r]$ with reward $(1 + X_j)/2$. Thus one episode reveals at most one task coordinate; RESET samples a new X, whereas CONTINUE preserves it. Summing the parity term in equation 3.1 over any unobserved coordinate shows that every proper marginal of X is exactly uniform under both values of σ. For target evaluation, we use the same MDP with adaptation budget $B = r - 1$. Let e_{full} query every coordinate once, define

$$d_{\text{par}}(\varrho) \coloneqq \varrho \log \frac{1 + \varrho}{1 - \varrho}, \tag{3.2}$$

and let $N_{\text{full}}(\mathsf{T})$ be the number of source tasks on which all r coordinate actions are queried before reset or source stopping.

Theorem 3.1 (Arbitrary-order target-relevant persistence)**.** *Fix $r \geq 2$, $\varrho \in (0, 1)$, $B = r - 1$, and $0 < \varepsilon < \varrho/2^{r+1}$. The family above satisfies the following.*

1. *Every possibly randomized, history-dependent same-task experiment that reveals at most $r - 1$ distinct coordinates has identical transcript law under $\sigma = +1$ and $\sigma = -1$. In contrast,*
$$D_{\text{KL}}\left(Q_+^{e_{\text{full}}} \middle\| Q_-^{e_{\text{full}}}\right) = d_{\text{par}}(\varrho),$$
and hence the pairwise persistence order is exactly r.

2. *The known-model target loss is*
$$\mathcal{L}_{r-1}^\star(\sigma) = \frac{1 - \varrho}{2^{r+1}},$$
and the two ε-optimal target-strategy sets are disjoint. Consequently, every source learner that almost surely leaves every source task incomplete obeys
$$\sup_{\sigma \in \{-1,+1\}} \mathbb{E}_\sigma^{\mathcal{A}} \left[\mathcal{L}_{r-1}(\sigma, \widehat{g}) - \mathcal{L}_{r-1}^\star(\sigma)\right] \geq \frac{\varrho}{2^{r+1}}. \tag{3.3}$$

3. *For each $\sigma \in \{-1, +1\}$, if the stopping time is finite $\mathbb{P}_\sigma^{\mathcal{A}}$-almost surely and $\mathbb{E}_\sigma^{\mathcal{A}}[N_{\text{full}}(\mathsf{T})] < \infty$, then the stopped source experiment satisfies*
$$D_{\text{KL}}\left(\mathbb{P}_\sigma^{\mathcal{A}} \middle\| \mathbb{P}_{-\sigma}^{\mathcal{A}}\right) = d_{\text{par}}(\varrho)\mathbb{E}_\sigma^{\mathcal{A}}\left[N_{\text{full}}(\mathsf{T})\right]. \tag{3.4}$$
Finite expected source cost implies this integrability condition.

Proof mechanism. The induction invariant is that every observed history revealing a proper subset of the current task coordinates has equal probability under both global models. Conditional on such a history, a safe action is deterministic, a repeated coordinate is already measurable, and every newly selected nonfinal coordinate is fair under both models; private randomization, resets, and stopping use the same history kernels. Only the final missing coordinate has a model-dependent conditional law. Induction first within a task and then across reset blocks proves stopped-transcript invariance. Information enters only when the final missing coordinate is queried: its conditional Bernoulli KL is exactly $d_{\text{par}}(\varrho)$, which yields equation 3.4 by the conditional KL chain rule. For the target problem, after $r - 1$ negative observations the missing bit has opposite parity-dependent biases under the two models; under their uniform mixture it is fair. This gives the target gap in equation 3.3. The complete construction and filtration-level argument appear in Section C; the final testing step is the standard binary change-of-measure method (Yu, 1997; Tsybakov, 2009).

Episodewise and finite-lookahead separation. The unperturbed informative parity experiment uses r coordinate episodes. The positive-rate perturbation first queries one additional weak source bit and therefore has maximum same-task depth $r + 1$. Add a task-fixed weak bit with reliability η that

is visible through one source action but absent from the target. Let $\Gamma^\star_{\mathrm{ep}}$ be the best information rate when every episode uses a fresh task. Then

$$\boxed{\Gamma^\star_{\mathrm{ep}} = 0 \quad \text{while} \quad \Gamma^\star_r > 0} \qquad (\eta = 0), \tag{3.5}$$

and, for positive $\eta \downarrow 0$, the persistent-to-episodewise rate ratio diverges. More strongly, for every fixed depth h of the zero-terminal-value receding-horizon information optimizer and every $M < \infty$, one can choose $r > h$ and $\eta > 0$ such that

$$\frac{\Gamma^\star_{r+1}}{\Gamma^\star_{\mathrm{RH}(h)}} \geq M. \tag{3.6}$$

The specified h-step rule replans after each observation, uses zero terminal value beyond its window, and breaks ties toward reset. It repeatedly takes the weak immediate signal and resets, because no h-episode continuation plan can complete the parity; the full Bellman policy accepts locally uninformative views to unlock the terminal signal. Appendix F gives the exact definitions and proofs. No reweighting of independently marginalized one-episode experiments can reproduce the persistent information.

Corollary 3.2 (Fixed-confidence bounds up to universal constants)**.** *Under the hypotheses of Theorem 3.1, let $\delta \in (0, 1/2)$. For each $\sigma \in \{-1, +1\}$,*

$$\mathfrak{C}^\star_{r-1,\varepsilon,\delta,\kappa}(\sigma) \geq (r+\kappa)\frac{\mathrm{kl}(1-\delta,\delta)}{d_{\mathrm{par}}(\varrho)}, \tag{3.7}$$

$$\mathfrak{C}^\star_{r-1,\varepsilon,\delta,\kappa}(\sigma) \leq (r+\kappa)\left\lceil \frac{2}{\varrho^2}\log\frac{1}{\delta}\right\rceil. \tag{3.8}$$

Hence, for $\varrho \leq 1/2$, as $\delta \downarrow 0$, the upper and lower bounds match up to universal constants:

$$\mathfrak{C}^\star_{r-1,\varepsilon,\delta,\kappa}(\sigma) = \Theta\left((r+\kappa)\varrho^{-2}\log\frac{1}{\delta}\right), \tag{3.9}$$

with universal implicit constants. The upper bound uses independent complete blocks and a majority test on their task parities. With n complete blocks, the same procedure has target excess loss at most

$$\frac{\varrho}{2^r}\exp\left(-\frac{n\varrho^2}{2}\right). \tag{3.10}$$

The hierarchy is qualitative as well as quantitative: infinitely many source episodes are useless if no task is completed, whereas each completed task contributes exactly one unit of global evidence with information $d_{\mathrm{par}}(\varrho)$. Parameterizing $\varrho = \rho^r$ yields the exponential family $\Theta((r+\kappa)\rho^{-2r}\log(1/\delta))$, while the exact results above remain stated in the intrinsic parity signal ϱ.

4 Bellman–Chernoff Complexity

We now solve the bounded reset-or-continue problem over the complete depth-L experiment-tree class. Fix a true observational class $\xi \in \overline{\Theta}_L$ with nonempty target-incompatible quotient set $\mathcal{J}_\xi$ from equation 2.4. For a tree distribution p, define

$$\Gamma^\star_L(\xi) \coloneqq \sup_{p\in\Delta(\mathfrak{E}_L)} \inf_{\xi'\in\mathcal{J}_\xi} \frac{\sum_e p_e D_{\mathrm{KL}}(Q^e_\xi \| Q'^e_\xi)}{\sum_e p_e c_\xi(e)}. \tag{4.1}$$

Theorem 4.1 (Exact complete-tree characteristic time)**.** *Assume finite primitive alphabets and physical horizon, a finite observational quotient $\overline{\Theta}_L$, maximum same-task depth L, the disjoint single-answer partition in Assumption 2.1, common domination and finite ordered tree KL, deterministic tie-breaking, and uniform conditional sub-gamma moments of centered tree log-likelihood ratios. Assume exact access to quotient tree likelihoods, a quotient MLE, and the Bellman–Chernoff saddle point. If $\Gamma^\star_L(\xi) > 0$, then every globally correct depth-L learner satisfies*

$$\liminf_{\delta\downarrow 0} \frac{\mathbb{E}_\xi \mathsf{C}_{T_\delta}}{\log(1/\delta)} \geq \frac{1}{\Gamma^\star_L(\xi)}, \tag{4.2}$$

and exact-design quotient Track-and-Stop attains equality. The lower bound ranges over all randomized history-dependent learners whose individual task blocks stop by depth L.

Theorem 4.2 (Certified approximate designs)**.** *Under the assumptions of Theorem 4.1, replace the exact design on nonforced block k by a predictable distribution p_k satisfying, after the quotient MLE has stabilized,*

$$\inf_{\xi' \in \mathcal{J}_\xi} \frac{\sum_e p_{k,e} d_e(\xi, \xi')}{\sum_e p_{k,e} c_\xi(e)} \geq \Gamma_L^\star(\xi) - \zeta_k, \qquad \frac{1}{n} \sum_{k=1}^{n} \zeta_k \longrightarrow 0. \tag{4.3}$$

Then the same GLR rule remains globally δ-correct and attains the exact characteristic constant. Master-LP and pricing errors may both contribute to ζ_k; correctness is sampling-rule independent, while only expected-cost optimality uses equation 4.3.

The domain statement is substantive. A model-independent private random tape turns any randomized within-task strategy—including correlated randomization across information sets—into a model-independent mixture over $\mathfrak{E}_L$, preserving the stopped transcript, random cost, source stopping, and recommendation. Observationally equivalent original models are then collapsed before learning; no unobservable nuisance distinction is charged.

Bellman–Chernoff representation. Fix a least-favorable mixture $\lambda \in \Delta(\mathcal{J}_\xi)$. For current same-task history h and next episodic policy π, let

$$\iota_{\xi,\lambda}(h, \pi) = \sum_{\xi'} \lambda_{\xi'} D_{\mathrm{KL}} \left(P_\xi^\pi(\cdot \mid h) \middle\| P_\xi'^\pi(\cdot \mid h) \right).$$

For candidate information rate γ, define

$$V_L^{\lambda,\gamma}(h) = 0, \tag{4.4}$$

$$V_t^{\lambda,\gamma}(h) = \max \left\{ 0, \max_\pi \left[\iota_{\xi,\lambda}(h, \pi) - \gamma + \mathbb{E}_\xi V_{t+1}^{\lambda,\gamma}(h \oplus (\pi, Y)) \right] \right\}. \tag{4.5}$$

The zero branch stops the current task; the other branch collects another trajectory. At a fresh-task root,

$$\Phi_{\xi,\lambda}(\gamma) = -\gamma\kappa + \max_\pi \left[\iota_{\xi,\lambda}(\varnothing, \pi) - \gamma + \mathbb{E}_\xi V_1^{\lambda,\gamma}((\pi, Y)) \right]. \tag{4.6}$$

Theorem 4.3 (Bellman–Chernoff theorem)**.** *For every λ, $\Phi_{\xi,\lambda}(\gamma)$ equals*

$$\max_{e \in \mathfrak{E}_L} \left\{ \sum_{\xi'} \lambda_{\xi'} D_{\mathrm{KL}}(Q_\xi^e \| Q_\xi'^e) - \gamma c_\xi(e) \right\}.$$

Its unique zero $\gamma_\lambda^\star$ is the best weighted information-to-cost ratio, and

$$\boxed{\Gamma_L^\star(\xi) = \min_{\lambda \in \Delta(\mathcal{J}_\xi)} \gamma_\lambda^\star.} \tag{4.7}$$

Optimal primal and dual mixtures exist. Moreover, an optimal primal design has support no larger than the number of linearly independent active alternative constraints.

Proof roadmap. Prefix absolute continuity gives a pathwise stopped-likelihood factorization; the KL chain rule writes each tree divergence as the expected sum of conditional increments. Backward induction proves equation 4.5; fractional programming identifies its zero. Cost-normalized tree frequencies form a finite zero-sum game, yielding equation 4.7 and the sparse-support statement. The lower bound is stopped change of measure on the complete tree laws. For the upper bound, sparse forced exploration identifies the quotient class, empirical designs track a saddle point, and conditional sub-gamma concentration controls wrong cells, evidence, realized cost, and the expected stopping tail. Appendices D– E give the complete argument.

Certified computation. The tree class is finite but exponential. Column generation solves a restricted information LP, extracts its least-favorable quotient mixture, and uses equation 4.5 to price the best omitted tree. If the pricing value exceeds the restricted dual by at most η, the returned value is η-optimal; exact pricing terminates after finitely many columns. Sparse support explains why few columns can suffice even when the implicit tree class is enormous. The complete algorithm and its exact and approximate guarantees are in Appendix E.

Algorithm 1 Quotient Bellman–Chernoff Track-and-Stop

for $k = 0, 1, \ldots$ **do**
  if forced exploration is due **then**
    $E_{k+1} \leftarrow$ next quotient-separating tree
  else
    compute quotient MLE $\widehat{\xi}_k$
    warm-start column generation at $\widehat{\xi}_k$ to certified gap η_k
    convert cost fractions $x_{k,e}$ to launch probabilities $p_{k,e} \propto x_{k,e}/c_{\widehat{\xi}_k}(e)$
    sample $E_{k+1} \sim p_k$
  end if
  reset once; execute E_{k+1} once to its built-in stopping time
  update quotient likelihoods and decision-cell statistics Z_{k+1}
  if $\max_i Z_{k+1}(i) \geq \log((|\overline{\Theta}_L| - 1)/\delta)$ **then**
    return the canonical strategy of a maximizing cell
  end if
end for

5 Breadth versus Correspondence

Let a compact group G act continuously on an anchor space $\mathcal{X}$. The global parameter is $g_\star \in G$ and each source task draws $X \sim \mu$. Two controlled policies produce views

$$T_0^g(X) = X, \qquad T_1^g(X) = gX.$$

Resetting draws a new anchor; continuation allows both views of the same anchor. Define the source support kernel

$$K_\mu = \bigcap_{x \in \mathrm{supp}(\mu)} \mathrm{Stab}_G(x).$$

Theorem 5.1 (Reset–continue complementarity)**.** *If μ is G-invariant, every adaptive reset-only transcript is independent of $g_\star$. The paired law $Q_g^{\mathrm{pair}} = (x \mapsto (x, gx))_{\#}\mu$ satisfies*

$$Q_g^{\mathrm{pair}} = Q_{g'}^{\mathrm{pair}} \iff g^{-1}g' \in K_\mu.$$

Given noiseless paired anchors $(x_i, g_\star x_i)_{i=1}^n$, the exact consistency set is

$$g_\star \bigcap_{i=1}^n \mathrm{Stab}_G(x_i).$$

For an isometric action and independent target $X_\star \sim \mu$, every predictor obeys the residual-symmetry lower bound

$$\inf_{\widehat{Y}} \sup_g \mathbb{E}_g d^2(\widehat{Y}, gX_\star) \geq \frac{1}{4}\mathbb{E} \max_{h \in \cap_i \mathrm{Stab}_G(X_i)} d^2(X_\star, hX_\star).$$

Thus continuation creates parameter-dependent correspondence, while resets shrink the residual stabilizer by supplying independent anchors. This is robust. If every reset-only one-view marginal has pairwise KL at most η_{marg}, then any reset-only learner using expected N episodes has total KL at most $\eta_{\mathrm{marg}} N$. There are binary group-action families with $\eta_{\mathrm{marg}} \to 0$ but constant paired information (Appendix W). Under source–target support shift, learning gK_μ suffices exactly when $K_\mu \subseteq K_\nu$, where K_ν is the target support kernel.

Orthogonal paired views. Take $G = \mathrm{O}(d)$, $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$, and observe

$$Y_{i,0} = X_i + \eta_{i,0}, \qquad Y_{i,1} = R_\star X_i + \eta_{i,1},$$

with isotropic effective variances τ_0^2, τ_1^2 and $\Delta_\tau = (1 + \tau_0^2)(1 + \tau_1^2) - 1$. Every unpaired marginal is independent of $R_\star$, while one paired task satisfies the exact geometry

$$D_{\mathrm{KL}}(P_R \| P_S) = \frac{\|R - S\|_{\mathrm{F}}^2}{2\Delta_\tau}. \tag{5.1}$$

The MLE is the polar factor of the empirical cross-moment. Let $p_d = d(d-1)/2$ and allow estimators to be ambient matrix valued.

Theorem 5.2 (Orthogonal correspondence and breadth). *For universal $c, C > 0$,*

$$\mathfrak{R}^{\mathrm{orth}}_{n,d}(\Delta_\tau) \geq \max\left\{ c\left(\frac{\Delta_\tau p_d}{n} \wedge d\right), (d-n)_+ \right\}.$$

For $n \geq 2d+2$,

$$\mathfrak{R}^{\mathrm{orth}}_{n,d}(\Delta_\tau) \leq C\left(\frac{\Delta_\tau p_d}{n} \wedge d\right).$$

In the noiseless experiment the risk is exactly $(d-n)_+$, whereas reset-only access has exact risk d.

The regular term measures noisy correspondence on the p_d-dimensional orthogonal manifold; the singular term measures missing anchor breadth. Appendices K–V contain the exact likelihood, finite-sample proofs, transition-regime upper bounds, and local asymptotic refinement. These tools are supporting analysis, not separate headline contributions.

6 Calibration and Fresh-Target Control

We instantiate the theory in a multi-step cross-modal calibration MDP. Each task draws an object orientation $X \in \{-1, +1\}$ and a latent sensing quality Q. One three-step corridor yields a noisy proprioceptive view of X, another yields a noisy visual-frame view of σX, and a cheap screening corridor produces a signal about Q. Continuing retains the same object; resetting draws a new one. A fresh target reveals $X_\star$, and the agent must execute the route labeled $\sigma X_\star$.

Conditional on screen outcome s, a completed proprioceptive–visual pair has parity reliability r_s and contributes $d_{\mathrm{par}}(r_s)$ information. Screening, followed by both sensing corridors, costs three episodes total.

Corollary 6.1 (Bellman-optimal calibration). *Let $\pi_s = \Pr(S = s)$, $\bar{r} = \sum_s \pi_s r_s$, and*

$$j_w = D_{\mathrm{KL}}\big(\mathcal{L}_+(S \mid W = w) \| \mathcal{L}_-(S \mid W = w)\big), \qquad p^+_w = \tfrac{1}{2}(1 + \bar{r}w).$$

The screen-first and pair-first Bellman values are the unique positive roots

$$\sum_s \pi_s [d_{\mathrm{par}}(r_s) - 2\gamma]_+ = (\kappa + 1)\gamma, \tag{6.1}$$

$$d_{\mathrm{par}}(\bar{r}) + \sum_{w \in \{-1,+1\}} p^+_w [j_w - \gamma]_+ = (\kappa + 2)\gamma. \tag{6.2}$$

Writing these roots as $\Gamma^\star_{S \to W}$ and $\Gamma^\star_{W \to S}$, the complete depth-three value is

$$\Gamma^\star_{\mathrm{cal}} = \max\{\Gamma^\star_{S \to W}, \Gamma^\star_{W \to S}\}, \qquad \mathbb{E}_\sigma \mathsf{C}_{T_\delta} \sim \frac{\log(1/\delta)}{\Gamma^\star_{\mathrm{cal}}}.$$

The screen-first tree completes the pair exactly when $d_{\mathrm{par}}(r_s) > 2\Gamma^\star_{S \to W}$; the pair-first tree acquires the quality signal exactly when $j_w > \Gamma^\star_{W \to S}$. Whenever the globally optimal order uses both stop and continue branches strictly, it outperforms every static depth choice.

This is a closed-form instance of Theorem 4.3, not a separate design principle. Appendix G specifies the finite-state MDP and proves the result.

Corollary 6.2 (Source–target decomposition for orthogonal LQ transfer). *In the Gaussian orthogonal specialization, let a fresh target have goal $G_\star = RX_\star$ and known linear dynamics with endpoint-loss matrix $W_H \succeq 0$. With B noisy target probes, write $\alpha_B = B/(B + \sigma_\star^2)$ and $v_B = \sigma_\star^2/(B + \sigma_\star^2)$, with the conventions in Appendix O. Then*

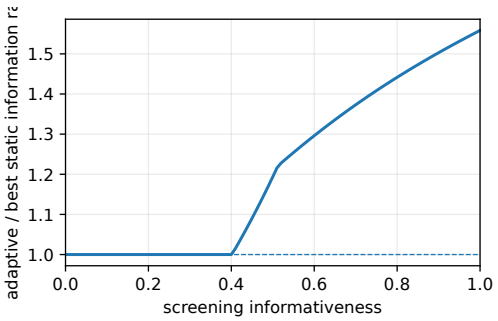
$$\mathfrak{R}^{\mathrm{oracle}}_{n,B} = v_B \operatorname{Tr}(W_H) + \mathfrak{R}^{\mathrm{meta}}_{n,B}.$$

For $n \geq 2d + 2$ and uniformly conditioned W_H,

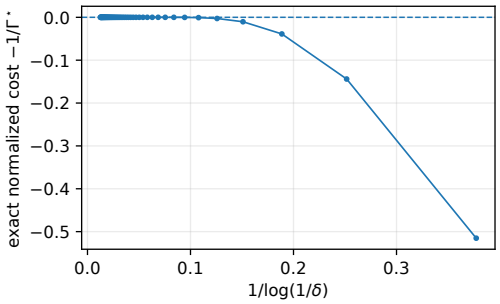
$$\mathfrak{R}^{\mathrm{meta}}_{n,B} \asymp \alpha_B \left(\frac{\Delta_\tau p_d}{n} \wedge d\right).$$

Reset-only source access instead has meta-excess $\alpha_B \operatorname{Tr}(W_H)$.

Thus persistence uncertainty changes an actual sequential target policy, not only parameter error.



(a) Observation-adaptive persistence has no advantage under uninformative screening and gains value as the signal separates useful from poor tasks.



(b) Exact finite-threshold residual $\mathbb{E}C_{T_\delta} / \log(1/\delta) - 1/\Gamma^\star_{\text{pair}}$ for the direct paired-view subexperiment, computed from the absorbing random-walk equations; lattice overshoot causes the visible finite-threshold oscillation.

Figure 2: Theory-calibrated behavior in the cross-modal calibration MDP.

7 Robustness, Scope, and Implications

The exact parity and invariance examples isolate the mechanism, but the phenomenon is quantitatively stable. For a pair of models define

$$I_m(\theta, \theta') := \sup_{e:\, \tau_e \le m} \frac{D_{\mathrm{KL}}(Q^e_\theta \| Q^e_{\theta'})}{c_\theta(e)}.$$

Theorem 7.1 (Quantitative persistence lower bound)**.** *If θ' is pairwise target incompatible with θ and $I_m(\theta, \theta') \le \eta$, every globally (B, ε, δ)-correct learner restricted to task depth m obeys*

$$\mathbb{E}_\theta C_T \ge \frac{\mathrm{kl}(1-\delta, \delta)}{\eta}.$$

For every $r \ge 2$ and every $\eta > 0$, there are finite persistent meta-MDPs with $I_{r-1} \le \eta$ while $I_r \ge d_{\mathrm{par}}(\varrho)/(r + \kappa)$.

The same argument gives $D_{\mathrm{KL}}(\mathbb{P}^{\mathcal{A}}_g \| \mathbb{P}^{\mathcal{A}}_{g'}) \le \eta_{\mathrm{marg}} \mathbb{E}_g T$ for every reset-only learner, although same-task paired information can stay order one as $\eta_{\mathrm{marg}} \downarrow 0$. In group-action transfer, the source quotient K_μ is target sufficient exactly when $K_\mu \subseteq K_\nu$; Appendix W proves both claims.

The reset rule. At a saddle point $(\lambda^\star, \gamma^\star)$, the continuation advantage at depth t and same-task history h is

$$A_t(h) = \max_\pi \left\{ \iota_{\xi, \lambda^\star}(h, \pi) - \gamma^\star + \mathbb{E}_\xi V^{\lambda^\star, \gamma^\star}_{t+1}(H_{t+1}) \right\}.$$

An optimal tree continues for $A_t(h) > 0$, stops for $A_t(h) < 0$, and may randomize only at ties. It can therefore accept zero immediate KL to unlock a complementary future view, or reject immediate information that is not worth delaying a reset. The rate $\gamma^\star = \Gamma^\star_L(\xi)$ also controls the first-order source cost.

Assumptions and limitations. Shared task latents create correspondence and fresh resets create breadth. Finite alphabets and bounded depth make the tree class finite; disjoint single-answer quotient cells make different labels target incompatible; and domination with conditional exponential moments supports the likelihood and cost concentration. The theorem does not cover unbounded task duration, noncompact policy classes, overlapping valid answers, heavy-tailed singular experiments, or function approximation. The finite-lookahead separation is only for the specified zero-terminal-value receding-horizon information rule, not arbitrary finite-memory methods. Bellman pricing is finite but can be exponential in reachable histories, although sparse support and certified column generation can avoid explicit enumeration.

Conclusion. Equal policies and episode counts need not yield equal information: preserving or resampling the hidden task changes the experiment. Bellman–Chernoff design assigns the exact bounded target-decision price to that choice.

AI USE STATEMENT

Generative AI systems were used for literature-search support, development and refinement of the conceptual framework, formulation and critical review of mathematical claims, assistance with proof ideas and proof exposition, manuscript organization and editing, LaTeX drafting, numerical verification code, and build debugging. The authors reviewed all AI-assisted text, code, claims, and citations; executable checks were run where applicable. The authors take full responsibility for every claim, proof, citation, and artifact in this submission. These executable checks support, but do not replace, independent mathematical proof review.

REPRODUCIBILITY STATEMENT

The source-access model, measurable spaces, filtrations, estimator classes, and edge-case conventions are specified in Section 2 and Appendix A. Complete proofs of the persistence hierarchy, complete-tree reduction, Bellman–Chernoff characterization, column-generation certificate, sub-exponential Track-and-Stop extension, robustness results, and structured examples appear in Appendices C, D–G, and F; reusable technical lemmas are collected in Appendix B, and assumption failures are audited in Appendix Q. The anonymous source archive is self-contained: `make all` regenerates every figure, runs all theorem sanity checks, and compiles the paper. The verification suite checks parity marginals and KL identities, behavioral-to-mixed equivalence on finite trees, Bellman and primal–dual values, column-generation certificates, noisy-screening likelihood sufficiency, near-invariance bounds, routing regret, Gaussian KL and Procrustes identities, ridge-risk formulas, and LQ loss identities.

REFERENCES

P.-A. Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008. ISBN 9780691132983.

Alekh Agarwal, Yuda Song, Wen Sun, Kaiwen Wang, Mengdi Wang, and Xuezhou Zhang. Provable benefits of representational transfer in reinforcement learning. In *Proceedings of the Thirty-Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 2114–2187. PMLR, 2023.

Maryam Aliakbarpour, Konstantina Bairaktari, Gavin Brown, Adam Smith, Nathan Srebro, and Jonathan Ullman. Metalearning with very few samples per task. In *Proceedings of the Thirty-Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 46–93. PMLR, 2024.

Brian D. O. Anderson and John B. Moore. *Optimal Control: Linear Quadratic Methods*. Dover Publications, Mineola, NY, 2007. ISBN 9780486457666.

T. W. Anderson. The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proceedings of the American Mathematical Society*, 6(2):170–176, 1955. doi: 10.1090/S0002-9939-1955-0069229-1.

Maryam Aziz, Jesse Anderton, Emilie Kaufmann, and Javed Aslam. Pure exploration in infinitely-armed bandit models with fixed-confidence. In *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pp. 3–24. PMLR, 2018.

Dimitri P. Bertsekas and Steven E. Shreve. *Stochastic Optimal Control: The Discrete-Time Case*. Athena Scientific, Belmont, MA, 1996. ISBN 9781886529038. Originally published by Academic Press in 1978.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24 (2):265–272, 1953. doi: 10.1214/aoms/1177729032.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013. ISBN 9780199535255. doi: 10.1093/acprof:oso/9780199535255.001.0001.

Emma Brunskill and Lihong Li. Sample complexity of multi-task reinforcement learning. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pp. 122–131. AUAI Press, 2013.

Emma Brunskill and Lihong Li. The online coupon-collector problem and its application to lifelong reinforcement learning. *arXiv preprint arXiv:1506.03379*, 2015.

Fan Chen, Constantinos Daskalakis, Noah Golowich, and Alexander Rakhlin. Near-optimal learning and planning in separated latent mdps. In *Proceedings of the Thirty-Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 995–1067. PMLR, 2024.

Yifang Chen, Kevin Jamieson, and Simon Du. Active multi-task representation learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 3271–3298. PMLR, 2022.

Yuan Cheng, Songtao Feng, Jing Yang, Hong Zhang, and Yingbin Liang. Provable benefit of multitask representation learning in reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 31741–31754, 2022.

Herman Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3): 755–770, 1959. doi: 10.1214/aoms/1177706205.

Alexandru Cioba, Michael Bromberg, Qian Wang, Ritwik Niyogi, Georgios Batzolis, Jezabel Garcia, Da-shan Shiu, and Alberto Bernacchia. How to distribute data across tasks for meta-learning? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6394–6401, 2022. doi: 10.1609/aaai.v36i6.20590.

Brian C. Dean, Michel X. Goemans, and Jan Vondrák. Approximating the stochastic knapsack problem: The benefit of adaptivity. *Mathematics of Operations Research*, 33(4):945–964, 2008. doi: 10.1287/moor.1080.0330.

Rémy Degenne and Wouter M. Koolen. Pure exploration with multiple correct answers. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Persi Diaconis. *Group Representations in Probability and Statistics*, volume 11 of *Institute of Mathematical Statistics Lecture Notes–Monograph Series*. Institute of Mathematical Statistics, Hayward, CA, 1988. ISBN 9780940600140. doi: 10.1214/lnms/1215467407.

John D. Dixon and Brian Mortimer. *Permutation Groups*, volume 163 of *Graduate Texts in Mathematics*. Springer, New York, 1996. ISBN 9780387945996. doi: 10.1007/978-1-4612-0731-3.

Morris L. Eaton. *Group Invariance Applications in Statistics*, volume 1 of *Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics and American Statistical Association, Hayward, CA and Alexandria, VA, 1989. ISBN 9780940600157.

Yonathan Efroni, Chi Jin, Akshay Krishnamurthy, and Sobhan Miryoosefi. Provable reinforcement learning with a short-term memory. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5832–5850. PMLR, 2022.

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

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 998–1027. PMLR, 2016.

Charles J. Geyer. On the asymptotics of constrained m-estimation. *The Annals of Statistics*, 22(4): 1993–2010, 1994. doi: 10.1214/aos/1176325768.

Dibya Ghosh, Jad Rahme, Aviral Kumar, Amy Zhang, Ryan P. Adams, and Sergey Levine. Why generalization in reinforcement learning is difficult: Epistemic pomdps and implicit partial observability. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Anupam Gupta, Viswanath Nagarajan, and Sahil Singla. Algorithms and adaptivity gaps for stochastic probing. In *Proceedings of the Twenty-Seventh Annual ACM–SIAM Symposium on Discrete Algorithms*, pp. 1731–1747. SIAM, 2016. doi: 10.1137/1.9781611974331.ch120.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 (2):217–288, 2011. doi: 10.1137/090771806.

Onésimo Hernández-Lerma and Jean B. Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, volume 30 of *Applications of Mathematics*. Springer, New York, 1996. doi: 10.1007/978-1-4612-0729-0.

Nicholas J. Higham. Computing the polar decomposition—with applications. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1160–1174, 1986. doi: 10.1137/0907079.

Jiachen Hu, Xiaoyu Chen, Chi Jin, Lihong Li, and Liwei Wang. Near-optimal representation learning for linear bandits and linear rl. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4349–4358. PMLR, 2021.

Mingjie Hu, Jian-Qiang Hu, and Enlu Zhou. Optimal data acquisition for reinforcement learning: A large deviations perspective. *arXiv preprint arXiv:2605.28675*, 2026.

Haque Ishfaq, Thanh Nguyen-Tang, Songtao Feng, Raman Arora, Mengdi Wang, Ming Yin, and Doina Precup. Offline multitask representation learning for reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Olav Kallenberg. *Foundations of Modern Probability*, volume 99 of *Probability Theory and Stochastic Modelling*. Springer, Cham, 3 edition, 2021. ISBN 9783030618711. doi: 10.1007/978-3-030-61871-1.

Emilie Kaufmann and Wouter M. Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22(246):1–44, 2021.

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):1–42, 2016.

Harold W. Kuhn. Extensive games and the problem of information. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games II*, volume 28 of *Annals of Mathematics Studies*, pp. 193–216. Princeton University Press, Princeton, NJ, 1953.

Jeongyeol Kwon, Shie Mannor, Constantine Caramanis, and Yonathan Efroni. Rl in latent mdps is tractable: Online guarantees via off-policy evaluation. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

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

Lucien Le Cam and Grace Lo Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer Series in Statistics. Springer, New York, 2 edition, 2000. ISBN 9780387950365. doi: 10.1007/978-1-4612-1166-2.

Zihao Li and Tianhao Liu. Ranking experiments under sequential sampling. *arXiv preprint arXiv:2608.19897*, 2026.

Rui Lu, Andrew Zhao, Simon S. Du, and Gao Huang. Provable general function class representation learning in multitask bandits and mdp. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Mohammad Naghshvar and Tara Javidi. Active sequential hypothesis testing. *The Annals of Statistics*, 41(6):2703–2738, 2013. doi: 10.1214/13-AOS1144.

Thomas Pumir, Amit Singer, and Nicolas Boumal. The generalized orthogonal procrustes problem in the high noise regime. *Information and Inference: A Journal of the IMA*, 2021. doi: 10.1093/imaiai/iaaa035.

Patrik Reizinger, Bálint Mucsányi, Siyuan Guo, Benjamin Eysenbach, Bernhard Schölkopf, and Wieland Brendel. Skill learning via policy diversity yields identifiable representations for reinforcement learning. In *International Conference on Learning Representations*, 2026.

Zohar Rimon, Aviv Tamar, and Gilad Adler. Meta reinforcement learning with finite training tasks: A density estimation approach. In *Advances in Neural Information Processing Systems*, volume 35, pp. 13640–13653, 2022.

Peter H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966. doi: 10.1007/BF02289451.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009. ISBN 9780387790510. doi: 10.1007/978-0-387-79052-7.

A. W. van der Vaart. *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998. ISBN 9780521784504. doi: 10.1017/CBO9780511802256.

Robert A. Vandermeulen and Clayton D. Scott. An operator theoretic approach to nonparametric mixture models. *The Annals of Statistics*, 47(5):2704–2733, 2019. doi: 10.1214/18-AOS1762.

George Vershinin, Asaf Cohen, and Omer Gurewitz. On cost-aware designs for sequential hypothesis testing. *arXiv preprint arXiv:2512.19067*, 2025. Version 3, revised 24 September 2026.

Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16 (2):117–186, 1945. doi: 10.1214/aoms/1177731118.

Abraham Wald and Jacob Wolfowitz. Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, 19(3):326–339, 1948. doi: 10.1214/aoms/1177730197.

Yifei Wang, Tavor Baharav, Yanjun Han, Jiantao Jiao, and David Tse. Beyond the best: Estimating distribution functionals in infinite-armed bandits. *arXiv preprint arXiv:2211.01743*, 2022.

Yiping Wang, Yifang Chen, Kevin Jamieson, and Simon Shaolei Du. Improved active multi-task representation learning via lasso. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35548–35578. PMLR, 2023.

Ziping Xu, Zifan Xu, Runxuan Jiang, Peter Stone, and Ambuj Tewari. Sample efficient myopic exploration through multitask reinforcement learning with diverse tasks. In *International Conference on Learning Representations*, 2024.

Haotian Ye, Xiaoyu Chen, Liwei Wang, and Simon Shaolei Du. On the power of pre-training for generalization in reinforcement learning: Provable benefits and hardness. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 39770–39800. PMLR, 2023.

Bin Yu. Assouad, fano, and le cam. In David Pollard, Erik Torgersen, and Grace L. Yang (eds.), *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pp. 423–435. Springer, New York, 1997. doi: 10.1007/978-1-4612-1880-7_29.

Xiangteng Zhang, Yang Guan, Bo Zhang, Hongyang Li, Ya-Qin Zhang, and Shengbo Eben Li. On the identifiability of controlled world models. *arXiv preprint arXiv:2607.22430*, 2026.

FLAGSHIP THEOREM PROOF MAP

The complete bounded reset-or-continue result is assembled in the order below. The map separates proof obligations from executable checks: the latter verify finite algebra, enumeration, and numerical certificates, but they do not replace measure-theoretic or asymptotic arguments.

Result	Role in the flagship theorem
Theorem D.2 and proposition D.4	Every randomized history-dependent depth-L learner is outcome-equivalent to predictable mixtures over the complete deterministic tree class.
Assumption D.5 and theorem D.6	Observational quotient, disjoint single-answer cells, learner-class lower bound, and equality between the learner domain and the complete tree game.
Lemmas E.2 and E.3	Prefix support, pathwise stopped-likelihood factorization, and exact conditional KL and cost decompositions.
Theorems E.4 and E.5	Bellman best response, fractional root, minimax duality, and sparse primal recovery.
Theorem E.6	Finite certified computation and conversion from cost fractions to actual tree-launch probabilities.
Lemmas E.10 and E.11 and proposition E.12	Finite expected identification prefix, exact allocation/cost tracking, and the eventwise uniform stopping tail.
Theorem E.13	Exact-design quotient Track-and-Stop and the sharp expected-cost constant without an additive transient debt.
Lemmas E.14 to E.16 and theorem E.17	Deterministic optimization debt, random predictable cost crossing, explicit post-crossing overshoot, and certified approximate-design optimality.
Theorems F.1 and F.2	Formal separation from episodewise acquisition and from the specified zero-terminal-value fixed-depth receding-horizon information rule.
Lemma G.1 and proposition G.2	Exhaustion of informative depth-three calibration trees and the closed-form Bellman stopping rule.

The scripts in `verification/` exhaustively check small tree games, quotient invariance, answer-cell disjointness, Bellman versus direct tree enumeration, primal–dual equality, column-generation certificates, parity identities, routing calculations, and structured Gaussian formulas. They do not certify optional-stopping hypotheses, asymptotic uniformity, or external geometric results; those remain proof and independent human-audit obligations.

A NOTATION, PROBABILITY SPACES, AND ACCESS MODELS

This appendix gives a complete measure-theoretic specification of the persistent meta-RL experiment used throughout the paper. The construction follows the standard stochastic-kernel formulation of finite-horizon control on Borel spaces (Bertsekas & Shreve, 1996; Hernández-Lerma & Lasserre, 1996); existence and uniqueness of the induced path measures follow from the Ionescu–Tulcea extension theorem (Kallenberg, 2021). The global parameter is frequentist and fixed throughout an experiment. Randomness arises from task sampling, controlled state transitions, rewards, and any randomization used by the learner.

A.1 MEASURABLE-SPACE CONVENTIONS

A measurable space $(\mathsf{X}, \mathfrak{X})$ is called *standard Borel* if it is measurably isomorphic to a Borel subset of a Polish space. Unless a sigma-field is displayed explicitly, every standard Borel space below is endowed with its Borel sigma-field. For a measurable space X, let $\mathcal{P}(\mathsf{X})$ denote the set of probability measures on X, equipped with the evaluation sigma-field generated by the maps $\nu \mapsto \nu(B)$ for measurable $B \subseteq \mathsf{X}$.

Definition A.1 (Stochastic kernel)**.** Let X and Y be measurable spaces. A stochastic kernel $K : \mathsf{X} \rightsquigarrow \mathsf{Y}$ is a map $(x, B) \mapsto K(B \mid x)$ such that $K(\cdot \mid x)$ is a probability measure on Y for every $x \in \mathsf{X}$, and $x \mapsto K(B \mid x)$ is measurable for every measurable $B \subseteq \mathsf{Y}$. We write

$$Y \mid X = x \sim K(\cdot \mid x)$$

when the conditional law of Y given $X = x$ is $K(\cdot \mid x)$.

For a finite sequence $x_{1:m}$, we use the shorthand $x_{1:0} = \varnothing$. Empty products are one, empty sums are zero, and $\inf \varnothing = +\infty$. If a supremum is taken over an empty set of nonnegative quantities, we set it equal to zero. All equalities between probability measures mean equality on the relevant measurable space.

A.2 Shared task law and source controlled interface

Fix a finite horizon $H \in \mathbb{N}$. Let

$$\Theta, \qquad \mathcal{Z}, \qquad \mathcal{S}, \qquad \mathcal{A}$$

be standard Borel spaces representing, respectively, the global meta-parameter, task-specific latent, physical state, and action. For the general finite-model theory, rewards take values in $\mathcal{R} = [0, 1]$. The restriction to $[0, 1]$ is a normalization; every bounded reward model can be reduced to this case by an affine rescaling. The fresh-target linear-quadratic specialization in Section O is a separate square-integrable cost model and does not invoke this bounded-reward assumption.

The shared fresh-task law and the source controlled interface are specified by the following jointly measurable stochastic kernels:

$$\mu : \Theta \rightsquigarrow \mathcal{Z}, \tag{A.1}$$

$$\rho : \Theta \times \mathcal{Z} \rightsquigarrow \mathcal{S}, \tag{A.2}$$

$$K_h : \Theta \times \mathcal{Z} \times \mathcal{S} \times \mathcal{A} \rightsquigarrow \mathcal{R} \times \mathcal{S}, \qquad h \in [H]. \tag{A.3}$$

Here, $\mu_\theta \coloneqq \mu(\cdot \mid \theta)$ is the distribution of a fresh task latent, $\rho_{\theta,z} \coloneqq \rho(\cdot \mid \theta, z)$ is the initial state law for a physical episode, and $K_h(\mathrm{d}r, \mathrm{d}s' \mid \theta, z, s, a)$ is the joint reward–next-state kernel at step h. Using a joint kernel avoids imposing an artificial conditional-independence assumption between reward and next state.

For $h \in [H]$, define the within-episode history space

$$\mathsf{H}_h \coloneqq \mathcal{S} \times (\mathcal{A} \times \mathcal{R} \times \mathcal{S})^{h-1}, \qquad \mathsf{H}_1 = \mathcal{S}, \tag{A.4}$$

and the complete trajectory space

$$\mathsf{T}_H \coloneqq \mathcal{S} \times (\mathcal{A} \times \mathcal{R} \times \mathcal{S})^{H}. \tag{A.5}$$

A generic trajectory is written

$$\tau = (s_1, a_1, r_1, s_2, \ldots, a_H, r_H, s_{H+1}) \in \mathsf{T}_H,$$

and its return is

$$G(\tau) \coloneqq \sum_{h=1}^{H} r_h \in [0, H]. \tag{A.6}$$

Definition A.2 (Episodic policy and trajectory law)**.** An episodic history-dependent randomized policy is a tuple $\pi = (\pi_h)_{h=1}^{H}$ of stochastic kernels

$$\pi_h : \mathsf{H}_h \rightsquigarrow \mathcal{A}.$$

For $(\theta, z) \in \Theta \times \mathcal{Z}$, the induced trajectory law $P^{\pi}_{\theta,z} \in \mathcal{P}(\mathsf{T}_H)$ is the unique measure satisfying the factorization

$$\begin{aligned} &P^{\pi}_{\theta,z}(\mathrm{d}s_1, \mathrm{d}a_1, \mathrm{d}r_1, \ldots, \mathrm{d}a_H, \mathrm{d}r_H, \mathrm{d}s_{H+1}) \\ &\quad = \rho(\mathrm{d}s_1 \mid \theta, z) \prod_{h=1}^{H} \pi_h(\mathrm{d}a_h \mid h_h) K_h(\mathrm{d}r_h, \mathrm{d}s_{h+1} \mid \theta, z, s_h, a_h), \end{aligned} \tag{A.7}$$

where $h_h \in \mathsf{H}_h$ denotes the realized history available before selecting a_h.

The Ionescu–Tulcea theorem guarantees that equation A.7 defines a unique probability measure. Let Π_H denote the class of all policies in Definition A.2. Define

$$V^{\pi}_{\theta,z} \coloneqq \int_{\mathsf{T}_H} G(\tau)\, P^{\pi}_{\theta,z}(\mathrm{d}\tau), \tag{A.8}$$

$$V^{\star}_{\theta,z} \coloneqq \sup_{\pi \in \Pi_H} V^{\pi}_{\theta,z}. \tag{A.9}$$

No existence of an optimal policy is assumed in the general model; the supremum in equation A.9 is always well-defined because returns are bounded.

A.3 Observed source histories and admissible acquisition rules

Let

$$\mathcal{C} \coloneqq \{\mathsf{reset}, \mathsf{continue}\}$$

with the discrete sigma-field. The learner observes its persistence actions and therefore knows exactly which source trajectories belong to the same task block. The task latent itself is not observed unless it is revealed through the controlled trajectory.

For $t \in \mathbb{N}_0$, define the observed and latent-augmented source-history spaces

$$\mathsf{O}_t \coloneqq (\mathcal{C} \times \mathsf{T}_H)^t, \qquad \mathsf{O}_0 = \{\varnothing\}, \tag{A.10}$$

$$\overline{\mathsf{O}}_t \coloneqq (\mathcal{C} \times \mathcal{Z} \times \mathsf{T}_H)^t, \qquad \overline{\mathsf{O}}_0 = \{\varnothing\}. \tag{A.11}$$

The corresponding random histories are

$$O_t \coloneqq (C_1, \boldsymbol{\tau}_1, \ldots, C_t, \boldsymbol{\tau}_t),$$
$$\overline{O}_t \coloneqq (C_1, Z_1, \boldsymbol{\tau}_1, \ldots, C_t, Z_t, \boldsymbol{\tau}_t).$$

The learner's observed filtration and the proof-only augmented filtration are

$$\mathcal{F}_t^{\mathrm{src}} \coloneqq \sigma(O_t), \qquad \overline{\mathcal{F}}_t^{\mathrm{src}} \coloneqq \sigma(\overline{O}_t). \tag{A.12}$$

We take both filtrations to be completed under the relevant source law only when invoking a theorem that requires completion; all algorithmic measurability statements refer to the raw canonical filtration.

Definition A.3 (Admissible source acquisition rule)**.** An admissible source acquisition rule is a collection

$$\mathfrak{a} = \big((\gamma_t)_{t\geq 1}, (\pi_{t,h})_{t\geq 1,\, h\in[H]}\big)$$

consisting of:

1. persistence kernels
$$\gamma_t : \mathsf{O}_{t-1} \rightsquigarrow \mathcal{C}, \qquad t \geq 1, \tag{A.13}$$
with $\gamma_1(\{\mathsf{reset}\} \mid \varnothing) = 1$; and
2. within-episode action kernels
$$\pi_{t,h} : \mathsf{O}_{t-1} \times \mathcal{C} \times \mathsf{H}_h \rightsquigarrow \mathcal{A}, \qquad t \geq 1,\ h \in [H]. \tag{A.14}$$

The kernels may be randomized. They may depend on the entire observed source history and on the current physical-episode history, but not on the latent Z_t except through observations contained in those histories.

Writing the learner directly as stochastic kernels subsumes auxiliary random seeds and avoids introducing a separate measurable space of policies. In particular, conditional on $O_{t-1} = o_{t-1}$ and $C_t = c_t$, the action law at step h is

$$\pi_{t,h}(\cdot \mid o_{t-1}, c_t, H_{t,h}),$$

where $H_{t,h} \in \mathsf{H}_h$ is the current within-episode history.

A.4 Persistence kernel and existence of the source process

For $t \geq 2$, define the task-persistence kernel

$$M_\theta(\mathrm{d}z' \mid z, c) \coloneqq \mathbf{1}\{c = \mathsf{reset}\}\, \mu(\mathrm{d}z' \mid \theta) + \mathbf{1}\{c = \mathsf{continue}\}\, \delta_z(\mathrm{d}z'). \tag{A.15}$$

At the first source episode, $C_1 = \mathsf{reset}$ almost surely and $Z_1 \sim \mu_\theta$.

For a fixed acquisition rule $\mathfrak{a}$, define the conditional episode kernel

$$\mathbf{P}^{\mathfrak{a}}_{\theta,z,t}(\mathrm{d}\tau \mid o_{t-1}, c) \in \mathcal{P}(\mathsf{T}_H) \tag{A.16}$$

by replacing π_h in equation A.7 with $\pi_{t,h}(\cdot \mid o_{t-1}, c, \cdot)$. Joint measurability of equation A.16 follows by repeated composition of the measurable kernels in equation A.2, equation A.3, and equation A.14.

Proposition A.4 (Well-posedness of the persistent source experiment)**.** *Fix $\theta \in \Theta$ and an admissible acquisition rule $\mathfrak{a}$. There exists a unique probability measure*

$$\mathbb{P}_\theta^{\mathfrak{a}} \in \mathcal{P}\left((\mathcal{C} \times \mathcal{Z} \times \mathsf{T}_H)^{\mathbb{N}}\right) \tag{A.17}$$

under which the canonical coordinates satisfy, almost surely,

$$\mathbb{P}_\theta^{\mathfrak{a}}(C_t \in \cdot \mid \overline{\mathcal{F}}_{t-1}^{\mathrm{src}}) = \gamma_t(\cdot \mid O_{t-1}), \tag{A.18}$$

$$\mathbb{P}_\theta^{\mathfrak{a}}(Z_1 \in \cdot \mid C_1) = \mu_\theta, \tag{A.19}$$

$$\mathbb{P}_\theta^{\mathfrak{a}}(Z_t \in \cdot \mid \overline{\mathcal{F}}_{t-1}^{\mathrm{src}}, C_t) = M_\theta(\cdot \mid Z_{t-1}, C_t), \qquad t \geq 2, \tag{A.20}$$

$$\mathbb{P}_\theta^{\mathfrak{a}}(\boldsymbol{\tau}_t \in \cdot \mid \overline{\mathcal{F}}_{t-1}^{\mathrm{src}}, C_t, Z_t) = \mathbf{P}_{\theta, Z_t, t}^{\mathfrak{a}}(\cdot \mid O_{t-1}, C_t). \tag{A.21}$$

Its marginal on $(\mathcal{C} \times \mathsf{T}_H)^{\mathbb{N}}$ is the observed source law.

Proof. The product space in equation A.17 is a countable product of standard Borel spaces. At episode 1, compose the kernels γ_1, μ_θ, and $\mathbf{P}_{\theta,z,1}^{\mathfrak{a}}$. For every $t \geq 2$, compose, in this order, the predictable persistence kernel $\gamma_t(\cdot \mid O_{t-1})$, the latent-update kernel $M_\theta(\cdot \mid Z_{t-1}, C_t)$, and the conditional episode kernel $\mathbf{P}_{\theta,Z_t,t}^{\mathfrak{a}}$. Each composition is a measurable stochastic kernel from the preceding augmented history to $\mathcal{C} \times \mathcal{Z} \times \mathsf{T}_H$. The Ionescu–Tulcea extension theorem therefore yields a unique measure with the asserted finite-dimensional factorization. The conditional identities equations (A.18)–(A.21) follow from that factorization, and the observed law is its coordinate marginal. □

Remark A.5 (Physical reset versus task reset). Every source episode begins from a fresh physical initial state drawn from ρ_{θ,Z_t}. The action $C_t = \mathsf{continue}$ preserves only the task latent Z_t; it does not concatenate physical trajectories across episode boundaries. This distinction is essential: continuation creates repeated controlled views of one task, not one longer physical episode.

A.5 Stopping, reset blocks, and source cost

A source stopping time is an $\mathbb{N}_0 \cup \{+\infty\}$-valued stopping time T with respect to $(\mathcal{F}_t^{\mathrm{src}})_{t \geq 0}$. Define the number of task blocks initiated through episode t by

$$N_{\mathsf{reset}}(t) \coloneqq \sum_{s=1}^{t} \mathbf{1}\{C_s = \mathsf{reset}\}, \qquad N_{\mathsf{reset}}(0) = 0, \tag{A.22}$$

and, for reset cost $\kappa \geq 0$, define

$$\mathsf{C}_t \coloneqq t + \kappa N_{\mathsf{reset}}(t), \qquad \mathsf{C}_{\mathsf{T}} \coloneqq \lim_{t \to \infty} \mathsf{C}_{\mathsf{T} \wedge t} \in [0, +\infty]. \tag{A.23}$$

Thus the compulsory first reset is charged. This convention makes the cost of every completed task block exactly additive. Declaring the first reset free subtracts the deterministic constant κ whenever $\mathsf{T} \geq 1$ and does not change any asymptotic characteristic time.

Define the reset times recursively by

$$\sigma_1 \coloneqq 1, \tag{A.24}$$

$$\sigma_{k+1} \coloneqq \inf\{t > \sigma_k : C_t = \mathsf{reset}\}, \tag{A.25}$$

with $\inf \varnothing = +\infty$. The kth task block is the random episode interval

$$\mathcal{I}_k \coloneqq \{\sigma_k, \ldots, \sigma_{k+1} - 1\}. \tag{A.26}$$

By equation A.15, the latent is constant on each $\mathcal{I}_k$, while the latents attached to distinct reset times are conditionally independent draws from μ_θ given θ.

The stopped observed history takes values in the countable disjoint union

$$\mathsf{O}_{\mathrm{stop}} \coloneqq \bigsqcup_{t \in \mathbb{N}_0} \left(\{t\} \times \mathsf{O}_t\right), \tag{A.27}$$

which is standard Borel. This provides a measurable domain for the learner's final recommendation.

A.6 FRESH-TARGET ADAPTATION AND DEPLOYMENT INTERFACES

The source, target-adaptation, and target-deployment interfaces may differ. They share the global parameter θ, task-latent space $\mathcal{Z}$, and fresh-task kernel μ_θ, but may have different horizons, state spaces, action spaces, observation or reward spaces, and controlled kernels. This includes ordinary meta-RL, where all interfaces coincide, and pretraining or transfer settings in which source observations, target probes, and deployment controls are different.

Fix finite horizons $H_{\mathrm{ad}}, H_{\mathrm{dep}} \in \mathbb{N}$ and standard Borel spaces

$$(\mathcal{S}_{\mathrm{ad}}, \mathcal{A}_{\mathrm{ad}}, \mathcal{R}_{\mathrm{ad}}), \qquad (\mathcal{S}_{\mathrm{dep}}, \mathcal{A}_{\mathrm{dep}}, \mathcal{R}_{\mathrm{dep}}).$$

The adaptation and deployment interfaces are specified by jointly measurable kernels

$$\rho^{\mathrm{ad}} : \Theta \times \mathcal{Z} \rightsquigarrow \mathcal{S}_{\mathrm{ad}}, \qquad K_h^{\mathrm{ad}} : \Theta \times \mathcal{Z} \times \mathcal{S}_{\mathrm{ad}} \times \mathcal{A}_{\mathrm{ad}} \rightsquigarrow \mathcal{R}_{\mathrm{ad}} \times \mathcal{S}_{\mathrm{ad}}, \qquad h \in [H_{\mathrm{ad}}], \tag{A.28}$$

$$\rho^{\mathrm{dep}} : \Theta \times \mathcal{Z} \rightsquigarrow \mathcal{S}_{\mathrm{dep}}, \quad K_h^{\mathrm{dep}} : \Theta \times \mathcal{Z} \times \mathcal{S}_{\mathrm{dep}} \times \mathcal{A}_{\mathrm{dep}} \rightsquigarrow \mathcal{R}_{\mathrm{dep}} \times \mathcal{S}_{\mathrm{dep}}, \quad h \in [H_{\mathrm{dep}}]. \tag{A.29}$$

For $q \in \{\mathrm{ad}, \mathrm{dep}\}$, define the corresponding history and trajectory spaces

$$\mathsf{H}_h^q := \mathcal{S}_q \times (\mathcal{A}_q \times \mathcal{R}_q \times \mathcal{S}_q)^{h-1}, \tag{A.30}$$

$$\mathsf{T}_q := \mathcal{S}_q \times (\mathcal{A}_q \times \mathcal{R}_q \times \mathcal{S}_q)^{H_q}. \tag{A.31}$$

Repeated kernel composition defines, for every episodic phase-q policy π, a trajectory law $P_{\theta,z}^{q,\pi}$ on T_q.

Let $G^{\mathrm{dep}} : \mathsf{T}_{\mathrm{dep}} \to \mathbb{R}$ be a Borel deployment-evaluation functional. In the bounded-reward theory we assume $0 \le G^{\mathrm{dep}} \le H_{\mathrm{dep}}$; the LQ specialization in Section O instead uses an integrable negative quadratic cost. Set

$$V_{\theta,z}^{\mathrm{dep},\pi} := \int G^{\mathrm{dep}}(\tau) P_{\theta,z}^{\mathrm{dep},\pi}(\mathrm{d}\tau), \qquad V_{\theta,z}^{\mathrm{tgt},\star} := \sup_\pi V_{\theta,z}^{\mathrm{dep},\pi}. \tag{A.32}$$

All target values below are assumed finite.

A fresh target consists of one latent $Z_\star \sim \mu_\theta$ conditionally independent of the complete source process given θ. The same $Z_\star$ persists through B adaptation episodes and one deployment episode; each physical episode is reinitialized according to its phase-specific initial-state kernel.

To avoid irrelevant function-space pathologies, fix a standard Borel strategy-code space $\mathcal{G}_B$. Every $g \in \mathcal{G}_B$ determines jointly measurable kernels

$$\psi_{b,h}^g : (\mathsf{T}_{\mathrm{ad}})^{b-1} \times \mathsf{H}_h^{\mathrm{ad}} \rightsquigarrow \mathcal{A}_{\mathrm{ad}}, \qquad b \in [B],\ h \in [H_{\mathrm{ad}}], \tag{A.33}$$

$$\psi_{\mathrm{dep},h}^g : (\mathsf{T}_{\mathrm{ad}})^B \times \mathsf{H}_h^{\mathrm{dep}} \rightsquigarrow \mathcal{A}_{\mathrm{dep}}, \qquad h \in [H_{\mathrm{dep}}]. \tag{A.34}$$

Joint measurability means, for example, that $(g, t_{1:b-1}, h) \mapsto \psi_{b,h}^g(A \mid t_{1:b-1}, h)$ is measurable for every measurable adaptation-action set A. Finite strategy classes and measurable finite-dimensional parameterizations are special cases.

For fixed (θ, z, g), repeated application of the phase-specific kernels and the strategy kernels defines a unique law

$$\mathbb{P}_{\theta,z}^{\mathrm{tgt},g,B} \in \mathcal{P}((\mathsf{T}_{\mathrm{ad}})^B \times \mathsf{T}_{\mathrm{dep}}) \tag{A.35}$$

for the adaptation and deployment trajectories. Let $\boldsymbol{\tau}_{\mathrm{dep}}$ denote the final coordinate and define

$$V_{\theta,z}^{\mathrm{tgt},g,B} := \mathbb{E}_{\theta,z}^{\mathrm{tgt},g,B}\left[G^{\mathrm{dep}}(\boldsymbol{\tau}_{\mathrm{dep}})\right]. \tag{A.36}$$

Adaptation rewards or observations do not enter the deployment criterion directly, but may carry information that changes the deployment policy.

Definition A.6 (Target loss and optimal B-shot loss)**.** For $g \in \mathcal{G}_B$, define

$$\mathcal{L}_B(\theta, g) := \int_{\mathcal{Z}} \left(V_{\theta,z}^{\mathrm{tgt},\star} - V_{\theta,z}^{\mathrm{tgt},g,B}\right) \mu(\mathrm{d}z \mid \theta), \tag{A.37}$$

$$\mathcal{L}_B^\star(\theta) := \inf_{g \in \mathcal{G}_B} \mathcal{L}_B(\theta, g). \tag{A.38}$$

The quantity $\mathcal{L}_B^\star(\theta)$ captures the irreducible loss of target adaptation even when the global model θ is known. In the bounded deployment-reward theory,

$$0 \le \mathcal{L}_B^\star(\theta) \le \mathcal{L}_B(\theta, g) \le H_{\mathrm{dep}}. \tag{A.39}$$

We impose the standing measurability condition that $(\theta, z) \mapsto V_{\theta,z}^{\mathrm{tgt},\star}$ is Borel measurable and that $\mathcal{G}_B$ is nonempty. This condition is automatic in every finite model used for the lower bounds and is verified directly for the structured continuous models analyzed later. Together with joint measurability of the strategy kernels, iterated kernel integration implies that $(\theta, g) \mapsto \mathcal{L}_B(\theta, g)$ is measurable. No attainment of the infimum in equation A.38 is assumed.

For $\varepsilon > 0$, define the nonempty set of ε-optimal target strategies

$$\mathcal{G}_{B,\varepsilon}(\theta) \coloneqq \{g \in \mathcal{G}_B : \mathcal{L}_B(\theta, g) \le \mathcal{L}_B^\star(\theta) + \varepsilon\}. \tag{A.40}$$

Nonemptiness follows directly from the definition of the infimum.

Definition A.7 (Pairwise target-incompatible alternatives)**.** For $\varepsilon > 0$, define

$$\mathrm{Alt}_{B,\varepsilon}(\theta) \coloneqq \{\theta' \in \Theta : \mathcal{G}_{B,\varepsilon}(\theta) \cap \mathcal{G}_{B,\varepsilon}(\theta') = \varnothing\}. \tag{A.41}$$

Elements of $\mathrm{Alt}_{B,\varepsilon}(\theta)$ are called *pairwise target-incompatible alternatives*. They are global models for which no single target strategy is ε-optimal under both models. The lower bounds in the paper use this pairwise set. A complete upper characterization with overlapping valid answers requires additional answer selection structure; Sections I and Q make this restriction explicit.

A.7 Complete source learners and statistical objectives

Definition A.8 (Source learner)**.** A source learner for target budget B is a tuple

$$\mathcal{A} = (\mathfrak{a}, \mathsf{T}, \Lambda),$$

where $\mathfrak{a}$ is an admissible acquisition rule, T is a stopping time with respect to $(\mathcal{F}_t^{\mathrm{src}})_{t\ge 0}$ satisfying $\mathbb{P}_\theta^{\mathfrak{a}}(\mathsf{T} < \infty) = 1$ for every $\theta \in \Theta$, and

$$\Lambda : \mathsf{O}_{\mathrm{stop}} \rightsquigarrow \mathcal{G}_B \tag{A.42}$$

is a recommendation kernel. Conditional on the stopped observed history, the output $\widehat{g}$ is sampled from Λ. Deterministic recommendations correspond to Dirac kernels.

The source law in Proposition A.4, the stopping rule, and equation A.42 induce a unique joint law of $(O_{\mathsf{T}}, \widehat{g})$. We write $\mathbb{P}_\theta^{\mathcal{A}}$ and $\mathbb{E}_\theta^{\mathcal{A}}$ for probability and expectation under this law. The target loss itself is already averaged over a fresh target task and fresh target trajectory randomness, so source-level expectations below are over source interaction and recommendation randomization.

For a deterministic source budget $T \ge 0$, define

$$\mathfrak{A}_\kappa(T) \coloneqq \left\{\mathcal{A} : \mathbb{P}_\theta^{\mathcal{A}}(\mathsf{C}_{\mathsf{T}} \le T) = 1 \text{ for every } \theta \in \Theta\right\}. \tag{A.43}$$

The fixed-budget minimax meta-excess risk is

$$\mathfrak{R}_{T,B,\kappa}^\star \coloneqq \inf_{\mathcal{A}\in\mathfrak{A}_\kappa(T)} \sup_{\theta\in\Theta} \mathbb{E}_\theta^{\mathcal{A}}\left[\mathcal{L}_B(\theta, \widehat{g}) - \mathcal{L}_B^\star(\theta)\right]. \tag{A.44}$$

The subtraction of $\mathcal{L}_B^\star(\theta)$ isolates the price of not knowing the global model from irreducible uncertainty about the fresh target latent.

Definition A.9 (Fixed-confidence correctness)**.** For $\varepsilon > 0$ and $\delta \in (0, 1)$, a source learner is (B, ε, δ)-correct on Θ if

$$\mathbb{P}_\theta^{\mathcal{A}}\left(\widehat{g} \in \mathcal{G}_{B,\varepsilon}(\theta)\right) \ge 1 - \delta \qquad \text{for every } \theta \in \Theta. \tag{A.45}$$

Let $\mathfrak{A}_{B,\varepsilon,\delta}$ denote the class of all such learners.

The instance-dependent optimal expected source cost is

$$\mathfrak{C}_{B,\varepsilon,\delta,\kappa}^\star(\theta) \coloneqq \inf_{\mathcal{A}\in\mathfrak{A}_{B,\varepsilon,\delta}} \mathbb{E}_\theta^{\mathcal{A}}\left[\mathsf{C}_{\mathsf{T}}\right], \tag{A.46}$$

with the convention that the infimum is $+\infty$ when no correct learner has finite expected cost. Importantly, the infimum is over learners correct for all $\theta' \in \Theta$, while the expectation in equation A.46 is evaluated at the particular instance θ.

A.8 DETERMINISTIC-LENGTH TASK-BLOCK EXPERIMENTS

The general source model above permits episode-by-episode reset decisions. The characteristic-time results in Section 4 initially use the following smaller, precisely controlled experiment class.

Definition A.10 (Task-block experiment)**.** A deterministic-length task-block experiment e consists of a length $\ell(e) \in \mathbb{N}$ and internal action kernels

$$\pi^e_{j,h} : \mathsf{T}_H^{j-1} \times \mathsf{H}_h \rightsquigarrow \mathcal{A}, \qquad j \in [\ell(e)],\ h \in [H]. \tag{A.47}$$

The experiment starts with a reset, holds the sampled task latent fixed for all $\ell(e)$ physical episodes, and permits episode j to depend on the previous within-block trajectories $\boldsymbol{\tau}^e_{1:j-1}$.

For fixed (θ, z, e), let $P^e_{\theta,z} \in \mathcal{P}(\mathsf{T}_H^{\ell(e)})$ be the unique conditional block-transcript law induced by equation A.47 and the controlled kernels equations (A.2)–(A.3). The marginalized block law is

$$Q^e_\theta(D) \coloneqq \int_{\mathcal{Z}} P^e_{\theta,z}(D)\, \mu(\mathrm{d}z \mid \theta), \qquad D \subseteq \mathsf{T}_H^{\ell(e)} \text{ measurable.} \tag{A.48}$$

The source cost of one such block is

$$c_\kappa(e) \coloneqq \ell(e) + \kappa. \tag{A.49}$$

A blockwise source algorithm selects an experiment E_k at the beginning of block k using only completed-block observations, resets the task latent, and executes the selected internal experiment without further dependence on pre-block history beyond the identity E_k. Let D_k be the resulting block transcript and let $\mathcal{B}_k$ be the sigma-field generated by $(E_1, D_1, \ldots, E_k, D_k)$.

Lemma A.11 (Fresh-block conditional law)**.** *For every blockwise source algorithm, every $k \geq 1$, every experiment e, and every measurable block-transcript set D,*

$$\mathbb{P}_\theta\left(D_k \in D \mid \mathcal{B}_{k-1}, E_k = e\right) = Q^e_\theta(D) \quad \textit{almost surely on } \{E_k = e\}. \tag{A.50}$$

Proof. At the beginning of block k, the reset action samples a latent $Z_k^{\mathrm{blk}} \sim \mu_\theta$ conditionally independently of $\mathcal{B}_{k-1}$ and of the experiment-selection randomization. Conditional on $(E_k = e, Z_k^{\mathrm{blk}} = z)$, the internal transcript law is $P^e_{\theta,z}$. Integrating over z gives equation A.48, which proves equation A.50. □

Remark A.12 (Hierarchy of characteristic-time theorems). Appendices H–I analyze a finite menu of deterministic blocks. Appendix T proves a reusable result for a supplied finite menu of bounded observation-adaptive stopped experiments. The flagship result in Appendices D–E then takes as its menu the complete finite class of all depth-L deterministic experiment trees. By Theorem D.2, this is outcome-equivalent to every randomized history-dependent within-task policy and stopping rule that terminates by depth L. Unbounded task duration and noncompact policy classes remain outside the theorem.

A.9 TARGET-RELEVANT PERSISTENCE ORDER

Let $\mathfrak{E}_m$ denote the collection of all task-block experiments with length at most m, and let $\mathfrak{E}_{<\infty} \coloneqq \bigcup_{m\geq 1} \mathfrak{E}_m$.

Definition A.13 (Pairwise persistence order)**.** For $\theta, \theta' \in \Theta$, define

$$r(\theta, \theta') \coloneqq \inf\left\{m \in \mathbb{N} : \exists e \in \mathfrak{E}_m \text{ such that } Q^e_\theta \neq Q^e_{\theta'}\right\}. \tag{A.51}$$

If no finite-length experiment separates the two models, set $r(\theta, \theta') = +\infty$.

Because two probability measures are equal if and only if their total variation distance is zero, equation A.51 may equivalently be defined by requiring positive total variation. Under common domination, it is also equivalent to requiring positive KL divergence in at least one direction.

For target budget B and tolerance $\varepsilon > 0$, define the instance-level target-relevant persistence order

$$r^\star_{B,\varepsilon}(\theta) \coloneqq \sup_{\theta' \in \mathrm{Alt}_{B,\varepsilon}(\theta)} r(\theta, \theta'). \tag{A.52}$$

If $\mathrm{Alt}_{B,\varepsilon}(\theta) = \varnothing$, then $r^\star_{B,\varepsilon}(\theta) = 0$ by convention. Thus persistence order is attached only to distinctions that can change the set of acceptable target decisions.

Table 2: Access-model conventions used throughout the paper.

Component	Convention
Global parameter	Fixed unknown θ shared by every source and target task.
Source task on reset	Fresh $Z \sim \mu_\theta$, conditionally independent of the past.
Source task on continue	The preceding task latent is preserved exactly.
Physical episode	State is reinitialized; task latent may persist.
Learner observations	Persistence actions and trajectories; not the latent itself.
Source policies	Arbitrary randomized history-dependent measurable kernels.
Target task	One fresh latent, shared by B adaptation episodes and deployment.
Target benchmark	Oracle target-interface value $V_{\theta,z}^{\mathrm{tgt},\star}$, with the irreducible B-shot loss subtracted in the meta-excess criterion.
General source cost	One unit per physical episode plus κ per initiated task block.
Characteristic-time scope	Finite observational quotient classes and the complete class of all randomized history-dependent experiments stopping by a fixed depth L.

A.10 Access-model summary and conventions

For clarity, the access model used by the principal results is summarized below.

Throughout the paper, statements indexed by θ are frequentist. A prior over Θ may be introduced inside a lower-bound proof, but it is then a proof device rather than part of the access model. Unless explicitly stated otherwise, all expectations involving a stopped learner are understood as extended nonnegative expectations; claims of finite expected source cost must be proved rather than inferred from almost-sure stopping.

B Technical Preliminaries

This appendix records the reusable probability, information-theoretic, martingale, Gaussian, matrix, group-invariance, and local-asymptotic tools invoked by the later proofs. We state each result in the precise form used in this paper. Model-specific likelihood factorizations, geometric reductions, and estimator analyses remain in their respective appendices. Standard references include Kallenberg (2021) for kernels and martingales, Tsybakov (2009); Yu (1997) for minimax reductions, Boucheron et al. (2013) for concentration inequalities, Halko et al. (2011); Higham (1986) for Gaussian pseudoinverses and polar factors, and van der Vaart (1998); Le Cam & Yang (2000) for local asymptotic theory.

Throughout, all logarithms are natural. For probability measures P, Q on a common measurable space, we use the extended-valued convention

$$D_{\mathrm{KL}}(P\|Q) = \begin{cases} \displaystyle\int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P, & P \ll Q, \\ +\infty, & \text{otherwise.} \end{cases}$$

For $p, q \in [0,1]$, define the binary relative entropy

$$\mathrm{kl}(p,q) \coloneqq p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}, \tag{B.1}$$

with the usual continuous-extension conventions at the boundary.

B.1 Relative entropy, testing, and minimax reductions

Lemma B.1 (Chain rule for relative entropy)**.** *Let P and Q be probability measures on the product of two standard Borel spaces $\mathsf{X} \times \mathsf{Y}$. Let P_X, Q_X be their first marginals, and let $P(\mathrm{d}y \mid x)$ and $Q(\mathrm{d}y \mid x)$ be regular conditional laws. Then, as an identity in $[0, +\infty]$,*

$$D_{\mathrm{KL}}(P\|Q) = D_{\mathrm{KL}}(P_X\|Q_X) + \int_{\mathsf{X}} D_{\mathrm{KL}}\left(P(\cdot \mid x)\|Q(\cdot \mid x)\right) P_X(\mathrm{d}x). \tag{B.2}$$

In particular, if the first-coordinate kernel is the same under the two models, then only the conditional second-coordinate divergence remains.

Proof. If $P \not\ll Q$, then either $P_X \not\ll Q_X$ or the conditional absolute continuity $P(\cdot \mid x) \ll Q(\cdot \mid x)$ fails on a set of positive P_X-measure; in either case the right-hand side is $+\infty$. Otherwise, the Radon–Nikodym derivative factorizes P-almost surely as

$$\frac{\mathrm{d}P}{\mathrm{d}Q}(x,y) = \frac{\mathrm{d}P_X}{\mathrm{d}Q_X}(x)\frac{\mathrm{d}P(\cdot \mid x)}{\mathrm{d}Q(\cdot \mid x)}(y).$$

Taking logarithms, integrating under P, and using conditional expectation gives equation B.2. The standard-Borel hypothesis guarantees the required disintegrations; see Kallenberg (2021). □

Corollary B.2 (Adaptive finite-horizon chain rule)**.** *Let $H_t = (A_1, Y_1, \ldots, A_t, Y_t)$ be generated sequentially. Conditional on H_{t-1}, the learner chooses A_t from a kernel $\alpha_t(\cdot \mid H_{t-1})$ that is identical under models θ and θ'. Conditional on (H_{t-1}, A_t), the observation laws are $P_{t,\theta}(\cdot \mid H_{t-1}, A_t)$ and $P_{t,\theta'}(\cdot \mid H_{t-1}, A_t)$. Then*

$$D_{\mathrm{KL}}\left(\mathbb{P}_\theta^{(m)} \middle\| \mathbb{P}_{\theta'}^{(m)}\right) = \mathbb{E}_\theta\left[\sum_{t=1}^{m} D_{\mathrm{KL}}\left(P_{t,\theta}(\cdot \mid H_{t-1}, A_t) \| P_{t,\theta'}(\cdot \mid H_{t-1}, A_t)\right)\right]. \tag{B.3}$$

Proof. Apply Lemma B.1 recursively to the action and observation at each stage. The conditional KL contribution of A_t is zero because the action-selection kernel α_t is shared by the two models. □

Lemma B.3 (Binary data processing)**.** *For probability measures P, Q and every measurable event A,*

$$D_{\mathrm{KL}}(P\|Q) \geq \mathrm{kl}(P(A), Q(A)). \tag{B.4}$$

Moreover, for $\delta \in (0, 1/2)$, if $P(A) \geq 1-\delta$ and $Q(A) \leq \delta$, then

$$D_{\mathrm{KL}}(P\|Q) \geq \mathrm{kl}(1-\delta, \delta). \tag{B.5}$$

Proof. Apply the data-processing inequality to the measurable map $x \mapsto \mathbf{1}\{x \in A\}$, yielding equation B.4. To obtain equation B.5, write $k(p,q) = \mathrm{kl}(p,q)$. On the region $p > q$,

$$\partial_p k(p,q) = \log\frac{p(1-q)}{q(1-p)} > 0, \qquad \partial_q k(p,q) = \frac{q-p}{q(1-q)} < 0.$$

Hence $k(p,q)$ is minimized over $p \geq 1-\delta$, $q \leq \delta$ at $(p,q) = (1-\delta, \delta)$. □

Lemma B.4 (Le Cam two-point reduction)**.** *Let (T, d) be a metric space, let $\vartheta_0, \vartheta_1 \in \mathsf{T}$, and let P_0, P_1 be the corresponding data laws. Set $\Delta = d(\vartheta_0, \vartheta_1)$. For every estimator $\widehat{\vartheta}$ taking values in T,*

$$\max_{i\in\{0,1\}} \mathbb{E}_i d^2(\widehat{\vartheta}, \vartheta_i) \geq \frac{\Delta^2}{8}\left(1 - \mathrm{TV}(P_0, P_1)\right). \tag{B.6}$$

Consequently, by Pinsker's inequality,

$$\max_i \mathbb{E}_i d^2(\widehat{\vartheta}, \vartheta_i) \geq \frac{\Delta^2}{8}\left(1 - \sqrt{\frac{1}{2}D_{\mathrm{KL}}(P_0\|P_1)}\right)_+. \tag{B.7}$$

Proof. Let $A = \{d(\widehat{\vartheta}, \vartheta_0) \leq d(\widehat{\vartheta}, \vartheta_1)\}$. By the triangle inequality, $d(\widehat{\vartheta}, \vartheta_0) \geq \Delta/2$ on A^c and $d(\widehat{\vartheta}, \vartheta_1) \geq \Delta/2$ on A. Therefore

$$\begin{aligned}
\max_i \mathbb{E}_i d^2(\widehat{\vartheta}, \vartheta_i) &\geq \frac{1}{2}\sum_{i=0}^{1} \mathbb{E}_i d^2(\widehat{\vartheta}, \vartheta_i) \\
&\geq \frac{\Delta^2}{8}\left(P_0(A^c) + P_1(A)\right) \\
&\geq \frac{\Delta^2}{8}\left(1 - \mathrm{TV}(P_0, P_1)\right).
\end{aligned}$$

Pinsker's inequality gives equation B.7. This is the standard two-point method; see Yu (1997); Tsybakov (2009). □

Lemma B.5 (Fano inequality with a reference measure)**.** *Let J be uniform on $[M]$, $M \geq 2$, and let $X \mid \{J = j\} \sim P_j$. For every decoder $\widehat{J} = \widehat{J}(X)$ and every probability measure P_0 dominating the P_j,*

$$\Pr(\widehat{J} \neq J) \geq 1 - \frac{M^{-1} \sum_{j=1}^{M} D_{\mathrm{KL}}(P_j \| P_0) + \log 2}{\log M}. \tag{B.8}$$

Suppose additionally that parameters $\vartheta_1, \ldots, \vartheta_M$ in a metric space satisfy $d(\vartheta_j, \vartheta_k) \geq 2s$ for $j \neq k$. Then every estimator $\widehat{\vartheta}$ satisfies

$$\sup_{1 \leq j \leq M} \mathbb{E}_j d^2(\widehat{\vartheta}, \vartheta_j) \geq s^2 \left[1 - \frac{M^{-1} \sum_{j=1}^{M} D_{\mathrm{KL}}(P_j \| P_0) + \log 2}{\log M} \right]. \tag{B.9}$$

Proof. Let $\overline{P} = M^{-1} \sum_j P_j$. The information-radius identity gives

$$\frac{1}{M} \sum_{j=1}^{M} D_{\mathrm{KL}}(P_j \| P_0) = I(J; X) + D_{\mathrm{KL}}(\overline{P} \| P_0) \geq I(J; X).$$

Fano's inequality then yields equation B.8. For the estimation claim, decode by the nearest packing point. If $d(\widehat{\vartheta}, \vartheta_J) < s$, the separation assumption forces the decoder to equal J. Hence $d^2(\widehat{\vartheta}, \vartheta_J) \geq s^2 \mathbf{1}\{\widehat{J} \neq J\}$, and averaging equation B.8 over J proves equation B.9. See Yu (1997); Tsybakov (2009) for standard variants. □

B.2 Finite information games and cost normalization

Lemma B.6 (Finite bilinear minimax theorem)**.** *Let $\mathcal{E}$ and $\mathcal{J}$ be finite nonempty sets, and let $A = (A_{e,j})_{e \in \mathcal{E}, j \in \mathcal{J}}$ be a real matrix. Then*

$$\max_{x \in \Delta(\mathcal{E})} \min_{j \in \mathcal{J}} \sum_{e \in \mathcal{E}} x_e A_{e,j} = \min_{\lambda \in \Delta(\mathcal{J})} \max_{e \in \mathcal{E}} \sum_{j \in \mathcal{J}} \lambda_j A_{e,j}. \tag{B.10}$$

Both extrema are attained.

Proof. The left-hand side is the optimal value of the linear program

$$\max_{x,t} \; t \quad \text{subject to} \quad \sum_e x_e A_{e,j} \geq t \; \forall j, \quad x \geq 0, \quad \sum_e x_e = 1.$$

Its dual is

$$\min_{\lambda,u} \; u \quad \text{subject to} \quad \sum_j \lambda_j A_{e,j} \leq u \; \forall e, \quad \lambda \geq 0, \quad \sum_j \lambda_j = 1.$$

Finite-dimensional linear-programming duality gives equality. Compactness of the simplices gives attainment. □

Lemma B.7 (Block-time and cost-time normalization)**.** *Let $c_e > 0$ for $e \in \mathcal{E}$, and suppose $w \in \mathbb{R}_+^{\mathcal{E}}$ satisfies $\sum_e c_e w_e = 1$. Set*

$$s = \sum_e w_e, \qquad p_e = \frac{w_e}{s}, \qquad \overline{c} = \sum_e p_e c_e.$$

Then

$$\overline{c} = \frac{1}{s}, \qquad \frac{p_e}{\overline{c}} = w_e \quad \forall e. \tag{B.11}$$

Conversely, every $p \in \Delta(\mathcal{E})$ induces a cost-normalized allocation $w_e = p_e / (\sum_f p_f c_f)$.

Proof. Direct calculation gives $\overline{c} = s^{-1} \sum_e c_e w_e = s^{-1}$. The remaining identities follow immediately. □

B.3 LIKELIHOOD-RATIO MARTINGALES AND CONCENTRATION

Lemma B.8 (Adaptive likelihood-ratio martingale)**.** *Let $(\mathcal{F}_t)_{t\geq 0}$ be a filtration. Conditional on $\mathcal{F}_{t-1}$, an observation Y_t has density $p_t(\cdot \mid \mathcal{F}_{t-1})$ under P and $q_t(\cdot \mid \mathcal{F}_{t-1})$ under Q, with $q_t \ll p_t$ almost surely. Then*

$$M_t \coloneqq \prod_{s=1}^{t} \frac{q_s(Y_s \mid \mathcal{F}_{s-1})}{p_s(Y_s \mid \mathcal{F}_{s-1})}, \qquad M_0 = 1, \tag{B.12}$$

is a nonnegative P-martingale.

Proof. By conditional integration,

$$\mathbb{E}_P\left[\frac{q_t(Y_t \mid \mathcal{F}_{t-1})}{p_t(Y_t \mid \mathcal{F}_{t-1})}\middle| \mathcal{F}_{t-1}\right] = 1.$$

Multiplying by the $\mathcal{F}_{t-1}$-measurable variable M_{t-1} proves the martingale property. □

Lemma B.9 (Ville's inequality)**.** *Let $(M_t)_{t\geq 0}$ be a nonnegative supermartingale with $\mathbb{E}[M_0] \leq 1$. Then, for every $u > 0$,*

$$\Pr\left(\sup_{t\geq 0} M_t \geq u\right) \leq \frac{1}{u}. \tag{B.13}$$

In particular, setting $u = 1/\delta$ gives an anytime-valid level-δ test.

Proof. Let $\tau = \inf\{t : M_t \geq u\}$. For every integer n, optional sampling at the bounded stopping time $\tau \wedge n$ gives $\mathbb{E}[M_{\tau\wedge n}] \leq 1$. Since $M_{\tau\wedge n} \geq u\mathbf{1}\{\tau \leq n\}$, $\Pr(\tau \leq n) \leq 1/u$. Letting $n \to \infty$ proves the claim. See also the mixture-martingale treatment of Kaufmann & Koolen (2021). □

Lemma B.10 (Azuma–Hoeffding)**.** *Let $(M_t, \mathcal{F}_t)_{t=0}^n$ be a real-valued martingale and suppose*

$$|M_t - M_{t-1}| \leq c_t \qquad \textit{almost surely}.$$

Then, for every $x > 0$,

$$\Pr(M_n - M_0 \geq x) \leq \exp\left(-\frac{x^2}{2\sum_{t=1}^n c_t^2}\right), \tag{B.14}$$

and the same bound holds for $\Pr(M_n - M_0 \leq -x)$. Consequently,

$$\Pr(|M_n - M_0| \geq x) \leq 2\exp\left(-\frac{x^2}{2\sum_{t=1}^n c_t^2}\right). \tag{B.15}$$

Proof. For each martingale difference $D_t = M_t - M_{t-1}$, conditional Hoeffding's lemma gives $\mathbb{E}[\exp(\lambda D_t) \mid \mathcal{F}_{t-1}] \leq \exp(\lambda^2 c_t^2/2)$. Iteration yields

$$\mathbb{E}\exp\left(\lambda(M_n - M_0) - \frac{\lambda^2}{2}\sum_t c_t^2\right) \leq 1.$$

Markov's inequality and optimization over $\lambda = x/(\sum_t c_t^2)$ prove equation B.14; apply the same argument to $-M_t$ for the lower tail. See Boucheron et al. (2013) for a general treatment. □

Corollary B.11 (Hoeffding bound for bounded averages)**.** *If $X_1, \ldots, X_n$ are independent, $X_i \in [a_i, b_i]$ almost surely, and $S_n = \sum_i (X_i - \mathbb{E}X_i)$, then*

$$\Pr(S_n \leq -x) \leq \exp\left(-\frac{2x^2}{\sum_i (b_i - a_i)^2}\right), \tag{B.16}$$

and analogously for the upper tail.

Lemma B.12 (Tail sums, Borel–Cantelli, and stretched-exponential tails)**.** *Let τ be a nonnegative integer-valued random variable. Then*

$$\mathbb{E}[\tau] = \sum_{k=0}^{\infty} \Pr(\tau > k) \tag{B.17}$$

with both sides possibly infinite. If events (A_k) satisfy $\sum_k \Pr(A_k) < \infty$, then A_k occurs only finitely often almost surely. Finally, for every $a, b, \beta > 0$,

$$\sum_{k \geq m} a \exp(-bk^{\beta}) = o(m^q) \qquad \textit{for every } q > 0. \tag{B.18}$$

Proof. Equation B.17 follows from $\tau = \sum_{k \geq 0} \mathbf{1}\{\tau > k\}$ and Tonelli's theorem. The second claim is the first Borel–Cantelli lemma. For the last claim, compare the sum with $ae^{-bm^{\beta}} + a \int_{m-1}^{\infty} e^{-bx^{\beta}} \, \mathrm{d}x$ and substitute $u = bx^{\beta}$. The resulting incomplete-gamma tail decays faster than every inverse polynomial. □

Lemma B.13 (Wald's first identity)**.** *Let $(X_t)_{t \geq 1}$ be i.i.d. with $\mathbb{E}|X_1| < \infty$, adapted to $\mathcal{F}_t = \sigma(X_1, \ldots, X_t)$, and let τ be a stopping time with $\mathbb{E}\tau < \infty$. Then*

$$\mathbb{E}\left[\sum_{t=1}^{\tau} X_t\right] = \mathbb{E}[\tau] \, \mathbb{E}[X_1]. \tag{B.19}$$

Proof. Since $\{\tau \geq t\} \in \mathcal{F}_{t-1}$ and X_t is independent of $\mathcal{F}_{t-1}$,

$$\mathbb{E}[X_t \mathbf{1}\{\tau \geq t\}] = \mathbb{E}[X_1] \Pr(\tau \geq t).$$

Moreover, $\mathbb{E} \sum_{t \geq 1} |X_t| \mathbf{1}\{\tau \geq t\} = \mathbb{E}|X_1| \, \mathbb{E}\tau < \infty$, so Fubini's theorem applies. Summing over t and using equation B.17 proves equation B.19. □

B.4 Gaussian and random-matrix identities

Lemma B.14 (Gaussian conditioning)**.** *Let (U, V) be jointly Gaussian with means m_U, m_V and covariance*

$$\begin{pmatrix} \Sigma_U & C \\ C^{\top} & \Sigma_V \end{pmatrix}, \qquad \Sigma_V \succ 0.$$

Then

$$U \mid \{V = v\} \sim \mathcal{N}\left(m_U + C\Sigma_V^{-1}(v - m_V), \Sigma_U - C\Sigma_V^{-1}C^{\top}\right). \tag{B.20}$$

The residual $U - m_U - C\Sigma_V^{-1}(V - m_V)$ is independent of V.

Proof. The residual is jointly Gaussian with V and has zero cross-covariance with V. Joint Gaussianity therefore implies independence, and direct covariance calculation gives equation B.20. □

Lemma B.15 (Isserlis bilinear identity)**.** *Let $U \in \mathbb{R}^m$ and $V \in \mathbb{R}^n$ be jointly centered Gaussian, with $\Sigma_U = \mathbb{E}[UU^{\top}]$, $\Sigma_V = \mathbb{E}[VV^{\top}]$, and $C = \mathbb{E}[UV^{\top}]$. For matrices $A, B \in \mathbb{R}^{m \times n}$,*

$$\begin{aligned} \mathbb{E}\left[(U^{\top}AV)(U^{\top}BV)\right] = \; & \operatorname{Tr}(A\Sigma_V B^{\top}\Sigma_U) + \operatorname{Tr}(AC^{\top}BC^{\top}) \\ & + \operatorname{Tr}(A^{\top}C)\operatorname{Tr}(B^{\top}C). \end{aligned} \tag{B.21}$$

In particular, if $\Sigma_U = bI_d$, $\Sigma_V = aI_d$, $C = I_d$, and $A = \Omega$, $B = \Lambda$ are skew symmetric, then

$$\mathbb{E}\left[(U^{\top}\Omega V)(U^{\top}\Lambda V)\right] = (ab - 1)\operatorname{Tr}(\Omega^{\top}\Lambda). \tag{B.22}$$

Proof. Expand both bilinear forms in coordinates and apply Isserlis' fourth-moment identity

$$\mathbb{E}[U_i V_j U_k V_\ell] = C_{ij}C_{k\ell} + (\Sigma_U)_{ik}(\Sigma_V)_{j\ell} + C_{i\ell}C_{kj}.$$

Summing the three contractions gives equation B.21. For skew-symmetric matrices, $\operatorname{Tr}(\Omega) = \operatorname{Tr}(\Lambda) = 0$ and $\operatorname{Tr}(\Omega\Lambda) = -\operatorname{Tr}(\Omega^{\top}\Lambda)$, which yields equation B.22. □

Lemma B.16 (Gaussian rank and pseudoinverse facts)**.** *Let $Z \in \mathbb{R}^{d\times n}$ have i.i.d. standard Gaussian entries. Then* $\operatorname{rank}(Z) = \min\{d, n\}$ *almost surely. If $n - d \geq 2$, then*

$$\mathbb{E}\left\|Z^\dagger\right\|_{\mathrm{F}}^2 = \frac{d}{n-d-1}. \tag{B.23}$$

If $d \geq 2$, $n = d + p$, and $p \geq 4$, then for every $u \geq 1$,

$$\Pr\left(\left\|Z^\dagger\right\|_{\mathrm{op}} \geq \frac{e\sqrt{n}}{p+1}u\right) \leq u^{-(p+1)}. \tag{B.24}$$

Proof. Almost-sure full rank follows because every maximal minor determinant is a polynomial and at least one such polynomial is nonzero. The exact Frobenius moment and the operator-norm tail are Propositions 10.2 and 10.4, respectively, of Halko et al. (2011); they follow from inverse-Wishart moments and lower-tail control of the smallest Gaussian singular value. □

Lemma B.17 (Laurent–Massart chi-square inequality)**.** *If $X \sim \chi_k^2$, then for every $t \geq 0$,*

$$\Pr\left(X - k \geq 2\sqrt{kt} + 2t\right) \leq e^{-t}, \tag{B.25}$$

$$\Pr\left(k - X \geq 2\sqrt{kt}\right) \leq e^{-t}. \tag{B.26}$$

Proof. These are the standard Laurent–Massart bounds (Laurent & Massart, 2000). They follow by applying Chernoff's method to the moment-generating function $\mathbb{E}e^{\lambda(X-k)} = e^{-k\lambda}(1-2\lambda)^{-k/2}$ and optimizing separately for positive and negative λ. □

B.5 Orthogonal projection and polar factors

Lemma B.18 (Orthogonal Procrustes projection)**.** *Let $A \in \mathbb{R}^{d\times d}$ have singular-value decomposition*

$$A = U\begin{pmatrix}\operatorname{diag}(s_1,\dots,s_q) & 0\\ 0 & 0\end{pmatrix}V^\top, \qquad s_1,\dots,s_q > 0.$$

Then

$$\operatorname*{arg\,min}_{Q\in\mathrm{O}(d)} \|A - Q\|_{\mathrm{F}} = \operatorname*{arg\,max}_{Q\in\mathrm{O}(d)} \operatorname{Tr}(Q^\top A) = \left\{U\begin{pmatrix}I_q & 0\\ 0 & W\end{pmatrix}V^\top : W \in \mathrm{O}(d-q)\right\}. \tag{B.27}$$

If A is nonsingular, the minimizer is unique and equals

$$\operatorname{Polar}(A) = UV^\top = A(A^\top A)^{-1/2}. \tag{B.28}$$

Proof. For orthogonal Q,

$$\|A - Q\|_{\mathrm{F}}^2 = \|A\|_{\mathrm{F}}^2 + d - 2\operatorname{Tr}(Q^\top A),$$

so the two optimization problems agree. Write $B = U^\top QV \in \mathrm{O}(d)$. Then

$$\operatorname{Tr}(Q^\top A) = \sum_{j=1}^{q} s_j B_{jj} \leq \sum_{j=1}^{q} s_j,$$

with equality if and only if $B_{jj} = 1$ for all $j \leq q$. Orthogonality then forces the first q rows and columns of B to equal those of the identity, leaving an arbitrary $W \in \mathrm{O}(d-q)$ on the null space. This is the classical orthogonal Procrustes solution (Schönemann, 1966; Higham, 1986). □

Corollary B.19 (Projection inequality for a polar estimator)**.** *Let $\widehat{Q}$ be any nearest orthogonal matrix to A in Frobenius norm. Then, for every $Q_\star \in \mathrm{O}(d)$,*

$$\left\|\widehat{Q} - Q_\star\right\|_{\mathrm{F}} \leq 2\left\|A - Q_\star\right\|_{\mathrm{F}}. \tag{B.29}$$

Proof. Nearest-point optimality gives $\left\|A - \widehat{Q}\right\|_{\mathrm{F}} \leq \|A - Q_\star\|_{\mathrm{F}}$. Apply the triangle inequality. □

Lemma B.20 (Matrix-exponential remainders)**.** *For every square matrix H,*

$$\left\|e^H - I - H\right\|_{\mathrm{F}} \le \frac{1}{2} e^{\|H\|_{\mathrm{op}}} \|H\|_{\mathrm{op}} \|H\|_{\mathrm{F}}, \tag{B.30}$$

$$\left\|e^H - I - H - \tfrac{1}{2}H^2\right\|_{\mathrm{F}} \le \frac{1}{6} e^{\|H\|_{\mathrm{op}}} \|H\|_{\mathrm{op}}^2 \|H\|_{\mathrm{F}}. \tag{B.31}$$

If $H^\top = -H$, then $e^H \in \mathrm{O}(d)$ and $\mathrm{Tr}(H^2) = -\|H\|_{\mathrm{F}}^2$.

Proof. Expand the exponential series and use $\left\|H^k\right\|_{\mathrm{F}} \le \|H\|_{\mathrm{op}}^{k-1} \|H\|_{\mathrm{F}}$. The skew-symmetric claims follow from $(e^H)^\top = e^{-H}$ and $\|H\|_{\mathrm{F}}^2 = \mathrm{Tr}(H^\top H) = -\mathrm{Tr}(H^2)$. □

B.6 Compact-group integration

Lemma B.21 (Quotient integration for compact groups)**.** *Let G be a compact second-countable group with normalized Haar probability λ_G, and let $H \le G$ be closed with normalized Haar probability λ_H. There is a unique G-invariant probability $\overline{\lambda}_{G/H}$ on the homogeneous space G/H such that every integrable $f : G \to \mathbb{R}$ satisfies*

$$\int_G f(g)\, \lambda_G(\mathrm{d}g) = \int_{G/H} \int_H f(gh)\, \lambda_H(\mathrm{d}h)\, \overline{\lambda}_{G/H}(\mathrm{d}(gH)). \tag{B.32}$$

The inner integral is independent of the representative chosen for the coset gH.

Proof. Right invariance of λ_H makes the inner integral well defined on G/H. The resulting positive linear functional on continuous functions is G-invariant and normalized; the Riesz representation theorem and uniqueness of invariant probability on the compact homogeneous space give $\overline{\lambda}_{G/H}$. The identity is the standard compact-group quotient integration formula; see Eaton (1989); Diaconis (1988). □

Lemma B.22 (Haar mean of an orthogonal transformation)**.** *If U has normalized Haar distribution on $\mathrm{O}(k)$ for $k \ge 1$, then*

$$\mathbb{E}[U] = 0. \tag{B.33}$$

Consequently, for every deterministic $x \in \mathbb{R}^k$, $\mathbb{E}[Ux] = 0$ and $\mathbb{E}\|Ux\|_2^2 = \|x\|_2^2$.

Proof. Since $-I_k \in \mathrm{O}(k)$ and Haar measure is left invariant, $U \overset{d}{=} (-I_k)U = -U$. Hence $\mathbb{E}U = -\mathbb{E}U$. The remaining claims follow from orthogonality. □

B.7 Local asymptotic theory

We finally record the exact classical asymptotic principles invoked after the paired-view model has been shown to be differentiable in quadratic mean. In this subsection, $h \in \mathbb{R}^p$ is a local Euclidean parameter and $I \succ 0$ is an information matrix.

Definition B.23 (Local asymptotic normality)**.** A sequence of experiments $\{P_{n,h} : h \in \mathbb{R}^p\}$ is locally asymptotically normal at $h = 0$ with information I if there exist random vectors Δ_n such that under $P_{n,0}$,

$$\log \frac{\mathrm{d}P_{n,h}}{\mathrm{d}P_{n,0}} = h^\top \Delta_n - \frac{1}{2} h^\top I h + o_{P_{n,0}}(1), \qquad \Delta_n \rightsquigarrow \mathcal{N}(0, I), \tag{B.34}$$

uniformly for h in every fixed compact set.

Proposition B.24 (DQM implies i.i.d. LAN)**.** *Suppose a dominated model $\{P_\eta : \eta \in \mathbb{R}^p\}$ is differentiable in quadratic mean at η_0 with score $\dot{\ell}_{\eta_0} \in L^2(P_{\eta_0})$ and nonsingular information $I = \mathbb{E}_{\eta_0}[\dot{\ell}_{\eta_0} \dot{\ell}_{\eta_0}^\top]$. Then the i.i.d. experiments with local parameters $\eta_n = \eta_0 + h/\sqrt{n}$ are LAN, with*

$$\Delta_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}_{\eta_0}(X_i). \tag{B.35}$$

Proposition B.25 (Le Cam's first and third lemmas)**.** *Assume equation B.34. For every fixed h, the sequences $(P_{n,h})$ and $(P_{n,0})$ are mutually contiguous. If, jointly under $P_{n,0}$,*

$$(T_n, \Delta_n) \rightsquigarrow (T, Z), \qquad Z \sim \mathcal{N}(0, I),$$

then under $P_{n,h}$, T_n converges to the exponentially tilted law characterized by

$$\mathbb{E}_h[f(T)] = \mathbb{E}\left[f(T)\exp\left(h^\top Z - \tfrac{1}{2}h^\top I h\right)\right] \tag{B.36}$$

for bounded continuous f. In particular, if $T = Z$, then under $P_{n,h}$, $\Delta_n \rightsquigarrow \mathcal{N}(Ih, I)$.

Theorem B.26 (Local asymptotic minimax and convolution principles)**.** *Assume a p-dimensional experiment is LAN at the origin with nonsingular information I. For $M < \infty$, let $\ell_M(u) = \min\{\|u\|_2^2, M\}$. Then*

$$\lim_{c\to\infty} \liminf_{n\to\infty} \inf_{T_n} \sup_{\|h\|_2 \le c} \mathbb{E}_{n,h}\ell_M(T_n - h) \ge \mathbb{E}\ell_M(I^{-1/2}Z), \qquad Z \sim \mathcal{N}(0, I_p). \tag{B.37}$$

Moreover, if T_n is regular at the origin and its centered local limit law is L, then the Hájek–Le Cam convolution theorem gives

$$L \overset{d}{=} I^{-1/2}Z + V, \qquad Z \perp V, \tag{B.38}$$

for some random vector V.

Reference. These are the standard DQM-to-LAN, contiguity, local asymptotic minimax, and convolution theorems; see van der Vaart (1998, Chs. 6–9) and Le Cam & Yang (2000, Chs. 3, 5–6). Appendix N verifies all model-specific hypotheses, performs the tangent-coordinate reduction from $\mathrm{O}(d)$ to $\mathbb{R}^{p_d}$, applies equation B.37 first to bounded truncated loss, and only then passes to unbounded squared Frobenius loss by monotone convergence. □

Scope. The statements above are tools, not additional claims about the persistent meta-RL model. Their use in this paper is restricted to fixed dimension and fixed positive effective noise. The singular noiseless limit is handled by the exact stabilizer and breadth arguments of Sections J, K and M rather than by regular Fisher-information asymptotics.

C Proofs for Persistence Hierarchies

This appendix constructs, for every integer $r \ge 2$, a finite persistent meta-MDP whose target-relevant global parameter is invisible to every same-task interaction that reveals fewer than r distinct task coordinates. The construction is a controlled analogue of the classical phenomenon that grouped observations may identify latent structure that is absent from every proper marginal (Vandermeulen & Scott, 2019). We prove invariance under fully adaptive coordinate selection, adaptive reset decisions, and random source stopping. We then derive matching fixed-confidence lower and upper bounds by a two-point change-of-measure argument (Yu, 1997; Tsybakov, 2009).

Throughout this appendix, fix an integer $r \ge 2$, a parity-bias parameter $\varrho \in (0, 1)$, and a reset cost $\kappa \ge 0$. We use $\sigma \in \{-1, +1\}$ for the global meta-parameter and write $\chi(x) := \prod_{j=1}^{r} x_j$ for the parity of $x \in \{-1, +1\}^r$.

C.1 The parity-tilted persistent meta-MDP

Definition C.1 (Parity-tilted task law)**.** Let

$$\Theta_r := \{-1, +1\}, \qquad \mathcal{Z}_r := \{-1, +1\}^r.$$

For $\sigma \in \Theta_r$, define the fresh-task distribution $\mu_\sigma^{(r,\varrho)} \in \mathcal{P}(\mathcal{Z}_r)$ by

$$\mu_\sigma^{(r,\varrho)}(x) := 2^{-r}\big(1 + \varrho\sigma\chi(x)\big), \qquad x \in \mathcal{Z}_r. \tag{C.1}$$

The mass function in equation C.1 is strictly positive and sums to one because $\varrho \in (0, 1)$ and exactly half of the hypercube has parity $+1$. The global parameter affects only the dependence among all r task coordinates; every proper marginal is uniform.

Lemma C.2 (Uniformity of every proper marginal)**.** *Let $J \subsetneq [r]$. Under either global model,*

$$\mu_\sigma^{(r,\varrho)}\big(X_J = x_J\big) = 2^{-|J|}, \qquad \sigma \in \{-1,+1\}, \quad x_J \in \{-1,+1\}^J. \tag{C.2}$$

Consequently, if $j \notin J$ and $|J| \le r-2$, then

$$\mathbb{P}_\sigma\big(X_j = x \mid X_J = x_J\big) = \frac{1}{2}, \qquad x \in \{-1,+1\}. \tag{C.3}$$

If $J = [r] \setminus \{j\}$, then

$$\mathbb{P}_\sigma\big(X_j = x \mid X_J = x_J\big) = \frac{1}{2}\left(1 + \varrho\sigma x \prod_{k\in J} x_k\right). \tag{C.4}$$

Proof. Fix $J \subsetneq [r]$ and $x_J \in \{-1,+1\}^J$. Summing equation C.1 over the unobserved coordinates gives

$$\begin{aligned}\mu_\sigma^{(r,\varrho)}(X_J = x_J) &= 2^{-r} \sum_{x_{J^c}\in\{-1,+1\}^{J^c}} \left(1 + \varrho\sigma \prod_{j\in J} x_j \prod_{k\in J^c} x_k\right) \\ &= 2^{-r}\, 2^{r-|J|} + 2^{-r}\varrho\sigma \prod_{j\in J} x_j \sum_{x_{J^c}} \prod_{k\in J^c} x_k.\end{aligned}$$

Because J^c is nonempty, the final sum is zero: flipping any one unobserved coordinate pairs terms of opposite sign. This proves equation C.2. Equation C.3 follows by taking the ratio of two uniform proper marginals. When exactly one coordinate remains unobserved, substituting equation C.1 into Bayes' rule gives equation C.4. □

We now embed this task law in a finite-state one-step MDP. Although the physical dynamics are deliberately minimal, the source interaction remains sequential at the meta level: each episode reveals at most one coordinate, and the learner controls both coordinate selection and task persistence.

Definition C.3 (Parity meta-MDP)**.** For every $(\sigma, x) \in \Theta_r \times \mathcal{Z}_r$, let $M_{\sigma,x}^{(r)}$ be the horizon-one MDP with

$$\mathcal{S}_r = \{s_\circ, s_\dagger\}, \qquad \mathcal{A}_r = \{0, 1, \dots, r\}.$$

The initial state is $s_\circ$ almost surely. Every action transitions to $s_\dagger$. Action 0 is a safe action with deterministic reward $1/2$, whereas coordinate action $j \in [r]$ has deterministic reward

$$R(j; x) \coloneqq \frac{1 + x_j}{2} \in \{0, 1\}. \tag{C.5}$$

A reset samples a fresh $X \sim \mu_\sigma^{(r,\varrho)}$; continuation preserves the same X.

The reward from coordinate action j reveals X_j exactly through $X_j = 2R(j; X) - 1$. Replaying an already queried coordinate or taking the safe action yields no new task information. For a within-task observed history h, let

$$J(h) \coloneqq \{j \in [r] : \text{coordinate action } j \text{ appears in } h\} \tag{C.6}$$

be the set of distinct revealed coordinates.

C.2 Adaptive subcritical invariance

We first prove that no adaptive experiment revealing a proper subset of the coordinates can distinguish the global models. The statement permits randomized action selection, repeated coordinates, safe actions, adaptive resets, and random stopping.

Lemma C.4 (Within-task adaptive invariance)**.** *Consider one task drawn from equation C.1. Let an arbitrary randomized, history-dependent rule select actions in $\mathcal{A}_r$ and stop at a within-task stopping time τ. Suppose*

$$|J(H_\tau)| \le r - 1 \qquad \textit{almost surely}, \tag{C.7}$$

where H_τ is the stopped action–reward history. Then the law of H_τ is identical under $\sigma = +1$ and $\sigma = -1$.

Proof. We first prove equality of every finite prefix law before the rth distinct coordinate is revealed. Assume a realized history h has positive probability and $|J(h)| \le r-1$. Conditional on h, the algorithm's next action kernel is the same under both global models because it is a fixed measurable function of the observed history and its own randomization.

If the selected action is 0, its reward is deterministically $1/2$ under both models. If it repeats a coordinate in $J(h)$, its reward is already determined by h. If it selects a new coordinate $j \notin J(h)$ while $|J(h)| \le r-2$, then equation C.3 implies that its reward is Bernoulli$(1/2)$ under both models. Hence the conditional law of the next action–reward pair is model independent until the action would reveal the rth distinct coordinate. Induction over the number of episodes gives equality of all such prefix laws.

The stopped history can be represented on the countable disjoint union of finite histories. For every stopped transcript h satisfying equation C.7, the event $\{H_\tau = h\}$ is determined by a subcritical prefix and the algorithm's stopping kernel. The prefix probability and stopping probability are equal under the two models. Summing over stopped transcripts proves equality of the stopped-history laws. Conditioning on an auxiliary random seed gives the same conclusion for randomized rules. □

Proposition C.5 (Full-source invariance below persistence order)**.** *Consider an arbitrary source learner in the parity meta-MDP. Suppose that, almost surely, no source task block reveals all r coordinate rewards before the learner resets or stops. Then the complete stopped observed source history has the same distribution under the two global models:*

$$\mathbb{P}_+^{\mathcal{A}}\big((\mathsf{T}, O_{\mathsf{T}}) \in \cdot\big) = \mathbb{P}_-^{\mathcal{A}}\big((\mathsf{T}, O_{\mathsf{T}}) \in \cdot\big). \tag{C.8}$$

Consequently, the recommendation $\widehat{g}$ has the same marginal law under $\sigma = +1$ and $\sigma = -1$.

Proof. Partition the source transcript into task blocks at reset times, using the construction in Section A.5. Conditional on the completed pre-block history, the reset samples a fresh task from $\mu_\sigma^{(r,\varrho)}$, and the learner then applies a possibly history-dependent within-task rule. By assumption the resulting stopped block reveals a proper subset of the coordinates, so Lemma C.4 shows that its conditional law is identical under the two global models. Induction over completed blocks, including the possibly incomplete final block, proves equation C.8. The recommendation kernel is the same measurable kernel under both models, so applying it to two equal stopped-history laws yields equal recommendation laws. □

C.3 Exact persistence order and complete-task information

Let e_{full} be the deterministic-length block experiment that queries coordinate actions $1, 2, \dots, r$ exactly once. Its transcript is in one-to-one correspondence with $X \in \{-1, +1\}^r$.

Proposition C.6 (Exact pairwise persistence order)**.** *For the parity meta-MDP,*

$$r(+1, -1) = r. \tag{C.9}$$

Moreover,

$$D_{\mathrm{KL}}\left(Q_+^{e_{\mathrm{full}}} \middle\| Q_-^{e_{\mathrm{full}}}\right) = d_{\mathrm{par}}(\varrho) := \varrho \log \frac{1+\varrho}{1-\varrho}, \tag{C.10}$$

$$\mathrm{TV}\left(Q_+^{e_{\mathrm{full}}}, Q_-^{e_{\mathrm{full}}}\right) = \varrho. \tag{C.11}$$

For $\varrho \in (0, 1/2]$,

$$2\varrho^2 \le d_{\mathrm{par}}(\varrho) \le \frac{8}{3}\varrho^2. \tag{C.12}$$

Proof. By Lemma C.4, every block experiment that reveals fewer than r distinct coordinates has the same law under the two models. In particular, no experiment of length below r separates them. The experiment e_{full} reveals X itself, and the two task laws in equation C.1 are distinct, proving equation C.9.

For the KL divergence, note that the likelihood ratio depends only on $W = \chi(X)$:

$$\frac{\mu_+^{(r,\varrho)}(x)}{\mu_-^{(r,\varrho)}(x)} = \frac{1+\varrho\chi(x)}{1-\varrho\chi(x)}.$$

Under model σ, the parity statistic satisfies

$$\mathbb{P}_\sigma(W = w) = \frac{1}{2}(1 + \varrho\sigma w), \qquad w \in \{-1, +1\}. \tag{C.13}$$

Therefore

$$\begin{aligned} D_{\mathrm{KL}}(\mu_+ \| \mu_-) &= \frac{1+\varrho}{2} \log \frac{1+\varrho}{1-\varrho} + \frac{1-\varrho}{2} \log \frac{1-\varrho}{1+\varrho} \\ &= \varrho \log \frac{1+\varrho}{1-\varrho}. \end{aligned}$$

The total-variation identity follows by direct summation. Finally,

$$d_{\mathrm{par}}(\varrho) = 2\varrho \operatorname{arctanh}(\varrho).$$

The bounds $\varrho \le \operatorname{arctanh}(\varrho) \le \varrho/(1-\varrho^2)$ imply equation C.12 when $\varrho \le 1/2$. □

The next lemma quantifies exactly where information enters an arbitrary adaptive source transcript. A task block is called *complete* if the learner queries every coordinate action at least once before resetting or stopping. Let $N_{\mathrm{full}}(\mathsf{T})$ be the number of complete task blocks by source stopping time T.

Lemma C.7 (Exact KL accounting by completed tasks)**.** *Let $\mathcal{A}$ be any source learner. For each $\sigma \in \{-1, +1\}$, suppose that the source stopping time T is finite $\mathbb{P}_\sigma^{\mathcal{A}}$-almost surely and*

$$\mathbb{E}_\sigma^{\mathcal{A}}\left[N_{\mathrm{full}}(\mathsf{T})\right] < \infty.$$

Then

$$D_{\mathrm{KL}}\left(\mathbb{P}_\sigma^{\mathcal{A}} \middle\| \mathbb{P}_{-\sigma}^{\mathcal{A}}\right) = d_{\mathrm{par}}(\varrho)\mathbb{E}_\sigma^{\mathcal{A}}\left[N_{\mathrm{full}}(\mathsf{T})\right]. \tag{C.14}$$

In particular, finite expected source cost implies the required integrability, because pathwise

$$(r+\kappa)N_{\mathrm{full}}(\mathsf{T}) \le \mathsf{C}_{\mathsf{T}}.$$

Proof. Fix $\sigma \in \{-1, +1\}$. Consider the observed process episode by episode, including the persistence action, within-episode action, and reward. Conditional on the observed past, the learner's persistence and action kernels are identical under the two models; only the conditional reward law can contribute to the KL chain rule. Safe actions and repeated coordinate actions are conditionally deterministic. By equation C.3, querying a new coordinate while at least two coordinates remain unobserved yields a fair bit under both models and contributes zero conditional KL.

Suppose instead that the learner queries the unique remaining coordinate j of the current task. If the already observed coordinate values are x_{-j}, then equation C.4 gives two Bernoulli laws with success parameters

$$\frac{1}{2}\left(1 + \varrho \prod_{k \ne j} x_k\right) \quad \text{and} \quad \frac{1}{2}\left(1 - \varrho \prod_{k \ne j} x_k\right).$$

Their KL divergence is exactly $d_{\mathrm{par}}(\varrho)$, independently of the realized sign of the observed product. This event occurs exactly once for every complete task block and never for an incomplete block.

For a bounded deterministic horizon, the conditional KL chain rule under $\mathbb{P}_\sigma^{\mathcal{A}}$ therefore gives equation C.14. For an almost surely finite stopping time, pad the transcript after stopping by a cemetery symbol, apply the bounded-horizon identity to $\mathsf{T} \wedge n$, and let $n \to \infty$. The stopped sigma-fields increase to the full stopped-history sigma-field, relative entropy is continuous along this increasing sequence, and monotone convergence applies to $N_{\mathrm{full}}(\mathsf{T} \wedge n)$. The assumed finite expectation makes both sides finite. If that expectation is infinite, the lower bounds below are immediate and no finite identity is required. □

C.4 Fresh-target decision geometry

We now show that the same parity family induces genuinely target-relevant alternatives under the target protocol of Section A.6. Set the target adaptation budget to

$$B = r - 1. \tag{C.15}$$

During adaptation and deployment, the target interacts with the same horizon-one MDP in Definition C.3. Adaptation rewards are observed but not included in the deployment objective, as specified in equation A.36.

For a realized task x, the oracle deployment value is

$$V^\star_{\sigma,x} = \begin{cases} 1, & \max_{j\in[r]} x_j = +1, \\ 1/2, & x_1 = \cdots = x_r = -1. \end{cases} \tag{C.16}$$

Lemma C.8 (Optimal known-model target loss)**.** *For every $\sigma \in \{-1, +1\}$,*

$$\mathcal{L}^\star_{r-1}(\sigma) = \frac{1-\varrho}{2^{r+1}}. \tag{C.17}$$

One optimal strategy $g^\star_\sigma$ queries any $r-1$ distinct coordinates during adaptation. At deployment, it selects an observed coordinate with value $+1$ whenever one exists. If all observed coordinates are -1, it selects the unique unobserved coordinate when $\sigma \prod_{j\in J} X_j = +1$ and selects the safe action otherwise.

Proof. Because adaptation rewards are not evaluated and the task latent does not change, querying an unobserved coordinate weakly Blackwell-dominates taking the safe action or repeating a previously queried coordinate: the learner can ignore the additional bit and simulate the less informative action. By backward induction over the finite adaptation horizon, there exists an optimal strategy that queries a new coordinate whenever no value-$+1$ coordinate has yet been found. Hence the only history at which nontrivial decision uncertainty remains after $r-1$ adaptation episodes is the event E that all $r-1$ revealed coordinates equal -1.

Proper-marginal uniformity gives $\mathbb{P}_\sigma(E) = 2^{-(r-1)}$. Conditional on a revealed set J of size $r-1$ and values x_J, the remaining coordinate $X_{j^\star}$ obeys equation C.4; hence

$$p_\sigma(x_J) \coloneqq \mathbb{P}_\sigma\left(X_{j^\star} = +1 \mid X_J = x_J\right) = \frac{1}{2}\left(1 + \varrho\sigma \prod_{j\in J} x_j\right). \tag{C.18}$$

If an observed coordinate equals $+1$, deploying it attains the oracle value one. On E, every observed coordinate has reward zero, the safe action has reward $1/2$, and the missing coordinate has posterior expected reward $p_\sigma(-\mathbf{1})$. Therefore the optimal deployment action is the missing coordinate when this posterior exceeds $1/2$ and the safe action when it is below $1/2$, proving the stated policy.

Conditional on E, the oracle expected value is $(1 + p_\sigma(-\mathbf{1}))/2$. If the missing coordinate is selected, the conditional excess loss is $(1 - p_\sigma(-\mathbf{1}))/2$; if the safe action is selected, it is $p_\sigma(-\mathbf{1})/2$. Since the two possible posterior probabilities are $(1 \pm \varrho)/2$, both cases give conditional loss $(1-\varrho)/4$. Multiplication by $\mathbb{P}_\sigma(E) = 2^{-(r-1)}$ yields equation C.17. □

Lemma C.9 (No common near-optimal target strategy)**.** *For every target strategy $g \in \mathcal{G}_{r-1}$,*

$$\frac{1}{2}\sum_{\sigma\in\{-1,+1\}} \left(\mathcal{L}_{r-1}(\sigma, g) - \mathcal{L}^\star_{r-1}(\sigma)\right) \geq \frac{\varrho}{2^{r+1}}. \tag{C.19}$$

Consequently, for every $0 < \varepsilon < \varrho/2^{r+1}$,

$$\mathcal{G}_{r-1,\varepsilon}(+1) \cap \mathcal{G}_{r-1,\varepsilon}(-1) = \varnothing. \tag{C.20}$$

Proof. Place the uniform prior on the global sign σ. The corresponding mixture task law is exactly uniform on $\{-1, +1\}^r$ because the parity tilt cancels:

$$\frac{1}{2}\mu^{(r,\varrho)}_+ + \frac{1}{2}\mu^{(r,\varrho)}_- = \mathrm{Unif}(\{-1, +1\}^r).$$

Under this mixture model the task coordinates are independent fair bits. As in the proof of Lemma C.8, querying a previously unobserved coordinate weakly dominates every noninformative adaptation action. Thus the smallest possible mixture target loss is attained by revealing $r-1$ distinct coordinates. If at least one revealed coordinate is $+1$, the oracle is attained. If all revealed coordinates are -1, an

unrevealed coordinate is still fair; both the safe action and that coordinate have expected reward $1/2$, while the conditional oracle value is $3/4$. This event has probability $2^{-(r-1)}$, so

$$\inf_{g\in\mathcal{G}_{r-1}} \frac{1}{2}\sum_{\sigma}\mathcal{L}_{r-1}(\sigma,g) = \frac{1}{2^{r+1}}. \tag{C.21}$$

Subtracting the average of equation C.17 proves equation C.19. If a single g were ε-optimal under both models, the left-hand side of equation C.19 would be at most ε, contradicting $\varepsilon < \varrho/2^{r+1}$. □

Combining source invariance with the preceding target geometry yields the arbitrary-order target-relevant persistence barrier.

Theorem C.10 (Arbitrary-order target-relevant persistence hierarchy)**.** *Fix $r \geq 2$, $\varrho \in (0,1)$, target budget $B = r-1$, and $0 < \varepsilon < \varrho/2^{r+1}$. The finite persistent meta-MDP in Definition C.3 satisfies:*

1. *its two global models are target-relevant alternatives and have exact pairwise persistence order r;*
2. *every source learner that almost surely reveals at most $r-1$ distinct coordinates from each source task obeys*

$$\sup_{\sigma\in\{-1,+1\}} \mathbb{E}_{\sigma}^{\mathcal{A}}\left[\mathcal{L}_{r-1}(\sigma,\widehat{g}) - \mathcal{L}_{r-1}^{\star}(\sigma)\right] \geq \frac{\varrho}{2^{r+1}}; \tag{C.22}$$

3. *observing all r coordinates of one fresh task produces exactly $d_{\mathrm{par}}(\varrho)$ nats of KL information about the global sign.*

Proof. Target relevance follows from equation C.20, and the exact persistence order follows from Proposition C.6. Under the subcritical source condition, Proposition C.5 shows that $\widehat{g}$ has one common marginal law, say ν, under both global models. Integrating equation C.19 with respect to $g \sim \nu$ gives

$$\frac{1}{2}\sum_{\sigma}\mathbb{E}_{\sigma}^{\mathcal{A}}\left[\mathcal{L}_{r-1}(\sigma,\widehat{g}) - \mathcal{L}_{r-1}^{\star}(\sigma)\right] \geq \frac{\varrho}{2^{r+1}}.$$

The maximum of the two risks is at least their average, yielding equation C.22. The final claim is equation C.10. □

C.5 Matching fixed-confidence source complexity

We next characterize the source cost required to recommend an ε-optimal target strategy for the parity family. The lower bound allows arbitrary adaptive reset and continuation rules, arbitrary random stopping, and arbitrary recommendation kernels.

Theorem C.11 (Fixed-confidence bounds for parity transfer)**.** *Fix the assumptions of Theorem C.10 and let $\delta \in (0, 1/2)$. Every source learner satisfying*

$$\mathbb{P}_{\sigma}^{\mathcal{A}}\left(\widehat{g} \in \mathcal{G}_{r-1,\varepsilon}(\sigma)\right) \geq 1-\delta, \qquad \sigma \in \{-1,+1\}, \tag{C.23}$$

obeys, for each σ,

$$\mathbb{E}_{\sigma}^{\mathcal{A}}\left[\mathsf{C}_{\mathsf{T}}\right] \geq (r+\kappa)\frac{\mathrm{kl}(1-\delta,\delta)}{d_{\mathrm{par}}(\varrho)}. \tag{C.24}$$

Conversely, for every integer $n \geq 1$, the complete-block procedure using n independent source tasks has cost $n(r+\kappa)$ and returns a target strategy with failure probability at most

$$\exp\left(-\frac{n\varrho^2}{2}\right). \tag{C.25}$$

In particular, when $\varrho \leq 1/2$, as $\delta \downarrow 0$ the optimal fixed-confidence source cost is

$$\Theta\left((r+\kappa)\varrho^{-2}\log\frac{1}{\delta}\right), \tag{C.26}$$

with universal implicit constants.

Proof. Because the two ε-optimal target-strategy sets are disjoint, define a test by declaring $\widehat{\sigma} = +1$ whenever $\widehat{g} \in \mathcal{G}_{r-1,\varepsilon}(+1)$ and declaring -1 otherwise. Correctness in equation C.23 implies

$$\mathbb{P}_+(\widehat{\sigma} = +1) \geq 1 - \delta, \qquad \mathbb{P}_-(\widehat{\sigma} = +1) \leq \delta.$$

Data processing for relative entropy therefore gives

$$D_{\mathrm{KL}}\left(\mathbb{P}_+^{\mathcal{A}} \middle\| \mathbb{P}_-^{\mathcal{A}}\right) \geq \mathrm{kl}(1-\delta, \delta). \tag{C.27}$$

Combining equation C.27 with Lemma C.7 yields

$$\mathbb{E}_+\left[N_{\mathrm{full}}(\mathsf{T})\right] \geq \frac{\mathrm{kl}(1-\delta,\delta)}{d_{\mathrm{par}}(\varrho)}.$$

Every complete task requires at least r source episodes and one charged reset. Pathwise,

$$\mathsf{C}_{\mathsf{T}} \geq (r+\kappa) N_{\mathrm{full}}(\mathsf{T}),$$

which proves equation C.24 under model $+$. The reverse-model argument is identical because the pairwise KL divergence is symmetric in this family.

For the upper bound, on each of n fresh tasks query every coordinate once and define

$$W_i \coloneqq \prod_{j=1}^{r} (2R_{i,j} - 1) = \chi(X_i).$$

By equation C.13, the variables σW_i are independent, take values in $\{-1, +1\}$, and have mean ϱ. Let

$$\widehat{\sigma}_n \in \operatorname*{arg\,max}_{s \in \{-1,+1\}} s \sum_{i=1}^{n} W_i,$$

with arbitrary tie breaking, and recommend $g^\star_{\widehat{\sigma}_n}$ from Lemma C.8. Hoeffding's inequality gives

$$\mathbb{P}_\sigma(\widehat{\sigma}_n \neq \sigma) \leq \mathbb{P}_\sigma\left(\frac{1}{n}\sum_{i=1}^{n} \sigma W_i - \varrho \leq -\varrho\right) \leq \exp\left(-\frac{n\varrho^2}{2}\right),$$

proving equation C.25. The rate equation C.26 follows from equation C.12 and the standard asymptotics $\mathrm{kl}(1-\delta,\delta) = \Theta(\log(1/\delta))$ as $\delta \downarrow 0$. □

Corollary C.12 (Plug-in target-regret bound)**.** *For the complete-block procedure in the proof of Theorem C.11,*

$$\sup_{\sigma \in \{-1,+1\}} \mathbb{E}_\sigma\left[\mathcal{L}_{r-1}(\sigma, g^\star_{\widehat{\sigma}_n}) - \mathcal{L}^\star_{r-1}(\sigma)\right] \leq \frac{\varrho}{2^r} \exp\left(-\frac{n\varrho^2}{2}\right). \tag{C.28}$$

Proof. If $\widehat{\sigma}_n = \sigma$, the recommended strategy is optimal. If $\widehat{\sigma}_n = -\sigma$, it differs from the optimal strategy only on the history where the $r-1$ revealed target coordinates are all -1, an event of probability $2^{-(r-1)}$. On that history, the expected deployment reward is smaller by $\varrho/2$. Thus the conditional excess target loss given a wrong sign is exactly $\varrho/2^r$. Multiplying by equation C.25 proves the claim. □

Remark C.13 (Exponential scaling as an order parameter). The primitive statistical signal in this construction is the parity bias ϱ. If one parameterizes it as $\varrho = \rho^r$ for a fixed $\rho \in (0,1)$, then equation C.26 becomes

$$\Theta\left((r+\kappa)\rho^{-2r} \log \frac{1}{\delta}\right).$$

This reparameterization is useful for exhibiting an exponentially worsening family of persistence orders, but all exact statements above are expressed in the intrinsic signal parameter ϱ.

D COMPLETE BOUNDED EXPERIMENT TREES

This appendix removes the supplied-menu restriction from Section T. Throughout this appendix the source interface is finite: the global-model class Θ, physical state, action, observation, and reward alphabets are finite; the physical episode horizon is H; and a single source task may be continued for at most L physical episodes. The complete class of deterministic bounded experiment trees is therefore finite. We prove that this class is outcome-equivalent to the entire class of randomized, history-dependent source experiments with the same depth bound. Consequently, the bounded-stopping characteristic time from Section T is the exact complexity of the complete bounded reset-or-continue problem, rather than of a hand-specified experiment menu.

D.1 FINITE INFORMATION HISTORIES AND DETERMINISTIC TREES

Let $\mathcal{Y}$ be the finite alphabet of complete physical-episode transcripts, including all actions, observations, rewards, and terminal states. Let Π_{det} be the finite set of deterministic behavioral policies for one physical episode. Finiteness follows because the physical horizon and all primitive alphabets are finite. A within-task history of depth t is

$$h_t = (\pi_1, y_1, \ldots, \pi_t, y_t) \in (\Pi_{\mathrm{det}} \times \mathcal{Y})^t, \qquad 0 \le t \le L,$$

with $h_0 = \varnothing$. We write $h \oplus (\pi, y)$ for concatenation. Under global model θ, the conditional law of the next trajectory after history h and policy π is denoted

$$P_\theta^\pi(\cdot \mid h).$$

This law integrates the posterior distribution of the task latent given the same-task history. On histories with zero P_θ-probability we fix an arbitrary version; all expectations below are under θ and hence do not depend on that choice.

Definition D.1 (Deterministic bounded experiment tree)**.** A deterministic depth-L experiment tree e assigns:

1. to the root h_0, a policy $\pi_e(h_0) \in \Pi_{\mathrm{det}}$;
2. to every history h_t, $1 \le t < L$, either the symbol stop or a policy $\pi_e(h_t) \in \Pi_{\mathrm{det}}$;
3. the symbol stop to every history of depth L.

The tree begins with a fresh task, executes the prescribed policies, and stops at

$$\tau_e := \inf\{t \ge 1 : e(h_t) = \mathsf{stop}\} \le L.$$

Its stopped transcript is $D_e = h_{\tau_e}$ and its source cost is $C_e = \kappa + \tau_e$. The finite set of all such trees is denoted $\mathfrak{E}_L$.

The law of (D_e, C_e) under model θ is Q_θ^e, with expected cost $c_\theta(e) = \mathbb{E}_\theta C_e$ and pairwise divergence

$$d_e(\theta, \theta') = D_{\mathrm{KL}}(Q_\theta^e \| Q_{\theta'}^e).$$

Notice that a tree may choose different physical policies after different same-task trajectories and may stop at different depths on different paths.

D.2 RANDOMIZED EXPERIMENTS AND PRIVATE RANDOM TAPES

A randomized bounded experiment is implemented by a private random tape U with a model-independent law, independent of the task and environment, together with deterministic measurable choice maps. Given $U = u$, those maps specify, at every within-task history and every primitive within-episode information set, the next action and whether to stop or continue. This representation allows arbitrary correlation among random choices at different information sets. In the finite setting it entails no loss: a finite family of stochastic kernels can be implemented by inverse-transform sampling from one sufficiently rich random tape.

Equivalently, one may describe fresh behavioral kernels

$$b_0(\cdot \mid h_0) \in \Delta(\Pi_{\mathrm{det}}), \qquad b_t(\cdot \mid h_t) \in \Delta(\{\mathsf{stop}\} \cup \Pi_{\mathrm{det}}),$$

but the random-tape formulation also covers mixed strategies and algorithms that deliberately correlate their private choices across histories.

Theorem D.2 (Constructive randomized-to-mixed equivalence)**.** *For every randomized bounded experiment, there exists a probability distribution $\nu \in \Delta(\mathfrak{E}_L)$, independent of the global model, such that simultaneously for every $\theta \in \Theta$,*

$$\mathcal{L}_\theta(D, C) = \sum_{e \in \mathfrak{E}_L} \nu(e) Q_\theta^e. \tag{D.1}$$

Conversely, every model-independent mixture over $\mathfrak{E}_L$ is realized by a randomized bounded experiment.

Proof. Fix one realization u of the private random tape. At every finite within-task history, including histories that are unreachable under some or all models, evaluate the experiment's deterministic choice map with tape u. At every primitive information set inside the chosen physical episode, evaluate the corresponding deterministic action map with the same tape. The result is one complete contingent plan and hence one deterministic tree $e(u) \in \mathfrak{E}_L$. Let ν be the pushforward of the model-independent law of U under $u \mapsto e(u)$.

Conditional on $U = u$, the randomized experiment and the pure tree $e(u)$ make exactly the same action, policy, and stop decisions on every realized path. They therefore have identical stopped transcript and cost, pathwise, under every global model. Integrating over U gives equation D.1. This argument preserves arbitrary correlations in the random tape and does not condition on positive-probability histories.

Conversely, draw $e \sim \nu$ privately at task initiation and execute its prescriptions. This is a valid randomized experiment and has law $\sum_e \nu(e) Q_\theta^e$ under every θ. □

Remark D.3 (Relation to perfect recall). The result is a constructive finite perfect-recall equivalence in the spirit of Kuhn's theorem (Kuhn, 1953). The random-tape proof is stronger than a product-of-local-kernels argument for the present purpose: it preserves correlated private randomization, defines actions on zero-probability histories, and preserves the full stopped transcript and random source cost.

D.3 Reduction of arbitrary bounded source learners

A depth-L source learner may use the complete history of previously completed tasks when selecting its next within-task behavioral experiment. It may randomize, choose physical policies adaptively inside the current task, and stop the current task at any observed history before depth L. It may only switch to another source task by ending the current experiment and resetting; it may not interrupt an experiment and later resume its latent task.

Proposition D.4 (Blockwise reduction to predictable pure-tree selection)**.** *For every depth-L randomized source learner $\mathcal{A}$, there is a learner $\widetilde{\mathcal{A}}$ using the same model-independent private random tape that, before each fresh task, selects a tree $E_k \in \mathfrak{E}_L$ predictably from the completed source history and its private tape and then executes that tree. Under every $\theta \in \Theta$, the two learners have the same joint law of the complete source transcript, total cost, source stopping decision, and target recommendation.*

Proof. Sample at time zero a private random tape U rich enough to implement every randomized choice of $\mathcal{A}$. Conditional on $U = u$, the learner is a deterministic function of its complete observed history. At the beginning of block k, fix the completed history and evaluate that deterministic function at every possible within-task continuation history up to depth L. This produces a pure tree $E_k(u) \in \mathfrak{E}_L$. The new learner uses the same u, selects this tree, and executes it. Conditional on $U = u$, the original and reduced learners agree pathwise throughout the block, including on the history at which the task and possibly the entire source procedure stop. Induction over blocks gives pathwise equality of the source transcript, cost, stopping decision, and recommendation. Integrating over the common tape gives the stated joint-law equality under every model. For likelihood arguments, the private tape is included in the augmented algorithm filtration; its model-independent law cancels from every likelihood ratio. □

D.4 Observational quotient and exact characteristic time

Define the depth-L observational equivalence relation

$$\theta \sim_L \theta' \quad \Longleftrightarrow \quad Q^e_\theta = Q^e_{\theta'}, \quad \forall e \in \mathfrak{E}_L, \tag{D.2}$$

and let $\overline{\Theta}_L = \Theta/\!\sim_L$. If $\xi = [\theta]$, set $Q^e_\xi = Q^e_\theta$ and $c_\xi(e) = c_\theta(e)$. These definitions are independent of the representative because cost is a measurable component of the stopped outcome. Define

$$d_e(\xi, \xi') = D_{\mathrm{KL}}(Q^e_\xi \| Q'^e_\xi).$$

Distinct quotient classes are separated by at least one tree by definition.

Assumption D.5 (Single-answer quotient partition)**.** There are a finite answer set $\mathcal{I}$, a surjective map $\bar{a} : \overline{\Theta}_L \to \mathcal{I}$, and canonical target strategies $(g_i)_{i\in\mathcal{I}}$ such that:

1. $g_{\bar{a}(\xi)}$ is ε-optimal for every original model $\theta \in \xi$;
2. whenever $\bar{a}(\xi) \neq \bar{a}(\xi')$, the acceptable strategy sets of every $\theta \in \xi$ and $\theta' \in \xi'$ are disjoint.

Then

$$\mathcal{J}_\xi = \{\xi' \in \overline{\Theta}_L : \bar{a}(\xi') \neq \bar{a}(\xi)\} \tag{D.3}$$

is exactly the set of pairwise target-incompatible quotient classes.

Define

$$\Gamma^\star_L(\xi) \coloneqq \sup_{p\in\Delta(\mathfrak{E}_L)} \inf_{\xi'\in\mathcal{J}_\xi} \frac{\sum_e p_e d_e(\xi,\xi')}{\sum_e p_e c_\xi(e)}. \tag{D.4}$$

Let $\mathfrak{C}^{L,\star}_{B,\varepsilon,\delta,\kappa}(\xi)$ be the infimum expected source cost over globally correct learners whose every task block stops by depth L, evaluated at any representative of ξ.

Theorem D.6 (Quotient-native complete-tree characteristic time)**.** *Assume: (i) the finite source interface and complete tree class above; (ii) the finite quotient $\overline{\Theta}_L$ and Assumption D.5; (iii) common domination and finite ordered KL divergences for every pure tree and quotient pair; (iv) deterministic tie-breaking and exact access to quotient tree likelihoods, the quotient MLE, and the finite Bellman–Chernoff design game; and (v) the uniform conditional sub-gamma condition of Assumption E.8. Then, for every ξ with nonempty $\mathcal{J}_\xi$ and $\Gamma^\star_L(\xi) > 0$,*

$$\lim_{\delta\downarrow 0} \frac{\mathfrak{C}^{L,\star}_{B,\varepsilon,\delta,\kappa}(\xi)}{\log(1/\delta)} = \frac{1}{\Gamma^\star_L(\xi)}. \tag{D.5}$$

The lower bound ranges over all randomized history-dependent depth-L learners, and the upper bound is attained by exact-design quotient Track-and-Stop. If every quotient class is a singleton, this reduces to the globally identifiable formulation.

Proof. By Proposition D.4, every admissible learner is outcome-equivalent under every original model to predictable randomized selection from $\mathfrak{E}_L$. Quotienting leaves the complete stopped source law, source cost, stopping event, and target validity unchanged. The stopped change-of-measure lower bound therefore applies to the finite quotient model and gives the lower half of equation D.5. Conversely, every pure tree is admissible. The exact-design result Theorem E.13, applied to quotient laws, attains the reciprocal information constant. The canonical strategy returned for a quotient cell is valid for every original representative by Assumption D.5. Hence the lower and upper optimization domains coincide exactly. □

The exponentially large class is accessed by Bellman pricing and column generation. Appendix E proves representative invariance of conditional laws, minimax duality, sparse primal recovery, and separate exact- and approximate-design Track-and-Stop results.

E Bellman–Chernoff Optimal Stopping and Computation

This appendix characterizes and computes the quotient-native complete-tree information in equation D.4. Original global models are denoted by θ, while observational quotient classes are denoted by $\xi, \zeta \in \overline{\Theta}_L$. All likelihoods, costs, answer cells, and Bellman values below are functions of quotient classes. The exact and certified approximate-design upper theorems are stated separately.

E.1 Representative invariance and stopped information

Fix a true class $\xi \in \overline{\Theta}_L$ and let

$$\mathcal{J}_\xi = \{\zeta \in \overline{\Theta}_L : \bar{a}(\zeta) \neq \bar{a}(\xi)\}.$$

By the single-answer partition, this is exactly the set of pairwise target-incompatible quotient classes.

Lemma E.1 (Representative invariance of reachable conditional laws)**.** *Let $\theta_1, \theta_2 \in \xi$. Fix a same-task history h that has positive probability under some deterministic tree and either representative, and fix a pure episodic policy π. Then*

$$P^\pi_{\theta_1}(\cdot \mid h) = P^\pi_{\theta_2}(\cdot \mid h).$$

Consequently every reachable conditional divergence, expected tree cost, Bellman value, and Bellman root depends only on ξ. On histories null under the class, we use the conditional kernel of a fixed canonical representative; these values enter no expectation under ξ.

Proof. Choose a deterministic tree that generates the policy sequence encoded by h, executes π after h, and then stops. Equality of every complete tree law for representatives in one quotient class implies equality of the probability of the prefix cylinder h and every extended cylinder $h \oplus (\pi, y)$. The common prefix mass is positive, so division yields the conditional-law identity. Cost is measurable from the stopped transcript. Backward induction in the Bellman recursion then gives invariance of all values and roots. □

For $\lambda \in \Delta(\mathcal{J}_\xi)$, history h, and $\pi \in \Pi_{\mathrm{det}}$, define

$$\iota_{\xi,\lambda}(h, \pi) \coloneqq \sum_{\zeta \in \mathcal{J}_\xi} \lambda_\zeta D_{\mathrm{KL}}\left(P^\pi_\xi(\cdot \mid h) \middle\| P^\pi_\zeta(\cdot \mid h)\right). \tag{E.1}$$

For a pure tree e, write

$$D_{\xi,\lambda}(e) \coloneqq \sum_{\zeta \in \mathcal{J}_\xi} \lambda_\zeta d_e(\xi, \zeta). \tag{E.2}$$

Lemma E.2 (Prefix absolute continuity and pathwise factorization)**.** *Fix a tree e and quotient classes ξ, ζ with $d_e(\xi, \zeta) < \infty$. Every tree prefix having positive Q^e_ξ-probability has positive Q^e_ζ-probability. Moreover, on a Q^e_ξ-full set,*

$$\log \frac{dQ^e_\xi}{dQ^e_\zeta}(D_e) = \sum_{t=0}^{\tau_e - 1} \log \frac{P^{\pi_e(H_t)}_\xi(Y_{t+1} \mid H_t)}{P^{\pi_e(H_t)}_\zeta(Y_{t+1} \mid H_t)}. \tag{E.3}$$

Proof. Finite KL implies $Q^e_\xi \ll Q^e_\zeta$. A positive-probability prefix is a positive-mass cylinder under Q^e_ξ and therefore also has positive Q^e_ζ-mass. Since the deterministic policy and stopping prescriptions are the same under both classes, their factors cancel from the stopped-path likelihood. Applying the finite chain rule along every supported prefix gives equation E.3. □

Lemma E.3 (Stopped conditional KL and cost decomposition)**.** *For every tree e and mixture λ,*

$$D_{\xi,\lambda}(e) = \mathbb{E}^e_\xi\left[\sum_{t=0}^{\tau_e - 1} \iota_{\xi,\lambda}(H_t, \pi_e(H_t))\right], \tag{E.4}$$

$$c_\xi(e) = \kappa + \mathbb{E}^e_\xi \tau_e. \tag{E.5}$$

Proof. Take expectations in equation E.3, use conditional KL at each prefix, and then average over λ. Bounded depth justifies finite interchange. The cost identity follows from $C_e = \kappa + \tau_e$. □

E.2 Bellman equation and fractional root

For $\gamma \in \mathbb{R}$, define

$$V_L^{\lambda,\gamma}(h) = 0, \tag{E.6}$$

$$V_t^{\lambda,\gamma}(h) = \max\left\{0, \max_{\pi\in\Pi_{\det}}\left[\iota_{\xi,\lambda}(h,\pi) - \gamma + \mathbb{E}_\xi V_{t+1}^{\lambda,\gamma}(h\oplus(\pi,Y))\right]\right\}, \quad 1 \le t < L, \tag{E.7}$$

and the fresh-task root value

$$\Phi_{\xi,\lambda}(\gamma) = -\gamma\kappa + \max_{\pi\in\Pi_{\det}}\left[\iota_{\xi,\lambda}(\varnothing,\pi) - \gamma + \mathbb{E}_\xi V_1^{\lambda,\gamma}((\pi,Y))\right]. \tag{E.8}$$

Theorem E.4 (Bellman representation and fractional root)**.** *For every λ and γ,*

$$\Phi_{\xi,\lambda}(\gamma) = \max_{e\in\mathfrak{E}_L}\{D_{\xi,\lambda}(e) - \gamma c_\xi(e)\}. \tag{E.9}$$

The function is continuous and strictly decreasing, with unique zero

$$\gamma_\lambda^\star = \max_{e\in\mathfrak{E}_L}\frac{D_{\xi,\lambda}(e)}{c_\xi(e)}. \tag{E.10}$$

A deterministic tree obtained from maximizing Bellman actions attains the maximum.

Proof. Backward induction gives the optimal conditional future weighted KL minus γ times future episode count; the root additionally pays reset cost $\gamma\kappa$. This proves equation E.9. Since the tree class is finite and $c_\xi(e) \ge 1+\kappa$, the maximum of the affine functions $c_\xi(e)(D_{\xi,\lambda}(e)/c_\xi(e) - \gamma)$ is continuous and strictly decreasing, positive below the largest ratio, zero at it, and negative above. □

E.3 Minimax duality and sparse optimal designs

Set $A_{e,\zeta} = d_e(\xi,\zeta)/c_\xi(e)$. Changing from tree-launch probabilities p to expected-cost fractions $x_e = p_e c_\xi(e)/(\sum_f p_f c_\xi(f))$ gives

$$\Gamma_L^\star(\xi) = \max_{x\in\Delta(\mathfrak{E}_L)}\min_{\zeta\in\mathcal{J}_\xi}\sum_e x_e A_{e,\zeta}, \tag{E.11}$$

$$= \min_{\lambda\in\Delta(\mathcal{J}_\xi)}\max_{e\in\mathfrak{E}_L}\sum_\zeta \lambda_\zeta A_{e,\zeta} = \min_\lambda \gamma_\lambda^\star. \tag{E.12}$$

The inverse transformation is $p_e \propto x_e/c_\xi(e)$.

Theorem E.5 (Bellman–Chernoff characterization and sparse support)**.** *The primal and dual values in equations (E.11)–(E.12) are equal and attained. Complementary slackness identifies the active trees and alternatives. There exists an optimal primal design supported on no more than the number of linearly independent active alternative constraints; in particular, support size is at most $|\mathcal{J}_\xi|$.*

Proof. Finite minimax duality gives equality and attainment. For sparsity, write the primal as an LP maximizing $\gamma \ge 0$ subject to $\sum_e x_e A_{e,\zeta} \ge \gamma$, $\sum_e x_e = 1$, and $x_e \ge 0$. Add one slack per alternative. At an optimal basic feasible solution, the positive variable γ consumes one basic coordinate, and each inactive alternative has a positive slack. Thus at most the rank of the active alternative constraints can be occupied by positive tree weights. The inverse cost-normalization preserves support. □

E.4 Certified column generation

Let $\mathcal{E}_k \subseteq \mathfrak{E}_L$ be a nonempty restricted column set. Solve the restricted primal and dual games exactly, obtaining a feasible cost-fraction vector x_k, restricted value v_k, and dual mixture $\lambda_k \in \Delta(\mathcal{J}_\xi)$. The full pricing value is

$$\rho_k := \max_{e\in\mathfrak{E}_L}\lambda_k^\top A_e.$$

A certified approximate Bellman oracle returns a tree e_k and value $\widetilde{\rho}_k = \lambda_k^\top A_{e_k}$ together with a deterministic certificate

$$\rho_k \le \widetilde{\rho}_k + \epsilon_k^{\mathrm{P}}. \tag{E.13}$$

Algorithm 2 Certified Bellman–Chernoff column generation

Require: Nonempty initial columns $\mathcal{E}_1$; certified pricing errors ϵ_k^{P}; target gaps η_k.

for $k = 1, 2, \dots$ **do**
  solve the restricted primal and dual games exactly, obtaining (x_k, λ_k, v_k)
  run certified Bellman pricing, obtaining $(e_k, \widetilde{\rho}_k, \epsilon_k^{\mathrm{P}})$ satisfying equation E.13
  if $\widetilde{\rho}_k + \epsilon_k^{\mathrm{P}} \le v_k + \eta_k$ **then**
    convert cost fractions to tree-launch probabilities

$$p_{k,e} = \frac{x_{k,e}/c_\xi(e)}{\sum_{f \in \mathcal{E}_k} x_{k,f}/c_\xi(f)} \qquad (e \in \mathcal{E}_k)$$

    return $(p_k, x_k, \lambda_k, v_k)$ and the certificate $0 \le \Gamma_L^\star(\xi) - v_k \le \eta_k$
  else if $e_k \in \mathcal{E}_k$ **then**
    decrease ϵ_k^{P} and rerun pricing
  else
    $\mathcal{E}_{k+1} \leftarrow \mathcal{E}_k \cup \{e_k\}$
  end if
end for

Theorem E.6 (Finite column generation and global certificate)**.** *At every iteration,*

$$v_k \le \Gamma_L^\star(\xi) \le \rho_k \le \widetilde{\rho}_k + \epsilon_k^{\mathrm{P}}. \tag{E.14}$$

Consequently, the displayed stopping test returns an η_k-optimal design. With exact pricing, every nonterminal iteration adds a new tree, so the algorithm terminates after at most $|\mathfrak{E}_L| - |\mathcal{E}_1|$ strict additions and returns an exact saddle point. One Bellman evaluation costs $O(|\mathcal{H}_{<L}|\,|\Pi_{\mathrm{det}}|\,|\mathcal{Y}|\,|\mathcal{J}_\xi|)$ arithmetic operations after conditional laws are tabulated; bisection adds a logarithmic factor in the pricing precision.

Proof. The restricted primal is feasible for the full primal and therefore gives the lower bound $v_k \le \Gamma_L^\star(\xi)$. The restricted dual mixture λ_k is feasible for the full dual; maximizing its payoff over all columns gives the full-dual upper bound ρ_k. The pricing certificate proves the last inequality in equation E.14. Hence the stopping condition implies $0 \le \Gamma_L^\star(\xi) - v_k \le \eta_k$. Under exact pricing, if the stopping condition fails, any maximizing tree not already in $\mathcal{E}_k$ violates the current restricted dual bound; if a maximizer is already present, the restricted primal-dual pair is already globally optimal. Thus every strict nonterminal iteration adds a new column, and finiteness gives termination. The complexity is one backward pass over every reachable history-policy-next-observation tuple and alternative quotient class. □

Remark E.7 (Approximately solved restricted masters)*.* Algorithm 2 uses an exact restricted master to keep the global certificate transparent. If an LP solver returns a primal feasibility residual r_k^{P}, a dual feasibility residual r_k^{D}, and a restricted primal-dual gap r_k^{M}, these quantities must be added explicitly to the returned global gap. An unused "master tolerance" is not a certificate.

E.5 Likelihood moments, separating exploration, and the online learner

Assumption E.8 (Uniform conditional sub-gamma likelihood ratios)**.** There exist $v, b < \infty$ such that, for every pure tree e, ordered quotient pair (ξ, ζ), and $|s| < 1/b$,

$$\log \mathbb{E}_\xi \exp\{s(\ell_e^{\xi,\zeta}(D_e) - d_e(\xi,\zeta))\} \le \frac{s^2 v^2}{2(1 - b|s|)}. \tag{E.15}$$

Every tree cost lies in $[c_{\min}, c_{\max}] = [1 + \kappa, L + \kappa]$.

Because E_k is selected predictably from the completed-block history and the next task block is conditionally independent of the past given E_k, the uniform treewise bound equation E.15 is also the conditional MGF bound required for the adaptive block sequence.

Algorithm 3 Quotient Complete-Tree Track-and-Stop

Require: δ, depth L, separating set $\mathcal{E}_{\mathrm{sep}}$, $\gamma \in (1/2, 1)$, and certified design gaps η_k.

for $k = 0, 1, \dots$ **do**
  if $\lfloor (k+1)^\gamma \rfloor > \lfloor k^\gamma \rfloor$ **then**
    $E_{k+1} \leftarrow$ next tree in a cyclic traversal of $\mathcal{E}_{\mathrm{sep}}$
  else
    compute the deterministic-tie-broken quotient MLE $\widehat{\xi}_k$
    warm-start column generation at $\widehat{\xi}_k$ to gap η_k
    convert its cost-fraction vector x_k to launch probabilities $p_{k,e} \propto x_{k,e}/c_{\widehat{\xi}_k}(e)$
    sample $E_{k+1} \sim p_k$
  end if
  reset once and execute E_{k+1} once to its built-in stopping time
  observe (D_{k+1}, C_{k+1}) and update quotient likelihoods
  compute quotient decision-cell GLRs $Z_{k+1}(i)$
  if $\max_i Z_{k+1}(i) \geq \log((|\overline{\Theta}_L| - 1)/\delta)$ **then**
    return the canonical target strategy of a maximizing cell
  end if
end for

For a completed sequence $(E_s, D_s, C_s)_{s \leq k}$ define the quotient log likelihood and answer-cell GLR

$$\mathsf{L}_k(\zeta) \coloneqq \sum_{s=1}^{k} \log q_\zeta^{E_s}(D_s), \tag{E.16}$$

$$Z_k(i) \coloneqq \max_{\zeta:\, \bar{a}(\zeta)=i} \mathsf{L}_k(\zeta) - \max_{\zeta:\, \bar{a}(\zeta)\neq i} \mathsf{L}_k(\zeta), \tag{E.17}$$

with $\max \varnothing = -\infty$. The quotient MLE is the first maximizer of L_k in a fixed total order.

For each ordered pair $\xi \neq \zeta$, choose $e_{\xi,\zeta} \in \arg\max_e d_e(\xi, \zeta)$ and set $\mathcal{E}_{\mathrm{sep}} = \{e_{\xi,\zeta} : \xi \neq \zeta\}$. Then $|\mathcal{E}_{\mathrm{sep}}| \leq |\overline{\Theta}_L|(|\overline{\Theta}_L| - 1)$.

Lemma E.9 (Adaptive Bernstein inequality)**.** *If martingale differences X_k satisfy the conditional form of equation E.15, then*

$$\mathbb{P}\left(\sum_{k=1}^{n} X_k \leq -x\right) \leq \exp\left\{-\frac{x^2}{2(nv^2 + bx)}\right\},$$

with the analogous upper-tail bound.

Proof. Exponentiate the conditional mgf bound to obtain a supermartingale, apply Markov's inequality, and optimize at $s = x/(nv^2 + bx)$. □

Lemma E.10 (Finite expected pre-identification cost)**.** *If $\mathbb{P}_\xi(\widehat{\xi}_k \neq \xi) \leq Ae^{-bk^\alpha}$, then for $\tau_{\mathrm{last}} = \sup\{k : \widehat{\xi}_k \neq \xi\}$,*

$$\mathbb{E}_\xi \tau_{\mathrm{last}} \leq \sum_{k \geq 1} k\, \mathbb{P}_\xi(\widehat{\xi}_k \neq \xi) < \infty, \qquad \mathbb{E}_\xi \mathsf{C}_{\tau_{\mathrm{last}}} < \infty.$$

Proof. Use the tail-sum formula, a union bound over future error times, and Fubini; then multiply by $c_{\max}$. □

E.6 Exact-design characteristic time

Let $p^\star(\xi)$ be a fixed deterministic-tie-broken optimal tree-launch distribution obtained from an optimal cost-fraction solution through $p_e^\star \propto x_e^\star / c_\xi(e)$. Define

$$\bar{c}_\xi^\star \coloneqq \sum_e p_e^\star c_\xi(e), \tag{E.18}$$

$$I_\xi(\zeta) \coloneqq \sum_e p_e^\star d_e(\xi, \zeta), \qquad I_\xi^\star \coloneqq \min_{\zeta \in \mathcal{J}_\xi} I_\xi(\zeta). \tag{E.19}$$

Then $I_\xi^\star = \Gamma_L^\star(\xi)\bar{c}_\xi^\star$. For completed blocks write $\mathsf{C}_k = \sum_{s=1}^k C_s$ and $N_e(k) = \sum_{s=1}^k \mathbf{1}\{E_s = e\}$.

Lemma E.11 (Exact-design allocation, information, and cost tracking)**.** *Under the assumptions of Theorem E.13, on the common infinite process,*

$$\frac{N_e(k)}{k} \longrightarrow p_e^\star(\xi) \qquad \text{for every } e, \tag{E.20}$$

$$\frac{1}{k}\sum_{s=1}^k d_{E_s}(\xi,\zeta) \longrightarrow I_\xi(\zeta) \qquad \text{for every } \zeta \in \mathcal{J}_\xi, \tag{E.21}$$

$$\frac{\mathsf{C}_k}{k} \longrightarrow \bar{c}_\xi^\star \qquad \text{almost surely and in } L^1. \tag{E.22}$$

Moreover, for every fixed tolerance in these displays, the corresponding failure probability at block k is bounded by $A\exp(-bk^{2\gamma-1})$ for all sufficiently large k.

Proof. Cyclic forced exploration uses $\lfloor k^\gamma \rfloor = o(k)$ blocks. Adaptive Bernstein concentration on the separating trees gives $\mathbb{P}_\xi(\widehat{\xi}_k \neq \xi) \le A_0 e^{-b_0 k^{2\gamma-1}}$; hence Lemma E.10 implies an almost surely finite last quotient-MLE error and finite expected pre-identification cost. After that last error, every nonforced tree is sampled from the fixed distribution $p^\star(\xi)$. For each tree, the centered selection indicators are bounded martingale differences, so Azuma–Hoeffding and the $o(k)$ forced prefix give equation E.20, with an exponential deviation conditional on quotient stabilization. The finite union over trees and the stretched- exponential last-error tail gives the stated unconditional rate.

Equation E.21 follows by multiplying the frequency limits by the finite numbers $d_e(\xi,\zeta)$. For cost, decompose

$$\mathsf{C}_k = \sum_{s=1}^k c_\xi(E_s) + \sum_{s=1}^k \{C_s - c_\xi(E_s)\}.$$

The first average converges to $\bar{c}_\xi^\star$ by tree-frequency tracking; the second is a bounded martingale average and converges almost surely, in L^1, and with an exponential tail. Since $\mathsf{C}_k/k \in [c_{\min}, c_{\max}]$, the almost-sure convergence also gives L^1 convergence by bounded convergence. □

Proposition E.12 (Uniform evidence growth and stopping-tail decomposition)**.** *Let $\alpha = 2\gamma - 1 \in (0,1)$. For every $\varepsilon_I \in (0, I_\xi^\star)$ and $\varepsilon_c > 0$, there exist $A, b > 0$ and $n_0 < \infty$ such that, for every $n \ge n_0$,*

$$\mathbb{P}_\xi\left\{\exists k \ge n : Z_k(\bar{a}(\xi)) < k(I_\xi^\star - \varepsilon_I)\right\} \le Ae^{-bn^\alpha}, \tag{E.23}$$

$$\mathbb{P}_\xi\left\{\exists k \ge n : \left|\frac{\mathsf{C}_k}{k} - \bar{c}_\xi^\star\right| > \varepsilon_c\right\} \le Ae^{-bn^\alpha}. \tag{E.24}$$

For each deterministic block index k, the failure of the simultaneous bounds is contained in the union of the following five events:

$$\begin{aligned}
E_k^{\mathrm{MLE}} &:= \{\tau_{\mathrm{last}} > \eta_0 k\},\\
E_k^{\mathrm{alloc}} &:= \left\{\min_{\zeta\in\mathcal{J}_\xi}\sum_{s=1}^k d_{E_s}(\xi,\zeta) < k(I_\xi^\star - \eta_d)\right\},\\
E_k^{\mathrm{LLR}} &:= \left\{\exists\zeta\in\mathcal{J}_\xi : \mathsf{L}_k(\xi) - \mathsf{L}_k(\zeta) - \sum_{s=1}^k d_{E_s}(\xi,\zeta) < -\eta_{\mathrm{llr}}k\right\},\\
E_k^{\mathrm{cost}} &:= \left\{\left|\mathsf{C}_k/k - \bar{c}_\xi^\star\right| > \varepsilon_c\right\},\\
E_k^{\mathrm{cell}} &:= \left\{\max_{\zeta:\,\bar{a}(\zeta)=\bar{a}(\xi)} \mathsf{L}_k(\zeta) > \mathsf{L}_k(\xi)\right\},
\end{aligned}$$

where the positive component tolerances are chosen so that $\eta_d + \eta_{\mathrm{llr}} < \varepsilon_I$ and η_0 is small enough for the allocation bound below. Each event has probability at most $A_0 e^{-b_0 k^\alpha}$ for all sufficiently large k, uniformly over the finite tree and quotient classes.

Proof. The MLE event is controlled by the stretched-exponential quotient-MLE tail:

$$\mathbb{P}_\xi(E_k^{\mathrm{MLE}}) \le \sum_{j>\eta_0 k} \mathbb{P}_\xi(\widehat{\xi}_j \neq \xi) \le A_1 e^{-b_1 k^\alpha}.$$

On its complement, at most $\eta_0 k + \lfloor k^\gamma \rfloor$ blocks are pre-identification or forced. Every such block has nonnegative directional KL. On the remaining blocks, the selected tree is sampled from $p^\star(\xi)$. The centered selected-divergence increments are bounded because the tree and quotient classes are finite. Choose η_0 and then k large enough that the lost optimal-design fraction contributes at most $\eta_d/2$ per block. Azuma–Hoeffding, followed by a finite union over $\zeta \in \mathcal{J}_\xi$, gives

$$\mathbb{P}_\xi(E_k^{\mathrm{alloc}} \cap (E_k^{\mathrm{MLE}})^c) \le A_2 e^{-b_2 k}.$$

For each wrong answer class, the centered likelihood increment

$$\ell^{\xi,\zeta}_{E_s}(D_s) - d_{E_s}(\xi,\zeta)$$

satisfies the conditional sub-gamma assumption after predictable tree selection. Lemma E.9 and a finite union therefore give $\mathbb{P}_\xi(E_k^{\mathrm{LLR}}) \le A_3 e^{-b_3 k}$. The bounded cost martingale, combined with the selection-frequency and last-error bounds from Lemma E.11, gives $\mathbb{P}_\xi(E_k^{\mathrm{cost}}) \le A_4 e^{-b_4 k^\alpha}$.

For a distinct quotient class ζ in the true answer cell, cyclic forced exploration includes a tree with positive directed divergence $d_e(\xi,\zeta)$. Its predictable cumulative drift is at least $c_{\xi,\zeta} k^\gamma$ for all large k. Applying Lemma E.9 at a deviation equal to half this drift yields

$$\mathbb{P}_\xi\{\mathsf{L}_k(\zeta) \ge \mathsf{L}_k(\xi)\} \le A_{\xi,\zeta} e^{-b_{\xi,\zeta} k^{2\gamma-1}}.$$

A finite union controls E_k^{cell}.

Outside the five events, ξ maximizes likelihood inside its answer cell and for every wrong-cell class ζ,

$$\mathsf{L}_k(\xi) - \mathsf{L}_k(\zeta) \ge k(I^\star_\xi - \eta_d - \eta_{\mathrm{llr}}) \ge k(I^\star_\xi - \varepsilon_I).$$

Hence the same lower bound holds for $Z_k(\bar{a}(\xi))$. Summing the five pointwise bounds over all $k \ge n$ and using $\sum_{k\ge n} e^{-bk^\alpha} \le A' e^{-b' n^\alpha}$ for large n proves equations (E.23)–(E.24). □

Theorem E.13 (Exact-design quotient Track-and-Stop)**.** *Assume the finite complete-tree interface, finite quotient $\overline{\Theta}_L$, Assumption D.5, common domination and finite ordered KL, Assumption E.8, deterministic tie-breaking, exact quotient likelihood and quotient-MLE access, and exact Bellman–Chernoff design solves. Then Algorithm 3 with exact designs is globally (B,ε,δ)-correct and, for every ξ with $\Gamma^\star_L(\xi) > 0$,*

$$\lim_{\delta\downarrow 0} \frac{\mathbb{E}_\xi \mathsf{C}_{T_\delta}}{\log(1/\delta)} = \frac{1}{\Gamma^\star_L(\xi)}. \tag{E.25}$$

Proof. All confidence levels are coupled on the one infinite completed-tree process obtained by omitting the stopping instruction. Conditional on the completed past and selected tree, the next stopped transcript has law $Q_\xi^{E_k}$. Every pairwise quotient likelihood ratio is therefore a nonnegative martingale. Ville's inequality and a union bound over at most $|\overline{\Theta}_L| - 1$ wrong classes give anytime wrong-cell probability at most δ for every predictable sampling rule.

Cyclic forced exploration and Lemma E.9 imply quotient-MLE stabilization with a stretched-exponential last-error tail. Thus Lemma E.10 gives finite expected source cost before permanent identification. Lemma E.11 gives the almost-sure and L^1 cost law and the predictable information rates. For a wrong answer cell, every class has strictly positive limiting directional KL, so its cell statistic has negative linear rate and finite pathwise supremum. For the true cell, Proposition E.12 gives positive linear evidence. Hence every fixed threshold is crossed almost surely, and on the common process all sufficiently high thresholds are crossed only by the true cell. This proves global fixed-confidence correctness and almost-sure termination for every fixed confidence level.

For the expected cost, fix $\varepsilon_I \in (0, I^\star_\xi)$ and define, without an additive transient,

$$k_\delta(\varepsilon_I) := \left\lceil \frac{\beta_\delta}{I^\star_\xi - \varepsilon_I} \right\rceil, \qquad \beta_\delta = \log\left(\frac{|\overline{\Theta}_L| - 1}{\delta}\right). \tag{E.26}$$

For every $k \geq k_\delta(\varepsilon_I)$, $\{T_\delta > k\}$ implies $Z_k(\bar{a}(\xi)) < k(I^\star_\xi - \varepsilon_I)$. Therefore Proposition E.12 yields

$$\sum_{k \geq k_\delta(\varepsilon_I)} \mathbb{P}_\xi(T_\delta > k) = o(\beta_\delta). \tag{E.27}$$

Using the common infinite process and positive bounded costs,

$$\mathbb{E}_\xi \mathsf{C}_{T_\delta} \leq \mathbb{E}_\xi \mathsf{C}_{k_\delta(\varepsilon_I)} + c_{\max} \sum_{k \geq k_\delta(\varepsilon_I)} \mathbb{P}_\xi(T_\delta > k). \tag{E.28}$$

The L^1 convergence in equation E.22 gives

$$\mathbb{E}_\xi \mathsf{C}_{k_\delta(\varepsilon_I)} = \bar{c}^\star_\xi k_\delta(\varepsilon_I) + o(k_\delta(\varepsilon_I)).$$

Combining with equation E.27,

$$\limsup_{\delta \downarrow 0} \frac{\mathbb{E}_\xi \mathsf{C}_{T_\delta}}{\beta_\delta} \leq \frac{\bar{c}^\star_\xi}{I^\star_\xi - \varepsilon_I}.$$

Let $\varepsilon_I \downarrow 0$, use $I^\star_\xi = \Gamma^\star_L(\xi)\bar{c}^\star_\xi$ and $\beta_\delta / \log(1/\delta) \to 1$, and combine with the complete-tree lower bound. □

E.7 Certified approximate designs

Let $\mathcal{G}_{k-1}$ be the sigma-field immediately before tree E_k is sampled. For every block k, let p_k be the $\mathcal{G}_{k-1}$-measurable launch distribution; on a forced-exploration block, p_k is the point mass at the prescribed separating tree. Define

$$\bar{c}_k := \sum_e p_{k,e} c_\xi(e), \tag{E.29}$$

$$m_k(\zeta) := \sum_e p_{k,e} d_e(\xi, \zeta), \qquad \zeta \in \mathcal{J}_\xi. \tag{E.30}$$

After quotient-MLE stabilization, a nonforced design is ζ_k-optimal if

$$\inf_{\zeta \in \mathcal{J}_\xi} \frac{m_k(\zeta)}{\bar{c}_k} \geq \Gamma^\star_L(\xi) - \zeta_k. \tag{E.31}$$

The certificate ζ_k may combine the returned global column-generation gap, pricing error, and any explicitly reported LP feasibility residuals. Conditionally on $\mathcal{G}_{k-1}$, the centered increment $\ell^{\xi,\zeta}_{E_k}(D_k) - m_k(\zeta)$ is sub-gamma with constants that are uniform in k: condition first on the sampled tree, apply Assumption E.8, and absorb the bounded finite-menu selection term $d_{E_k}(\xi, \zeta) - m_k(\zeta)$ into the variance and scale parameters.

Lemma E.14 (Deterministic certificate debt)**.** *Suppose* $0 \leq \zeta_k \leq \bar{\zeta}_k$ *and* $n^{-1} \sum_{k=1}^n \bar{\zeta}_k \to 0$*. For every* $\varepsilon > 0$*, define*

$$D_\varepsilon := \sup_{n \geq 1} \left[c_{\max} \sum_{k=1}^n \bar{\zeta}_k - \varepsilon c_{\min} n \right]_+ . \tag{E.32}$$

Then $D_\varepsilon < \infty$ *and, for every predictable cost sequence* $c_{\min} \leq \bar{c}_k \leq c_{\max}$*,*

$$\sum_{k=1}^n \bar{c}_k \zeta_k \leq \varepsilon \sum_{k=1}^n \bar{c}_k + D_\varepsilon \qquad (n \geq 1). \tag{E.33}$$

Proof. Cesàro convergence implies that for all sufficiently large n, $c_{\max} \sum_{k \leq n} \bar{\zeta}_k \leq \varepsilon c_{\min} n$, so the supremum in equation E.32 is over a finite effective prefix and is finite. Furthermore,

$$\sum_{k \leq n} \bar{c}_k \zeta_k \leq c_{\max} \sum_{k \leq n} \bar{\zeta}_k \leq \varepsilon c_{\min} n + D_\varepsilon \leq \varepsilon \sum_{k \leq n} \bar{c}_k + D_\varepsilon.$$

□

Lemma E.15 (Predictable-cost comparison time)**.** *Let $\bar{c}_k$ be measurable before block k and satisfy $c_{\min} \le \bar{c}_k \le c_{\max}$. For $B > 0$, define*

$$\nu(B) \coloneqq \inf\left\{ n : \sum_{k=1}^{n} \bar{c}_k \ge B \right\}. \tag{E.34}$$

Then $\nu(B)$ is a bounded stopping time for the completed-block filtration,

$$\frac{B}{c_{\max}} \le \nu(B) \le \frac{B}{c_{\min}} + 1, \qquad \mathbb{E}_\xi \sum_{k=1}^{\nu(B)} (C_k - \bar{c}_k) = 0, \tag{E.35}$$

and, for arbitrary events E_n,

$$\mathbb{P}_\xi\{\exists n \ge \nu(B) : E_n\} \le \sum_{n \ge \lfloor B/c_{\max} \rfloor} \mathbb{P}_\xi(E_n). \tag{E.36}$$

Moreover,

$$\mathbb{E}_\xi \mathsf{C}_{\nu(B)} = \mathbb{E}_\xi \sum_{k=1}^{\nu(B)} \bar{c}_k \le B + c_{\max}. \tag{E.37}$$

Proof. The deterministic bounds follow from the first crossing and the uniform cost bounds; in particular $\nu(B)$ is bounded by a deterministic integer. Since $C_k - \bar{c}_k$ is a bounded martingale difference, optional sampling at this bounded stopping time gives the expectation identity. At the crossing, the predictable cumulative cost overshoots B by at most $c_{\max}$, proving equation E.37. Finally, $\nu(B) \ge B/c_{\max}$ gives the event inclusion underlying equation E.36. □

Lemma E.16 (Uniform approximate-design evidence)**.** *Let the hypotheses of Theorem E.17 hold. Fix $\varepsilon > 0$ and set*

$$C_\xi \coloneqq 1 + \frac{2\Gamma^\star_L(\xi) c_{\max}}{c_{\min}}, \qquad r_\varepsilon \coloneqq \Gamma^\star_L(\xi) - C_\xi \varepsilon. \tag{E.38}$$

For ε small enough that $r_\varepsilon > 0$, there exist $A, b, n_0 > 0$ such that, for every $n \ge n_0$,

$$\mathbb{P}_\xi\left\{ \exists m \ge n : Z_m(\bar{a}(\xi)) < r_\varepsilon \sum_{k=1}^{m} \bar{c}_k - D_\varepsilon \right\} \le A e^{-b n^{2\gamma - 1}}. \tag{E.39}$$

Proof. Let $F(m) = \lfloor m^\gamma \rfloor$ be the number of forced blocks and let τ_{last} be the final quotient-MLE error. On the event $\tau_{\text{last}} \le \varepsilon m$ and for all sufficiently large m with $F(m) \le \varepsilon m$, at most $2\varepsilon m$ blocks are forced or use a pre-stabilization design. Their directional KL is nonnegative, while their predictable cost is at most $2\varepsilon c_{\max} m$. Since $\sum_{k \le m} \bar{c}_k \ge c_{\min} m$, removing those blocks loses at most $2\Gamma^\star_L(\xi) c_{\max} \varepsilon / c_{\min}$ times the total predictable cost. On every remaining block, equation E.31 and Lemma E.14 give, simultaneously for all incompatible ζ,

$$\sum_{k=1}^{m} m_k(\zeta) \ge \left(\Gamma^\star_L(\xi) - \frac{2\Gamma^\star_L(\xi) c_{\max}}{c_{\min}} \varepsilon - \varepsilon \right) \sum_{k=1}^{m} \bar{c}_k - D_\varepsilon.$$

The coefficient is r_ε. Conditional sub-gamma concentration controls centered likelihood fluctuations uniformly over the finite alternative set; cyclic separation controls quotient-MLE and within-cell maximization errors with exponent $2\gamma - 1$. The same five-event decomposition as in Proposition E.12, followed by a union bound over all $m \ge n$, proves equation E.39. □

Theorem E.17 (Certified approximate-design asymptotic optimality)**.** *Under the assumptions of Theorem E.13, let every nonforced design used after quotient-MLE stabilization satisfy equation E.31, with deterministic certificates $0 \le \zeta_k \le \bar{\zeta}_k$ and $n^{-1} \sum_{k=1}^{n} \bar{\zeta}_k \to 0$. Extend both certificate sequences by zero on forced and pre-stabilization blocks; those blocks are handled separately in Lemma E.16. Then the same GLR rule is globally (B, ε, δ)-correct and*

$$\lim_{\delta \downarrow 0} \frac{\mathbb{E}_\xi \mathsf{C}_{T_\delta}}{\log(1/\delta)} = \frac{1}{\Gamma^\star_L(\xi)}. \tag{E.40}$$

Proof. Anytime wrong-cell control is valid for every predictable tree-selection rule, so the error guarantee is unaffected by design approximation. Fix $\varepsilon > 0$ small enough that $r_\varepsilon > 0$ and define

$$B_\delta(\varepsilon) \coloneqq \frac{\beta_\delta + D_\varepsilon}{r_\varepsilon}, \qquad \nu_\delta \coloneqq \nu(B_\delta(\varepsilon)). \tag{E.41}$$

By Lemma E.16, on the complementary good event $Z_{\nu_\delta}(\bar{a}(\xi)) \geq \beta_\delta$, so the procedure has stopped by ν_δ. More generally, for every integer $m \geq 0$, since $\nu_\delta + m \geq \lfloor B_\delta / c_{\max} \rfloor + m$,

$$\mathbb{P}_\xi(T_\delta > \nu_\delta + m) \leq A \sum_{n \geq \lfloor B_\delta / c_{\max} \rfloor + m} e^{-bn^{2\gamma-1}}. \tag{E.42}$$

Therefore, writing $n_\delta = \lfloor B_\delta / c_{\max} \rfloor$,

$$\begin{aligned} \mathbb{E}_\xi[(T_\delta - \nu_\delta)_+] &= \sum_{m \geq 0} \mathbb{P}_\xi(T_\delta > \nu_\delta + m) \\ &\leq A \sum_{n \geq n_\delta} (n - n_\delta + 1) e^{-bn^{2\gamma-1}} = o(\beta_\delta). \end{aligned} \tag{E.43}$$

The last equality follows from stretched-exponential summability and $B_\delta = \Theta(\beta_\delta)$. Using bounded block cost and Lemma E.15,

$$\begin{aligned} \mathbb{E}_\xi \mathsf{C}_{T_\delta} &\leq \mathbb{E}_\xi \mathsf{C}_{\nu_\delta} + c_{\max} \mathbb{E}_\xi[(T_\delta - \nu_\delta)_+] \\ &\leq B_\delta(\varepsilon) + c_{\max} + o(\beta_\delta). \end{aligned}$$

Consequently,

$$\limsup_{\delta \downarrow 0} \frac{\mathbb{E}_\xi \mathsf{C}_{T_\delta}}{\beta_\delta} \leq \frac{1}{r_\varepsilon}.$$

Let $\varepsilon \downarrow 0$ and combine with the complete-tree lower bound. Almost-sure stopping follows from equation E.39; therefore the stated fixed-confidence correctness includes termination. □

Remark E.18 (Finite Gaussian tree families). If every trajectory kernel in a finite tree is a nonsingular Gaussian law from a finite parameter and policy set, each stopped log-likelihood ratio is a bounded-depth sum of affine-quadratic Gaussian increments. A common neighborhood of the origin has finite centered mgf, yielding Assumption E.8. This covers finite discretizations of the Gaussian calibration examples; it does not assert a general continuously parameterized composite-testing theorem.

F Non-Tensorization and Fixed-Lookahead Separations

This appendix isolates two formal reasons why the complete Bellman problem is not equivalent to ordinary episodewise or short-horizon experiment design.

F.1 Parity with a weak fresh-task channel

Fix $r \geq 2$, $\varrho \in (0, 1)$, and $\eta \in [0, 1)$. A task draws the parity latent $X \sim \mu_\sigma^{(r,\varrho)}$ from equation 3.1 and, independently conditional on the global sign σ, a task-fixed weak bit

$$\Pr_\sigma(B = b) = \tfrac{1}{2}(1 + \eta\sigma b), \qquad b \in \{-1, +1\}.$$

The source actions are $a_{\mathrm{w}}, a_1, \ldots, a_r$: a_{w} reveals B, and a_j reveals X_j. Repeating an action returns the same task-fixed symbol. The target problem is the parity target of Theorem 3.1; the weak bit is not observed at target time. Set

$$d_{\mathrm{w}}(\eta) = \eta \log \frac{1+\eta}{1-\eta}, \qquad d_{\mathrm{p}}(\varrho) = \varrho \log \frac{1+\varrho}{1-\varrho}.$$

Let $\Gamma^\star_{\mathrm{ep}}(\sigma)$ be the best information per source cost when every episode is run on a fresh task. Let $\Gamma^\star_{r+1}(\sigma)$ be the complete-tree value at depth $r + 1$.

Theorem F.1 (Non-tensorization gap)**.** *For both signs,*

$$\Gamma^{\star}_{\mathrm{ep}}(\sigma) = \frac{d_{\mathrm{w}}(\eta)}{1+\kappa}, \tag{F.1}$$

$$\Gamma^{\star}_{r+1}(\sigma) \geq \frac{d_{\mathrm{w}}(\eta) + d_{\mathrm{p}}(\varrho)}{r+1+\kappa}. \tag{F.2}$$

At $\eta = 0$, episodewise information is zero while persistent information is positive. For every $M < \infty$, some $\eta > 0$ makes both rates positive and

$$\frac{\Gamma^{\star}_{r+1}(\sigma)}{\Gamma^{\star}_{\mathrm{ep}}(\sigma)} \geq M. \tag{F.3}$$

Proof. Every reset-only coordinate observation is uniform under both signs by proper marginal cancellation. The weak action has binary KL $d_{\mathrm{w}}(\eta)$, and each fresh episode costs $1+\kappa$, proving equation F.1. The same-task tree that queries the weak bit and all r coordinates has product information $d_{\mathrm{w}}(\eta) + d_{\mathrm{p}}(\varrho)$ and cost $r+1+\kappa$, proving equation F.2. The ratio diverges as $d_{\mathrm{w}}(\eta) \downarrow 0$. □

Thus no reweighting of independently marginalized one-episode experiments can reproduce the information in the persistent joint experiment.

F.2 Failure of zero-terminal-value finite lookahead

Fix an integer $h \geq 1$. Define the receding-horizon h-step information rule as follows. At each task boundary or within-task history, it maximizes expected KL information per expected source cost over plans that may execute at most h additional episodes, assigns zero terminal value beyond that window, executes the first decision of a maximizing plan, and replans after the observation. Ties are resolved in favor of reset and then the weak action. Denote its long-run information rate by $\Gamma_{\mathrm{RH}(h)}(\sigma)$.

Theorem F.2 (Arbitrarily bad zero-terminal-value finite lookahead)**.** *For every fixed h and every $M < \infty$, there exist $r > h$, $\eta > 0$, and a weak-channel parity instance such that*

$$\frac{\Gamma^{\star}_{r+1}(\sigma)}{\Gamma_{\mathrm{RH}(h)}(\sigma)} \geq M \qquad \textit{for both } \sigma \in \{-1, +1\}. \tag{F.4}$$

Indeed, $\Gamma_{\mathrm{RH}(h)}(\sigma) = d_{\mathrm{w}}(\eta)/(1+\kappa)$.

Proof. At a fresh task, querying the weak bit yields positive information $d_{\mathrm{w}}(\eta)$. Every plan of at most $h < r$ coordinate queries reveals a proper parity marginal and therefore has zero information. The receding horizon rule queries the weak bit. After that observation, any continuation plan of at most h additional coordinate episodes still reveals at most $h < r$ coordinates and has zero information, whereas resetting and querying a new weak bit has positive information. Hence the rule resets and repeats, giving the stated rate. The Bellman lower bound is equation F.2; choosing η sufficiently small makes the ratio exceed M. □

The construction shows that delayed information, rather than merely a poor one-step heuristic, can defeat the specified zero-terminal-value receding-horizon information optimizer for every fixed planning depth. It does not assert failure of arbitrary finite-memory rules or planners equipped with a learned terminal-value approximation.

G Cross-Modal Sensor Calibration Benchmark

We give a natural finite-horizon environment in which the reset-or-continue decision has the interpretation of retaining the same object configuration to acquire a complementary sensing modality. The model is deliberately small enough that the Bellman–Chernoff value, optimal stopping tree, and fixed-confidence constant can be computed exactly.

G.1 Environment

The global calibration parameter is $\sigma \in \{-1, +1\}$. Fix sensor error probabilities $0 < \epsilon_{vq} < 1/2$ for $v \in \{0, 1\}$ and $q \in \{\mathsf{g}, \mathsf{b}\}$. Every source task draws an object orientation $X \sim \mathrm{Unif}\{-1, +1\}$ and a quality state $Q \in \{\mathsf{g}, \mathsf{b}\}$ with $\Pr(Q = \mathsf{g}) = p$. At task creation it also draws one screening symbol $S \sim K(\cdot \mid Q)$ and fixed sensor corruptions N_0, N_1. The complete latent tuple (X, Q, S, N_0, N_1) persists under continuation and is replaced independently under reset. Re-entering a corridor returns the same stored reading, so repeating one modality carries no additional information.

Each physical episode has horizon $H = 3$. From the start state the agent chooses one of three disjoint diagnostic corridors:

$$\mathsf{screen}, \qquad \mathsf{prop}, \qquad \mathsf{visual}.$$

At the second state it must execute the corridor-specific activation action; any other action reaches a failure state with an uninformative terminal symbol. Correct activation reveals the corresponding task-fixed terminal symbol:

$$S \qquad \text{in the screening corridor,} \tag{G.1}$$

$$Y_0 = XN_0, \qquad \Pr(N_0 = -1 \mid Q = q) = \epsilon_{0q}, \qquad \text{in the proprioceptive corridor,} \tag{G.2}$$

$$Y_1 = \sigma X N_1, \qquad \Pr(N_1 = -1 \mid Q = q) = \epsilon_{1q}, \qquad \text{in the visual corridor.} \tag{G.3}$$

The task-fixed corruptions N_0, N_1 are conditionally independent given Q and independent of S given Q. Re-entering any corridor returns the same task-fixed symbol. No physical episode can visit two corridors. At deployment, a fresh object orientation $X_\star$ is revealed and the agent must traverse one of two terminal routing corridors; it receives reward one exactly when the selected terminal label equals $\sigma X_\star$.

Reset-only one-modality trajectories are independent of σ, since X is uniform. A same-task proprioceptive–visual pair reveals the correlation

$$W = Y_0 Y_1 = \sigma N_0 N_1.$$

For a screen outcome s, let

$$q_s = \Pr(Q = \mathsf{g} \mid S = s)$$

and define

$$r_s \coloneqq q_s(1 - 2\epsilon_{0\mathsf{g}})(1 - 2\epsilon_{1\mathsf{g}}) + (1 - q_s)(1 - 2\epsilon_{0\mathsf{b}})(1 - 2\epsilon_{1\mathsf{b}}). \tag{G.4}$$

Then

$$\Pr_\sigma(W = w \mid S = s) = \frac{1}{2}(1 + r_s \sigma w),$$

so a completed pair after screen outcome s contributes $d_{\mathrm{par}}(r_s)$ nats about σ. Without screening, a direct proprioceptive–visual pair has reliability

$$\bar{r} \coloneqq \mathbb{E}[(1 - 2\epsilon_{0Q})(1 - 2\epsilon_{1Q})] = \sum_s \pi_s r_s, \qquad \pi_s = \Pr(S = s),$$

and contributes $d_{\mathrm{par}}(\bar{r})$ nats.

The direct paired-view subexperiment has information rate

$$\Gamma^\star_{\mathrm{pair}} \coloneqq \frac{d_{\mathrm{par}}(\bar{r})}{\kappa + 2}. \tag{G.5}$$

Lemma G.1 (One-modality conditional irrelevance)**.** *A single proprioceptive observation Y_0 or visual observation Y_1 is jointly independent of (σ, Q). Consequently, every randomized branch decision measurable with respect to one such observation is a model-independent mixture of deterministic branches and cannot improve on its best deterministic branch for any information-to-cost objective.*

Proof. Conditional on $Q = q$, multiplication by the independent uniform sign X makes both XN_0 and $\sigma X N_1$ uniform on $\{-1, +1\}$, independently of q and σ. Hence the posterior law of (σ, Q) is unchanged after one modality. A randomized branch kernel based only on that symbol therefore has the same mixing weights under every model. Its expected information and cost are affine in those weights, so its ratio cannot exceed the best pure branch appearing in the mixture. □

There are two potentially useful observation orders. A screen-first tree may condition acquisition of the complementary pair on S. A pair-first tree may condition acquisition of the quality signal on the observed parity W. For $w \in \{-1, +1\}$, define

$$p_w^+ \coloneqq \Pr_{\sigma=+1}(W = w) = \frac{1}{2}(1 + \bar{r}w), \tag{G.6}$$

$$j_w \coloneqq D_{\mathrm{KL}}\left(\mathcal{L}_+(S \mid W = w) \| \mathcal{L}_-(S \mid W = w)\right), \tag{G.7}$$

where

$$\Pr_{\sigma}(S = s \mid W = w) = \frac{\pi_s(1 + \sigma r_s w)}{1 + \sigma \bar{r} w}.$$

Proposition G.2 (Bellman-optimal calibration tree)**.** *Allow every deterministic experiment tree of depth at most three. Let $\Gamma^\star_{S\to W}$ and $\Gamma^\star_{W\to S}$ be the unique positive solutions of*

$$\sum_s \pi_s [d_{\mathrm{par}}(r_s) - 2\gamma]_+ = (\kappa + 1)\gamma, \tag{G.8}$$

$$d_{\mathrm{par}}(\bar{r}) + \sum_{w\in\{-1,+1\}} p_w^+ [j_w - \gamma]_+ = (\kappa + 2)\gamma, \tag{G.9}$$

respectively. Then the complete-tree information value is

$$\Gamma^\star_{\mathrm{cal}} = \max\left\{\Gamma^\star_{S\to W}, \Gamma^\star_{W\to S}\right\}. \tag{G.10}$$

The screen-first best response completes the pair after s iff

$$d_{\mathrm{par}}(r_s) > 2\Gamma^\star_{S\to W}, \tag{G.11}$$

while the pair-first best response acquires the screen after w iff $j_w > \Gamma^\star_{W\to S}$; equality permits randomization. The exact fixed-confidence expected source cost is

$$\mathbb{E}_\sigma \mathsf{C}_{T_\delta} \sim \frac{\log(1/\delta)}{\Gamma^\star_{\mathrm{cal}}}. \tag{G.12}$$

Proof. A tree that never obtains both modalities has a σ-independent transcript and zero information. Repeating a corridor returns the same stored symbol and is dominated. Suppose the first informative corridor is proprioceptive or visual. By Lemma G.1, branching on that observation is a model-independent mixture and cannot improve on its best deterministic branch. Up to exchanging the two modalities, every undominated informative depth-three tree is consequently of one of two forms: screen first and optionally acquire the pair; or acquire the pair and optionally screen. This exhausts the complete tree class.

For a screen-first tree, completing the pair after s costs two additional episodes and yields $d_{\mathrm{par}}(r_s)$ information. The Bellman continuation advantage is $d_{\mathrm{par}}(r_s) - 2\gamma$; charging one root episode and reset cost κ gives equation G.8. For a pair-first tree, the pair itself costs two episodes and yields $d_{\mathrm{par}}(\bar{r})$. Conditional on $W = w$, screening costs one additional episode and yields conditional KL j_w, giving equation G.9. The fractional-root theorem proves both threshold rules. Maximizing the two root values gives equation G.10, and the complete-tree characteristic time theorem yields equation G.12. □

Corollary G.3 (Observation-adaptive acquisition can strictly beat every static tree)**.** *Among trees whose full episode sequence is fixed before any task observation, the best information rate is*

$$\Gamma_{\mathrm{static}} = \max\left\{\frac{d_{\mathrm{par}}(\bar{r})}{\kappa + 2}, \frac{\sum_s \pi_s d_{\mathrm{par}}(r_s)}{\kappa + 3}\right\}. \tag{G.13}$$

Whenever either Bellman root in equations (G.8)–(G.9) uses both its stop and continue branches with positive probability and its value exceeds the other observation order, then

$$\Gamma^\star_{\mathrm{cal}} > \Gamma_{\mathrm{static}}. \tag{G.14}$$

Proof. The complete tree can imitate either static sequence. If an optimal root uses both branches strictly, then at the corresponding static ratio, deleting the positive-probability negative-net-value continuations while retaining the positive-net-value continuations strictly increases the root net value. Its unique fractional root is therefore strictly larger than the rate of the associated always-continue static sequence. If that root also dominates the other observation order, it exceeds both terms in equation G.13. □

Table 3: Information-rate baselines in the calibration benchmark. The quantities are computed exactly by the supplied script for the parameters of Figure 2.

Acquisition rule	information per expected cost
Reset only / screen only	0
One modality only, at any depth	0
Myopic one-step KL	0
Direct complementary pair	$d_{\mathrm{par}}(\bar{r})/(\kappa+2)$
Screen, then always pair	$\sum_s \pi_s d_{\mathrm{par}}(r_s)/(\kappa+3)$
Best static tree	Γ_{static}
Screen-first Bellman tree	$\Gamma^\star_{S\to W}$
Pair-first Bellman tree	$\Gamma^\star_{W\to S}$
Complete Bellman–Chernoff tree	$\Gamma^\star_{\mathrm{cal}}$

G.2 Reproducible calibration experiments

The supplementary script `experiments/generate_calibration_figures.py` computes the exact Bellman root, enumerates all static policies, and simulates the sequential likelihood-ratio test. It produces:

1. the adaptivity gain as screening quality varies;
2. a breadth–depth phase diagram as reset cost and sensing noise vary;
3. the robust persistence-order and near-invariance curves in Appendix W;
4. the exact finite-threshold residual $\mathbb{E}C_{T_\delta}/\log(1/\delta) - 1/\Gamma^\star_{\mathrm{pair}}$ for a direct paired-view subexperiment, computed by solving the absorbing random-walk equations rather than by Monte Carlo.

The accompanying JSON file records every parameter, exact expected block count, achieved wrong-boundary probability, static and adaptive information rates, and the generated figure values.

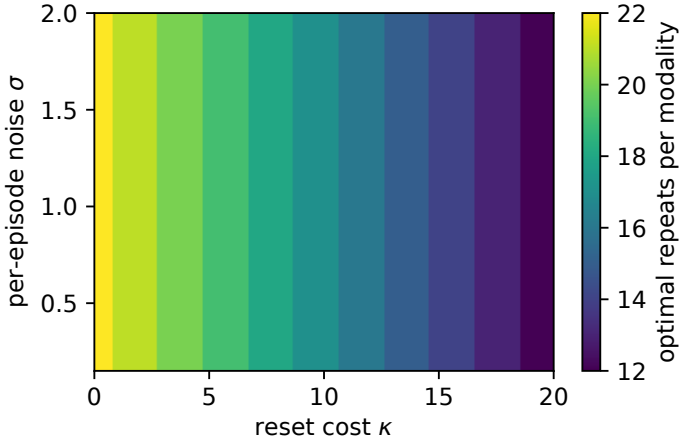


Figure 3: Static breadth–depth phase diagram in the orthogonal calibration specialization. Higher sensing noise and higher reset cost favor longer continuation, whereas the explicit breadth constraint prevents spending the entire budget on one latent task.

The benchmark is not intended as a performance leaderboard; it verifies the predicted optimal-stopping rule, phase transition, and characteristic constant in a recognizably sequential source-to-target control problem.

H PERSISTENCE–CHERNOFF LOWER BOUND

This appendix proves the information-theoretic lower half of the Persistence–Chernoff characterization. The argument belongs to the classical line of sequential experimental design initiated by Chernoff (1959) and to the modern change-of-measure methodology for adaptive identification (Naghshvar & Javidi, 2013; Kaufmann et al., 2016; Garivier & Kaufmann, 2016). The new feature is the experiment class: one source experiment is an internally adaptive, multi-episode block whose trajectories share an unobserved task latent. The proof is otherwise deliberately exact: experiment-selection randomization cancels from the likelihood ratio, stopping is handled at the level of the stopped transcript, and only global-model distinctions that force incompatible target decisions enter the lower bound.

H.1 FINITE BLOCKWISE EXPERIMENT CLASS

Throughout this appendix, let the global model class Θ and the block-experiment class $\mathcal{E}$ be finite and nonempty. Every $e \in \mathcal{E}$ is a deterministic-length experiment in the sense of Definition A.10, with cost

$$c_\kappa(e) = \ell(e) + \kappa > 0$$

and marginalized transcript law Q_θ^e on a measurable space $(\mathsf{D}_e, \mathcal{D}_e)$. To avoid notationally inessential product-space bookkeeping, we regard the disjoint union

$$\mathsf{D} := \bigsqcup_{e \in \mathcal{E}} \big(\{e\} \times \mathsf{D}_e\big)$$

as the common marked-block observation space.

Assumption H.1 (Common domination and finite directional information)**.** For every $e \in \mathcal{E}$, there is a sigma-finite measure ν_e such that

$$Q_\theta^e(\mathrm{d}d) = q_\theta^e(d)\, \nu_e(\mathrm{d}d) \qquad (\theta \in \Theta),$$

and, for all $\theta, \theta' \in \Theta$,

$$d_e(\theta, \theta') := D_{\mathrm{KL}}\left(Q_\theta^e \| Q_{\theta'}^e\right) < \infty. \tag{H.1}$$

Because the assumption is imposed for every ordered pair, all block laws for a fixed experiment are mutually absolutely continuous. In particular, the block log-likelihood ratios below are well defined almost surely under every model. Finite KL also implies absolute integrability of each log-likelihood ratio: if $L = \mathrm{d}P/\mathrm{d}Q$, then

$$\mathbb{E}_P[(-\log L)_+] = \int_{\{L<1\}} -L \log L \,\mathrm{d}Q \le e^{-1},$$

while $\mathbb{E}_P[(\log L)_+] < \infty$ follows from $D_{\mathrm{KL}}(P\|Q) < \infty$.

A blockwise learner chooses experiments only between completed blocks. More formally, after the history

$$H_{k-1} = (E_1, D_1, \ldots, E_{k-1}, D_{k-1}),$$

it samples E_k from a model-independent stochastic kernel $\alpha_k(\cdot \mid H_{k-1})$ on $\mathcal{E}$, resets to a fresh task, and observes

$$D_k \mid (H_{k-1}, E_k = e) \sim Q_\theta^e.$$

The learner stops at an almost surely finite stopping time τ for the block filtration

$$\mathcal{B}_k = \sigma(E_1, D_1, \ldots, E_k, D_k)$$

and then draws a target recommendation $\widehat{g}$ from a model-independent kernel $\Lambda(\cdot \mid H_\tau)$. Define

$$N_e(\tau) := \sum_{k=1}^{\tau} \mathbf{1}\{E_k = e\}, \qquad \mathsf{C}_\tau^{\mathrm{blk}} := \sum_{k=1}^{\tau} c_\kappa(E_k) = \sum_{e \in \mathcal{E}} c_\kappa(e) N_e(\tau). \tag{H.2}$$

The algorithmic randomization used by α_k, τ, and Λ is part of the learner and has the same law under every global model.

Let $\mathfrak{A}^{\mathrm{blk}}_{B,\varepsilon,\delta}(\mathcal{E})$ be the blockwise learners that are (B,ε,δ)-correct on Θ in the sense of Definition A.9. The restricted instance-dependent complexity is

$$\mathfrak{C}^{\mathrm{blk},\star}_{B,\varepsilon,\delta,\kappa}(\theta;\mathcal{E}) \coloneqq \inf_{\mathcal{A}\in\mathfrak{A}^{\mathrm{blk}}_{B,\varepsilon,\delta}(\mathcal{E})} \mathbb{E}^{\mathcal{A}}_{\theta}\left[\mathsf{C}^{\mathrm{blk}}_{\tau}\right]. \tag{H.3}$$

This is the object characterized by the finite-experiment Persistence–Chernoff theory. It should not be confused with the unrestricted episode-by-episode complexity in equation A.46; the first exact characteristic-time theorem deliberately restricts the learner to the block class $\mathcal{E}$.

H.2 The cost-normalized information game

Define the cost-normalized allocation simplex

$$\mathcal{W}_{\kappa}(\mathcal{E}) \coloneqq \left\{ w \in \mathbb{R}^{\mathcal{E}}_{+} : \sum_{e\in\mathcal{E}} c_{\kappa}(e) w_e = 1 \right\}. \tag{H.4}$$

Here w_e has the units "expected number of type-e blocks per unit source cost." For a fixed instance θ, target budget B, and tolerance ε, define

$$\Gamma^{\star}_{\mathrm{pers}}(\theta;B,\varepsilon,\kappa,\mathcal{E}) \coloneqq \sup_{w\in\mathcal{W}_{\kappa}(\mathcal{E})} \inf_{\theta'\in\mathrm{Alt}_{B,\varepsilon}(\theta)} \sum_{e\in\mathcal{E}} w_e d_e(\theta,\theta'). \tag{H.5}$$

We adopt

$$\inf\varnothing = +\infty, \qquad \frac{1}{+\infty} = 0, \qquad \frac{1}{0} = +\infty.$$

Thus no source information is required when there is no target-relevant alternative, whereas zero information rate against a nonempty alternative set corresponds to infinite fixed-confidence cost.

Proposition H.2 (Identifiability and positivity)**.** *Suppose* $\mathrm{Alt}_{B,\varepsilon}(\theta) \neq \varnothing$*. Then*

$$\Gamma^{\star}_{\mathrm{pers}}(\theta;B,\varepsilon,\kappa,\mathcal{E}) > 0$$

if and only if, for every $\theta' \in \mathrm{Alt}_{B,\varepsilon}(\theta)$*, there exists* $e \in \mathcal{E}$ *such that* $d_e(\theta,\theta') > 0$*.*

Proof. If some alternative θ' satisfies $d_e(\theta,\theta') = 0$ for every experiment, then the inner infimum in equation H.5 is zero for every allocation. Conversely, suppose every alternative is separated by at least one experiment. Choose any allocation with $w_e > 0$ for every $e \in \mathcal{E}$ and normalize it to belong to $\mathcal{W}_{\kappa}(\mathcal{E})$. For each alternative, the finite sum in equation H.5 is strictly positive. The alternative set is finite, so its minimum is strictly positive. □

The next representation is useful for the upper-bound design problem in Section I. Let $\Delta(S)$ denote the probability simplex on a finite set S.

Proposition H.3 (Dual adversarial-mixture representation)**.** *If* $\mathrm{Alt}_{B,\varepsilon}(\theta)$ *is nonempty, then*

$$\begin{aligned}&\Gamma^{\star}_{\mathrm{pers}}(\theta;B,\varepsilon,\kappa,\mathcal{E})\\ &\quad= \inf_{\lambda\in\Delta(\mathrm{Alt}_{B,\varepsilon}(\theta))} \max_{e\in\mathcal{E}} \frac{\sum_{\theta'\in\mathrm{Alt}_{B,\varepsilon}(\theta)} \lambda_{\theta'} d_e(\theta,\theta')}{c_{\kappa}(e)}.\end{aligned} \tag{H.6}$$

Proof. For $w \in \mathcal{W}_{\kappa}(\mathcal{E})$, set $x_e = c_{\kappa}(e) w_e$. Then $x \in \Delta(\mathcal{E})$ and

$$\sum_e w_e d_e(\theta,\theta') = \sum_e x_e \frac{d_e(\theta,\theta')}{c_{\kappa}(e)}.$$

The infimum over a finite set of alternatives equals the infimum over its probability simplex because the payoff is linear in the alternative mixture. Von Neumann's finite minimax theorem therefore permits interchange of the maximum over x and the minimum over λ. Maximizing a linear function over $\Delta(\mathcal{E})$ selects a pure experiment, yielding equation H.6. □

The dual representation in equation H.6 has a direct interpretation: the alternative player selects a least-favorable mixture of target-relevant models, and the learner selects the task-block experiment with the largest average KL information per unit source cost. This is analogous to the information games appearing in sequential hypothesis testing and fixed-confidence identification (Chernoff, 1959; Naghshvar & Javidi, 2013; Garivier & Kaufmann, 2016), but the actions here are persistent-task blocks rather than one-step sensing actions or bandit pulls.

H.3 ADAPTIVE LIKELIHOOD FACTORIZATION

Fix $\theta, \theta' \in \Theta$ and write

$$\ell_e^{\theta,\theta'}(d) \coloneqq \log \frac{q_\theta^e(d)}{q_{\theta'}^e(d)}. \tag{H.7}$$

Let $\mathbb{P}_\vartheta^{(m)}$ denote the law of H_m under global model ϑ. For a block history of deterministic length m, define

$$L_m(\theta, \theta') = \sum_{k=1}^{m} \ell_{E_k}^{\theta,\theta'}(D_k).$$

Lemma H.4 (Adaptive block-likelihood ratio)**.** *For every deterministic $m \geq 1$, the likelihood ratio of the first m selected experiments and block transcripts satisfies*

$$\frac{\mathrm{d}\mathbb{P}_\theta^{(m)}}{\mathrm{d}\mathbb{P}_{\theta'}^{(m)}}(H_m) = \exp\big(L_m(\theta, \theta')\big) \qquad \mathbb{P}_{\theta'}^{(m)}\text{-a.s.} \tag{H.8}$$

Moreover,

$$\mathbb{E}_\theta\left[\ell_{E_k}^{\theta,\theta'}(D_k)\middle|\mathcal{B}_{k-1}, E_k\right] = d_{E_k}(\theta, \theta') \quad \text{a.s.} \tag{H.9}$$

Proof. The joint law of H_m under model ϑ factors as

$$\prod_{k=1}^{m} \alpha_k(\mathrm{d}e_k \mid h_{k-1}) Q_\vartheta^{e_k}(\mathrm{d}d_k).$$

The experiment-selection kernels α_k are algorithmic and therefore identical under θ and θ'. They cancel from the Radon–Nikodym derivative, leaving the product of block density ratios in equation H.8. Conditional on $(\mathcal{B}_{k-1}, E_k = e)$, the fresh-block law is Q_θ^e by equation A.50; taking the conditional expectation of equation H.7 gives equation H.9. □

The next theorem is the central technical identity. It includes random stopping and the randomized recommendation kernel rather than treating them informally.

Theorem H.5 (Stopped block information identity)**.** *Let $\mathcal{A}$ be a blockwise learner with almost surely finite stopping time τ. Fix $\theta, \theta' \in \Theta$. If*

$$\mathbb{E}_\theta^{\mathcal{A}}[\mathsf{C}_\tau^{\mathrm{blk}}] < \infty,$$

then the joint stopped law of $(H_\tau, \widehat{g})$ satisfies

$$\begin{aligned} D_{\mathrm{KL}}\left(\mathbb{P}_\theta^{\mathcal{A}} \middle\| \mathbb{P}_{\theta'}^{\mathcal{A}}\right) &= \mathbb{E}_\theta^{\mathcal{A}}\left[\sum_{k=1}^{\tau} \ell_{E_k}^{\theta,\theta'}(D_k)\right] \\ &= \sum_{e \in \mathcal{E}} \mathbb{E}_\theta^{\mathcal{A}}[N_e(\tau)] d_e(\theta, \theta'). \end{aligned} \tag{H.10}$$

Proof. Because the costs are strictly positive and $\mathcal{E}$ is finite, finite expected source cost implies $\mathbb{E}_\theta[\tau] < \infty$. For every e, the log-likelihood increment $\ell_e^{\theta,\theta'}$ is absolutely integrable under Q_θ^e by the observation following Assumption H.1. Let

$$b_{\theta,\theta'} \coloneqq \max_{e \in \mathcal{E}} \int \left|\ell_e^{\theta,\theta'}(d)\right| Q_\theta^e(\mathrm{d}d) < \infty.$$

Then the fresh-block conditional law and Tonelli's theorem give

$$\mathbb{E}_\theta\left[\sum_{k=1}^{\tau} \left|\ell_{E_k}^{\theta,\theta'}(D_k)\right|\right] \leq b_{\theta,\theta'} \mathbb{E}_\theta[\tau] < \infty. \tag{H.11}$$

Thus all subsequent sums and conditional expectations are justified by Fubini's theorem.

The random stopped history takes values in the disjoint union

$$\mathsf{H}_{\mathrm{stop}} = \bigsqcup_{m=0}^{\infty} (\mathcal{E} \times \mathsf{D})^m .$$

On the stratum $\{\tau = m\}$, the joint density contains: the product of the experiment-selection kernels; the product of the block densities; and the indicator or randomized kernel specifying stopping at m. Every algorithmic factor is identical under the two global models. Therefore the Radon–Nikodym derivative of the stopped history is

$$\exp\left(\sum_{k=1}^{\tau} \ell_{E_k}^{\theta,\theta'}(D_k)\right).$$

The recommendation kernel $\Lambda(\cdot \mid H_\tau)$ is also the same under both models. The conditional KL chain rule therefore adds zero recommendation divergence, establishing the first equality in equation H.10.

For the second equality, use absolute integrability, expand the stopped sum, and apply equation H.9:

$$\begin{aligned}
\mathbb{E}_\theta\left[\sum_{k=1}^{\tau} \ell_{E_k}^{\theta,\theta'}(D_k)\right] &= \sum_{k\geq 1} \mathbb{E}_\theta\left[\mathbf{1}\{\tau \geq k\}\ell_{E_k}^{\theta,\theta'}(D_k)\right] \\
&= \sum_{k\geq 1} \mathbb{E}_\theta\left[\mathbf{1}\{\tau \geq k\} d_{E_k}(\theta,\theta')\right] \\
&= \sum_{e\in\mathcal{E}} \mathbb{E}_\theta[N_e(\tau)] d_e(\theta,\theta').
\end{aligned}$$

Here $\{\tau \geq k\} \in \mathcal{B}_{k-1}$, as required for the conditional expectation step. □

Remark H.6 (Why blockwise adaptivity costs no extra KL). The identity does not require a non-adaptive experiment schedule. Experiment selection may depend arbitrarily on every completed source block. Its conditional kernel appears under both models and therefore contributes zero to the likelihood ratio. All model-dependent information is carried by the selected block laws Q_θ^e.

H.4 Decision change of measure

For $p, q \in [0,1]$, write

$$\mathrm{kl}(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q},$$

with the usual boundary conventions.

Lemma H.7 (Binary data processing)**.** *For probability measures P, Q on the same measurable space and every measurable event A,*

$$D_{\mathrm{KL}}(P\|Q) \geq \mathrm{kl}(P(A), Q(A)). \tag{H.12}$$

If $\delta \in (0, 1/2)$, $P(A) \geq 1-\delta$, and $Q(A) \leq \delta$, then

$$D_{\mathrm{KL}}(P\|Q) \geq \mathrm{kl}(1-\delta, \delta). \tag{H.13}$$

Proof. Apply the data-processing inequality to the measurable map $x \mapsto \mathbf{1}\{x \in A\}$ to obtain equation H.12. On the rectangle $p \in [1-\delta, 1]$ and $q \in [0, \delta]$, the binary KL is nondecreasing in p and nonincreasing in q, because

$$\frac{\partial}{\partial p}\mathrm{kl}(p,q) = \log\frac{p(1-q)}{q(1-p)} \geq 0, \qquad \frac{\partial}{\partial q}\mathrm{kl}(p,q) = \frac{q-p}{q(1-q)} \leq 0.$$

This proves equation H.13; boundary cases follow by continuity or infinite divergence. □

Lemma H.8 (Target-decision transportation inequality)**.** *Let $\mathcal{A}$ be (B, ε, δ)-correct on Θ, with $\delta \in (0, 1/2)$. Then for every $\theta \in \Theta$ and every $\theta' \in \mathrm{Alt}_{B,\varepsilon}(\theta)$,*

$$\sum_{e\in\mathcal{E}} \mathbb{E}_\theta^{\mathcal{A}}[N_e(\tau)] d_e(\theta,\theta') \geq \mathrm{kl}(1-\delta, \delta), \tag{H.14}$$

provided the expected source cost under θ is finite.

Proof. Let

$$A_\theta = \{\widehat{g} \in \mathcal{G}_{B,\varepsilon}(\theta)\}.$$

Correctness under θ gives $\mathbb{P}_\theta^{\mathcal{A}}(A_\theta) \geq 1-\delta$. Because $\theta' \in \mathrm{Alt}_{B,\varepsilon}(\theta)$, the two ε-optimal strategy sets are disjoint. Correctness under θ' therefore implies

$$\mathbb{P}_{\theta'}^{\mathcal{A}}(A_\theta) \leq \mathbb{P}_{\theta'}^{\mathcal{A}}\left(\widehat{g} \notin \mathcal{G}_{B,\varepsilon}(\theta')\right) \leq \delta.$$

Apply Lemma H.7 to the joint stopped experiment and then use Theorem H.5. □

This is the persistent-task analogue of the transportation inequalities used in adaptive best-arm identification (Kaufmann et al., 2016). The difference is that the count $N_e(\tau)$ indexes complete same-task block experiments and its divergence already integrates over the persistent task latent.

H.5 Instance-dependent lower bound

Theorem H.9 (Finite-block target-decision lower bound)**.** *Fix $B \in \mathbb{N}_0$, $\varepsilon > 0$, $\delta \in (0, 1/2)$, and $\theta \in \Theta$. Under Assumption H.1, every $\mathcal{A} \in \mathfrak{A}^{\mathrm{blk}}_{B,\varepsilon,\delta}(\mathcal{E})$ satisfies*

$$\mathbb{E}_\theta^{\mathcal{A}}\left[\mathsf{C}_\tau^{\mathrm{blk}}\right] \geq \frac{\mathrm{kl}(1-\delta,\delta)}{\Gamma^\star_{\mathrm{pers}}(\theta; B, \varepsilon, \kappa, \mathcal{E})}. \tag{H.15}$$

Consequently,

$$\mathfrak{C}^{\mathrm{blk},\star}_{B,\varepsilon,\delta,\kappa}(\theta; \mathcal{E}) \geq \frac{\mathrm{kl}(1-\delta,\delta)}{\Gamma^\star_{\mathrm{pers}}(\theta; B, \varepsilon, \kappa, \mathcal{E})}. \tag{H.16}$$

In extended-real arithmetic this includes the following cases: if the target-relevant alternative set is empty, the right-hand side is zero; if the alternative set is nonempty and the Persistence–Chernoff information is zero, no globally correct blockwise learner has finite expected source cost under θ.

Proof. The statement is immediate if the expected cost is infinite. If $\mathrm{Alt}_{B,\varepsilon}(\theta) = \varnothing$, the right-hand side is zero by convention, so the claim is also immediate. Hence assume the alternative set is nonempty and the expected cost is finite. Write

$$C_\theta = \mathbb{E}_\theta^{\mathcal{A}}[\mathsf{C}_\tau^{\mathrm{blk}}].$$

By Lemma H.8, at least one expected block count is positive, so $C_\theta > 0$. Define the realized average allocation

$$\overline{w}_e = \frac{\mathbb{E}_\theta^{\mathcal{A}}[N_e(\tau)]}{C_\theta}.$$

By equation H.2, $\overline{w} \in \mathcal{W}_\kappa(\mathcal{E})$. Applying Lemma H.8 to every target-relevant alternative yields

$$\begin{aligned}\mathrm{kl}(1-\delta,\delta) &\leq C_\theta \inf_{\theta' \in \mathrm{Alt}_{B,\varepsilon}(\theta)} \sum_e \overline{w}_e d_e(\theta,\theta') \\ &\leq C_\theta \Gamma^\star_{\mathrm{pers}}(\theta; B, \varepsilon, \kappa, \mathcal{E}).\end{aligned}$$

Rearranging proves equation H.15. Taking the infimum over all globally correct blockwise learners proves equation H.16. The empty-alternative convention is immediate. If the alternative set is nonempty and the information rate is zero, the displayed inequality contradicts finite expected cost because $\mathrm{kl}(1-\delta,\delta) > 0$. □

Corollary H.10 (Asymptotic characteristic-time lower bound)**.** *For every fixed θ, B, ε, κ, and finite $\mathcal{E}$,*

$$\liminf_{\delta \downarrow 0} \frac{\mathfrak{C}^{\mathrm{blk},\star}_{B,\varepsilon,\delta,\kappa}(\theta; \mathcal{E})}{\log(1/\delta)} \geq \frac{1}{\Gamma^\star_{\mathrm{pers}}(\theta; B, \varepsilon, \kappa, \mathcal{E})}. \tag{H.17}$$

Proof. Use Theorem H.9 and

$$\mathrm{kl}(1-\delta,\delta) = (1-2\delta)\log\frac{1-\delta}{\delta},$$

whose ratio to $\log(1/\delta)$ tends to one. □

H.6 Recovery of the parity hierarchy

The abstract information game recovers the exact lower bound of Section 3 without losing the persistence mechanism. Consider the two parity models from Theorem 3.1, and let $\mathcal{E}$ be any finite collection of deterministic same-task experiments.

Proposition H.11 (Parity-family characteristic information)**.** *For $B = r - 1$ and $0 < \varepsilon < \varrho/2^{r+1}$,*

$$\Gamma^{\star}_{\mathrm{pers}}(+; r-1, \varepsilon, \kappa, \mathcal{E}) \le \frac{d_{\mathrm{par}}(\varrho)}{r+\kappa}. \tag{H.18}$$

If $e_{\mathrm{full}} \in \mathcal{E}$, equality holds.

Proof. There is one target-relevant alternative. Every experiment that leaves a coordinate unrevealed has zero divergence by Theorem 3.1. Every experiment that reveals all coordinates has cost at least $r + \kappa$. Its transcript is generated from the complete latent vector by a model-independent Markov kernel, so data processing and Theorem 3.1 give divergence at most $d_{\mathrm{par}}(\varrho)$. Hence every experiment has information per cost at most $d_{\mathrm{par}}(\varrho)/(r+\kappa)$. The complete block attains this ratio exactly. With a single alternative, the game in equation H.5 reduces to maximizing information per unit cost over experiments. □

Combining Proposition H.11 with Theorem H.9 recovers equation 3.7. This sanity check also clarifies what the general characteristic time measures: it is not the number of physical episodes alone, but the largest worst-case rate at which complete persistent task experiments separate target-incompatible global models.

H.7 Scope of the lower bound

Three restrictions are deliberate.

1. The result concerns the finite blockwise class $\mathfrak{A}^{\mathrm{blk}}_{B,\varepsilon,\delta}(\mathcal{E})$. It does not claim a lower bound for every episode-by-episode source learner in equation A.46 unless that learner can be represented by the stated experiment class.
2. Block lengths and costs are deterministic. If the decision to stop within a task is itself adaptive, the experiment law and cost become a controlled renewal object; the exact information-per-cost game then requires a separate analysis.
3. The lower bound is information-theoretic. It does not assume that the target-optimal strategy, the likelihoods, or the max–min allocation can be computed efficiently. Computational issues are separate from the statistical limit proved here.

Within the decision-partition subclass and additional regularity conditions stated in Section I, Theorem H.9 is exact at the leading $\log(1/\delta)$ scale. Outside that subclass, the lower bound remains valid, but a matching upper theorem may require an answer-specific information game.

I Persistence Track-and-Stop

This appendix proves the algorithmic upper half of the finite-block Persistence–Chernoff characterization. The construction follows the classical program of sequential experimental design (Chernoff, 1959), and the Track-and-Stop principle of tracking an information-optimal design and stopping with a likelihood-ratio certificate (Garivier & Kaufmann, 2016). The statistical actions here are not one-step measurements: every action is a deterministic-length, internally adaptive source block whose physical trajectories share one unobserved task latent.

The lower bound in Section H is valid for arbitrary ε-optimal strategy sets. A matching upper bound requires a recommendation structure: pairwise overlap of acceptable-strategy sets does not, by itself, guarantee that a whole confidence set admits one common recommendation. We therefore state the first exact upper theorem under a finite *decision partition*. This is the single-answer analogue of pure-exploration identification. Problems with genuinely multiple correct answers require an answer-specific information game and a sticky selection rule, as in Degenne & Koolen (2019); that extension is outside the theorem proved here.

I.1 Decision cells and additional regularity

We retain the finite model class Θ, finite deterministic block class $\mathcal{E}$, costs $c_\kappa(e) > 0$, densities q_θ^e, and directional divergences $d_e(\theta, \theta')$ from Section H.1. Fix the target budget B and tolerance $\varepsilon > 0$ throughout this appendix.

Assumption I.1 (Finite decision partition)**.** There exist a finite nonempty answer set $\mathcal{I}$, a surjective map $a : \Theta \to \mathcal{I}$, and canonical target strategies $(g_i)_{i\in\mathcal{I}} \subseteq \mathcal{G}_B$ such that:

1. for every $\theta \in \Theta$, $g_{a(\theta)} \in \mathcal{G}_{B,\varepsilon}(\theta)$;
2. whenever $a(\theta) \neq a(\theta')$,
$$\mathcal{G}_{B,\varepsilon}(\theta) \cap \mathcal{G}_{B,\varepsilon}(\theta') = \varnothing.$$

Write
$$\Theta_i := \{\theta \in \Theta : a(\theta) = i\}.$$
The first condition supplies one strategy that is valid throughout each cell; the second says that distinct cells correspond to incompatible target decisions.

Lemma I.2 (Decision cells equal target-relevant equivalence classes)**.** *Under Assumption I.1, for every $\theta \in \Theta$,*
$$\mathrm{Alt}_{B,\varepsilon}(\theta) = \{\theta' \in \Theta : a(\theta') \neq a(\theta)\}. \tag{I.1}$$

Proof. If $a(\theta') = a(\theta) = i$, the common canonical strategy g_i belongs to both ε-optimal sets, so θ' is not an alternative. If the labels differ, the sets are disjoint by assumption, so θ' is an alternative. □

Assumption I.3 (Global block identifiability)**.** For every distinct $\theta, \theta' \in \Theta$, there exists $e \in \mathcal{E}$ with
$$d_e(\theta, \theta') > 0.$$

This condition is stronger than target-relevant identifiability: it is used to make the model maximum-likelihood estimator eventually exact. One can weaken it by quotienting observationally equivalent models and tracking a set-valued oracle, but that would obscure the main persistence argument.

Assumption I.4 (Uniformly bounded block log-likelihood ratios)**.** There is $L < \infty$ such that, for every ordered pair $\theta, \theta' \in \Theta$, every $e \in \mathcal{E}$, and ν_e-almost every d,
$$\left|\log \frac{q_\theta^e(d)}{q_{\theta'}^e(d)}\right| \leq L. \tag{I.2}$$

The boundedness assumption gives the stretched-exponential concentration needed to control *expected* stopping cost. It holds, for example, for finite block-transcript spaces with model probabilities uniformly bounded away from zero. It may be replaced by a uniform exponential-moment condition; time-uniform mixture-martingale methods provide sharper thresholds under more structured models (Kaufmann & Koolen, 2021).

Let
$$M := |\Theta|, \qquad m := |\mathcal{E}|, \qquad c_{\min} := \min_e c_\kappa(e), \qquad c_{\max} := \max_e c_\kappa(e).$$
Under global identifiability and finiteness,
$$\underline{d} := \min_{\theta\neq\theta'} \max_{e\in\mathcal{E}} d_e(\theta, \theta') > 0. \tag{I.3}$$

I.2 Oracle designs in block and cost time

For every θ whose alternative set is nonempty, choose once and for all an optimizer
$$w^\star(\theta) \in \operatorname*{arg\,max}_{w\in\mathcal{W}_\kappa(\mathcal{E})} \inf_{\theta'\in\mathrm{Alt}_{B,\varepsilon}(\theta)} \sum_{e\in\mathcal{E}} w_e d_e(\theta, \theta'). \tag{I.4}$$
The optimizer exists because the cost-normalized simplex is compact and the objective is continuous. A deterministic ordering of $\mathcal{E}$ is used to select one optimizer when the set is not a singleton. Define
$$s_\theta := \sum_{e\in\mathcal{E}} w_e^\star(\theta), \qquad p_e^\star(\theta) := \frac{w_e^\star(\theta)}{s_\theta}, \qquad \bar{c}_\theta := \sum_e p_e^\star(\theta) c_\kappa(e). \tag{I.5}$$

Lemma I.5 (Block-time and cost-time representations)**.** *For every θ with a nonempty alternative set,*

$$\bar{c}_\theta = \frac{1}{s_\theta}, \qquad \frac{p_e^\star(\theta)}{\bar{c}_\theta} = w_e^\star(\theta), \tag{I.6}$$

and

$$\min_{\theta' \in \mathrm{Alt}_{B,\varepsilon}(\theta)} \sum_e p_e^\star(\theta) d_e(\theta, \theta') = \bar{c}_\theta \Gamma^\star_{\mathrm{pers}}(\theta; B, \varepsilon, \kappa, \mathcal{E}). \tag{I.7}$$

Proof. Because $w^\star(\theta) \in \mathcal{W}_\kappa(\mathcal{E})$, $\sum_e c_\kappa(e) w_e^\star(\theta) = 1$. Substitution into equation I.5 gives $\bar{c}_\theta = 1/s_\theta$ and hence the first display. Multiplying the cost-time objective by $1/s_\theta = \bar{c}_\theta$ gives equation I.7. □

I.3 Persistence Track-and-Stop

Fix a forced-exploration exponent

$$\gamma \in (1/2, 1)$$

and an ordering $\mathcal{E} = \{e_1, \dots, e_m\}$. Let

$$F(k) \coloneqq \lfloor k^\gamma \rfloor, \qquad k \in \mathbb{N}_0.$$

Block k is forced when $F(k) > F(k-1)$; forced blocks cycle deterministically through $e_1, \dots, e_m$. Thus the total number of forced blocks through block k is exactly $F(k)$.

For a completed block history H_k, define the model log likelihood

$$\mathsf{L}_k(\vartheta) \coloneqq \sum_{s=1}^{k} \log q_\vartheta^{E_s}(D_s), \qquad \vartheta \in \Theta, \tag{I.8}$$

and let $\widehat{\theta}_k$ be a maximum-likelihood model, with ties broken according to a fixed total order on Θ.

For an answer $i \in \mathcal{I}$, define the decision-cell generalized likelihood ratio

$$Z_k(i) \coloneqq \max_{\vartheta \in \Theta_i} \mathsf{L}_k(\vartheta) - \max_{\lambda \in \Theta \setminus \Theta_i} \mathsf{L}_k(\lambda), \tag{I.9}$$

with $\max \varnothing = -\infty$. If $|\mathcal{I}| = 1$, no source data are needed and the unique canonical strategy is returned immediately. Otherwise set

$$\beta_\delta \coloneqq \log\left(\frac{M-1}{\delta}\right). \tag{I.10}$$

Algorithm 4 Persistence Track-and-Stop

Require: confidence $\delta \in (0,1)$, exponent $\gamma \in (1/2, 1)$
Set $k \leftarrow 0$ and $N_e(0) \leftarrow 0$ for all e.
if $|\mathcal{I}| = 1$ **then**
  return the unique canonical target strategy.
end if
while true **do**
  Compute $\widehat{\theta}_k \in \arg\max_\vartheta \mathsf{L}_k(\vartheta)$.
  if $F(k+1) > F(k)$ **then**
    Set $j \leftarrow 1 + ((F(k+1) - 1) \bmod m)$ and $E_{k+1} \leftarrow e_j$.
  else
    Draw $E_{k+1} \sim p^\star(\widehat{\theta}_k)$.
  end if
  Reset to a fresh source task, execute E_{k+1}, and observe D_{k+1}.
  Update all model likelihoods and set $k \leftarrow k+1$.
  if $\max_{i \in \mathcal{I}} Z_k(i) \geq \beta_\delta$ **then**
    Choose the first maximizing $\widehat{i}_k$ in a fixed order and **return** $g_{\widehat{i}_k}$.
  end if
end while

The source cost at stopping is

$$\mathsf{C}^{\mathrm{blk}}_{\tau_\delta} = \sum_{s=1}^{\tau_\delta} c_\kappa(E_s).$$

The procedure is information-theoretic: it assumes exact block likelihoods, exact solution of equation I.4, and access to the canonical target strategies.

Common infinite sampling process. For the asymptotic statements below, construct a single infinite, unstopped process by running the experiment-selection and observation steps of Algorithm 4 forever and omitting only its stopping instruction. The resulting block sequence does not depend on δ. For each confidence level δ, define τ_δ as the first block index on this common process at which the threshold β_δ is crossed. The actually stopped algorithm is the truncation of this common process at τ_δ. This coupling makes every almost-sure statement as $\delta \downarrow 0$ pathwise meaningful.

I.4 Time-uniform correctness

Lemma I.6 (Adaptive pairwise likelihood-ratio martingale)**.** *Fix $\theta, \lambda \in \Theta$. Under global model θ,*

$$\mathcal{M}_k(\lambda, \theta) := \exp\left(\mathsf{L}_k(\lambda) - \mathsf{L}_k(\theta)\right), \qquad k \geq 0, \tag{I.11}$$

is a nonnegative martingale with initial value one. Consequently, for every $x > 0$,

$$\mathbb{P}_\theta\left(\sup_{k\geq 0}\{\mathsf{L}_k(\lambda) - \mathsf{L}_k(\theta)\} \geq x\right) \leq e^{-x}. \tag{I.12}$$

Proof. Conditionally on $(\mathcal{B}_{k-1}, E_k = e)$, the block law is Q^e_θ. Hence

$$\mathbb{E}_\theta\left[\frac{q^e_\lambda(D_k)}{q^e_\theta(D_k)}\middle|\mathcal{B}_{k-1}, E_k = e\right] = 1.$$

The predictable experiment-selection kernel therefore preserves the martingale property. The time-uniform bound is Ville's inequality for a nonnegative martingale. □

Theorem I.7 (Anytime wrong-cell control)**.** *Assume Assumption I.1 and $M \geq 2$. Fix any true model $\theta \in \Theta$, any $\delta \in (0, 1)$, and any predictable block-selection rule, not necessarily the rule in Algorithm 4. Then*

$$\mathbb{P}_\theta\left(\exists k \geq 1,\ \exists i \neq a(\theta) : Z_k(i) \geq \beta_\delta\right) \leq \delta. \tag{I.13}$$

Consequently, for any procedure that recommends g_i only when $Z_k(i) \geq \beta_\delta$, the unconditional probability of returning a strategy outside $\mathcal{G}_{B,\varepsilon}(\theta)$ is at most δ. This theorem does not assert that the procedure stops.

Proof. Fix the true model θ and write $i_\theta = a(\theta)$. If, at some block k, a wrong cell $i \neq i_\theta$ satisfies $Z_k(i) \geq \beta_\delta$, choose $\lambda_k \in \Theta_i$ attaining the first maximum in equation I.9. Because $\theta \notin \Theta_i$, the second maximum is at least $\mathsf{L}_k(\theta)$, and therefore

$$\mathsf{L}_k(\lambda_k) - \mathsf{L}_k(\theta) \geq \beta_\delta.$$

Thus the wrong-cell event is contained in

$$\bigcup_{\lambda\in\Theta\setminus\{\theta\}} \left\{\sup_{k\geq 0}\big(\mathsf{L}_k(\lambda) - \mathsf{L}_k(\theta)\big) \geq \beta_\delta\right\}.$$

Applying equation I.12, a union bound over at most $M - 1$ models, and the definition $\beta_\delta = \log\big((M-1)/\delta\big)$ gives

$$\mathbb{P}_\theta(\text{some wrong cell crosses}) \leq (M-1)e^{-\beta_\delta} = \delta.$$

If the returned cell is i_θ, its canonical strategy is ε-optimal under θ by Assumption I.1. No stopping claim has been used. □

The proof uses a constant threshold because the global model class is finite and every wrong recommendation implies that one simple likelihood ratio has crossed the boundary. More refined mixture-martingale thresholds become useful for composite or continuous model classes (Kaufmann & Koolen, 2021).

I.5 Forced exploration and maximum-likelihood consistency

Let $N_e^{\mathrm{F}}(k)$ be the number of forced selections of experiment e through block k.

Lemma I.8 (Deterministic forced counts)**.** *For every $e \in \mathcal{E}$ and every $k \geq 1$,*

$$N_e^{\mathrm{F}}(k) \geq \left\lfloor \frac{F(k)}{m} \right\rfloor, \qquad \sum_e N_e^{\mathrm{F}}(k) = F(k). \tag{I.14}$$

In particular, $N_e^{\mathrm{F}}(k) \to \infty$ and $F(k)/k \to 0$.

Proof. The $F(k)$ forced blocks are assigned cyclically to the m experiments, so every experiment receives either $\lfloor F(k)/m \rfloor$ or $\lceil F(k)/m \rceil$ forced selections. Since $\gamma < 1$, $F(k) = o(k)$. □

For $\theta \neq \lambda$, define the log-likelihood advantage of the true model

$$\Delta_k(\theta, \lambda) := \mathsf{L}_k(\theta) - \mathsf{L}_k(\lambda), \tag{I.15}$$

and its predictable drift

$$A_k(\theta, \lambda) := \sum_{s=1}^{k} d_{E_s}(\theta, \lambda). \tag{I.16}$$

Then $\Delta_k(\theta, \lambda) - A_k(\theta, \lambda)$ is a martingale with increments bounded in absolute value by $2L$.

Lemma I.9 (Stretched-exponential MLE error)**.** *Let*

$$\rho_\gamma := 2\gamma - 1 \in (0, 1).$$

Under Assumptions I.3 and I.4, for every true model θ there are constants $A_\theta < \infty$, $b_\theta > 0$, and $k_\theta < \infty$ such that

$$\mathbb{P}_\theta(\widehat{\theta}_k \neq \theta) \leq A_\theta \exp(-b_\theta k^{\rho_\gamma}) \qquad (k \geq k_\theta). \tag{I.17}$$

Consequently, $\widehat{\theta}_k = \theta$ for all sufficiently large k almost surely, and

$$\mathbb{E}_\theta \left[\sum_{k \geq 0} \mathbf{1}\{\widehat{\theta}_k \neq \theta\} \right] < \infty. \tag{I.18}$$

Proof. Fix $\lambda \neq \theta$ and choose an experiment with divergence $d_e(\theta, \lambda) = \max_f d_f(\theta, \lambda)$. By equation I.14 and equation I.3, for all large k,

$$A_k(\theta, \lambda) \geq \underline{d} \left\lfloor \frac{F(k)}{m} \right\rfloor \geq \frac{\underline{d}}{2m} k^\gamma.$$

Azuma–Hoeffding applied to the centered log-likelihood martingale gives

$$\begin{aligned} \mathbb{P}_\theta\left(\Delta_k(\theta, \lambda) \leq 0\right) &\leq \exp\left(-\frac{A_k(\theta, \lambda)^2}{8L^2 k}\right) \\ &\leq \exp(-bk^{2\gamma - 1}) \end{aligned}$$

for a constant $b > 0$. If the MLE is not θ, then some $\lambda \neq \theta$ has likelihood at least that of θ. A finite union bound proves equation I.17. The series $\sum_k \exp(-bk^{\rho_\gamma})$ is finite, so Borel–Cantelli and Tonelli's theorem give the two conclusions. □

I.6 Tracking the cost-optimal design

Fix the true model θ and abbreviate

$$p_e^\star = p_e^\star(\theta), \qquad w_e^\star = w_e^\star(\theta), \qquad \bar{c} = \bar{c}_\theta.$$

Let

$$\mathsf{C}_k^{\mathrm{blk}} = \sum_{s=1}^{k} c_\kappa(E_s)$$

denote cost after a deterministic number k of blocks, extending the sampling rule beyond its eventual stopping time when needed for analysis.

Lemma I.10 (Allocation concentration)**.** *For every $\eta > 0$, there exist constants $A_\eta < \infty$, $b_\eta > 0$, and $k_\eta < \infty$ such that*

$$\mathbb{P}_\theta\left(\max_{e\in\mathcal{E}}\left|\frac{N_e(k)}{k} - p_e^\star\right| > \eta\right) \le A_\eta \exp(-b_\eta k^{\rho_\gamma}) \qquad (k \ge k_\eta). \tag{I.19}$$

Consequently,

$$\frac{N_e(k)}{k} \longrightarrow p_e^\star, \qquad \frac{\mathsf{C}_k^{\mathrm{blk}}}{k} \longrightarrow \bar{c}, \qquad \frac{N_e(k)}{\mathsf{C}_k^{\mathrm{blk}}} \longrightarrow w_e^\star \tag{I.20}$$

almost surely for every e. The convergence of cost per block also holds in L^1.

Proof. Fix $\eta > 0$ and let $q_k = \lfloor \eta k/8 \rfloor$. By Lemma I.9 and a standard integral bound for stretched exponentials,

$$\mathbb{P}_\theta\left(\exists s \ge q_k : \widehat{\theta}_s \ne \theta\right) \le A\exp(-bk^{\rho_\gamma})$$

for adjusted constants. On the complementary event, every nonforced choice after block q_k is drawn from the fixed distribution $p^\star$.

For each e, the centered randomized-selection sum over nonforced blocks

$$\sum_{s=1}^{k} \mathbf{1}\{s \text{ nonforced}\}\left(\mathbf{1}\{E_s = e\} - p_e^\star(\widehat{\theta}_{s-1})\right)$$

is a martingale with increments bounded by one. The first q_k blocks can alter a count by at most q_k, and the forced schedule contributes at most $2F(k)$ to the discrepancy from $kp_e^\star$. For all sufficiently large k, $q_k + 2F(k) \le \eta k/2$. Azuma–Hoeffding and a union bound over the finite experiment class therefore yield equation I.19.

The probability bound is summable, so the almost-sure frequency convergence follows by Borel–Cantelli, first on a countable sequence of η values and then by monotonicity. The cost convergence follows by the finite weighted sum

$$\frac{\mathsf{C}_k^{\mathrm{blk}}}{k} = \sum_e c_\kappa(e)\frac{N_e(k)}{k}.$$

The final limit uses equation I.6. Since $\mathsf{C}_k^{\mathrm{blk}}/k \in [c_{\min}, c_{\max}]$, bounded convergence also gives L^1 convergence. □

I.7 Growth of target-relevant evidence

For $\lambda \ne \theta$, define the oracle block-time information

$$I_\theta(\lambda) := \sum_{e\in\mathcal{E}} p_e^\star(\theta) d_e(\theta, \lambda). \tag{I.21}$$

By Lemma I.5,

$$\min_{\lambda\in \mathrm{Alt}_{B,\varepsilon}(\theta)} I_\theta(\lambda) = \bar{c}_\theta \Gamma^\star_{\mathrm{pers}}(\theta; B, \varepsilon, \kappa, \mathcal{E}). \tag{I.22}$$

Lemma I.11 (Uniform evidence concentration)**.** *Assume the target-relevant alternative set of θ is nonempty, and write*

$$\Gamma_\theta := \Gamma^\star_{\mathrm{pers}}(\theta; B, \varepsilon, \kappa, \mathcal{E}) > 0.$$

For every $\eta \in (0, \Gamma_\theta)$ there exist constants $A_\eta < \infty$, $b_\eta > 0$, and $k_\eta < \infty$ such that

$$\mathbb{P}_\theta\left(Z_k(a(\theta)) < (\Gamma_\theta - \eta)\mathsf{C}_k^{\mathrm{blk}}\right) \le A_\eta \exp(-b_\eta k^{\rho_\gamma}) \qquad (k \ge k_\eta). \tag{I.23}$$

Moreover, almost surely,

$$\frac{Z_k(a(\theta))}{\mathsf{C}_k^{\mathrm{blk}}} \longrightarrow \Gamma_\theta. \tag{I.24}$$

Proof. For each $\lambda \neq \theta$,

$$\Delta_k(\theta, \lambda) - \sum_e N_e(k) d_e(\theta, \lambda)$$

is a martingale with increments bounded by $2L$. Azuma–Hoeffding, the allocation bound equation I.19, and a finite union bound over λ show that, for every fixed $\zeta > 0$,

$$\mathbb{P}_\theta \left(\max_{\lambda \neq \theta} \left| \frac{\Delta_k(\theta, \lambda)}{k} - I_\theta(\lambda) \right| > \zeta \right) \leq A_\zeta \exp(-b_\zeta k^{\rho_\gamma}). \tag{I.25}$$

The same allocation event controls $\mathsf{C}_k^{\mathrm{blk}}/k - \bar{c}_\theta$.

With probability at least the complement of a stretched-exponential event, θ is the unique maximum-likelihood model and all pairwise rates and the cost rate are uniformly close to their limits. On that event,

$$Z_k(a(\theta)) = \mathsf{L}_k(\theta) - \max_{\lambda \in \mathrm{Alt}_{B,\varepsilon}(\theta)} \mathsf{L}_k(\lambda) = \min_{\lambda \in \mathrm{Alt}_{B,\varepsilon}(\theta)} \Delta_k(\theta, \lambda).$$

Using equation I.22, and choosing the intermediate accuracy ζ sufficiently small relative to η, proves equation I.23. Summability and Borel–Cantelli give the almost-sure pairwise information rates; division by the positive cost limit yields equation I.24. □

Lemma I.12 (One-block cell-GLR increment)**.** *Under Assumption I.4, for every cell $i \in \mathcal{I}$ and every $k \geq 0$,*

$$|Z_{k+1}(i) - Z_k(i)| \leq L \qquad \textit{almost surely}. \tag{I.26}$$

Proof. For the realized block (E_{k+1}, D_{k+1}), define

$$r_\vartheta := \log q_\vartheta^{E_{k+1}}(D_{k+1}).$$

If $A_k = \max_{\vartheta \in \Theta_i} \mathsf{L}_k(\vartheta)$, then

$$\min_{\vartheta \in \Theta_i} r_\vartheta \leq A_{k+1} - A_k \leq \max_{\vartheta \in \Theta_i} r_\vartheta.$$

The analogous statement holds for the maximum over $\Theta \setminus \Theta_i$. Hence the increment of their difference lies between the minimum and maximum pairwise differences $r_\vartheta - r_\lambda$. Assumption I.4 therefore gives equation I.26. Notice that the constant is L, not $2L$, because the assumption bounds every pairwise log-density difference directly. □

Proposition I.13 (Almost-sure stopping and pathwise characteristic time)**.** *Assume Assumptions H.1, I.1, I.3, and I.4. Fix a true model θ with $\mathrm{Alt}_{B,\varepsilon}(\theta) \neq \varnothing$ and $\Gamma_\theta > 0$. Then, on the common infinite process defined above, $\tau_\delta < \infty$ $\mathbb{P}_\theta$-almost surely for every $\delta \in (0, 1)$. Moreover, as $\delta \downarrow 0$,*

$$\frac{\mathsf{C}_{\tau_\delta}^{\mathrm{blk}}}{\beta_\delta} \longrightarrow \frac{1}{\Gamma_\theta} \qquad \mathbb{P}_\theta\textit{-almost surely}. \tag{I.27}$$

Proof. Write $i_\theta = a(\theta)$. We first rule out wrong-cell crossings at all sufficiently large thresholds. For any $i \neq i_\theta$, because $\theta \notin \Theta_i$,

$$Z_k(i) \leq \max_{\lambda \in \Theta_i} \{\mathsf{L}_k(\lambda) - \mathsf{L}_k(\theta)\}. \tag{I.28}$$

By the allocation and likelihood-ratio laws of large numbers proved above, for every $\lambda \neq \theta$,

$$\frac{\mathsf{L}_k(\lambda) - \mathsf{L}_k(\theta)}{\mathsf{C}_k^{\mathrm{blk}}} \longrightarrow -\frac{I_\theta(\lambda)}{\bar{c}_\theta} \qquad \mathbb{P}_\theta\text{-a.s.}$$

If λ lies in a wrong cell, then $\lambda \in \mathrm{Alt}_{B,\varepsilon}(\theta)$ and $I_\theta(\lambda) \geq \bar{c}_\theta \Gamma_\theta > 0$. Since every wrong cell is finite, equation I.28 implies

$$\limsup_{k \to \infty} \frac{Z_k(i)}{\mathsf{C}_k^{\mathrm{blk}}} \leq -\Gamma_\theta < 0 \qquad \mathbb{P}_\theta\text{-a.s.} \tag{I.29}$$

for every $i \neq i_\theta$. Hence $\sup_{k \geq 1} Z_k(i) < \infty$ almost surely. Because there are finitely many wrong cells, the random quantity

$$B_{\mathrm{wrong}} := \max_{i \neq i_\theta} \sup_{k \geq 1} Z_k(i)$$

is finite almost surely.

For the true cell, equation I.24 gives

$$\frac{Z_k(i_\theta)}{\mathsf{C}_k^{\mathrm{blk}}} \longrightarrow \Gamma_\theta > 0,$$

so $Z_k(i_\theta) \to +\infty$ and every fixed threshold is crossed in finite time. This proves $\tau_\delta < \infty$ almost surely. On every sample path for which the preceding conclusions hold, all sufficiently large thresholds $\beta > B_{\mathrm{wrong}}$ can be crossed only by the true cell. Furthermore, for every fixed integer K,

$$\max_{1\le k\le K} \max_{i\in\mathcal{I}} Z_k(i) < \infty.$$

Therefore the first crossing time of a threshold $\beta \to \infty$ cannot remain in a bounded set of block indices. In particular, $\tau_\delta \to \infty$ as $\delta \downarrow 0$ on the common infinite process.

For all sufficiently small δ, the stopping cell is i_θ. By definition of the first crossing and Lemma I.12,

$$\beta_\delta \le Z_{\tau_\delta}(i_\theta) < \beta_\delta + L.$$

Since $\tau_\delta \to \infty$ and the true-cell evidence-to-cost ratio converges to Γ_θ,

$$\frac{Z_{\tau_\delta}(i_\theta)}{\mathsf{C}_{\tau_\delta}^{\mathrm{blk}}} \longrightarrow \Gamma_\theta.$$

Combining the last two displays and using $L/\beta_\delta \to 0$ proves equation I.27. □

Corollary I.14 (Global fixed-confidence correctness of Persistence Track-and-Stop)**.** *Under the assumptions of Proposition I.13, Algorithm 4 stops almost surely and satisfies*

$$\mathbb{P}_\theta\left(g_{\widehat{i}_{\tau_\delta}} \notin \mathcal{G}_{B,\varepsilon}(\theta)\right) \le \delta.$$

If $|\mathcal{I}| = 1$, the same conclusion holds with zero source cost.

Proof. Almost-sure stopping is Proposition I.13. By Theorem I.7, the probability that any wrong cell crosses the threshold is at most δ. On the complementary event, the returned cell is $a(\theta)$ and its canonical strategy is ε-optimal by Assumption I.1. The one-cell case is immediate from the same assumption. □

I.8 Expected-cost asymptotic optimality

Pathwise convergence does not by itself imply convergence of expected source cost. The stretched-exponential evidence bound supplies the required uniform-integrability control.

Theorem I.15 (Asymptotically optimal expected source cost)**.** *Suppose Assumptions H.1, I.1, I.3, and I.4 hold. Fix a true model θ with $\mathrm{Alt}_{B,\varepsilon}(\theta) \neq \varnothing$ and $\Gamma_\theta > 0$. Then Algorithm 4 stops almost surely, is (B,ε,δ)-correct, and*

$$\limsup_{\delta\downarrow 0} \frac{\mathbb{E}_\theta[\mathsf{C}_{\tau_\delta}^{\mathrm{blk}}]}{\log(1/\delta)} \le \frac{1}{\Gamma_\theta}. \tag{I.30}$$

Together with Theorem H.9,

$$\lim_{\delta\downarrow 0} \frac{\mathfrak{C}_{B,\varepsilon,\delta,\kappa}^{\mathrm{blk},\star}(\theta;\mathcal{E})}{\log(1/\delta)} = \frac{1}{\Gamma_{\mathrm{pers}}^\star(\theta;B,\varepsilon,\kappa,\mathcal{E})}. \tag{I.31}$$

Proof. Almost-sure stopping and correctness are Corollary I.14. Write $\Gamma = \Gamma_\theta$ and $\bar{c} = \bar{c}_\theta$. Fix

$$0 < \eta < \frac{1}{2}\min\{\Gamma, \bar{c}\}$$

and define

$$k_\delta(\eta) := \left\lceil \frac{\beta_\delta}{(\Gamma-\eta)(\bar{c}-\eta)} \right\rceil. \tag{I.32}$$

For every $k \geq k_\delta(\eta)$, on the event

$$\mathsf{C}_k^{\mathrm{blk}} \geq (\bar{c} - \eta)k \quad \text{and} \quad Z_k(a(\theta)) \geq (\Gamma - \eta)\mathsf{C}_k^{\mathrm{blk}},$$

the true-cell statistic is at least β_δ, so $\tau_\delta \leq k$. The cost concentration in Lemma I.10 and the evidence concentration in Lemma I.11 therefore imply constants $A, b > 0$, independent of sufficiently small δ, such that

$$\mathbb{P}_\theta(\tau_\delta > k) \leq A \exp(-bk^{\rho_\gamma}) \qquad \text{for all } k \geq k_\delta(\eta). \tag{I.33}$$

Extend the experiment-selection process after stopping using the same rule. Because costs are positive,

$$\mathsf{C}_{\tau_\delta}^{\mathrm{blk}} \leq \mathsf{C}_{k_\delta(\eta)}^{\mathrm{blk}} + c_{\max}(\tau_\delta - k_\delta(\eta))_+.$$

Taking expectations and using the tail-sum formula gives

$$\mathbb{E}_\theta[\mathsf{C}_{\tau_\delta}^{\mathrm{blk}}] \leq \mathbb{E}_\theta[\mathsf{C}_{k_\delta(\eta)}^{\mathrm{blk}}] \tag{I.34}$$

$$+ c_{\max} \sum_{k \geq k_\delta(\eta)} \mathbb{P}_\theta(\tau_\delta > k). \tag{I.35}$$

The second term is $o(\beta_\delta)$ by equation I.33. By the L^1 cost convergence in Lemma I.10,

$$\frac{\mathbb{E}_\theta[\mathsf{C}_{k_\delta(\eta)}^{\mathrm{blk}}]}{k_\delta(\eta)} \longrightarrow \bar{c}.$$

Hence

$$\limsup_{\delta \downarrow 0} \frac{\mathbb{E}_\theta[\mathsf{C}_{\tau_\delta}^{\mathrm{blk}}]}{\beta_\delta} \leq \frac{\bar{c}}{(\Gamma - \eta)(\bar{c} - \eta)}.$$

Letting $\eta \downarrow 0$ gives $1/\Gamma$. Finally, $\beta_\delta / \log(1/\delta) \to 1$, proving equation I.30. The lower bound in Theorem H.9 and $\mathrm{kl}(1-\delta, \delta)/\log(1/\delta) \to 1$ give the matching limit equation I.31. □

I.9 Consistency with the parity hierarchy

For the parity family of Section 3, let the available experiment class contain the complete r-episode block and any collection of subcritical blocks. For the two decision cells indexed by $\sigma \in \{-1, +1\}$,

$$\Gamma^\star_{\mathrm{pers}}(\sigma; r-1, \varepsilon, \kappa, \mathcal{E}) = \frac{d_{\mathrm{par}}(\varrho)}{r + \kappa}$$

by Proposition H.11. Therefore Theorem I.15 gives

$$\mathbb{E}_\sigma[\mathsf{C}_{\tau_\delta}^{\mathrm{blk}}] \sim \frac{r + \kappa}{d_{\mathrm{par}}(\varrho)} \log \frac{1}{\delta}. \tag{I.36}$$

This matches the exact lower bound from Corollary 3.2.

I.10 Scope of the upper theorem

The theorem above establishes the exact leading characteristic time for a finite decision-partition problem with finite deterministic block experiments. Three extensions require additional ideas.

1. *Multiple correct answers.* If acceptable-strategy sets overlap without forming a partition, the correct complexity is answer-specific and ordinary tracking can be suboptimal. Sticky Track-and-Stop resolves the analogous issue in bandit pure exploration (Degenne & Koolen, 2019).
2. *Random within-task stopping.* If a block may decide online when to reset, both its law and cost are controlled stopping-time objects. The information game must then be formulated over persistent-task renewal experiments.
3. *Continuous global models.* Finite union bounds and eventual exact MLE recovery no longer apply. One needs confidence regions, local alternatives, and time-uniform composite likelihood control, for which mixture martingales provide a natural route (Kaufmann & Koolen, 2021).

J GROUP-ACTION IDENTIFIABILITY

This appendix develops the algebraic and decision-theoretic foundations of the structured source model studied in Section 5. The classical statistical theory of group invariance asks how a group action transforms a statistical experiment and its decision rules (Eaton, 1989; Diaconis, 1988). Our setting introduces an additional control variable: the learner decides whether two controlled observations are generated from independent task anchors or from the same persistent anchor. The former operation exposes new points of the group action, whereas the latter reveals correspondences between two views of one point. Stabilizers and bases are standard objects in the theory of group actions (Dixon & Mortimer, 1996); here they provide exact source-task identifiability and target-risk criteria.

We separate three notions that are sometimes conflated. First, *distributional identifiability* asks whether two group parameters induce the same law for an indefinitely repeated experiment. Second, *samplewise exact identification* asks whether one finite noiseless sample has a singleton consistency class. Third, *target relevance* asks whether any residual symmetry changes a fresh-task decision. The paired experiment is distributionally identifiable modulo the action kernel after one block type is available, but finite-sample exact identification may require several independent task anchors.

J.1 TOPOLOGICAL GROUP-ACTION MODEL

Throughout this appendix, let G be a second-countable compact Hausdorff group with identity element e and normalized Haar probability measure λ_G. Let $(\mathcal{X}, d)$ be a Polish metric space, and let

$$G \times \mathcal{X} \longrightarrow \mathcal{X}, \qquad (g, x) \longmapsto g \cdot x,$$

be a continuous left action. Thus $e \cdot x = x$ and $g \cdot (h \cdot x) = (gh) \cdot x$. Let $\mu \in \mathcal{P}(\mathcal{X})$ be a G-invariant Borel probability measure:

$$\mu(g^{-1} \cdot A) = \mu(A) \qquad \text{for every } g \in G \text{ and } A \in \mathcal{B}(\mathcal{X}), \tag{J.1}$$

where $g^{-1} \cdot A = \{x : g \cdot x \in A\}$. Compactness is not needed for the elementary consistency-set identities below, but it guarantees normalized Haar measures on G and its closed subgroups and permits the posterior and Bayes-risk analysis in Section J.4.

For $x \in \mathcal{X}$, define its stabilizer

$$\mathrm{Stab}_G(x) \coloneqq \{g \in G : g \cdot x = x\}. \tag{J.2}$$

For a tuple $x_{1:n} = (x_1, \ldots, x_n)$, define the pointwise stabilizer

$$G_{x_{1:n}} \coloneqq \bigcap_{i=1}^{n} \mathrm{Stab}_G(x_i), \qquad G_{x_{1:0}} \coloneqq G. \tag{J.3}$$

Every stabilizer is closed because the action is continuous and $\mathcal{X}$ is Hausdorff.

The source distribution need not have full support. The statistically unavoidable ambiguity is therefore the kernel of the action on $\mathrm{supp}(\mu)$ rather than on all of $\mathcal{X}$.

Definition J.1 (Support kernel)**.** Define

$$K_\mu \coloneqq \bigcap_{x \in \mathrm{supp}(\mu)} \mathrm{Stab}_G(x). \tag{J.4}$$

We call the action *μ-faithful* when $K_\mu = \{e\}$.

Lemma J.2 (Basic properties of the support kernel)**.** *The subgroup K_μ is closed and normal in G. Moreover, for every $g \in G$, the following are equivalent:*

1. *$g \in K_\mu$;*
2. *$g \cdot x = x$ for μ-almost every x;*
3. *$(g \cdot X, X)$ is concentrated on the diagonal when $X \sim \mu$.*

Consequently, the action factors faithfully through the quotient group G/K_μ on $\mathrm{supp}(\mu)$.

Proof. Closedness follows from equation J.4 because each point stabilizer is closed. The support is G-invariant: if $x \in \mathrm{supp}(\mu)$ and U is an open neighborhood of $g \cdot x$, then $g^{-1} \cdot U$ is an open neighborhood of x, and invariance gives $\mu(U) = \mu(g^{-1} \cdot U) > 0$. Hence, for $k \in K_\mu$, $a \in G$, and $x \in \mathrm{supp}(\mu)$,

$$(aka^{-1}) \cdot x = a \cdot \big(k \cdot (a^{-1} \cdot x)\big) = x,$$

so $aka^{-1} \in K_\mu$ and K_μ is normal.

The implication (i)⇒(ii) is immediate because $X \in \mathrm{supp}(\mu)$ almost surely. Conversely, the fixed-point set $F_g = \{x : g \cdot x = x\}$ is closed. If $\mu(F_g) = 1$, then every point in $\mathrm{supp}(\mu)$ belongs to F_g: otherwise the open set F_g^c would contain a support point and have positive μ-measure. Thus (ii) implies (i). The equivalence of (ii) and (iii) is merely a restatement. Normality then makes the quotient action well defined, and its kernel is trivial by construction. □

J.2 A PERSISTENT GROUP-ACTION META-MDP

We next realize the two-view experiment as a valid episodic MDP. Let the global parameter be $g_\star \in G$ and the task latent be an anchor $X \sim \mu$. Define the horizon-one MDP with action set $\mathcal{A} = \{0, 1\}$, zero reward, initial state $s_\circ$, and terminal state space $\{0, 1\} \times \mathcal{X}$. Conditional on anchor x, action v transitions deterministically to

$$\big(v, T_v^{g_\star}(x)\big), \qquad T_0^g(x) \coloneqq x, \qquad T_1^g(x) \coloneqq g \cdot x. \tag{J.5}$$

Every physical episode restarts at $s_\circ$. A task reset draws an independent $X \sim \mu$, while continuation preserves the current anchor. Thus a two-episode block using actions $(0, 1)$ observes the paired views $(X, g_\star \cdot X)$. This minimal MDP is only an exact realization of the statistical interface; all results below also apply whenever two policies in a larger MDP admit trajectory statistics with the laws in equation J.5.

A reset-only learner sees at most one view of every independent anchor. The following proposition permits arbitrary adaptation across tasks and random source stopping.

Proposition J.3 (Reset-only nonidentifiability)**.** *Consider any possibly randomized source learner that resets before every source episode in the MDP above. Its view action at episode t may be an arbitrary measurable function of the previously observed view actions and terminal states, and it may stop at any almost surely finite observed-history stopping time. Then the complete stopped observed transcript has the same law for every $g_\star \in G$. In particular, for all $g, g' \in G$,*

$$D_{\mathrm{KL}}\left(\mathbb{P}_g^{\mathcal{A}} \middle\| \mathbb{P}_{g'}^{\mathcal{A}}\right) = 0. \tag{J.6}$$

Under an arbitrary prior on $g_\star$, the reset-only transcript is independent of $g_\star$.

Proof. Condition on an observed history before a source episode. The learner's next view action has the same conditional kernel under every global parameter. After a reset, $X \sim \mu$ is independent of the past. If action 0 is chosen, the observation has law μ. If action 1 is chosen, its law is the pushforward $(g_\star)_\# \mu$, which equals μ by equation J.1. Thus the conditional law of the next action–observation pair is independent of $g_\star$. Induction gives equality of every finite-prefix law. Padding a stopped transcript by a cemetery symbol, or equivalently summing over stopped finite histories, gives the same conclusion under random stopping. Equality of the transcript laws implies equation J.6; independence under a prior follows by integrating a parameter-independent conditional law. □

The paired experiment has the same two one-view marginals but a different coupling. Write

$$Q_g^{\mathrm{pair}} \coloneqq \big(x \mapsto (x, g \cdot x)\big)_\# \mu, \qquad Q_g^{\mathrm{fresh}} \coloneqq \mu \otimes \mu. \tag{J.7}$$

The second law corresponds to observing view 0 on one task and view 1 on an independently reset task.

Proposition J.4 (Distributional identifiability of the paired law)**.** *For every $g, g' \in G$,*

$$Q_g^{\mathrm{pair}} = Q_{g'}^{\mathrm{pair}} \quad \Longleftrightarrow \quad g^{-1} g' \in K_\mu. \tag{J.8}$$

In contrast, $Q_g^{\mathrm{fresh}} = \mu \otimes \mu$ for every g. Therefore repeated paired blocks identify the quotient parameter $gK_\mu \in G/K_\mu$ at the level of the statistical experiment, whereas arbitrarily many unpaired fresh views contain no information about it.

Proof. If $g^{-1}g' \in K_\mu$, then $g \cdot x = g' \cdot x$ for μ-almost every x by Lemma J.2, so the pair laws agree. Conversely, consider the Borel graph

$$\mathsf{Graph}(g) \coloneqq \{(x, y) \in \mathcal{X}^2 : y = g \cdot x\}.$$

It is closed because the action is continuous. Under Q_g^{pair} this graph has probability one. Equality of the pair laws therefore gives

$$1 = Q_{g'}^{\mathrm{pair}}\big(\mathsf{Graph}(g)\big) = \mu\big(\{x : g' \cdot x = g \cdot x\}\big).$$

Hence $g^{-1}g'$ fixes μ-almost every anchor and belongs to K_μ. The statement for independent fresh views follows from invariance of μ. □

J.3 Consistency cosets and source-task bases

The preceding proposition concerns identifiability from the law of infinitely many paired blocks. We now characterize exactly what one finite noiseless sample determines. For deterministic anchors $x_{1:n}$ and paired outputs $y_i = g_\star \cdot x_i$, define the consistency set

$$\mathcal{C}(x_{1:n}, y_{1:n}) \coloneqq \{g \in G : g \cdot x_i = y_i \text{ for every } i \in [n]\}. \tag{J.9}$$

Theorem J.5 (Exact consistency coset)**.** *For every $g_\star \in G$ and every anchor tuple $x_{1:n} \in \mathcal{X}^n$,*

$$\mathcal{C}\big(x_{1:n}, g_\star \cdot x_{1:n}\big) = g_\star G_{x_{1:n}}. \tag{J.10}$$

If $x_i \in \mathrm{supp}(\mu)$ for every i, then $K_\mu \subseteq G_{x_{1:n}}$. Hence the quotient parameter $g_\star K_\mu$ is uniquely determined by the sample if and only if

$$G_{x_{1:n}} = K_\mu. \tag{J.11}$$

Under a μ-faithful action, exact identification of $g_\star$ is equivalent to $G_{x_{1:n}} = \{e\}$.

Proof. A parameter g is consistent with the paired sample precisely when $g \cdot x_i = g_\star \cdot x_i$ for every i. Applying $g_\star^{-1}$ on the left shows that this is equivalent to $g_\star^{-1}g \in G_{x_{1:n}}$, or $g \in g_\star G_{x_{1:n}}$. This proves equation J.10. Every element of K_μ fixes every support point and hence every x_i, giving $K_\mu \subseteq G_{x_{1:n}}$. Because K_μ is normal, all elements of the consistency set represent one quotient parameter exactly when the residual subgroup equals K_μ, proving equation J.11. □

The usual group-theoretic notion of a base asks for a tuple with trivial pointwise stabilizer. Since only $\mathrm{supp}(\mu)$ is sampled and K_μ is unavoidable, the appropriate statistical version is relative to the support kernel.

Definition J.6 (Support base, essential base size, and identification profile)**.** A tuple $x_{1:n} \in \mathrm{supp}(\mu)^n$ is a *μ-base* if $G_{x_{1:n}} = K_\mu$. Define the algebraic support-base size

$$b_{\mathrm{alg}}(G, \mu) \coloneqq \inf\{n \geq 0 : \text{there exists a } \mu\text{-base of length } n\}, \tag{J.12}$$

and the essential base size

$$b_{\mathrm{ess}}(G, \mu) \coloneqq \inf\left\{n \geq 0 : \mu^{\otimes n}\big(G_{X_{1:n}} = K_\mu\big) = 1\right\}. \tag{J.13}$$

We set the infimum of the empty set to $+\infty$. The exact-identification profile is

$$p_n(G, \mu) \coloneqq \mu^{\otimes n}\big(G_{X_{1:n}} = K_\mu\big). \tag{J.14}$$

The event in equation J.14 is Borel. Indeed, because G is compact metrizable, its nonempty compact subsets form a Polish hyperspace under the Hausdorff metric, and the compact-valued correspondence $x_{1:n} \mapsto G_{x_{1:n}}$ is upper hemicontinuous: its graph is closed by continuity of the action, and compactness of G upgrades closed graph to upper hemicontinuity. Hence the correspondence is Borel measurable and the preimage of the singleton $\{K_\mu\}$ is Borel.

Proposition J.7 (Finite-sample exact-identification profile)**.** *The sequence $p_n(G, \mu)$ is nondecreasing, and*

$$b_{\mathrm{alg}}(G, \mu) \leq b_{\mathrm{ess}}(G, \mu). \tag{J.15}$$

With n independent noiseless paired tasks, the quotient parameter is samplewise exactly identified with probability $p_n(G, \mu)$. In particular, if $b_{\mathrm{ess}}(G, \mu) < \infty$, then this parameter is recovered exactly almost surely from $b_{\mathrm{ess}}(G, \mu)$ paired tasks, while every smaller sample size fails with positive probability.

Proof. Since $G_{x_{1:n+1}} = G_{x_{1:n}} \cap \mathrm{Stab}_G(x_{n+1})$ and every support-point stabilizer contains K_μ, once the intersection equals K_μ it remains equal to K_μ after adding anchors. Thus the events in equation J.14 are increasing in n, proving monotonicity. If the event has probability one for some n, it is nonempty, so a μ-base of length n exists; this proves equation J.15. The exact-recovery statement follows directly from Theorem J.5. Finally, minimality in equation J.13 implies that for every smaller n the identification event has probability strictly below one. □

J.4 Haar posterior and residual-symmetry risk

The consistency coset is not merely an algebraic set: under the invariant prior it is the exact posterior support, with uniform Haar uncertainty along the residual stabilizer. This yields a general target lower bound. The required quotient integration formula is standard for compact groups and closed subgroups; see, for example, Eaton (1989); Diaconis (1988).

Fix deterministic anchors $x_{1:n}$ and write $H = G_{x_{1:n}}$. Let λ_H denote normalized Haar probability measure on the compact subgroup H. Let the global parameter $G_\star$ have prior λ_G, and observe noiseless pairs $Y_i = G_\star \cdot x_i$.

Proposition J.8 (Regular conditional Haar posterior on the consistency coset)**.** *Fix deterministic anchors $x_{1:n}$, put $H = G_{x_{1:n}}$, and let*

$$\mathcal{O}_x \coloneqq \{g \cdot x_{1:n} : g \in G\} \subseteq \mathcal{X}^n.$$

There exists a Borel Markov kernel $K_x : \mathcal{X}^n \rightsquigarrow G$ that is a regular conditional distribution of $G_\star \sim \lambda_G$ given $Y_{1:n} = G_\star \cdot x_{1:n}$ and that satisfies, for every orbit point $y = g_0 \cdot x_{1:n}$,

$$K_x(y, \cdot) = (h \mapsto g_0 h)_\# \lambda_H. \tag{J.16}$$

The right-hand side is independent of the representative g_0 of the coset $g_0 H$.

Proof. Let $q : G \to G/H$ be the quotient map. The orbit map $\phi_x(g) = g \cdot x_{1:n}$ is constant on the right cosets of H, and hence factors as

$$\phi_x = \bar{\phi}_x \circ q, \qquad \bar{\phi}_x : G/H \to \mathcal{O}_x.$$

By Theorem J.5, $\bar{\phi}_x$ is one-to-one; it is surjective by definition. Since G/H is compact and $\mathcal{O}_x$ is a compact subspace of the Hausdorff space $\mathcal{X}^n$, the continuous bijection $\bar{\phi}_x$ is a homeomorphism.

For a coset $c = gH \in G/H$, define

$$\nu_c \coloneqq (h \mapsto gh)_\# \lambda_H \in \mathcal{P}(G).$$

This does not depend on the representative: if $g' = gh_0$, then right invariance of λ_H gives $(h \mapsto g'h)_\# \lambda_H = \nu_c$. Moreover, the map $c \mapsto \nu_c$ is continuous for the weak topology on $\mathcal{P}(G)$. Indeed, for every $f \in C(G)$,

$$c = gH \longmapsto \int_G f(u)\, \nu_c(\mathrm{d}u) = \int_H f(gh)\, \lambda_H(\mathrm{d}h)$$

is continuous in g and constant on cosets, so it descends continuously to G/H.

Fix any probability measure $\nu_0 \in \mathcal{P}(G)$ and define the $\mathcal{P}(G)$-valued map

$$\mathsf{K}_x(y) = \begin{cases} \nu_{\bar{\phi}_x^{-1}(y)}, & y \in \mathcal{O}_x, \\ \nu_0, & y \notin \mathcal{O}_x. \end{cases}$$

The orbit is compact and therefore Borel, so K_x is Borel. Because G is compact metrizable, the evaluation maps $\nu \mapsto \nu(A)$ are Borel on $\mathcal{P}(G)$ for every Borel $A \subseteq G$. Hence $K_x(y, A) \coloneqq \mathsf{K}_x(y)(A)$ is a Borel Markov kernel.

Let $\bar{\lambda}_{G/H} = q_\# \lambda_G$. The compact-group quotient integration formula states that, for every bounded Borel $f : G \to \mathbb{R}$,

$$\int_G f(g)\, \lambda_G(\mathrm{d}g) = \int_{G/H} \int_H f(gh)\, \lambda_H(\mathrm{d}h)\, \bar{\lambda}_{G/H}(\mathrm{d}(gH)). \tag{J.17}$$

Therefore, for every bounded Borel $b : \mathcal{X}^n \to \mathbb{R}$ and $f : G \to \mathbb{R}$,

$$\begin{aligned}\mathbb{E}\big[b(Y_{1:n})f(G_\star)\big] &= \int_{G/H} b\big(\bar{\phi}_x(c)\big)\left[\int_G f(g)\,\nu_c(\mathrm{d}g)\right]\bar{\lambda}_{G/H}(\mathrm{d}c)\\ &= \int_{\mathcal{X}^n} b(y)\left[\int_G f(g)K_x(y,\mathrm{d}g)\right]\mathbb{P}_Y(\mathrm{d}y).\end{aligned}$$

Thus K_x is a regular conditional distribution of $G_\star$ given $Y_{1:n}$. On every orbit point it has the form equation J.16, completing the proof. □

For the next theorem, additionally assume that the action is isometric:

$$d(g\cdot x, g\cdot x') = d(x,x') \qquad \text{for all } g\in G,\quad x,x'\in\mathcal{X}, \tag{J.18}$$

and that $d^2(x_0, X)$ is integrable for one, hence every, x_0 in the relevant orbit. A fresh target anchor $X_\star \sim \mu$ is observed, and the prediction target is $G_\star \cdot X_\star$. An estimator may use the paired source data and $X_\star$.

Theorem J.9 (Residual-stabilizer target lower bound)**.** *Under the Haar prior on $G_\star$, condition on source anchors $X_{1:n} = x_{1:n}$, their noiseless paired outputs, and a target anchor $X_\star = x_\star$. Let $H = G_{x_{1:n}}$. Every estimator $\widehat{Y}$ satisfies*

$$\mathbb{E}\left[d^2\big(\widehat{Y}, G_\star\cdot x_\star\big)\Big| X_{1:n}, Y_{1:n}, X_\star\right] \geq \frac{1}{4}\max_{h\in H} d^2(x_\star, h\cdot x_\star). \tag{J.19}$$

Consequently, the minimax prediction risk dominates the Bayes risk and obeys

$$\inf_{\widehat{Y}}\sup_{g\in G}\mathbb{E}_g\left[d^2\big(\widehat{Y}, g\cdot X_\star\big)\right] \geq \frac{1}{4}\mathbb{E}\left[\max_{h\in G_{X_{1:n}}} d^2(X_\star, h\cdot X_\star)\right]. \tag{J.20}$$

The maximum in both displays exists by compactness and continuity.

Proof. Fix the conditioned data and write $A = \widehat{Y}$. By Proposition J.8, the posterior law of $G_\star$ is right H-invariant. Hence, for every $h \in H$,

$$\begin{aligned}R &\coloneqq \mathbb{E}\left[d^2(A, G_\star\cdot x_\star)\big| X_{1:n}, Y_{1:n}, X_\star\right]\\ &= \mathbb{E}\left[d^2(A, G_\star h\cdot x_\star)\big| X_{1:n}, Y_{1:n}, X_\star\right].\end{aligned}$$

For every realization of $G_\star$, the triangle inequality and $(a+b)^2 \leq 2(a^2+b^2)$ give

$$\begin{aligned}&d^2(A, G_\star\cdot x_\star) + d^2(A, G_\star h\cdot x_\star)\\ &\qquad\geq \frac{1}{2}d^2\big(G_\star\cdot x_\star, G_\star h\cdot x_\star\big) = \frac{1}{2}d^2(x_\star, h\cdot x_\star),\end{aligned}$$

where the last equality uses equation J.18. Taking the posterior expectation yields $2R \geq \frac{1}{2}d^2(x_\star, h\cdot x_\star)$, or $R \geq \frac{1}{4}d^2(x_\star, h\cdot x_\star)$. Maximizing over the compact subgroup H proves equation J.19.

Integrating over the source anchors and target anchor gives the Bayes-risk lower bound on the right-hand side of equation J.20. The minimax risk is at least the Bayes risk under every prior, in particular the normalized Haar prior, proving the claim. Measurability of the maximum follows from the measurable maximum theorem because the stabilizer correspondence has closed graph and compact values and the objective is continuous. □

Corollary J.10 (Target relevance of source breadth)**.** *Suppose there is positive probability that the residual stabilizer $G_{X_{1:n}}$ contains an element moving the fresh target anchor by a nonzero distance. Then the minimax target-prediction risk after n noiseless paired source tasks is strictly positive. Conversely, if $G_{X_{1:n}} = K_\mu$ almost surely, then every residual parameter ambiguity acts trivially on $X_\star$ almost surely.*

Proof. The first statement follows from equation J.20. For the second, every $h \in K_\mu$ fixes $\mathrm{supp}(\mu)$ pointwise by definition, and $X_\star \in \mathrm{supp}(\mu)$ almost surely. □

J.5 Structural consequence for reset and continuation

The preceding results isolate two complementary source resources in one model. Resetting generates independent anchors and thereby intersects pointwise stabilizers; continuation creates paired views and thereby exposes the action of the global parameter on each anchor.

Theorem J.11 (Reset–continue complementarity)**.** *In the persistent group-action meta-MDP:*

1. *every reset-only source learner receives a parameter-independent experiment;*
2. *a paired same-task block is distributionally identifiable modulo K_μ;*
3. *after n paired tasks, the exact finite-sample ambiguity is the right coset $g_\star G_{X_{1:n}}$;*
4. *the probability of exact quotient recovery is $p_n(G, \mu)$, and any residual stabilizer yields the target lower bound equation J.20.*

Thus continuation is necessary to create parameter-dependent couplings, while task breadth is necessary to shrink the remaining stabilizer. The orthogonal Gaussian specialization beginning in Section K computes the exact likelihood geometry; matching noisy minimax bounds are developed in Sections L and M.

Proof. Parts (i)–(iv) are respectively Propositions J.3, J.4 and J.7 and theorems J.5 and J.9. □

K The Noisy Orthogonal Model

This appendix specializes the group-action experiment of Section J to the standard orthogonal action of $\mathrm{O}(d)$ on $\mathbb{R}^d$. The specialization is deliberately chosen so that every quantity relevant to persistence can be computed exactly: the marginal paired law is Gaussian, the likelihood reduces to the classical orthogonal Procrustes objective, pairwise Kullback–Leibler divergence is an exact multiple of squared Frobenius distance, and the Fisher information is isotropic on the tangent space of $\mathrm{O}(d)$. The Procrustes and polar-factor formulas are classical (Schönemann, 1966; Higham, 1986); the manifold notation follows standard treatments of matrix manifolds (Absil et al., 2008). We include all derivations because the exact constants are used in the minimax upper and lower bounds of Sections L and M.

K.1 Persistent Gaussian two-view experiment

Fix $d \geq 2$. The unknown global parameter is $R_\star \in \mathrm{O}(d)$, and each fresh source task draws an independent anchor

$$X_i \sim \mathcal{N}(0, I_d). \tag{K.1}$$

A continuation action preserves X_i, while a reset samples a new anchor. For a fixed task i, a type-0 source episode and a type-1 source episode produce, respectively,

$$Y_{i,0}^{(k)} = X_i + \varepsilon_{i,0}^{(k)}, \qquad \varepsilon_{i,0}^{(k)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_0^2 I_d), \tag{K.2}$$

$$Y_{i,1}^{(\ell)} = R_\star X_i + \varepsilon_{i,1}^{(\ell)}, \qquad \varepsilon_{i,1}^{(\ell)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_1^2 I_d), \tag{K.3}$$

where all noises are mutually independent and independent of every anchor. This is a valid horizon-one persistent meta-MDP: the source action chooses the view, the terminal state records the noisy vector, every physical episode restarts, and only the task anchor persists.

Suppose a paired source task receives $m_0 \geq 1$ type-0 episodes and $m_1 \geq 1$ type-1 episodes. Define the sample means

$$\overline{Y}_{i,v} := \frac{1}{m_v} \sum_{k=1}^{m_v} Y_{i,v}^{(k)}, \qquad \tau_v^2 := \frac{\sigma_v^2}{m_v}, \qquad v \in \{0, 1\}. \tag{K.4}$$

Then

$$\overline{Y}_{i,0} = X_i + \eta_{i,0}, \qquad \overline{Y}_{i,1} = R_\star X_i + \eta_{i,1}, \qquad \eta_{i,v} \sim \mathcal{N}(0, \tau_v^2 I_d). \tag{K.5}$$

The next proposition shows that no information about $R_\star$ is lost by replacing all repeated episodes on one task with their two sample means.

Proposition K.1 (Exact reduction to averaged paired views)**.** *For fixed positive integers (m_0, m_1), the complete raw same-task experiment in equations (K.2)–(K.3) is statistically equivalent, for inference on $R_\star$, to observing the pair $(\overline{Y}_{i,0}, \overline{Y}_{i,1})$ in equation K.5. More precisely, there is an invertible orthogonal transformation of the raw observations under which the transformed data decompose as*

$$(\overline{Y}_{i,0}, \overline{Y}_{i,1}, U_i),$$

where U_i is independent of the two sample means and has a distribution that does not depend on $R_\star$. Consequently, the sample-mean pair is a sufficient statistic for $R_\star$; in particular, every f-divergence between two raw block laws equals the corresponding f-divergence between their sample-mean pair laws.

Proof. For $v \in \{0, 1\}$, choose an orthogonal matrix $A_v \in \mathbb{R}^{m_v \times m_v}$ whose first row is $m_v^{-1/2}(1, \dots, 1)$. Apply $A_v \otimes I_d$ to the stacked vector of all type-v observations. The first transformed coordinate is $\sqrt{m_v}\,\overline{Y}_{i,v}$. Every remaining coordinate is an orthogonal contrast whose coefficients sum to zero, so the common signal X_i or $R_\star X_i$ cancels and only transformed Gaussian noise remains. Orthogonality of A_v preserves the independent isotropic Gaussian noise law. Moreover, the mean coordinate and all contrast coordinates are independent because they are jointly Gaussian and orthogonal. Performing this transformation separately for the two views yields the asserted factorization. The transformation is invertible, and the contrast law is parameter independent, which proves statistical equivalence and sufficiency. Equality of f-divergences follows from this product factorization, or equivalently by applying data processing in both directions. □

A reset-only learner receives at most one view from each independent anchor. Each such marginal is rotationally invariant:

$$\overline{Y}_{i,0} \sim \mathcal{N}(0, (1 + \tau_0^2) I_d), \qquad \overline{Y}_{i,1} \sim \mathcal{N}(0, (1 + \tau_1^2) I_d), \tag{K.6}$$

independently of $R_\star$. Thus Proposition J.3 applies exactly: resetting without pairing yields zero information about the orthogonal transformation.

K.2 Exact marginal paired law

For compactness, write

$$a := 1 + \tau_0^2, \qquad b := 1 + \tau_1^2, \qquad \Delta_\tau := ab - 1 = \tau_0^2 + \tau_1^2 + \tau_0^2 \tau_1^2. \tag{K.7}$$

The regular noisy model is the regime $\Delta_\tau > 0$; the singular noiseless case is treated separately in Section K.5.

Lemma K.2 (Whitening equivalence and reduction to Δ_τ)**.** *Assume $\Delta_\tau > 0$ and define the parameter-independent invertible map*

$$U = \frac{\overline{Y}_0}{\sqrt{a}}, \qquad V = \sqrt{\frac{a}{\Delta_\tau}}\,\overline{Y}_1. \tag{K.8}$$

Under $R \in \mathrm{O}(d)$,

$$\mathrm{Cov}(U) = I_d, \qquad \mathrm{Cov}(V) = \left(1 + \frac{1}{\Delta_\tau}\right) I_d, \qquad \mathrm{Cov}(V, U) = \frac{R}{\sqrt{\Delta_\tau}}. \tag{K.9}$$

Hence every positive-noise paired experiment with the same Δ_τ is statistically equivalent for inference on R. Consequently, every likelihood-based information quantity and every minimax risk over the same estimator class depends on (τ_0, τ_1) only through Δ_τ.

Proof. From the averaged model, $\mathrm{Cov}(\overline{Y}_0) = aI_d$, $\mathrm{Cov}(\overline{Y}_1) = bI_d$, and $\mathrm{Cov}(\overline{Y}_1, \overline{Y}_0) = R$. Applying equation K.8 gives

$$\mathrm{Cov}(V) = \frac{ab}{\Delta_\tau} I_d = \left(1 + \frac{1}{\Delta_\tau}\right) I_d, \qquad \mathrm{Cov}(V, U) = \frac{R}{\sqrt{\Delta_\tau}},$$

and the remaining identity is immediate. The map is invertible and does not depend on R. Its centered Gaussian output law is determined by equation K.9, which contains the noise levels only through Δ_τ. Applying the inverse transformation gives a Markov kernel in the reverse direction, establishing equivalence of the statistical experiments. □

Proposition K.3 (Gaussian paired-view law)**.** *Under global parameter $R \in \mathrm{O}(d)$, one averaged paired task has law*

$$\begin{pmatrix} \overline{Y}_0 \\ \overline{Y}_1 \end{pmatrix} \sim \mathcal{N}(0, \Sigma_R), \qquad \Sigma_R := \begin{pmatrix} aI_d & R^\top \\ R & bI_d \end{pmatrix}. \tag{K.10}$$

If $\Delta_\tau > 0$, then Σ_R is positive definite and

$$\det(\Sigma_R) = \Delta_\tau^d, \tag{K.11}$$

$$\Sigma_R^{-1} = \frac{1}{\Delta_\tau} \begin{pmatrix} bI_d & -R^\top \\ -R & aI_d \end{pmatrix}. \tag{K.12}$$

Consequently, its Lebesgue density is

$$p_R(y_0, y_1) = (2\pi)^{-d} \Delta_\tau^{-d/2} \exp\left\{ -\frac{b\,\|y_0\|_2^2 + a\,\|y_1\|_2^2 - 2y_1^\top R y_0}{2\Delta_\tau} \right\}. \tag{K.13}$$

The corresponding conditional laws are

$$\overline{Y}_1 \mid \overline{Y}_0 = y_0 \sim \mathcal{N}\left(\frac{1}{a} R y_0, \frac{\Delta_\tau}{a} I_d \right), \tag{K.14}$$

$$\overline{Y}_0 \mid \overline{Y}_1 = y_1 \sim \mathcal{N}\left(\frac{1}{b} R^\top y_1, \frac{\Delta_\tau}{b} I_d \right). \tag{K.15}$$

Proof. The covariance blocks follow immediately from equation K.5:

$$\mathrm{Cov}(\overline{Y}_0) = aI_d, \qquad \mathrm{Cov}(\overline{Y}_1) = bI_d, \qquad \mathrm{Cov}(\overline{Y}_1, \overline{Y}_0) = R.$$

Using $R^\top R = RR^\top = I_d$, direct block multiplication gives

$$\begin{pmatrix} aI_d & R^\top \\ R & bI_d \end{pmatrix} \begin{pmatrix} bI_d & -R^\top \\ -R & aI_d \end{pmatrix} = \Delta_\tau I_{2d},$$

which proves equation K.12. The determinant follows from the Schur complement:

$$\det(\Sigma_R) = a^d \det\left(bI_d - a^{-1} R R^\top \right) = (ab - 1)^d.$$

The density is the centered Gaussian density with this inverse and determinant. The conditional formulas are the standard Gaussian Schur complements, or follow by completing the square in equation K.13. □

K.3 Exact divergence geometry

The special covariance structure makes the full global divergence geometry explicit rather than merely local.

Proposition K.4 (Exact likelihood ratio and pairwise KL divergence)**.** *Assume $\Delta_\tau > 0$. For every $R, S \in \mathrm{O}(d)$,*

$$\log \frac{p_R(y_0, y_1)}{p_S(y_0, y_1)} = \frac{1}{\Delta_\tau} y_1^\top (R - S) y_0, \tag{K.16}$$

$$D_{\mathrm{KL}}(P_R \| P_S) = D_{\mathrm{KL}}(P_S \| P_R) = \frac{d - \mathrm{Tr}(S^\top R)}{\Delta_\tau} = \frac{\|R - S\|_{\mathrm{F}}^2}{2\Delta_\tau}. \tag{K.17}$$

For n independent paired tasks, the KL divergence is n times the quantity in equation K.17. In particular, the regular paired experiment is globally identifiable: $P_R = P_S$ if and only if $R = S$.

Proof. All terms in equation K.13 except the cross term are independent of R, which gives equation K.16. Under P_R, $\mathbb{E}_R[\overline{Y}_0 \overline{Y}_1^\top] = R^\top$, and hence

$$\begin{aligned} D_{\mathrm{KL}}(P_R \| P_S) &= \frac{1}{\Delta_\tau} \mathrm{Tr}\left((R - S) R^\top \right) \\ &= \frac{d - \mathrm{Tr}(S R^\top)}{\Delta_\tau}. \end{aligned}$$

Cyclicity of trace gives $\mathrm{Tr}(SR^\top) = \mathrm{Tr}(S^\top R)$, while orthogonality gives $\|R - S\|_{\mathrm{F}}^2 = 2d - 2\,\mathrm{Tr}(S^\top R)$. This proves the formula and its symmetry. Additivity over independent tasks follows from the KL chain rule. □

Remark K.5 (An exact statistical embedding of chordal geometry). Equation K.17 says that the map $R \mapsto P_R$ embeds the orthogonal group into the statistical experiment with KL divergence equal to $(2\Delta_\tau)^{-1}$ times squared chordal Frobenius distance. Thus no local approximation is needed for the pairwise information calculations used later in the minimax lower bound.

K.4 Sufficient statistic and exact maximum likelihood estimator

Let $(\overline{Y}_{i,0}, \overline{Y}_{i,1})_{i=1}^n$ be independent paired tasks and define the empirical cross-moment

$$C_n := \sum_{i=1}^n \overline{Y}_{i,1} \overline{Y}_{i,0}^\top. \tag{K.18}$$

Proposition K.6 (Sufficiency and Procrustes maximum likelihood)**.** *Assume $\Delta_\tau > 0$. The log likelihood satisfies*

$$\ell_n(R) = \text{const} + \frac{1}{\Delta_\tau} \operatorname{Tr}(R^\top C_n), \tag{K.19}$$

so C_n is sufficient for R. Let

$$C_n = U \begin{pmatrix} \operatorname{diag}(s_1, \dots, s_q) & 0 \\ 0 & 0 \end{pmatrix} V^\top$$

be a singular value decomposition with $s_1 \geq \cdots \geq s_q > 0$ and $q = \operatorname{rank}(C_n)$. Then the complete MLE set is

$$\arg\max_{R \in \mathrm{O}(d)} \ell_n(R) = \left\{ U \begin{pmatrix} I_q & 0 \\ 0 & W \end{pmatrix} V^\top : W \in \mathrm{O}(d-q) \right\}. \tag{K.20}$$

If C_n is nonsingular, the MLE is unique and equals its orthogonal polar factor

$$\widehat{R}_n = \operatorname{Polar}(C_n) = UV^\top = C_n (C_n^\top C_n)^{-1/2}. \tag{K.21}$$

Moreover, when $n \geq d$, C_n is nonsingular almost surely.

Proof. Summing equation K.13 gives equation K.19; the Fisher–Neyman factorization criterion proves sufficiency. For $Q = U^\top R V \in \mathrm{O}(d)$,

$$\operatorname{Tr}(R^\top C_n) = \sum_{j=1}^q s_j Q_{jj} \leq \sum_{j=1}^q s_j.$$

Equality holds if and only if $Qe_j = e_j$ for every $j \leq q$, which is equivalent to $Q = \operatorname{diag}(I_q, W)$ for some $W \in \mathrm{O}(d-q)$. This proves equation K.20. If $q = d$, the maximizing matrix is unique and is the polar factor in equation K.21; this is the classical orthogonal Procrustes solution (Schönemann, 1966; Higham, 1986).

It remains to prove nonsingularity. When $\Delta_\tau > 0$, the complete sample has a strictly positive Lebesgue density on $\mathbb{R}^{2dn}$. The polynomial

$$(y_{1,0}, y_{1,1}, \dots, y_{n,0}, y_{n,1}) \longmapsto \det \left(\sum_{i=1}^n y_{i,1} y_{i,0}^\top \right)$$

is not identically zero for $n \geq d$: take the first d pairs to be (e_i, e_i) and all remaining pairs to be zero. The zero set of a nonzero polynomial has Lebesgue measure zero, so $\det(C_n) \neq 0$ almost surely. □

K.5 Noiseless breadth threshold and exact prediction risk

Set $\tau_0 = \tau_1 = 0$. For n source tasks, write

$$X_{1:n} := (X_1 \quad \cdots \quad X_n), \qquad Y_{1:n} = R_\star X_{1:n}, \qquad S_n := \operatorname{span}(X_1, \dots, X_n). \tag{K.22}$$

Because Gaussian vectors are in general position,

$$\dim(S_n) = \min\{n, d\} \qquad \text{almost surely.} \tag{K.23}$$

Proposition K.7 (Exact noiseless consistency set and breadth threshold)**.** *Conditionally on the noiseless source data in equation K.22, the set of consistent transformations is*

$$R_\star \{H \in \mathrm{O}(d) : Hx = x \text{ for every } x \in S_n\}. \tag{K.24}$$

The residual subgroup is isomorphic to $\mathrm{O}(d - \dim S_n)$*. Consequently, the algebraic and essential base sizes of the Gaussian orthogonal action are both* d*. If* $n \geq d$*, then*

$$R_\star = Y_{1:n} X_{1:n}^\top (X_{1:n} X_{1:n}^\top)^{-1} \qquad \textit{almost surely}. \tag{K.25}$$

If $n < d$*, exact recovery is impossible even with infinitely many repeated noiseless observations of those same* n *tasks.*

Proof. The consistency-set identity is the specialization of Theorem J.5: a candidate R is consistent if and only if $R_\star^\top R$ fixes every observed anchor, equivalently every vector in S_n. With respect to the orthogonal decomposition $\mathbb{R}^d = S_n \oplus S_n^\perp$, such a transformation has block form $I_{S_n} \oplus W$ with $W \in \mathrm{O}(S_n^\perp)$, proving the subgroup isomorphism. Gaussian general position gives equation K.23. Hence the residual subgroup is trivial almost surely exactly when $n \geq d$, proving both base-size claims. When $n \geq d$, $X_{1:n} X_{1:n}^\top$ is invertible and equation K.25 follows from $Y_{1:n} = R_\star X_{1:n}$. □

The residual ambiguity has an exact decision-theoretic value for a fresh target anchor. The following theorem strengthens the general lower bound of Theorem J.9 to an equality.

Theorem K.8 (Exact noiseless fresh-anchor prediction risk)**.** *After observing* n *independent noiseless source pairs, let an independent fresh anchor* $X_\star \sim \mathcal{N}(0, I_d)$ *be revealed, and consider prediction of* $R_\star X_\star$ *under squared Euclidean loss. Then*

$$\inf_{\widehat{Y}} \sup_{R_\star \in \mathrm{O}(d)} \mathbb{E}_{R_\star} \left[\left\| \widehat{Y} - R_\star X_\star \right\|_2^2 \right] = (d - n)_+. \tag{K.26}$$

Proof. Let P_{S_n} denote orthogonal projection onto S_n. The source pairs determine the restriction of $R_\star$ to S_n exactly. Let $X_{1:n}^\dagger$ be the Moore–Penrose pseudoinverse of the anchor matrix in equation K.22. The observable estimator

$$\widehat{Y} = Y_{1:n} X_{1:n}^\dagger X_\star = R_\star P_{S_n} X_\star$$

has conditional squared error $\left\| P_{S_n^\perp} X_\star \right\|_2^2$. Taking expectation and using equation K.23 gives the upper bound $(d - n)_+$.

For the lower bound, the case $n \geq d$ again follows from nonnegativity because $S_n = \mathbb{R}^d$ almost surely. Suppose therefore that $n < d$. Place normalized Haar prior on $R_\star$. By Proposition J.8, conditional on the source pairs the posterior is Haar measure on the residual coset $R_0 H_n$, where H_n fixes S_n pointwise. Decompose $X_\star = s + u$ with $s \in S_n$ and $u \in S_n^\perp$. Since $d - \dim S_n = d - n \geq 1$, the Haar average of Hu over $H_n \cong \mathrm{O}(d - \dim S_n)$ is zero: multiplication by the element $I_{S_n} \oplus (-I_{S_n^\perp})$ leaves Haar measure invariant and changes that average's sign. Thus the posterior mean of $R_\star X_\star$ is $R_0 s$, and its conditional Bayes risk is exactly $\|u\|_2^2$. Integrating gives $d - \dim(S_n) = d - n$. The minimax risk is at least this Bayes risk, matching the upper bound. □

K.6 Score and Fisher information on the orthogonal group

The orthogonal group is a smooth compact manifold with tangent space

$$T_R \mathrm{O}(d) = \{R\Omega : \Omega \in \mathfrak{so}(d)\}, \qquad \mathfrak{so}(d) := \{\Omega \in \mathbb{R}^{d \times d} : \Omega^\top = -\Omega\}. \tag{K.27}$$

We use the Frobenius metric $\langle \Omega, \Lambda \rangle_{\mathrm{F}} = \mathrm{Tr}(\Omega^\top \Lambda)$ on $\mathfrak{so}(d)$; its dimension is

$$p_d := \dim \mathfrak{so}(d) = \frac{d(d-1)}{2}. \tag{K.28}$$

Proposition K.9 (Exact score and Fisher information)**.** *Assume* $\Delta_\tau > 0$*. For a tangent path* $R_t = R \exp(t\Omega)$ *with* $\Omega \in \mathfrak{so}(d)$*, the one-pair score is*

$$\dot{\ell}_R[\Omega] = \frac{\mathrm{d}}{\mathrm{d}t} \log p_{R_t}(\overline{Y}_0, \overline{Y}_1) \Big|_{t=0} = \frac{1}{\Delta_\tau} \overline{Y}_1^\top R \Omega \overline{Y}_0. \tag{K.29}$$

For every $\Omega, \Lambda \in \mathfrak{so}(d)$*,*

$$\mathcal{I}_R(\Omega, \Lambda) \coloneqq \mathbb{E}_R\left[\dot{\ell}_R[\Omega]\dot{\ell}_R[\Lambda]\right] = \frac{1}{\Delta_\tau}\operatorname{Tr}(\Omega^\top \Lambda). \tag{K.30}$$

Thus one paired task induces the isotropic Fisher metric $\Delta_\tau^{-1} I_{p_d}$ *in every Frobenius-orthonormal tangent basis, and* n *independent paired tasks induce* $n\Delta_\tau^{-1} I_{p_d}$*. Moreover,*

$$D_{\mathrm{KL}}\left(P_R \middle\| P_{R\exp(t\Omega)}\right) = \frac{t^2}{2\Delta_\tau}\|\Omega\|_{\mathrm{F}}^2 + o(t^2), \qquad t \to 0. \tag{K.31}$$

Proof. Differentiating equation K.13 along the tangent path gives equation K.29. Put $U = R^\top \overline{Y}_1$ and $V = \overline{Y}_0$. Under P_R,

$$\operatorname{Cov}(U) = bI_d, \qquad \operatorname{Cov}(V) = aI_d, \qquad \operatorname{Cov}(U, V) = I_d.$$

Since $\operatorname{Tr}(\Omega) = 0$, the score has mean zero. Isserlis' identity for the jointly Gaussian pair (U, V) gives

$$\begin{aligned}\mathbb{E}\left[(U^\top \Omega V)(U^\top \Lambda V)\right] &= ab\operatorname{Tr}(\Omega^\top \Lambda) + \operatorname{Tr}(\Omega\Lambda) \\ &= (ab-1)\operatorname{Tr}(\Omega^\top \Lambda) \\ &= \Delta_\tau \operatorname{Tr}(\Omega^\top \Lambda),\end{aligned}$$

where the second equality uses $\operatorname{Tr}(\Omega\Lambda) = -\operatorname{Tr}(\Omega^\top \Lambda)$ for skew-symmetric matrices. Dividing by Δ_τ^2 proves equation K.30. Finally, Proposition K.4 gives

$$D_{\mathrm{KL}}\left(P_R \middle\| P_{R\exp(t\Omega)}\right) = \frac{1}{2\Delta_\tau}\left\|I_d - \exp(t\Omega)\right\|_{\mathrm{F}}^2,$$

and the matrix-exponential expansion yields $\|I_d - \exp(t\Omega)\|_{\mathrm{F}}^2 = t^2\|\Omega\|_{\mathrm{F}}^2 + o(t^2)$. □

K.7 Information supplied by repeated continuation

Combining Propositions K.1, K.4 and K.9 gives an exact information-versus-cost curve for a static paired source design. For raw per-episode noise variances σ_0^2, σ_1^2 satisfying $\sigma_0^2 + \sigma_1^2 > 0$ and positive repeat counts (m_0, m_1), define

$$\Delta(m_0, m_1) \coloneqq \frac{\sigma_0^2}{m_0} + \frac{\sigma_1^2}{m_1} + \frac{\sigma_0^2\sigma_1^2}{m_0 m_1}. \tag{K.32}$$

Corollary K.10 (Exact information of a repeated paired task)**.** *Assume* $\sigma_0^2 + \sigma_1^2 > 0$*. For one source task receiving* m_0 *type-*0 *episodes and* m_1 *type-*1 *episodes,*

$$D_{\mathrm{KL}}\left(P_R^{(m_0,m_1)} \middle\| P_S^{(m_0,m_1)}\right) = \frac{\|R - S\|_{\mathrm{F}}^2}{2\Delta(m_0, m_1)}, \tag{K.33}$$

$$\mathcal{I}_R^{(m_0,m_1)}(\Omega, \Lambda) = \frac{\operatorname{Tr}(\Omega^\top \Lambda)}{\Delta(m_0, m_1)}. \tag{K.34}$$

If a fresh-task reset costs κ*, then the Fisher information per unit source cost in tangent direction* Ω *is*

$$\frac{\mathcal{I}_R^{(m_0,m_1)}(\Omega, \Omega)}{m_0 + m_1 + \kappa} = \frac{\|\Omega\|_{\mathrm{F}}^2}{(m_0 + m_1 + \kappa)\Delta(m_0, m_1)}. \tag{K.35}$$

Proof. By Proposition K.1, the raw repeated-view experiment is statistically equivalent to the averaged model with $\tau_v^2 = \sigma_v^2/m_v$. Substitution into equation K.17 and equation K.30 proves the first two displays. The last display divides the directional information by the additive block cost. □

Remark K.11 (Regular and singular limits). For every fixed positive effective noise level, the paired family is a regular p_d-dimensional model with full isotropic Fisher information. As $m_0, m_1 \to \infty$, $\Delta(m_0, m_1) \downarrow 0$ and the regular information metric diverges. The limiting noiseless experiment is not a regular full-dimensional Gaussian model: one realized anchor leaves a nontrivial sample-dependent stabilizer. This is why the exact noiseless breadth theorem and the positive-noise Fisher calculation must be treated as distinct regimes rather than connected by an unqualified limiting argument.

K.8 Consequences and scope

The orthogonal specialization gives three exact structural facts. First, unpaired reset-only observations are parameter independent by equation K.6. Second, paired tasks identify the global transformation distributionally and induce the exact divergence geometry equation K.17. Third, in the noiseless finite-sample problem, d independent anchors are necessary and sufficient for exact recovery, and the unresolved breadth has exact fresh-anchor prediction cost $(d-n)_+$.

The remaining appendices convert these identities into nonasymptotic estimation rates. Section L analyzes the polar MLE, while Section M proves matching minimax lower bounds using the exact KL geometry. No finite-sample noisy minimax claim is made in the present appendix.

L Upper Bounds for Orthogonal Transfer

This appendix proves nonasymptotic upper bounds for estimation of the global orthogonal transformation in the noisy paired-view experiment of Section K. We analyze two estimators. The first is an orthogonalized conditional least-squares estimator whose mean-squared risk can be computed through an inverse-Wishart identity. The second is the Procrustes maximum-likelihood estimator from Proposition K.6. A weighted curvature argument shows that the latter attains the same rate. The only external random-matrix facts we use are standard Gaussian pseudoinverse moments and tails (Halko et al., 2011) and a chi-square tail inequality (Laurent & Massart, 2000); all model-specific reductions and deterministic inequalities are proved below.

Throughout, let

$$R_\star \in \mathrm{O}(d), \qquad d \geq 2, \qquad \Delta_\tau = \tau_0^2 + \tau_1^2 + \tau_0^2\tau_1^2 > 0,$$

and let $(\overline{Y}_{i,0}, \overline{Y}_{i,1})_{i=1}^n$ be independent paired source tasks following equation K.5. We write

$$Y_0 \coloneqq \begin{pmatrix} \overline{Y}_{1,0} & \cdots & \overline{Y}_{n,0} \end{pmatrix}, \qquad Y_1 \coloneqq \begin{pmatrix} \overline{Y}_{1,1} & \cdots & \overline{Y}_{n,1} \end{pmatrix}, \tag{L.1}$$

and recall $a = 1 + \tau_0^2$ and $C_n = Y_1 Y_0^\top$.

L.1 Conditional Gaussian regression and equivariance

The conditional law in equation K.14 converts the paired experiment into a random-design multivariate regression with an exactly known noise level.

Lemma L.1 (Whitened conditional-regression representation)**.** *Conditionally on Y_0,*

$$Y_1 = \frac{1}{a} R_\star Y_0 + \Xi, \qquad \Xi \mid Y_0 \sim \mathcal{N}_{d\times n}\left(0, \frac{\Delta_\tau}{a} I_d \otimes I_n\right), \tag{L.2}$$

where Ξ is independent of Y_0. Equivalently, there exist independent standard Gaussian matrices $Z, G \in \mathbb{R}^{d\times n}$ such that

$$Y_0 = \sqrt{a}\, Z, \tag{L.3}$$

$$R_\star^\top C_n = ZZ^\top + \sqrt{\Delta_\tau}\, GZ^\top. \tag{L.4}$$

Proof. Equation K.14 gives, independently for each source task,

$$\overline{Y}_{i,1} = \frac{1}{a} R_\star \overline{Y}_{i,0} + \xi_i, \qquad \xi_i \sim \mathcal{N}\left(0, \frac{\Delta_\tau}{a} I_d\right),$$

with ξ_i independent of $\overline{Y}_{i,0}$. Stacking the columns proves equation L.2. Since $\overline{Y}_{i,0} \sim \mathcal{N}(0, aI_d)$, we may write $Y_0 = \sqrt{a}\, Z$ with Z standard Gaussian. Writing $\Xi = \sqrt{\Delta_\tau / a}\, R_\star G$ with an independent standard Gaussian G and multiplying equation L.2 by $Y_0^\top$ gives equation L.4. □

Both estimators studied below are orthogonally equivariant. In particular, their Frobenius risks do not depend on the true value of $R_\star$. Indeed, left multiplication of Y_1 by $U \in \mathrm{O}(d)$ changes the parameter from $R_\star$ to $UR_\star$ and left-multiplies either estimator by U. We nevertheless retain $R_\star$ in the formulas to keep the statistical interpretation visible.

L.2 PROJECTED CONDITIONAL LEAST SQUARES

Assume $n \geq d$. Then $Y_0 Y_0^\top$ is invertible almost surely. Define the scaled unconstrained conditional least-squares estimator

$$\widetilde{R}_n \coloneqq a Y_1 Y_0^\top (Y_0 Y_0^\top)^{-1}, \tag{L.5}$$

and its orthogonal projection

$$\check{R}_n \coloneqq \mathrm{Polar}(\widetilde{R}_n). \tag{L.6}$$

For $\Delta_\tau > 0$, $\widetilde{R}_n$ is nonsingular almost surely, so the polar factor is unique. Indeed, conditional on any full-row-rank realization of Z, each row of $GZ^\dagger$ is a nondegenerate Gaussian vector with covariance $(ZZ^\top)^{-1}$; hence $I_d + \sqrt{\Delta_\tau} GZ^\dagger$ has a Lebesgue density on $\mathbb{R}^{d\times d}$, while the singular matrices form the zero set of the nonzero determinant polynomial. On the resulting null event of singularity, any measurable nearest orthogonal matrix may be selected.

The next theorem gives an explicit expected-risk bound. It is already sufficient for the minimax upper bound matched in Section M.

Theorem L.2 (Projected least-squares upper bound)**.** *If $n \geq d+2$, then*

$$\sup_{R_\star \in \mathrm{O}(d)} \mathbb{E}_{R_\star}\left[\left\|\check{R}_n - R_\star\right\|_{\mathrm{F}}^2\right] \leq \left(\frac{4\Delta_\tau d^2}{n-d-1}\right) \wedge 4d. \tag{L.7}$$

Consequently, if $n \geq 2d+2$ and $p_d = d(d-1)/2$, then

$$\sup_{R_\star \in \mathrm{O}(d)} \mathbb{E}_{R_\star}\left[\left\|\check{R}_n - R_\star\right\|_{\mathrm{F}}^2\right] \leq 32\left(\frac{\Delta_\tau p_d}{n} \wedge d\right). \tag{L.8}$$

Proof. Let $Z^\dagger = Z^\top (ZZ^\top)^{-1}$ denote the right Moore–Penrose pseudoinverse of the full-row-rank matrix Z. Substituting equation L.2 and equation L.3 into equation L.5 yields the exact error representation

$$R_\star^\top \widetilde{R}_n - I_d = \sqrt{\Delta_\tau}\, GZ^\dagger. \tag{L.9}$$

Conditionally on Z, the rows of G are independent standard Gaussian vectors, and therefore

$$\mathbb{E}\left[\left\|GZ^\dagger\right\|_{\mathrm{F}}^2 \mid Z\right] = d\left\|Z^\dagger\right\|_{\mathrm{F}}^2. \tag{L.10}$$

The matrix $ZZ^\top$ has the Wishart law $W_d(n, I_d)$. For $n \geq d+2$, the standard inverse-Wishart identity gives (Halko et al., 2011, Prop. 10.2)

$$\mathbb{E}\left\|Z^\dagger\right\|_{\mathrm{F}}^2 = \mathbb{E}\,\mathrm{Tr}\left((ZZ^\top)^{-1}\right) = \frac{d}{n-d-1}. \tag{L.11}$$

Combining equations (L.9)–(L.11) gives

$$\mathbb{E}\left\|\widetilde{R}_n - R_\star\right\|_{\mathrm{F}}^2 = \frac{\Delta_\tau d^2}{n-d-1}. \tag{L.12}$$

The orthogonal polar factor is a nearest orthogonal matrix in Frobenius norm (Higham, 1986). Since $R_\star$ is itself orthogonal,

$$\left\|\check{R}_n - \widetilde{R}_n\right\|_{\mathrm{F}} \leq \left\|R_\star - \widetilde{R}_n\right\|_{\mathrm{F}}.$$

The triangle inequality therefore implies

$$\left\|\check{R}_n - R_\star\right\|_{\mathrm{F}} \leq 2\left\|\widetilde{R}_n - R_\star\right\|_{\mathrm{F}}.$$

Taking squared expectations and using equation L.12 proves the first term in equation L.7. The deterministic diameter bound $\|Q-R\|_{\mathrm{F}}^2 \leq 4d$ for $Q, R \in \mathrm{O}(d)$ gives the truncation. Finally, if $n \geq 2d+2$, then $n-d-1 \geq n/2$ and, for $d \geq 2$, $d^2 \leq 4p_d$. This proves equation L.8. □

Remark L.3 (Why the conditional estimator is useful). The estimator $\check{R}_n$ is not generally identical to the joint maximum-likelihood estimator $\widehat{R}_n = \mathrm{Polar}(C_n)$ because the empirical design covariance $Y_0 Y_0^\top$ need not be proportional to the identity. Its purpose is twofold: it gives a short, exact finite-sample upper bound, and it isolates the intrinsic source of difficulty through the Gaussian pseudoinverse. The next subsection proves that the actual Procrustes MLE has the same rate.

L.3 A DETERMINISTIC CURVATURE INEQUALITY FOR PROCRUSTES ESTIMATION

The MLE proof is driven by a weighted strong-concavity inequality for the orthogonal Procrustes objective. Unlike a generic perturbation bound for the polar factor, the inequality below charges only the component of the noise that can rotate the optimizer relative to the empirical design.

Lemma L.4 (Weighted Procrustes curvature)**.** *Let $S \in \mathbb{R}^{d\times d}$ be symmetric positive definite, let $N \in \mathbb{R}^{d\times d}$, and let*

$$\widehat{Q} \in \operatorname*{arg\,max}_{Q\in \mathrm{O}(d)} \operatorname{Tr}\left(Q^\top (S+N)\right). \tag{L.13}$$

Then

$$\left\|(\widehat{Q}-I_d)S^{1/2}\right\|_{\mathrm{F}} \le 2\left\|NS^{-1/2}\right\|_{\mathrm{F}}, \tag{L.14}$$

$$\left\|\widehat{Q}-I_d\right\|_{\mathrm{F}}^2 \le 4\left\|NS^{-1/2}\right\|_{\mathrm{F}}^2 \left\|S^{-1}\right\|_{\mathrm{op}}. \tag{L.15}$$

Proof. Optimality of $\widehat{Q}$ relative to the feasible point I_d gives

$$0 \le \operatorname{Tr}\left((\widehat{Q}-I_d)^\top (S+N)\right).$$

Because S is symmetric and $\widehat{Q}$ is orthogonal,

$$\begin{aligned}\operatorname{Tr}(S)-\operatorname{Tr}(\widehat{Q}^\top S) &= \frac{1}{2}\operatorname{Tr}\left((\widehat{Q}-I_d)^\top(\widehat{Q}-I_d)S\right)\\ &= \frac{1}{2}\left\|(\widehat{Q}-I_d)S^{1/2}\right\|_{\mathrm{F}}^2.\end{aligned}$$

Hence

$$\frac{1}{2}\left\|(\widehat{Q}-I_d)S^{1/2}\right\|_{\mathrm{F}}^2 \le \operatorname{Tr}\left((\widehat{Q}-I_d)^\top N\right).$$

By cyclicity of trace, the right-hand side equals

$$\operatorname{Tr}\left(\left((\widehat{Q}-I_d)S^{1/2}\right)^\top NS^{-1/2}\right),$$

which is at most the product of the two Frobenius norms. Cancelling the first norm proves equation L.14; if that norm is zero, the conclusion is immediate. Finally,

$$\left\|\widehat{Q}-I_d\right\|_{\mathrm{F}}^2 \le \left\|(\widehat{Q}-I_d)S^{1/2}\right\|_{\mathrm{F}}^2 \left\|S^{-1}\right\|_{\mathrm{op}},$$

which gives equation L.15. □

L.4 FINITE-SAMPLE RISK OF THE PROCRUSTES MLE

Let

$$\widehat{R}_n = \operatorname{Polar}(C_n) \tag{L.16}$$

be the unique MLE from Proposition K.6. For the theorem below, set

$$p \coloneqq n-d. \tag{L.17}$$

Theorem L.5 (Finite-sample upper bound for the Procrustes MLE)**.** *Assume $n \ge 2d+2$. Then*

$$\sup_{R_\star\in\mathrm{O}(d)} \mathbb{E}_{R_\star}\left[\left\|\widehat{R}_n-R_\star\right\|_{\mathrm{F}}^2\right] \le \left(\frac{48e^2\Delta_\tau d^2}{n}\right)\wedge 4d. \tag{L.18}$$

In particular,

$$\sup_{R_\star\in\mathrm{O}(d)} \mathbb{E}_{R_\star}\left[\left\|\widehat{R}_n-R_\star\right\|_{\mathrm{F}}^2\right] \le 192e^2\left(\frac{\Delta_\tau p_d}{n}\wedge d\right). \tag{L.19}$$

Moreover, for every $t \ge 0$ and $u \ge 1$, with probability at least $1-e^{-t}-u^{-(p+1)}$,

$$\left\|\widehat{R}_n-R_\star\right\|_{\mathrm{F}}^2 \le \frac{4e^2\Delta_\tau n}{(p+1)^2}u^2\left(d^2+2d\sqrt{t}+2t\right). \tag{L.20}$$

Proof. By orthogonal equivariance, it is enough to analyze

$$\widehat{Q}_n \coloneqq R_\star^\top \widehat{R}_n.$$

The decomposition equation L.4 and the Procrustes characterization of the MLE imply

$$\widehat{Q}_n \in \operatorname*{arg\,max}_{Q \in \mathrm{O}(d)} \operatorname{Tr}\left(Q^\top \left(ZZ^\top + \sqrt{\Delta_\tau}\, GZ^\top\right)\right).$$

Since $n \geq d$, the Gaussian matrix Z has full row rank almost surely. Apply Lemma L.4 with

$$S = ZZ^\top, \qquad N = \sqrt{\Delta_\tau}\, GZ^\top.$$

Let

$$P_Z \coloneqq Z^\top (ZZ^\top)^{-1/2}.$$

Then $P_Z^\top P_Z = I_d$, and the deterministic inequality becomes

$$\left\|\widehat{Q}_n - I_d\right\|_{\mathrm{F}}^2 \leq 4\Delta_\tau \left\|GP_Z\right\|_{\mathrm{F}}^2 \left\|Z^\dagger\right\|_{\mathrm{op}}^2. \tag{L.21}$$

Conditionally on Z, the matrix

$$H \coloneqq GP_Z$$

is a standard $d \times d$ Gaussian matrix. Its conditional law does not depend on Z, so H is independent of Z. Therefore

$$\mathbb{E}\left\|H\right\|_{\mathrm{F}}^2 = d^2. \tag{L.22}$$

We next control the inverse smallest singular value of Z. Since Z is a $d \times (d+p)$ standard Gaussian matrix and $p \geq 4$, the Gaussian pseudoinverse tail bound (Halko et al., 2011, Prop. 10.4) gives, for every $u \geq 1$,

$$\mathbb{P}\left(\left\|Z^\dagger\right\|_{\mathrm{op}} \geq \frac{e\sqrt{n}}{p+1} u\right) \leq u^{-(p+1)}. \tag{L.23}$$

Integrating this tail yields

$$\begin{aligned}\mathbb{E}\left\|Z^\dagger\right\|_{\mathrm{op}}^2 &\leq \frac{e^2 n}{(p+1)^2}\left(1 + \frac{2}{p-1}\right) \\ &\leq \frac{3e^2 n}{(p+1)^2} \leq \frac{12e^2}{n},\end{aligned} \tag{L.24}$$

where the final inequality uses $p + 1 = n - d + 1 \geq n/2$. Taking expectations in equation L.21, using the independence of H and Z, and applying equations (L.22)–(L.24) proves

$$\mathbb{E}\left\|\widehat{R}_n - R_\star\right\|_{\mathrm{F}}^2 \leq \frac{48e^2 \Delta_\tau d^2}{n}.$$

The diameter bound $4d$ gives equation L.18, and $d^2 \leq 4p_d$ gives equation L.19.

For the high-probability statement, $\|H\|_{\mathrm{F}}^2$ has the $\chi^2_{d^2}$ distribution. The Laurent–Massart inequality (Laurent & Massart, 2000) yields

$$\mathbb{P}\left(\|H\|_{\mathrm{F}}^2 > d^2 + 2d\sqrt{t} + 2t\right) \leq e^{-t}. \tag{L.25}$$

Intersecting the events in equation L.23 and equation L.25, and substituting their bounds into equation L.21, proves equation L.20. □

Remark L.6 (Why the rate depends on Δ_τ). A direct bound on $\|C_n - nR_\star\|_{\mathrm{F}}$ would contain random anchor-covariance fluctuations of order d^2/n even when the observation noise vanishes. Such fluctuations are symmetric in the coordinates aligned with $R_\star$ and do not rotate the population polar factor. The conditional-regression representation and Lemma L.4 separate these design fluctuations from the rotation-inducing residual noise. This is why the upper bounds vanish with the effective noise level Δ_τ.

L.5 Minimax-rate upper bound and acquisition cost

The preceding theorems show that both a simple projected regression estimator and the joint maximum-likelihood estimator attain the Fisher-predicted intrinsic rate.

Corollary L.7 (Uniform orthogonal upper rate)**.** *There is a universal constant $C < \infty$ such that, for every $d \geq 2$, $n \geq 2d + 2$, and $\Delta_\tau > 0$, each of the estimators $\check{R}_n$ and $\widehat{R}_n$ satisfies*

$$\sup_{R_\star \in \mathrm{O}(d)} \mathbb{E}_{R_\star} \left[\left\| \widehat{R} - R_\star \right\|_{\mathrm{F}}^2 \right] \leq C \left(\frac{\Delta_\tau p_d}{n} \wedge d \right). \tag{L.26}$$

Now return to the raw repeated-view model with $m_0, m_1 \geq 1$ and per-episode noise variances σ_0^2, σ_1^2. Recall from equation K.32 that

$$\Delta(m_0, m_1) = \frac{\sigma_0^2}{m_0} + \frac{\sigma_1^2}{m_1} + \frac{\sigma_0^2 \sigma_1^2}{m_0 m_1}.$$

If every paired task has acquisition cost $m_0 + m_1 + \kappa$ and n such tasks are collected, then the total source cost is

$$T = n(m_0 + m_1 + \kappa).$$

Substituting $n = T/(m_0 + m_1 + \kappa)$ into equation L.26 gives the following direct connection between the static persistence design and estimation risk.

Corollary L.8 (Static reset–continue design bound)**.** *For any positive integers m_0, m_1 and any number of paired tasks $n \geq 2d + 2$, there is an orthogonal estimator satisfying*

$$\sup_{R_\star \in \mathrm{O}(d)} \mathbb{E}_{R_\star} \left[\left\| \widehat{R} - R_\star \right\|_{\mathrm{F}}^2 \right] \leq C \left[\frac{p_d}{T} (m_0 + m_1 + \kappa) \Delta(m_0, m_1) \wedge d \right], \tag{L.27}$$

where $T = n(m_0 + m_1 + \kappa)$ and C is universal.

The factor

$$(m_0 + m_1 + \kappa) \Delta(m_0, m_1) \tag{L.28}$$

is the reciprocal, up to the tangent norm, of the exact Fisher information per unit source cost in Corollary K.10. Thus the same persistence-design quantity controls both local information and a global finite-sample upper bound. Appendix M establishes the matching minimax lower rate.

M Minimax Lower Bounds for Orthogonal Transfer

This appendix proves minimax lower bounds for the noisy orthogonal paired-view experiment introduced in Section K. Combined with the upper bounds of Section L, the result identifies the global finite-sample rate, up to universal constants, as

$$\left(\frac{\Delta_\tau p_d}{n} \wedge d \right), \qquad p_d \coloneqq \dim \mathrm{O}(d) = \frac{d(d-1)}{2}.$$

The proof uses three ingredients. First, source-task breadth creates a separate singular obstruction that remains even when observation noise vanishes. Second, an explicit coding construction produces a localized Frobenius packing whose logarithmic cardinality is proportional to the intrinsic dimension. Third, the paired-view experiment has the exact divergence geometry

$$D_{\mathrm{KL}}(P_R^{\otimes n} \| P_S^{\otimes n}) = \frac{n}{2\Delta_\tau} \|R - S\|_{\mathrm{F}}^2$$

from Proposition K.4. A Fano reduction then converts the metric packing into the desired estimation lower bound (Yu, 1997; Tsybakov, 2009).

Throughout, the parameter lies in $\mathrm{O}(d)$ but estimators may take values in the ambient matrix space $\mathbb{R}^{d \times d}$. More precisely, after adjoining an independent private seed $U \sim \mathrm{Unif}[0, 1]$, the infimum below ranges over all Borel maps

$$\widehat{R} : ((\mathbb{R}^d \times \mathbb{R}^d)^n \times [0, 1]) \longrightarrow \mathbb{R}^{d \times d}.$$

Restricting the codomain to $\mathrm{O}(d)$ defines a different decision problem and can change exact constants: the reset-only upper bound uses $\widehat{R} \equiv 0$, and the optimal noiseless estimator below is generally nonorthogonal. For $n \geq 1$ and $\Delta_\tau > 0$, define

$$\mathfrak{R}^{\mathrm{orth}}_{n,d}(\Delta_\tau) \coloneqq \inf_{\widehat{R}} \sup_{R \in \mathrm{O}(d)} \mathbb{E}_R \left[\left\| \widehat{R} - R \right\|_{\mathrm{F}}^2 \right], \tag{M.1}$$

where the estimator observes the n independent paired vectors $(\overline{Y}_{i,0}, \overline{Y}_{i,1})_{i=1}^n$ from equation K.10 and the independent seed U.

M.1 The singular source-breadth obstruction

The regular positive-noise family is locally identifiable from a single paired task, but the limit $\Delta_\tau \downarrow 0$ is singular. In the noiseless experiment, $n < d$ source anchors reveal the action of R only on their span. The next proposition gives the exact matrix-estimation risk of that experiment.

Proposition M.1 (Exact noiseless matrix-estimation risk)**.** *Let*

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d), \qquad Y_i = RX_i, \qquad R \in \mathrm{O}(d),$$

and let an estimator be any Borel map of $(X_{1:n}, Y_{1:n}, U)$ into $\mathbb{R}^{d \times d}$, where $U \sim \mathrm{Unif}[0,1]$ is independent. Then

$$\inf_{\widehat{R}} \sup_{R \in \mathrm{O}(d)} \mathbb{E}_R \left[\left\| \widehat{R} - R \right\|_{\mathrm{F}}^2 \right] = (d-n)_+. \tag{M.2}$$

An optimal estimator is

$$\widehat{R}_{\mathrm{span}} = Y_{1:n} X_{1:n}^\dagger = R P_{S_n}, \qquad S_n \coloneqq \mathrm{span}(X_1, \ldots, X_n). \tag{M.3}$$

Proof. By Gaussian general position, $\dim S_n = \min\{n, d\}$ almost surely; see Proposition K.7. Since $X_{1:n} X_{1:n}^\dagger = P_{S_n}$, equation M.3 is measurable from the data and satisfies

$$\left\| \widehat{R}_{\mathrm{span}} - R \right\|_{\mathrm{F}}^2 = \| R(P_{S_n} - I_d) \|_{\mathrm{F}}^2 = \| I_d - P_{S_n} \|_{\mathrm{F}}^2 = d - \dim S_n = (d-n)_+.$$

This proves the upper bound.

For the converse, if $n \geq d$, Gaussian general position gives $S_n = \mathbb{R}^d$ almost surely, and the claimed lower bound is the trivial nonnegativity bound zero. It remains to consider $n < d$. Place normalized Haar measure on $R \in \mathrm{O}(d)$. Conditional on the noiseless source data, the posterior is translated Haar measure on the consistency coset

$$R H_{S_n}, \qquad H_{S_n} = \{ H \in \mathrm{O}(d) : Hx = x \text{ for every } x \in S_n \},$$

by Proposition J.8. In an orthonormal decomposition $\mathbb{R}^d = S_n \oplus S_n^\perp$, a posterior draw has the form

$$R \begin{pmatrix} I_{\dim S_n} & 0 \\ 0 & U \end{pmatrix}, \qquad U \sim \mathrm{Haar}(\mathrm{O}(d - \dim S_n)),$$

where now $d - \dim S_n = d - n \geq 1$. The Haar mean of U is zero, so the posterior mean of R is exactly RP_{S_n}. Under squared Frobenius loss, the posterior mean is Bayes optimal, and its conditional Bayes risk is

$$\mathbb{E}\left[\| R - RP_{S_n} \|_{\mathrm{F}}^2 \,\middle|\, X_{1:n}, Y_{1:n} \right] = d - \dim S_n.$$

A Bayes risk lower bounds the minimax risk. Averaging over the anchors proves equation M.2. □

The noisy paired-view experiment is a parameter-independent garbling of the noiseless experiment: given (X_i, RX_i), one obtains the noisy pair by adding the prescribed independent Gaussian noises. Data processing therefore gives the following breadth lower bound for every positive noise level.

Corollary M.2 (Breadth cannot be replaced by repeated continuation)**.** *For every $d \geq 2$, $n \geq 1$, and $\Delta_\tau > 0$,*

$$\mathfrak{R}^{\mathrm{orth}}_{n,d}(\Delta_\tau) \geq (d-n)_+. \tag{M.4}$$

In particular, arbitrarily accurate repeated measurements on fewer than d independent source tasks cannot recover the full orthogonal transformation.

Proof. Given a noisy-data estimator, simulate the same noise from the noiseless source pairs and apply the estimator. This produces an estimator in the more informative noiseless experiment with identical risk. Hence the noisy minimax risk is no smaller than the noiseless risk in Proposition M.1. □

At the opposite extreme, if the learner never pairs two views of the same anchor, the complete source transcript is independent of R by Proposition J.3. The corresponding squared Frobenius minimax risk is exactly the full ambient radius.

Proposition M.3 (Exact reset-only minimax risk)**.** *Under any source protocol whose transcript law is independent of $R \in \mathrm{O}(d)$,*

$$\inf_{\widehat{R}} \sup_{R \in \mathrm{O}(d)} \mathbb{E}_R \left[\left\| \widehat{R} - R \right\|_{\mathrm{F}}^2 \right] = d. \tag{M.5}$$

Proof. The estimator $\widehat{R} \equiv 0$ has risk $\|R\|_{\mathrm{F}}^2 = d$, proving the upper bound. For the lower bound, place Haar prior on R. Because the data law is parameter independent, R remains Haar and is independent of $\widehat{R}$ conditionally on the transcript. Haar invariance implies $\mathbb{E}[R] = 0$. Therefore

$$\mathbb{E} \left\| \widehat{R} - R \right\|_{\mathrm{F}}^2 = \mathbb{E} \left\| \widehat{R} \right\|_{\mathrm{F}}^2 + d \geq d.$$

□

M.2 A SELF-CONTAINED LOCALIZED PACKING OF THE ORTHOGONAL GROUP

The minimax lower bound requires exponentially many alternatives that are both mutually separated and contained in a Frobenius neighborhood of the identity. Rather than invoke a global metric-entropy theorem and translate between several metric normalizations, we prove the precise packing needed below directly. The construction combines a binary Hamming code, a Rademacher-matrix spectral-norm estimate, and a quantitative local bi-Lipschitz property of the matrix exponential.

Lemma M.4 (Matrix exponential on a small skew-symmetric ball)**.** *Let $A, B \in \mathbb{R}^{d \times d}$ be skew-symmetric and suppose $\max\{\|A\|_{\mathrm{op}}, \|B\|_{\mathrm{op}}\} \leq \rho < 1$. Then*

$$(1 - \rho) \|A - B\|_{\mathrm{F}} \leq \left\| e^A - e^B \right\|_{\mathrm{F}} \leq \|A - B\|_{\mathrm{F}}. \tag{M.6}$$

In particular, $e^A, e^B \in \mathrm{SO}(d)$.

Proof. For skew-symmetric matrices, every exponential is orthogonal and $\det(e^A) = e^{\mathrm{Tr}(A)} = 1$. The Duhamel identity gives

$$e^A - e^B = \int_0^1 e^{(1-t)A}(A - B)e^{tB} \, \mathrm{d}t.$$

Orthogonal invariance of the Frobenius norm immediately yields the upper bound. Moreover,

$$\begin{aligned} &\left\| e^{(1-t)A}(A - B)e^{tB} - (A - B) \right\|_{\mathrm{F}} \\ &\quad \leq \left\| e^{(1-t)A} - I \right\|_{\mathrm{op}} \|A - B\|_{\mathrm{F}} + \|A - B\|_{\mathrm{F}} \left\| e^{tB} - I \right\|_{\mathrm{op}}. \end{aligned}$$

Since $e^{sA} - I = \int_0^s e^{uA} A \, \mathrm{d}u$ and e^{uA} is orthogonal,

$$\left\| e^{sA} - I \right\|_{\mathrm{op}} \leq s \|A\|_{\mathrm{op}}.$$

Thus the preceding display is at most $[(1 - t)\|A\|_{\mathrm{op}} + t\|B\|_{\mathrm{op}}] \|A - B\|_{\mathrm{F}} \leq \rho \|A - B\|_{\mathrm{F}}$. Integrating and using the reverse triangle inequality proves the lower bound. □

Lemma M.5 (Many spectrally regular sign matrices)**.** *Let $m \geq 1$, let $N = m^2$, and let $\mathcal{B}_m = \{-1, +1\}^{m \times m}$. There is a subset $\mathcal{C}_m \subseteq \mathcal{B}_m$ such that*

$$\|B\|_{\mathrm{op}} \leq 12\sqrt{m}, \qquad B \in \mathcal{C}_m, \tag{M.7}$$

$$d_{\mathrm{H}}(B, B') > \frac{N}{4}, \qquad B \neq B',\ B, B' \in \mathcal{C}_m, \tag{M.8}$$

and, whenever $N \geq 16$*,*

$$\log |\mathcal{C}_m| \geq \frac{\log 2 - h_2(1/4)}{2}\, N, \qquad h_2(u) \coloneqq -u \log u - (1-u)\log(1-u). \tag{M.9}$$

In addition, $|\mathcal{C}_m| \geq 4$ *for every* $m \geq 4$*.*

Proof. Let B have independent Rademacher entries. A Euclidean $1/4$-net of the unit sphere in $\mathbb{R}^m$ can be chosen with cardinality at most 9^m. The standard two-net argument gives

$$\|B\|_{\mathrm{op}} \leq 2 \max_{x,y \in \mathcal{N}} |y^\top B x|.$$

For fixed unit vectors x, y, $y^\top B x = \sum_{i,j} y_i x_j B_{ij}$ is centered and satisfies

$$\mathbb{E} e^{\lambda y^\top B x} \leq \exp(\lambda^2/2),$$

because $\sum_{i,j} y_i^2 x_j^2 = 1$. Therefore

$$\Pr\big(\|B\|_{\mathrm{op}} > 12\sqrt{m}\big) \leq 2\, 9^{2m} e^{-18m} < \frac{1}{2}.$$

Hence at least 2^{N-1} sign matrices satisfy equation M.7.

Choose a maximal subset of these good matrices with pairwise Hamming distance strictly larger than $N/4$. Maximality implies that Hamming balls of radius $\lfloor N/4 \rfloor$ around the chosen matrices cover the good set. The standard entropy bound

$$\sum_{j=0}^{\lfloor N/4 \rfloor} \binom{N}{j} \leq \exp\big(N h_2(1/4)\big)$$

therefore yields

$$\log |\mathcal{C}_m| \geq N \log 2 - \log 2 - N h_2(1/4).$$

For $N \geq 16$, the right-hand side is at least $\frac{1}{2}[\log 2 - h_2(1/4)]N$, proving equation M.9; the same exact bound exceeds $\log 4$ for $N \geq 16$, proving the final assertion for $m \geq 4$. □

Lemma M.6 (Explicit localized Frobenius packing of $\mathrm{SO}(d)$)**.** *There exist universal constants* $c_{\mathrm{rad}}, c_{\mathrm{ent}} > 0$ *and* $A = 64$ *such that, for every* $d \geq 2$ *and every*

$$0 < \eta \leq c_{\mathrm{rad}} \sqrt{d}, \tag{M.10}$$

there are matrices $R_1, \ldots, R_M \in \mathrm{SO}(d)$ *satisfying*

$$\|R_j - I_d\|_{\mathrm{F}} \leq A\eta, \qquad j \in [M], \tag{M.11}$$

$$\|R_j - R_k\|_{\mathrm{F}} > 4\eta, \qquad j \neq k, \tag{M.12}$$

$$\log M \geq c_{\mathrm{ent}} p_d, \qquad p_d = \frac{d(d-1)}{2}, \tag{M.13}$$

and $M \geq 4$*. All constants are independent of* d *and* η*.*

Proof. We first treat $d \geq 8$. Put $m = \lfloor d/2 \rfloor$ and, for $B \in \{-1, +1\}^{m \times m}$, define the skew-symmetric $d \times d$ matrix

$$K(B) \coloneqq \begin{pmatrix} 0 & B \\ -B^\top & 0 \end{pmatrix} \oplus 0_{d-2m}.$$

Then

$$\|K(B)\|_{\mathrm{F}} = \sqrt{2}\, m, \qquad \|K(B)\|_{\mathrm{op}} = \|B\|_{\mathrm{op}}.$$

If B, B' differ in h entries, then

$$\|K(B) - K(B')\|_{\mathrm{F}}^2 = 8h.$$

Let $\mathcal{C}_m$ be the code from Lemma M.5 and set

$$H_B \coloneqq \frac{4\sqrt{2}\,\eta}{m} K(B), \qquad R_B \coloneqq e^{H_B}.$$

For every $B \in \mathcal{C}_m$,

$$\|H_B\|_{\mathrm{F}} = 8\eta, \qquad \|H_B\|_{\mathrm{op}} \le 48\sqrt{2}\,\frac{\eta}{\sqrt{m}} \le 48\sqrt{6}\,\frac{\eta}{\sqrt{d}}.$$

Choose, for example, $c_{\mathrm{rad}} = (256\sqrt{6})^{-1}$, so the last quantity is at most $3/16 < 1/4$. Moreover, by equation M.8, distinct codewords satisfy

$$\|H_B - H_{B'}\|_{\mathrm{F}} > 8\eta.$$

Applying Lemma M.4 with $\rho = 1/4$ gives

$$\|R_B - I_d\|_{\mathrm{F}} \le 8\eta \le A\eta, \qquad \|R_B - R_{B'}\|_{\mathrm{F}} > 6\eta > 4\eta.$$

If $d = 2m$, then $p_d = m(2m-1) \le 2m^2$; if $d = 2m+1$, then $p_d = m(2m+1) \le 3m^2$. Hence $m^2 \ge p_d/3$, and $m^2 \ge 16$ because $d \ge 8$. The code bound therefore implies

$$\log|\mathcal{C}_m| \ge \frac{\log 2 - h_2(1/4)}{6}\, p_d.$$

It also gives $|\mathcal{C}_m| \ge 4$.

For $2 \le d \le 7$, let $J_2 = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ and put

$$\theta_j = 5\eta\left(j - \frac{7}{2}\right), \qquad R_j = \exp(\theta_j J_2) \oplus I_{d-2}, \qquad j = 0, \dots, 7.$$

Our choice of c_{rad} ensures $|\theta_j| < \pi/2$. For the planar rotation $R(\theta) = e^{\theta J_2}$,

$$\|R(\theta) - R(\phi)\|_{\mathrm{F}} = 2\sqrt{2}\left|\sin\frac{\theta - \phi}{2}\right|.$$

All half-angle differences lie in $[0, \pi/2]$; hence the adjacent pair is the least separated. Using $\sin x \ge 2x/\pi$ on this interval shows that all distinct matrices are more than 4η apart, while $\|R_j - I_d\|_{\mathrm{F}} \le \sqrt{2}|\theta_j| < 25\eta < A\eta$. Thus $M = 8$ works. Taking

$$c_{\mathrm{ent}} = \min\left\{\frac{\log 2 - h_2(1/4)}{6}, \frac{\log 8}{21}\right\}$$

proves the claim in every dimension. □

M.3 A Fano reduction for matrix estimation

We record the precise reduction used below. It is a standard form of Fano's method, included to make the constants and loss conversion explicit.

Lemma M.7 (Fano reduction under squared Frobenius loss)**.** *Let $\{R_1, \dots, R_M\} \subset \mathbb{R}^{d\times d}$ satisfy*

$$\|R_j - R_k\|_{\mathrm{F}} > 4\eta \qquad (j \ne k),$$

and let $P_1, \dots, P_M$ be the corresponding data laws. Suppose there is a reference probability measure P_0 such that

$$\frac{1}{M}\sum_{j=1}^{M} D_{\mathrm{KL}}(P_j \| P_0) \le \alpha \log M \tag{M.14}$$

for some $\alpha \in (0,1)$. Then every estimator $\widehat{R}$ satisfies

$$\frac{1}{M}\sum_{j=1}^{M} \mathbb{E}_j\left[\left\|\widehat{R} - R_j\right\|_{\mathrm{F}}^2\right] \ge 4\eta^2\left(1 - \alpha - \frac{\log 2}{\log M}\right). \tag{M.15}$$

Proof. Let V be uniform on $[M]$, and conditionally on $V = j$, draw data according to P_j. Given $\widehat{R}$, define the minimum-distance decoder

$$\widehat{V} \in \operatorname*{arg\,min}_{k\in[M]} \left\|\widehat{R} - R_k\right\|_{\mathrm{F}}.$$

If $\left\|\widehat{R} - R_V\right\|_{\mathrm{F}} < 2\eta$, then the 4η separation implies $\widehat{V} = V$. Hence

$$\mathbb{E}\left\|\widehat{R} - R_V\right\|_{\mathrm{F}}^2 \geq 4\eta^2 \Pr(\widehat{V} \neq V).$$

Fano's inequality gives

$$\Pr(\widehat{V} \neq V) \geq 1 - \frac{I(V;\mathsf{X}) + \log 2}{\log M}.$$

The information radius identity implies

$$I(V;\mathsf{X}) = \inf_{Q} \frac{1}{M}\sum_{j=1}^{M} D_{\mathrm{KL}}(P_j \| Q) \leq \frac{1}{M}\sum_{j=1}^{M} D_{\mathrm{KL}}(P_j \| P_0).$$

Substituting equation M.14 proves equation M.15. □

M.4 Global noisy minimax lower bound

We now combine the exact KL geometry with the localized packing. The argument simultaneously covers the local Fisher regime and the diameter-saturated regime.

Theorem M.8 (Global noisy orthogonal minimax lower bound)**.** *There exists a universal constant $c > 0$ such that, for every $d \geq 2$, $n \geq 1$, and $\Delta_\tau > 0$,*

$$\mathfrak{R}^{\mathrm{orth}}_{n,d}(\Delta_\tau) \geq c\left(\frac{\Delta_\tau p_d}{n} \wedge d\right), \qquad p_d = \frac{d(d-1)}{2}. \tag{M.16}$$

Together with the singular breadth obstruction,

$$\mathfrak{R}^{\mathrm{orth}}_{n,d}(\Delta_\tau) \geq \max\left\{c\left(\frac{\Delta_\tau p_d}{n} \wedge d\right), (d-n)_+\right\}. \tag{M.17}$$

Proof. Let $A, c_{\mathrm{rad}}, c_{\mathrm{ent}}$ be the constants from Lemma M.6. Set

$$q \coloneqq \frac{\Delta_\tau p_d}{n} \wedge d, \qquad \eta^2 \coloneqq \gamma q, \tag{M.18}$$

where $\gamma > 0$ is a sufficiently small universal constant satisfying

$$\gamma \leq c_{\mathrm{rad}}^2 \qquad \text{and} \qquad \frac{A^2\gamma}{2} \leq \frac{c_{\mathrm{ent}}}{16}. \tag{M.19}$$

Since $q \leq d$, the first condition ensures $\eta \leq c_{\mathrm{rad}}\sqrt{d}$. Apply Lemma M.6 to obtain $R_1, \ldots, R_M \in \mathrm{SO}(d)$ satisfying equations (M.11)–(M.13).

Let $P_j = P_{R_j}^{\otimes n}$ and use $P_0 = P_{I_d}^{\otimes n}$ as reference. By the exact KL identity equation K.17,

$$D_{\mathrm{KL}}(P_j \| P_0) = \frac{n}{2\Delta_\tau}\|R_j - I_d\|_{\mathrm{F}}^2 \leq \frac{nA^2\eta^2}{2\Delta_\tau}. \tag{M.20}$$

The definition of q implies in both regimes that

$$\frac{nq}{\Delta_\tau} \leq p_d. \tag{M.21}$$

Indeed, equality holds when $q = \Delta_\tau p_d/n$; if $q = d$, then $d \leq \Delta_\tau p_d/n$. Combining equations (M.18)–(M.21) gives

$$D_{\mathrm{KL}}(P_j \| P_0) \leq \frac{A^2\gamma}{2} p_d \leq \frac{c_{\mathrm{ent}}}{16} p_d \leq \frac{1}{16}\log M.$$

Thus Lemma M.7 applies with $\alpha = 1/16$. Moreover, Lemma M.6 gives $M \geq 4$, so $\log 2/\log M \leq 1/2$. Hence every estimator satisfies

$$\sup_{R \in \mathrm{O}(d)} \mathbb{E}_R\left\|\widehat{R} - R\right\|_{\mathrm{F}}^2 \geq 4\eta^2\left(1 - \frac{1}{16} - \frac{1}{2}\right) \geq \eta^2.$$

Substituting $\eta^2 = \gamma q$ proves equation M.16 with $c = \gamma$. Equation M.17 follows by combining this bound with Corollary M.2. □

Remark M.9 (Why one packing proves both regimes). When $\Delta_\tau p_d/n \le d$, the packing radius in equation M.18 is $\eta^2 \asymp \Delta_\tau p_d/n$, and the lower bound is the regular intrinsic-dimension rate predicted by the Fisher information in Proposition K.9. When $\Delta_\tau p_d/n > d$, the radius is of order $\sqrt{d}$; the packing then contains exponentially many globally separated orthogonal transformations, and Fano's inequality yields the diameter-scale lower bound. The same exact KL formula controls both cases.

M.5 Matching rate and static persistence design

Combining the preceding lower bound with the estimators constructed in Section L gives the complete finite-sample minimax characterization in the breadth-sufficient regime.

Corollary M.10 (Matching minimax rate for noisy orthogonal transfer)**.** *There exist universal constants $0 < c \le C < \infty$ such that, for every $d \ge 2$, $n \ge 2d+2$, and $\Delta_\tau > 0$,*

$$c\left(\frac{\Delta_\tau p_d}{n} \wedge d\right) \le \mathfrak{R}^{\mathrm{orth}}_{n,d}(\Delta_\tau) \le C\left(\frac{\Delta_\tau p_d}{n} \wedge d\right). \tag{M.22}$$

The upper bound is attained by both the projected conditional least-squares estimator and the Procrustes maximum-likelihood estimator from Corollary L.7.

Return now to the repeated-view source design of equations (K.2)–(K.3). For positive integers m_0, m_1, recall

$$\Delta(m_0, m_1) = \frac{\sigma_0^2}{m_0} + \frac{\sigma_1^2}{m_1} + \frac{\sigma_0^2\sigma_1^2}{m_0 m_1}.$$

If n paired tasks are collected at total source cost $T = n(m_0 + m_1 + \kappa)$, then the next corollary evaluates that fixed repeated-view allocation; it does not optimize over (m_0, m_1).

Corollary M.11 (Minimax rate for a fixed static repeated-view allocation)**.** *There are universal constants $0 < c \le C < \infty$ such that, for every $d \ge 2$, $n \ge 2d+2$, and positive integers m_0, m_1,*

$$\begin{aligned} c\left[\frac{p_d}{T}(m_0+m_1+\kappa)\Delta(m_0,m_1) \wedge d\right] &\le \inf_{\widehat{R}} \sup_{R\in \mathrm{O}(d)} \mathbb{E}_R \left\|\widehat{R} - R\right\|_{\mathrm{F}}^2 \\ &\le C\left[\frac{p_d}{T}(m_0+m_1+\kappa)\Delta(m_0,m_1) \wedge d\right]. \end{aligned} \tag{M.23}$$

The infimum ranges over estimators based on the complete raw repeated-view data. In addition, for arbitrary n, the risk is at least $(d-n)_+$, independently of m_0 and m_1.

Proof. By Proposition K.1, the raw experiment is statistically equivalent to the averaged paired experiment with effective noise $\Delta_\tau = \Delta(m_0, m_1)$. Substitute $n = T/(m_0+m_1+\kappa)$ into Theorem M.8 and corollary L.8. The breadth term follows from Corollary M.2. □

The result separates the two non-substitutable source resources. Increasing m_0 or m_1 reduces the effective noise in each paired task, but it cannot remove the breadth term $(d-n)_+$. Conversely, collecting arbitrarily many fresh one-view tasks gives no information about R at all by Proposition M.3. Thus minimax transfer requires both continued interaction to form correspondences and resets to expose new independent directions.

N Local Asymptotic Normality and Efficiency

This appendix establishes the sharp local asymptotic theory of the regular positive-noise orthogonal experiment from Section K. Throughout this appendix, the dimension $d \ge 2$ is fixed, the effective noise level satisfies $\Delta_\tau > 0$, and the number n of independent paired source tasks tends to infinity. The parameter is locally restricted to the connected component of a fixed $R_0 \in \mathrm{O}(d)$; this is without loss for local theory, because the two components of $\mathrm{O}(d)$ are separated by a positive Frobenius distance. We prove differentiability in quadratic mean, a uniform local asymptotic normality expansion, the local asymptotic minimax lower bound, and asymptotic efficiency of the Procrustes maximum-likelihood estimator. The general LAN, convolution, and local minimax principles used below are classical (Le Cam & Yang, 2000; van der Vaart, 1998); our contribution here is to verify their hypotheses and identify all constants explicitly for the persistent paired-view experiment. Related constrained M-estimation theory is developed by Geyer (1994).

N.1 Local coordinates and Gaussian tangent notation

Let

$$\mathfrak{g}_d \coloneqq \mathfrak{so}(d) = \{\Omega \in \mathbb{R}^{d\times d} : \Omega^\top = -\Omega\}$$

with Frobenius inner product $\langle \Omega, \Lambda\rangle_{\mathrm{F}} = \mathrm{Tr}(\Omega^\top \Lambda)$. Its dimension is $p_d = d(d-1)/2$. Fix a Frobenius-orthonormal basis $E_1, \dots, E_{p_d}$ of $\mathfrak{g}_d$ and define the linear isometry

$$\mathsf{V} : \mathbb{R}^{p_d} \to \mathfrak{g}_d, \qquad \mathsf{V}(h) \coloneqq \sum_{j=1}^{p_d} h_j E_j. \tag{N.1}$$

Thus $\|\mathsf{V}(h)\|_{\mathrm{F}} = \|h\|_2$. For $h \in \mathbb{R}^{p_d}$ and $n \geq 1$, define the local alternative

$$R_{n,h} \coloneqq R_0 \exp\left(\frac{\mathsf{V}(h)}{\sqrt{n}}\right). \tag{N.2}$$

The exponential map is a smooth coordinate chart on a sufficiently small neighborhood of the identity within $\mathrm{SO}(d)$, and hence equation N.2 parametrizes all local directions in the connected component of R_0; see, e.g., Absil et al. (2008).

For one paired observation $Y = (\overline{Y}_0, \overline{Y}_1)$ under P_{R_0}, set

$$A(Y) \coloneqq R_0^\top \overline{Y}_1 \overline{Y}_0^\top, \qquad S(Y) \coloneqq \mathrm{skew}(A(Y)) = \frac{A(Y) - A(Y)^\top}{2}. \tag{N.3}$$

By Proposition K.9,

$$\mathbb{E}_{R_0}[S(Y)] = 0, \tag{N.4}$$

$$\mathbb{E}_{R_0}\left[\langle S(Y), \Omega\rangle_{\mathrm{F}} \langle S(Y), \Lambda\rangle_{\mathrm{F}}\right] = \Delta_\tau \langle \Omega, \Lambda\rangle_{\mathrm{F}}, \qquad \Omega, \Lambda \in \mathfrak{g}_d. \tag{N.5}$$

For independent observations $Y_1, \dots, Y_n \sim P_{R_0}$, define the matrix-valued central statistic

$$W_n \coloneqq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S(Y_i) \in \mathfrak{g}_d. \tag{N.6}$$

We write $W \sim \mathcal{N}_{\mathfrak{g}_d}(0, \Delta_\tau I)$ if, in any Frobenius-orthonormal basis of $\mathfrak{g}_d$, its coordinates are independent $\mathcal{N}(0, \Delta_\tau)$ random variables. The multivariate central limit theorem and equation N.5 give

$$W_n \rightsquigarrow W \sim \mathcal{N}_{\mathfrak{g}_d}(0, \Delta_\tau I). \tag{N.7}$$

N.2 Differentiability in quadratic mean

We first verify the regularity property underlying the LAN expansion. Let p_h denote the one-pair density under $R(h) = R_0 \exp(\mathsf{V}(h))$, and let $p_0 = p_{R_0}$.

Proposition N.1 (Differentiability in quadratic mean)**.** *The map $h \mapsto P_{R(h)}$ is differentiable in quadratic mean at $h = 0$. Its score vector is*

$$\dot{\ell}_{R_0}(Y) = \frac{1}{\Delta_\tau} \begin{pmatrix} \langle S(Y), E_1\rangle_{\mathrm{F}} \\ \vdots \\ \langle S(Y), E_{p_d}\rangle_{\mathrm{F}} \end{pmatrix}, \tag{N.8}$$

meaning that

$$\int \left(\sqrt{p_h} - \sqrt{p_0} - \frac{1}{2} h^\top \dot{\ell}_{R_0} \sqrt{p_0}\right)^2 = o(\|h\|_2^2), \qquad h \to 0. \tag{N.9}$$

The Fisher information matrix is

$$I_{R_0} = \mathbb{E}_{R_0}\left[\dot{\ell}_{R_0} \dot{\ell}_{R_0}^\top\right] = \Delta_\tau^{-1} I_{p_d}. \tag{N.10}$$

Proof. By the exact likelihood ratio in Proposition K.4,

$$L_h(Y) \coloneqq \log \frac{p_h(Y)}{p_0(Y)} = \frac{1}{\Delta_\tau} \langle A(Y), \exp(\mathsf{V}(h)) - I_d\rangle_{\mathrm{F}}. \tag{N.11}$$

Since $\mathsf{V}(h)$ is skew-symmetric,

$$\langle A(Y), \mathsf{V}(h)\rangle_{\mathrm{F}} = \langle S(Y), \mathsf{V}(h)\rangle_{\mathrm{F}}.$$

The matrix-exponential expansion gives, for $\|h\|_2$ sufficiently small,

$$L_h(Y) = h^\top \dot{\ell}_{R_0}(Y) + r_h(Y), \qquad |r_h(Y)| \le C\,\|h\|_2^2 \left\|\overline{Y}_0\right\|_2 \left\|\overline{Y}_1\right\|_2, \tag{N.12}$$

where C depends only on a fixed neighborhood of the origin and on Δ_τ.

We verify the $L^2(P_{R_0})$ Taylor expansion of the square-root likelihood ratio. The elementary inequality

$$|e^{x/2} - 1 - x/2| \le Cx^2 e^{|x|/2}$$

holds for all real x with a universal constant C after enlarging C on a compact interval. Moreover, for sufficiently small $t > 0$,

$$\mathbb{E}_{R_0} \exp\left(t \left\|\overline{Y}_0\right\|_2 \left\|\overline{Y}_1\right\|_2\right) < \infty. \tag{N.13}$$

Indeed, $2\,\|u\|_2\,\|v\|_2 \le \|u\|_2^2 + \|v\|_2^2$, and the jointly Gaussian vector $(\overline{Y}_0, \overline{Y}_1)$ has a finite quadratic exponential moment in a neighborhood of zero. Using $\sqrt{p_h/p_0} = e^{L_h/2}$, equation N.12, and equation N.13, dominated convergence yields

$$\begin{aligned}
&\mathbb{E}_{R_0}\left[\left(e^{L_h/2} - 1 - \frac{1}{2}h^\top \dot{\ell}_{R_0}\right)^2\right] \\
&\quad \le 2\mathbb{E}_{R_0}\left[\left(e^{L_h/2} - 1 - L_h/2\right)^2\right] + \frac{1}{2}\mathbb{E}_{R_0}[r_h^2] = o(\|h\|_2^2).
\end{aligned}$$

This proves equation N.9. Finally, equation N.10 follows from equation N.5 and the orthonormality of $E_1, \dots, E_{p_d}$. □

N.3 Uniform local asymptotic normality

The exact likelihood ratio permits a direct LAN proof with a remainder that is uniform over compact tangent sets.

Theorem N.2 (Uniform LAN on the orthogonal group)**.** *Let $K \subset \mathfrak{g}_d$ be compact. Under $P_{R_0}^{\otimes n}$, uniformly over $H \in K$,*

$$\log \frac{\mathrm{d}P_{R_0 \exp(H/\sqrt{n})}^{\otimes n}}{\mathrm{d}P_{R_0}^{\otimes n}} = \frac{1}{\Delta_\tau}\langle H, W_n\rangle_{\mathrm{F}} - \frac{1}{2\Delta_\tau}\|H\|_{\mathrm{F}}^2 + o_{P_{R_0}}(1). \tag{N.14}$$

Together with equation N.7, this is LAN with information operator $\Delta_\tau^{-1} I$ on $\mathfrak{g}_d$. Equivalently, the limit experiment is the Gaussian shift

$$\mathcal{W} = H + \sqrt{\Delta_\tau}\,\mathcal{Z}, \qquad \mathcal{Z} \sim \mathcal{N}_{\mathfrak{g}_d}(0, I). \tag{N.15}$$

Proof. For each observation set $A_i = A(Y_i) = R_0^\top \overline{Y}_{i,1} \overline{Y}_{i,0}^\top$. The exact likelihood ratio gives

$$L_n(H) = \frac{1}{\Delta_\tau}\sum_{i=1}^n \left\langle A_i, \exp(H/\sqrt{n}) - I_d\right\rangle_{\mathrm{F}}. \tag{N.16}$$

For H in a fixed compact set,

$$\exp(H/\sqrt{n}) - I_d = \frac{H}{\sqrt{n}} + \frac{H^2}{2n} + \mathcal{R}_n(H), \qquad \sup_{H\in K}\|\mathcal{R}_n(H)\|_{\mathrm{F}} \le C_K n^{-3/2}. \tag{N.17}$$

The first-order term in equation N.16 is

$$\frac{1}{\sqrt{n}\Delta_\tau}\sum_{i=1}^n \langle A_i, H\rangle_{\mathrm{F}} = \frac{1}{\Delta_\tau}\langle W_n, H\rangle_{\mathrm{F}},$$

because H is skew-symmetric. Let $\overline{A}_n = n^{-1}\sum_i A_i$. Since $\mathbb{E}_{R_0}[A_i] = I_d$ and A_i has finite moments of every order, $\overline{A}_n \to I_d$ in probability. Therefore, uniformly over $H \in K$,

$$\begin{aligned}\frac{1}{2\Delta_\tau}\left\langle \overline{A}_n, H^2\right\rangle_{\mathrm{F}} &\to \frac{1}{2\Delta_\tau}\operatorname{Tr}(H^2)\\ &= -\frac{1}{2\Delta_\tau}\|H\|_{\mathrm{F}}^2,\end{aligned}$$

where $\operatorname{Tr}(H^2) = -\|H\|_{\mathrm{F}}^2$ for skew-symmetric H. Finally,

$$\sup_{H\in K}\left|\frac{1}{\Delta_\tau}\sum_{i=1}^n \langle A_i, \mathcal{R}_n(H)\rangle_{\mathrm{F}}\right| \le \frac{C_K}{\Delta_\tau\sqrt{n}}\left(\frac{1}{n}\sum_{i=1}^n \|A_i\|_{\mathrm{F}}\right) = o_{P_{R_0}}(1).$$

Combining the three terms proves equation N.14. The Gaussian-shift representation follows from equation N.7. □

Corollary N.3 (Contiguity and shifted central sequence)**.** *For every fixed $H \in \mathfrak{g}_d$, the sequences $P_{R_0\exp(H/\sqrt{n})}^{\otimes n}$ and $P_{R_0}^{\otimes n}$ are mutually contiguous. Under $P_{R_0\exp(H/\sqrt{n})}^{\otimes n}$,*

$$W_n \rightsquigarrow \mathcal{N}_{\mathfrak{g}_d}(H, \Delta_\tau I). \tag{N.18}$$

Proof. The LAN likelihood ratio in equation N.14 has under the null the nondegenerate Gaussian limit

$$\frac{1}{\Delta_\tau}\langle H, W\rangle_{\mathrm{F}} - \frac{1}{2\Delta_\tau}\|H\|_{\mathrm{F}}^2,$$

whose exponential has expectation one. Le Cam's first and third lemmas then give mutual contiguity and the mean shift in equation N.18; see van der Vaart (1998, Chs. 6–7) or Le Cam & Yang (2000, Chs. 5–6). □

N.4 Local asymptotic minimax lower bound

The Gaussian shift in equation N.15 has minimax squared-error risk $\Delta_\tau p_d$. The next theorem transfers this limit exactly to squared Frobenius loss on the orthogonal group.

Lemma N.4 (Minimax risk in a Gaussian shift under truncated quadratic loss)**.** *Let $\Sigma \succ 0$, observe $W = h + \Sigma^{1/2}Z$ with $h \in \mathbb{R}^p$ and $Z \sim \mathcal{N}(0, I_p)$, and fix $M < \infty$. For $\ell_M(u) = \min\{\|u\|_2^2, M\}$,*

$$\inf_{\delta:\mathbb{R}^p\to\mathbb{R}^p}\sup_{h\in\mathbb{R}^p}\mathbb{E}_h\ell_M(\delta(W) - h) = \mathbb{E}\ell_M(\Sigma^{1/2}Z). \tag{N.19}$$

The estimator $\delta(W) = W$ attains the upper bound.

Proof. The estimator $\delta(W) = W$ has risk $\mathbb{E}\ell_M(\Sigma^{1/2}Z)$ for every h, proving the upper bound. For the converse, put the Gaussian prior $H \sim \mathcal{N}(0, \tau^2 I_p)$. The posterior law is

$$H \mid W \sim \mathcal{N}(m_\tau(W), V_\tau), \qquad V_\tau = (\tau^{-2}I_p + \Sigma^{-1})^{-1}.$$

For any action a, translation gives the conditional risk

$$\mathbb{E}[\ell_M(a - H) \mid W] = \mathbb{E}\ell_M\big(a - m_\tau(W) - V_\tau^{1/2}Z\big).$$

A centered Gaussian assigns maximal probability to every Euclidean ball when the ball is centered at the origin: for every $b \in \mathbb{R}^p$ and $r \ge 0$,

$$\Pr\Big(\Big\|V_\tau^{1/2}Z - b\Big\|_2 \le r\Big) \le \Pr\Big(\Big\|V_\tau^{1/2}Z\Big\|_2 \le r\Big),$$

which is Anderson's inequality for symmetric convex sets (Anderson, 1955). Since

$$\mathbb{E}\ell_M(X) = \int_0^M \Pr(\|X\|_2^2 > t)\,\mathrm{d}t,$$

the posterior risk is minimized at $a = m_\tau(W)$ and the Bayes risk equals $\mathbb{E}\ell_M(V_\tau^{1/2}Z)$. Every minimax risk is at least every Bayes risk, so

$$\inf_\delta\sup_h \mathbb{E}_h\ell_M(\delta(W) - h) \ge \mathbb{E}\ell_M(V_\tau^{1/2}Z).$$

As $\tau \to \infty$, $V_\tau \to \Sigma$ in operator norm. Coupling all variables through the same Z and using bounded convergence gives $\mathbb{E}\ell_M(V_\tau^{1/2}Z) \to \mathbb{E}\ell_M(\Sigma^{1/2}Z)$, proving the lower bound. □

Theorem N.5 (Sharp local asymptotic minimax lower bound)**.** *For every fixed $R_0 \in \mathrm{O}(d)$,*

$$\lim_{c\to\infty} \liminf_{n\to\infty} \inf_{\widehat{R}_n} \sup_{\substack{H\in\mathfrak{g}_d \\ \|H\|_{\mathrm{F}}\le c}} \mathbb{E}_{R_0\exp(H/\sqrt{n})} \left[n \left\| \widehat{R}_n - R_0 \exp(H/\sqrt{n}) \right\|_{\mathrm{F}}^2 \right] \ge \Delta_\tau p_d, \tag{N.20}$$

where the infimum is over all measurable $\mathbb{R}^{d\times d}$-valued estimators.

Proof. Let $\mathrm{vec}_{\mathfrak{g}} : \mathfrak{g}_d \to \mathbb{R}^{p_d}$ denote coordinates in the basis $E_1, \dots, E_{p_d}$. Given an arbitrary matrix-valued estimator $\widehat{R}_n$, define the induced tangent estimator

$$T_n := \sqrt{n}\ \mathrm{vec}_{\mathfrak{g}} \left(\mathrm{skew}(R_0^\top \widehat{R}_n) \right). \tag{N.21}$$

For $R_{n,H} = R_0 \exp(H/\sqrt{n})$,

$$\sup_{\|H\|_{\mathrm{F}}\le c} \left\| \sqrt{n}\ \mathrm{skew}\left(R_0^\top R_{n,H}\right) - H \right\|_{\mathrm{F}} = O_c(n^{-1}). \tag{N.22}$$

Because skew is an orthogonal projection under the Frobenius norm,

$$\|T_n - \mathrm{vec}_{\mathfrak{g}}(H)\|_2 \le \sqrt{n} \left\| \widehat{R}_n - R_{n,H} \right\|_{\mathrm{F}} + O_c(n^{-1}). \tag{N.23}$$

For every finite M, the truncated loss $u \mapsto \min\{\|u\|_2^2, M\}$ is $2\sqrt{M}$-Lipschitz in $\|u\|_2$. Hence the uniform $O_c(n^{-1})$ term in equation N.23 changes the corresponding truncated risks by $o(1)$ uniformly over $\|H\|_{\mathrm{F}} \le c$.

Apply the local asymptotic minimax theorem to the LAN experiment of Theorem N.2, first with the bounded continuous losses

$$\ell_M(u) = \min\{\|u\|_2^2, M\}, \qquad M < \infty.$$

The limit Gaussian shift observes $\mathcal{W} = H + \sqrt{\Delta_\tau}\mathcal{Z}$. By Lemma N.4, its minimax risk under ℓ_M is

$$\mathbb{E} \min\left\{ \Delta_\tau \chi^2_{p_d}, M \right\}.$$

The local asymptotic minimax theorem therefore gives, using equation N.23,

$$\begin{aligned} \lim_{c\to\infty} \liminf_{n\to\infty} \inf_{\widehat{R}_n} \sup_{\|H\|_{\mathrm{F}}\le c} \mathbb{E}_{R_{n,H}} \min\left\{ n \left\| \widehat{R}_n - R_{n,H} \right\|_{\mathrm{F}}^2, M \right\} \\ \ge \mathbb{E} \min\left\{ \Delta_\tau \chi^2_{p_d}, M \right\}. \end{aligned}$$

Finally let $M \to \infty$ and use monotone convergence to obtain $\Delta_\tau p_d$. The local asymptotic minimax theorem in this form is standard; see van der Vaart (1998, Chs. 8–9) and Le Cam & Yang (2000, Chs. 3, 6). □

N.5 First-order expansion of the Procrustes MLE

We now show that the MLE attains the lower bound. The key deterministic fact is the derivative of the polar-factor map at the identity.

Lemma N.6 (Second-order expansion of the polar factor)**.** *There exist universal constants $c, C > 0$ such that, whenever $E \in \mathbb{R}^{d\times d}$ satisfies $\|E\|_{\mathrm{op}} \le c$,*

$$\mathrm{Polar}(I_d + E) = I_d + \mathrm{skew}(E) + \mathcal{R}(E), \qquad \|\mathcal{R}(E)\|_{\mathrm{F}} \le C \|E\|_{\mathrm{op}} \|E\|_{\mathrm{F}}. \tag{N.24}$$

Proof. For $\|E\|_{\mathrm{op}}$ sufficiently small, $I_d + E$ is nonsingular and

$$\mathrm{Polar}(I_d + E) = (I_d + E)\left((I_d + E)^\top (I_d + E)\right)^{-1/2}.$$

Write

$$(I_d + E)^\top (I_d + E) = I_d + (E + E^\top) + E^\top E.$$

The analytic functional calculus for $X \mapsto X^{-1/2}$ in a neighborhood of I_d gives

$$(I_d + F)^{-1/2} = I_d - \frac{1}{2}F + \mathcal{S}(F), \qquad \|\mathcal{S}(F)\|_{\mathrm{F}} \le C \|F\|_{\mathrm{op}} \|F\|_{\mathrm{F}}.$$

Substituting $F = E + E^\top + E^\top E$ and multiplying by $I_d + E$ yields the first-order term

$$E - \frac{1}{2}(E + E^\top) = \mathrm{skew}(E),$$

while every remaining term is bounded by $C \|E\|_{\mathrm{op}} \|E\|_{\mathrm{F}}$. This proves equation N.24. The differentiability of the polar map and the same derivative are classical; see Higham (1986). □

Whenever a matrix logarithm appears below, it denotes the smooth local inverse of the exponential map on a fixed neighborhood of I_d in $\mathrm{SO}(d)$. Since the relative MLE converges to I_d, this event has probability tending to one; define the logarithm arbitrarily on its complement, which does not affect any asserted weak limit.

Let

$$C_n = \sum_{i=1}^{n} \overline{Y}_{i,1} \overline{Y}_{i,0}^\top, \qquad \widehat{R}_n = \mathrm{Polar}(C_n), \tag{N.25}$$

which is the unique MLE almost surely for all $n \ge d$ by Proposition K.6. Under R_0, define

$$E_n \coloneqq \frac{1}{n} R_0^\top C_n - I_d = \frac{1}{n} \sum_{i=1}^{n} \big(A(Y_i) - I_d\big). \tag{N.26}$$

Proposition N.7 (Asymptotic linearity and risk convergence)**.** *Under* $P_{R_0}^{\otimes n}$*,*

$$\sqrt{n}\left(R_0^\top \widehat{R}_n - I_d\right) = W_n + o_{L^2}(1), \tag{N.27}$$

$$\sqrt{n}\ \mathrm{skew}\left(R_0^\top \widehat{R}_n - I_d\right) = W_n + o_{L^2}(1), \tag{N.28}$$

$$n\, \mathbb{E}_{R_0}\left[\left\|\widehat{R}_n - R_0\right\|_{\mathrm{F}}^2\right] \longrightarrow \Delta_\tau p_d. \tag{N.29}$$

In particular,

$$\sqrt{n}\ \mathrm{skew}\left(R_0^\top \widehat{R}_n - I_d\right) \rightsquigarrow \mathcal{N}_{\mathfrak{g}_d}(0, \Delta_\tau I). \tag{N.30}$$

Proof. By left equivariance of the polar factor,

$$R_0^\top \widehat{R}_n = \mathrm{Polar}\left(\frac{1}{n} R_0^\top C_n\right) = \mathrm{Polar}(I_d + E_n).$$

The random matrices $A(Y_i) - I_d$ are centered and have moments of every order. Since the ambient dimension is fixed, standard moment inequalities for sums of independent finite-dimensional random vectors give, for each fixed integer $q \ge 2$,

$$\mathbb{E}_{R_0} \|E_n\|_{\mathrm{F}}^q = O_q(n^{-q/2}). \tag{N.31}$$

On the event $\{\|E_n\|_{\mathrm{op}} \le c\}$, apply Lemma N.6:

$$\sqrt{n}\left(R_0^\top \widehat{R}_n - I_d\right) = \sqrt{n}\,\mathrm{skew}(E_n) + \sqrt{n}\mathcal{R}(E_n) = W_n + \sqrt{n}\mathcal{R}(E_n).$$

Moreover,

$$n\mathbb{E}\left[\|\mathcal{R}(E_n)\|_{\mathrm{F}}^2\, \mathbf{1}\{\|E_n\|_{\mathrm{op}} \le c\}\right] \le Cn\mathbb{E}\,\|E_n\|_{\mathrm{F}}^4 = O(n^{-1}).$$

On the complement, write $D_n = \sqrt{n}(R_0^\top \widehat{R}_n - I_d) - W_n$ and use $\left\|R_0^\top \widehat{R}_n - I_d\right\|_{\mathrm{F}} \le 2\sqrt{d}$ to obtain

$$\mathbb{E}\left[\|D_n\|_{\mathrm{F}}^2\, \mathbf{1}\{\|E_n\|_{\mathrm{op}} > c\}\right] \le 8nd\ \Pr(\|E_n\|_{\mathrm{op}} > c) + 2\mathbb{E}\left[\|W_n\|_{\mathrm{F}}^2\, \mathbf{1}\{\|E_n\|_{\mathrm{op}} > c\}\right].$$

By Markov's inequality and equation N.31, the first term is $o(1)$ after choosing an even $q > 2$. The second is $o(1)$ by Cauchy–Schwarz, bounded fourth moments of W_n, and the same probability bound. This proves equation N.27; applying the contractive linear map skew gives equation N.28. Finally, equation N.5 gives

$$\mathbb{E}_{R_0} \|W_n\|_{\mathrm{F}}^2 = \Delta_\tau p_d$$

for every n. The L^2 expansion implies equation N.29, and equation N.30 follows from the central limit theorem in equation N.7. □

N.6 Regularity, convolution, and asymptotic efficiency

The model and the estimator are left-equivariant. If data generated under R are transformed by $(y_0, y_1) \mapsto (y_0, R^\top y_1)$, their law becomes the law under I_d, and

$$R^\top \widehat{R}_n = \mathrm{Polar}(R^\top C_n).$$

Consequently, the distribution of the relative MLE error $R^\top \widehat{R}_n$ is independent of R.

Theorem N.8 (Regularity and asymptotic efficiency of the Procrustes MLE)**.** *For every bounded sequence $h_n \in \mathbb{R}^{p_d}$ with $h_n \to h$, let $H_n = \mathsf{V}(h_n)$ and $R_{n,h_n} = R_0 \exp(H_n/\sqrt{n})$. Then*

$$\sqrt{n}\ \mathrm{skew}\left(R_{n,h_n}^\top \widehat{R}_n - I_d\right) \rightsquigarrow \mathcal{N}_{\mathfrak{g}_d}(0, \Delta_\tau I), \tag{N.32}$$

$$\sqrt{n}\ \mathrm{skew}\left(R_0^\top \widehat{R}_n - I_d\right) \rightsquigarrow \mathcal{N}_{\mathfrak{g}_d}(H, \Delta_\tau I), \qquad H = \mathsf{V}(h). \tag{N.33}$$

Furthermore, for every fixed $c < \infty$,

$$\lim_{n\to\infty} \sup_{\|H\|_{\mathrm{F}} \le c} \left| n\mathbb{E}_{R_0 \exp(H/\sqrt{n})} \left\| \widehat{R}_n - R_0 \exp(H/\sqrt{n}) \right\|_{\mathrm{F}}^2 - \Delta_\tau p_d \right| = 0. \tag{N.34}$$

Hence the MLE attains the local asymptotic minimax lower bound in equation N.20.

For the convolution statement, restrict attention to estimator sequences $\widetilde{R}_n$ taking values in $\mathrm{O}(d)$. Call such a sequence regular *at R_0 if, for every fixed $H \in \mathfrak{g}_d$, under $P^{\otimes n}_{R_0 \exp(H/\sqrt{n})}$,*

$$\left\| (R_0 \exp(H/\sqrt{n}))^\top \widetilde{R}_n - I_d \right\|_{\mathrm{F}} \longrightarrow 0 \quad \textit{in probability}, \tag{N.35}$$

and the centered tangent error

$$T_{n,H}(\widetilde{R}_n) \coloneqq \sqrt{n}\ \mathrm{skew}\left((R_0 \exp(H/\sqrt{n}))^\top \widetilde{R}_n - I_d\right) \tag{N.36}$$

converges to a law L on $\mathfrak{g}_d$ that does not depend on H. For every regular estimator,

$$L \overset{d}{=} Z + V, \qquad Z \sim \mathcal{N}_{\mathfrak{g}_d}(0, \Delta_\tau I), \qquad Z \perp V, \tag{N.37}$$

for some $\mathfrak{g}_d$-valued random variable V. The expansion in Proposition N.7 verifies equation N.35 for the Procrustes MLE; it has $V = 0$ and is asymptotically efficient.

Proof. The paired-view model and the Procrustes estimator are left equivariant. In particular, the distribution of $R^\top \widehat{R}_n$ under parameter R is independent of R. Therefore equation N.32 and its second-moment analogue follow from Proposition N.7 applied at the identity.

Write $A_n = R_{n,h_n}^\top \widehat{R}_n - I_d$. The relative expansion gives $\sqrt{n}A_n = O_P(1)$. Since $R_0^\top R_{n,h_n} = I_d + H_n/\sqrt{n} + O(n^{-1})$ uniformly for bounded H_n,

$$R_0^\top \widehat{R}_n - I_d = \frac{H_n}{\sqrt{n}} + A_n + O_P(n^{-1}).$$

Taking skew parts and multiplying by $\sqrt{n}$ proves equation N.33. Exact equivariance also makes the relative Frobenius risk independent of the true parameter, so equation N.34 follows from equation N.29.

Now let $\widetilde{R}_n$ be regular and define the Euclidean-coordinate estimator

$$\widehat{h}_n = \sqrt{n}\, \mathsf{V}^{-1}\, \mathrm{skew}\left(R_0^\top \widetilde{R}_n - I_d\right).$$

Fix a local parameter h and put $H = \mathsf{V}(h)$, $R_{n,H} = R_0 e^{H/\sqrt{n}}$, and $E_n = R_{n,H}^\top \widetilde{R}_n - I_d$. By equation N.35, $E_n = o_P(1)$, while regularity makes $\sqrt{n}\ \mathrm{skew}(E_n)$ tight. Orthogonality gives

$$E_n + E_n^\top + E_n^\top E_n = 0.$$

Hence, on the event $\|E_n\|_{\mathrm{op}} \le 1/2$,

$$\|E_n\|_{\mathrm{F}} \le \frac{4}{3}\|\mathrm{skew}(E_n)\|_{\mathrm{F}}, \qquad \|\mathrm{sym}(E_n)\|_{\mathrm{F}} = O_P(n^{-1}),$$

so $E_n = O_P(n^{-1/2})$. Expanding $R_0^\top \widetilde{R}_n = e^{H/\sqrt{n}}(I_d + E_n)$ therefore yields

$$\widehat{h}_n - h = \mathsf{V}^{-1} T_{n,H}(\widetilde{R}_n) + o_P(1).$$

Thus $\widehat{h}_n$ is a regular estimator in the Euclidean LAN experiment, with centered limit $\mathsf{V}^{-1}L$. The Hájek–Le Cam convolution theorem applied to the LAN experiment of Theorem N.2 yields

$$\mathsf{V}^{-1}L \overset{d}{=} \mathcal{N}(0, \Delta_\tau I_{p_d}) + V',$$

with independent summands. Mapping back through the isometry V proves equation N.37. The MLE's limit in equation N.32 is exactly the efficient Gaussian component, so its convolution remainder is zero. □

Corollary N.9 (Asymptotic chi-square pivot)**.** *For every fixed $R \in \mathrm{O}(d)$,*

$$\frac{n}{\Delta_\tau}\left\|\widehat{R}_n - R\right\|_{\mathrm{F}}^2 \rightsquigarrow \chi^2_{p_d}, \tag{N.38}$$

and

$$\frac{n}{\Delta_\tau}\mathbb{E}_R\left\|\widehat{R}_n - R\right\|_{\mathrm{F}}^2 \longrightarrow p_d. \tag{N.39}$$

Proof. By equivariance it is enough to take $R = R_0$. The first conclusion follows from equation N.27 and equation N.7; the second is equation N.29 divided by Δ_τ. □

Remark N.10 (Scope of the asymptotic statement). All limits in this appendix hold for fixed d and fixed positive Δ_τ. They do not cover a growing-dimensional regime or the singular limit $\Delta_\tau \downarrow 0$. The latter is governed instead by the finite-source breadth obstruction of Proposition M.1 and corollary M.2. In particular, divergence of the regular Fisher information as $\Delta_\tau \downarrow 0$ does not imply that one noiseless task identifies an arbitrary orthogonal transformation.

O Fresh-Target Linear-Quadratic Control

This appendix converts the source-side orthogonal estimation theory of Sections K to M into an operational fresh-target control guarantee. The target problem is a finite-horizon linear-quadratic endpoint-tracking problem with known dynamics and an unknown goal generated by the same global orthogonal transformation as the source tasks. Linear-quadratic tracking is classical (Anderson & Moore, 2007); all identities needed below are derived directly.

The specialization uses square-integrable quadratic costs rather than the bounded rewards imposed for the general finite-model theory in Section A. No bounded-reward assertion is used in this appendix. Conditional on the fixed global parameter $R \in \mathrm{O}(d)$, source data and the fresh target task are independent.

Proposition O.1 (Embedding in the source–target kernel framework)**.** *The Gaussian paired-view source experiment and the target probe–LQ experiment below are a specialization of the measurable source, adaptation, and deployment interfaces in Sections A.2 and A.6.*

Proof. Take global parameter space $\Theta = \mathrm{O}(d)$, task-latent space $\mathcal{Z} = \mathbb{R}^d$, and fresh-task kernel $\mu_R = \mathcal{N}(0, I_d)$. The source interface is the horizon-one paired-view controlled process of Section K.1: action $v \in \{0,1\}$ produces terminal observation $X + \varepsilon_0$ or $RX + \varepsilon_1$, respectively. Its kernels are Borel because they are Gaussian translations with means jointly continuous in (R, X).

For target adaptation, take horizon one, a singleton action space, and state space $\{s_\circ\} \sqcup \mathbb{R}^d$. The initial state is $s_\circ$ and the unique transition sends it to $Z = X + \xi$, $\xi \sim \mathcal{N}(0, \sigma_\star^2 I_d)$, with zero reward. Thus each adaptation trajectory reveals exactly one probe in equation O.2.

For deployment, use the standard-Borel stage-augmented state and action spaces

$$\mathcal{S}_{\mathrm{dep}} \coloneqq \bigsqcup_{h=0}^{H} \left(\{h\} \times \mathbb{R}^d \times \prod_{j=0}^{h-1} \mathbb{R}^{m_j} \right),$$
$$\mathcal{A}_{\mathrm{dep}} \coloneqq \bigsqcup_{h=0}^{H-1} \left(\{h\} \times \mathbb{R}^{m_h} \right).$$

The initial state is $(0, 0, \varnothing)$. At a stage-h state $(h, s, u_{0:h-1})$, a stage-compatible action (h, u_h) deterministically produces

$$(h+1, \mathsf{A}_h s + \mathsf{B}_h u_h, u_{0:h}).$$

Admissible deployment policies select stage-compatible actions; the kernels may be defined arbitrarily on incompatible state–action pairs. Rewards are zero before the final transition and, on the final transition, equal

$$-\left(\|s_H - RX\|_2^2 + U^\top \Lambda U \right), \qquad U = (u_0^\top, \dots, u_{H-1}^\top)^\top.$$

The augmented state stores the complete control history, so this terminal reward is a Borel function of the current state, action, (R, X), and next state. It is integrable under every strategy considered below. Maximizing deployment value is therefore exactly equivalent to minimizing $J_H(U; RX)$ in equation O.8. The source task, all target adaptation episodes, and deployment share the same latent X law and global transformation R, while their interfaces differ exactly as allowed by Section A.6. □

O.1 Source experiment and fresh-target probes

Let D_n denote n independent paired source observations from the orthogonal model in equation K.5, with effective noise parameter $\Delta_\tau > 0$. A fresh target task draws

$$X_\star \sim \mathcal{N}(0, I_d), \qquad G_\star \coloneqq RX_\star, \tag{O.1}$$

independently of D_n. Before deployment, the learner receives $B \in \mathbb{N}_0$ probes

$$Z_b = X_\star + \xi_b, \qquad \xi_b \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\star^2 I_d), \qquad b \in [B], \tag{O.2}$$

where $\sigma_\star^2 > 0$. For $B = 0$, the probe sigma-field is trivial. Define

$$\alpha_B \coloneqq \frac{B}{B + \sigma_\star^2}, \qquad v_B \coloneqq \frac{\sigma_\star^2}{B + \sigma_\star^2}, \qquad \alpha_B + v_B = 1. \tag{O.3}$$

Proposition O.2 (Gaussian posterior decomposition)**.** *Let $\mathcal{Z}_B = \sigma(Z_1, \dots, Z_B)$ and use the convention that an empty sum is zero. Then*

$$M_B \coloneqq \mathbb{E}[X_\star \mid \mathcal{Z}_B] = \frac{1}{B + \sigma_\star^2} \sum_{b=1}^{B} Z_b, \tag{O.4}$$

$$X_\star = M_B + E_B, \qquad M_B \sim \mathcal{N}(0, \alpha_B I_d), \qquad E_B \sim \mathcal{N}(0, v_B I_d), \tag{O.5}$$

where M_B and E_B are independent. When $B \geq 1$, the centered probe contrasts are independent of $(X_\star, M_B)$ and contain no additional information about $X_\star$.

Proof. For $B \geq 1$, the conditional density of $X_\star$ given $Z_{1:B} = z_{1:B}$ is, up to normalization,

$$\exp\left\{ -\frac{1}{2}\|x\|_2^2 - \frac{1}{2\sigma_\star^2} \sum_{b=1}^{B} \|z_b - x\|_2^2 \right\}.$$

Completing the square gives posterior covariance $v_B I_d$ and posterior mean equation O.4. The Gaussian orthogonality principle yields $\mathrm{Cov}(M_B, E_B) = 0$; since (M_B, E_B) is jointly Gaussian, the two variables are independent. The law of total covariance gives $\mathrm{Cov}(M_B) = I_d - v_B I_d = \alpha_B I_d$.

Finally, apply an orthogonal transformation in the probe index to $(Z_1, \dots, Z_B)$. Its first coordinate is proportional to the sample mean, and the remaining coordinates are centered contrasts. The shared signal $X_\star$ cancels from every contrast, and Gaussian orthogonality makes the contrasts independent of both the sample mean and $X_\star$. The case $B = 0$ follows directly from $M_B = 0$ and $E_B = X_\star$. □

O.2 FINITE-HORIZON ENDPOINT TRACKING

Consider known, possibly time-varying linear dynamics

$$S_{h+1} = \mathsf{A}_h S_h + \mathsf{B}_h U_h, \qquad h = 0, \dots, H-1, \qquad S_0 = 0, \tag{O.6}$$

where $S_h \in \mathbb{R}^d$ and $U_h \in \mathbb{R}^{m_h}$. For $0 \le s \le t \le H$, let

$$\Phi_{t,s} \coloneqq \begin{cases} \mathsf{A}_{t-1} \cdots \mathsf{A}_s, & s < t, \\ I_d, & s = t. \end{cases}$$

Stack the controls into $U = (U_0^\top, \dots, U_{H-1}^\top)^\top \in \mathbb{R}^{m_{\mathrm{tot}}}$ and define the endpoint controllability matrix

$$\mathsf{C}_H \coloneqq (\Phi_{H,1}\mathsf{B}_0 \quad \Phi_{H,2}\mathsf{B}_1 \quad \cdots \quad \Phi_{H,H}\mathsf{B}_{H-1}), \qquad S_H = \mathsf{C}_H U. \tag{O.7}$$

Let $\Lambda \in \mathbb{R}^{m_{\mathrm{tot}} \times m_{\mathrm{tot}}}$ be symmetric positive definite. For a goal $g \in \mathbb{R}^d$, the deployment cost is

$$J_H(U; g) \coloneqq \|\mathsf{C}_H U - g\|_2^2 + U^\top \Lambda U. \tag{O.8}$$

Set

$$\mathsf{M}_H \coloneqq \mathsf{C}_H^\top \mathsf{C}_H + \Lambda, \qquad \mathsf{L}_H \coloneqq \mathsf{M}_H^{-1} \mathsf{C}_H^\top, \qquad W_H \coloneqq \mathsf{C}_H \mathsf{M}_H^{-1} \mathsf{C}_H^\top. \tag{O.9}$$

We assume throughout this appendix that C_H has full row rank. Then W_H is symmetric positive definite. Write

$$\underline{w}_H \coloneqq \lambda_{\min}(W_H), \qquad \overline{w}_H \coloneqq \lambda_{\max}(W_H). \tag{O.10}$$

Proposition O.3 (Exact endpoint-control identities)**.** *For every goal $g \in \mathbb{R}^d$:*

1. *the unique optimal control is*
$$U^\star(g) = \mathsf{L}_H g; \tag{O.11}$$
2. *for every control U,*
$$J_H(U; g) - J_H(U^\star(g); g) = \|U - U^\star(g)\|_{\mathsf{M}_H}^2; \tag{O.12}$$
3. *for every estimated goal $\widehat{g}$,*
$$J_H(U^\star(\widehat{g}); g) - J_H(U^\star(g); g) = (\widehat{g} - g)^\top W_H (\widehat{g} - g). \tag{O.13}$$

Here $\|u\|_{\mathsf{M}_H}^2 = u^\top \mathsf{M}_H u$.

Proof. Expanding equation O.8 gives

$$J_H(U; g) = U^\top \mathsf{M}_H U - 2 g^\top \mathsf{C}_H U + \|g\|_2^2.$$

Since $\mathsf{M}_H \succ 0$, setting the gradient to zero gives equation O.11. Completing the square yields equation O.12. Substituting $U^\star(\widehat{g}) - U^\star(g) = \mathsf{L}_H(\widehat{g} - g)$ into that identity and using

$$\mathsf{L}_H^\top \mathsf{M}_H \mathsf{L}_H = \mathsf{C}_H \mathsf{M}_H^{-1} \mathsf{C}_H^\top = W_H$$

proves equation O.13. □

The next lemma shows that arbitrary control rules cannot improve on goal-estimation followed by the optimal linear controller.

Lemma O.4 (Control-to-estimation reduction)**.** *For $U \in \mathbb{R}^{m_{\mathrm{tot}}}$, define*

$$\widehat{g}(U) \coloneqq W_H^{-1} \mathsf{C}_H U, \qquad \Pi_H U \coloneqq \mathsf{L}_H \widehat{g}(U). \tag{O.14}$$

Then, for every $g \in \mathbb{R}^d$,

$$\begin{aligned} J_H(U; g) - J_H(U^\star(g); g) &= \|U - \Pi_H U\|_{\mathsf{M}_H}^2 + \big(\widehat{g}(U) - g\big)^\top W_H \big(\widehat{g}(U) - g\big) \\ &\ge \big(\widehat{g}(U) - g\big)^\top W_H \big(\widehat{g}(U) - g\big). \end{aligned} \tag{O.15}$$

Consequently, pointwise projection $U \mapsto \Pi_H U$ never increases deployment cost for any goal.

Proof. The range of L_H is a d-dimensional subspace of the control space. Since

$$\mathsf{L}_H^\top \mathsf{M}_H = \mathsf{C}_H \qquad \text{and} \qquad \mathsf{C}_H \mathsf{L}_H = W_H,$$

we have

$$\mathsf{L}_H^\top \mathsf{M}_H (U - \Pi_H U) = \mathsf{C}_H U - W_H W_H^{-1} \mathsf{C}_H U = 0.$$

Thus $U - \Pi_H U$ is M_H-orthogonal to $\mathrm{range}(\mathsf{L}_H)$. Since

$$\Pi_H U - U^\star(g) = \mathsf{L}_H(\widehat{g}(U) - g),$$

the Pythagorean identity in the M_H inner product, followed by Proposition O.3, gives equation O.15. □

O.3 Exact reduction to random-design prediction

A randomized target controller is a Borel map of $(\mathsf{D}_n, Z_{1:B}, U_0)$ into the stacked control space, where $U_0 \sim \mathrm{Unif}[0,1]$ is independent private randomization. The oracle benchmark knows both R and the realized target anchor $X_\star$, whereas the meta-learning benchmark knows R but receives only the same B target probes as the learner. Define

$$\mathfrak{R}_{n,B}^{\mathrm{oracle}}(\Delta_\tau; W_H) \coloneqq \inf_{\widehat{U}} \sup_{R \in \mathrm{O}(d)} \mathbb{E}_R \left[J_H(\widehat{U}; RX_\star) - J_H(U^\star(RX_\star); RX_\star) \right], \tag{O.16}$$

$$\mathfrak{R}_{n,B}^{\mathrm{meta}}(\Delta_\tau; W_H) \coloneqq \inf_{\widehat{U}} \sup_{R \in \mathrm{O}(d)} \mathbb{E}_R \left[J_H(\widehat{U}; RX_\star) - J_H(U^\star(RM_B); RX_\star) \right]. \tag{O.17}$$

For $\alpha \in [0,1]$ and a positive-definite matrix W, define the random-design prediction risk

$$\mathfrak{P}_{n,d}(\Delta_\tau, \alpha; W) \coloneqq \inf_{\psi} \sup_{R \in \mathrm{O}(d)} \mathbb{E}_R \left[\big(\psi(\mathsf{D}_n, M) - RM\big)^\top W \big(\psi(\mathsf{D}_n, M) - RM\big) \right], \tag{O.18}$$

where $M \sim \mathcal{N}(0, \alpha I_d)$ is independent of the source data and the infimum is over all jointly measurable predictors.

Theorem O.5 (Exact source–target benchmark decomposition)**.** *For every n, B, d and $\Delta_\tau > 0$,*

$$\mathfrak{R}_{n,B}^{\mathrm{oracle}}(\Delta_\tau; W_H) = v_B \,\mathrm{Tr}(W_H) + \mathfrak{P}_{n,d}(\Delta_\tau, \alpha_B; W_H), \tag{O.19}$$

$$\mathfrak{R}_{n,B}^{\mathrm{meta}}(\Delta_\tau; W_H) = \mathfrak{P}_{n,d}(\Delta_\tau, \alpha_B; W_H). \tag{O.20}$$

Moreover,

$$\mathfrak{P}_{n,d}(\Delta_\tau, \alpha; W) = \alpha \, \mathfrak{P}_{n,d}(\Delta_\tau, 1; W) \qquad \forall \alpha \in [0,1]. \tag{O.21}$$

If the global transformation R is known, the optimal B-probe target loss relative to the full-information oracle is

$$\mathcal{L}_B^\star(R) = v_B \,\mathrm{Tr}(W_H). \tag{O.22}$$

Hence equation O.20, rather than the total oracle risk, is the source-induced excess corresponding to the benchmark in Section 2.

Proof. By Lemma O.4, every control rule can be projected, pointwise in the observed data, onto a rule of the form $U^\star(\widehat{G})$ without increasing cost. Conversely, every measurable goal estimate $\widehat{G}$ induces such a control rule. Hence the oracle risk on the left-hand side of equation O.19 equals

$$\inf_{\widehat{G}} \sup_{R} \mathbb{E}_R \left[(\widehat{G} - RX_\star)^\top W_H (\widehat{G} - RX_\star) \right].$$

The centered probe contrasts are independent of $(X_\star, M_B, \mathsf{D}_n)$ by Proposition O.2. Conditional Jensen therefore shows that replacing any $\widehat{G}(\mathsf{D}_n, Z_{1:B})$ by its conditional expectation given (D_n, M_B) cannot increase the quadratic risk. We may therefore write $\widehat{G} = \psi(\mathsf{D}_n, M_B)$. From equation O.5,

$$RX_\star = RM_B + RE_B,$$

where E_B is independent of (D_n, M_B) and has mean zero. The cross term vanishes, while orthogonality of R gives

$$\mathbb{E}\left[(RE_B)^\top W_H (RE_B) \right] = v_B \,\mathrm{Tr}(W_H).$$

Taking the minimax infimum proves equation O.19 and equation O.22. The known-R B-probe controller uses the posterior-mean goal RM_B and has exactly the irreducible excess $v_B \operatorname{Tr}(W_H)$; subtracting this benchmark from the oracle-risk identity gives equation O.20.

For equation O.21, the claim is immediate when $\alpha = 0$. When $\alpha > 0$, write $M = \sqrt{\alpha}\,U$ with $U \sim \mathcal{N}(0, I_d)$. The bijection

$$\phi(\mathsf{D}_n, U) = \alpha^{-1/2}\psi(\mathsf{D}_n, \sqrt{\alpha}\,U)$$

between predictors shows that every risk is multiplied by α. □

O.4 Prediction lower and upper bounds

For brevity, write

$$\mathfrak{P}^{\mathrm{E}}_{n,d}(\Delta_\tau, \alpha) \coloneqq \mathfrak{P}_{n,d}(\Delta_\tau, \alpha; I_d).$$

Theorem O.6 (Minimax random-design prediction)**.** *There are universal constants $0 < c \le C < \infty$ such that, for every $d \ge 2$, $n \ge 1$, $\Delta_\tau > 0$, and $\alpha \in [0, 1]$,*

$$\mathfrak{P}^{\mathrm{E}}_{n,d}(\Delta_\tau, \alpha) \ge \alpha \max\left\{ c\left(\frac{\Delta_\tau p_d}{n} \wedge d\right), (d-n)_+ \right\}. \tag{O.23}$$

If $n \ge 2d + 2$, then

$$\mathfrak{P}^{\mathrm{E}}_{n,d}(\Delta_\tau, \alpha) \le C\alpha\left(\frac{\Delta_\tau p_d}{n} \wedge d\right). \tag{O.24}$$

In the noiseless source experiment,

$$\mathfrak{P}^{\mathrm{E}}_{n,d}(0, \alpha) = \alpha(d-n)_+. \tag{O.25}$$

If the complete source transcript law is independent of R—in particular, under reset-only one-view acquisition—then for every positive-definite W,

$$\mathfrak{P}_{\mathrm{reset}}(\alpha; W) = \alpha \operatorname{Tr}(W). \tag{O.26}$$

Proof. By equation O.21, it is enough to prove the statements for $\alpha = 1$.

For the upper bound, let $\widehat{R}$ be either estimator from Theorems L.2 and L.5 and use $\psi(\mathsf{D}_n, M) = \widehat{R}M$. Since $M \sim \mathcal{N}(0, I_d)$ is independent of the source data,

$$\mathbb{E}_R\|\widehat{R}M - RM\|_2^2 = \mathbb{E}_R\left\|\widehat{R} - R\right\|_{\mathrm{F}}^2.$$

The upper bounds in Section L prove equation O.24.

For the regular lower bound, use the localized packing $R_1, \dots, R_M$ from Lemma M.6. Regard

$$f_j(u) = R_j u$$

as elements of the Hilbert space $L^2(\gamma_d; \mathbb{R}^d)$, where $\gamma_d = \mathcal{N}(0, I_d)$. Their squared Hilbert distance is

$$\|f_j - f_k\|^2_{L^2(\gamma_d)} = \mathbb{E}_{U\sim\gamma_d}\|(R_j - R_k)U\|_2^2 = \|R_j - R_k\|_{\mathrm{F}}^2.$$

Given any predictor ψ, and conditioning on its source transcript and internal random seed, interpret $u \mapsto \psi(\mathsf{D}_n, u)$ as an estimated function in this Hilbert space. Decode the nearest f_j in integrated squared distance, breaking ties by the smallest index. The Fano argument of Lemma M.7, with the same exact source KL divergence as in Section M, therefore gives

$$\inf_{\psi}\sup_{R}\mathbb{E}_R\|\psi(\mathsf{D}_n, U) - RU\|_2^2 \ge c\left(\frac{\Delta_\tau p_d}{n} \wedge d\right).$$

The independent random design contributes no KL divergence.

The noisy source experiment is a parameter-independent garbling of the noiseless source pairs. Hence its prediction risk is no smaller than the noiseless prediction risk. The exact identity equation O.25 for $\alpha = 1$ is Theorem K.8; scaling gives the general case and establishes the breadth term in equation O.23.

Finally, suppose the source transcript is independent of R. The zero predictor has risk

$$\mathbb{E}_R[(RM)^\top W(RM)] = \alpha \operatorname{Tr}(W)$$

for every fixed R, proving the upper bound in equation O.26. For the converse, place normalized Haar prior on R. Conditional on M, the Haar mean of RM is zero, so the Bayes estimator is zero and its Bayes risk is

$$\mathbb{E}_M\left[\frac{\|M\|_2^2}{d}\operatorname{Tr}(W)\right] = \alpha \operatorname{Tr}(W).$$

A Bayes risk lower bounds the minimax risk. □

O.5 Source-to-target control complexity

Theorem O.7 (Source–target decomposition for orthogonal LQ goal transfer)**.** *There are universal constants $0 < c \le C < \infty$ such that, for every $d \ge 2$, $n \ge 1$, $B \ge 0$, and $\Delta_\tau > 0$,*

$$\mathfrak{R}^{\mathrm{meta}}_{n,B}(\Delta_\tau; W_H) \ge c\alpha_B \underline{w}_H \max\left\{\frac{\Delta_\tau p_d}{n} \wedge d, (d-n)_+\right\}, \tag{O.27}$$

$$\mathfrak{R}^{\mathrm{oracle}}_{n,B}(\Delta_\tau; W_H) \ge v_B \operatorname{Tr}(W_H) + c\alpha_B \underline{w}_H \max\left\{\frac{\Delta_\tau p_d}{n} \wedge d, (d-n)_+\right\}. \tag{O.28}$$

If $n \ge 2d+2$, then

$$\mathfrak{R}^{\mathrm{meta}}_{n,B}(\Delta_\tau; W_H) \le C\alpha_B \overline{w}_H \left(\frac{\Delta_\tau p_d}{n} \wedge d\right), \tag{O.29}$$

$$\mathfrak{R}^{\mathrm{oracle}}_{n,B}(\Delta_\tau; W_H) \le v_B \operatorname{Tr}(W_H) + C\alpha_B \overline{w}_H \left(\frac{\Delta_\tau p_d}{n} \wedge d\right). \tag{O.30}$$

Thus, under a uniformly conditioned target system and $n \ge 2d+2$, the source-induced meta-excess is minimax equivalent to

$$\alpha_B \left(\frac{\Delta_\tau p_d}{n} \wedge d\right),$$

while the total oracle excess adds the irreducible term $v_B \operatorname{Tr}(W_H)$. Under any reset-only one-view source protocol whose transcript law is independent of R,

$$\mathfrak{R}^{\mathrm{meta}}_{\mathrm{reset},B}(W_H) = \alpha_B \operatorname{Tr}(W_H), \qquad \mathfrak{R}^{\mathrm{oracle}}_{\mathrm{reset},B}(W_H) = \operatorname{Tr}(W_H). \tag{O.31}$$

Proof. For every vector u,

$$\underline{w}_H \|u\|_2^2 \le u^\top W_H u \le \overline{w}_H \|u\|_2^2.$$

Apply these inequalities to equation O.19 and invoke Theorem O.6. Under reset-only access, equation O.26 gives the meta-excess $\alpha_B \operatorname{Tr}(W_H)$; adding the known-$R$ benchmark $v_B \operatorname{Tr}(W_H)$ and using $\alpha_B + v_B = 1$ gives the oracle excess $\operatorname{Tr}(W_H)$. □

Corollary O.8 (Exact isotropic integrator consequences)**.** *Consider the fully actuated integrator*

$$S_{h+1} = S_h + U_h, \qquad S_0 = 0, \qquad U_h \in \mathbb{R}^d, \tag{O.32}$$

with control penalty $\lambda \sum_{h=0}^{H-1} \|U_h\|_2^2$, where $\lambda > 0$. Then

$$W_H = w_H I_d, \qquad w_H := \frac{H}{H+\lambda}. \tag{O.33}$$

Consequently:

1. *under noiseless paired source observations,*

$$\mathfrak{R}^{\mathrm{oracle}}_{n,B}(0; w_H I_d) = w_H\left[dv_B + \alpha_B(d-n)_+\right]; \tag{O.34}$$

2. *under reset-only one-view source acquisition,*

$$\mathfrak{R}^{\mathrm{oracle}}_{\mathrm{reset},B}(w_H I_d) = w_H d, \tag{O.35}$$

independently of the number of reset-only source episodes and of B;

3. for positive source noise and $n \geq 2d + 2$,

$$\mathfrak{R}^{\mathrm{oracle}}_{n,B}(\Delta_\tau; w_H I_d) \asymp w_H \left[d v_B + \alpha_B \left(\frac{\Delta_\tau p_d}{n} \wedge d \right) \right]. \tag{O.36}$$

Proof. For equation O.32,

$$\mathsf{C}_H = (I_d \quad I_d \quad \cdots \quad I_d), \qquad \mathsf{C}_H \mathsf{C}_H^\top = H I_d, \qquad \Lambda = \lambda I_{dH}.$$

The push-through identity gives

$$W_H = \mathsf{C}_H (\mathsf{C}_H^\top \mathsf{C}_H + \lambda I)^{-1} \mathsf{C}_H^\top = \mathsf{C}_H \mathsf{C}_H^\top (\mathsf{C}_H \mathsf{C}_H^\top + \lambda I_d)^{-1} = \frac{H}{H + \lambda} I_d.$$

Equation O.34 follows from Theorem O.5 and equation O.25. Equation O.35 follows from equation O.26 and $\alpha_B + v_B = 1$. The final claim is Theorem O.7 specialized to $W_H = w_H I_d$. □

O.6 Interpretation and scope

The oracle-benchmark decomposition

$$\mathfrak{R}^{\mathrm{oracle}}_{n,B} = \underbrace{v_B \operatorname{Tr}(W_H)}_{\text{fresh-task uncertainty}} + \underbrace{\mathfrak{P}_{n,d}(\Delta_\tau, \alpha_B; W_H)}_{\text{source uncertainty about } R}$$

separates two statistically distinct resources, while the meta-learning benchmark subtracts the first term and leaves $\mathfrak{R}^{\mathrm{meta}}_{n,B} = \mathfrak{P}_{n,d}$. Target probes shrink the unavoidable fresh-task term through v_B, but they simultaneously increase the amount of target-anchor energy on which source knowledge of R can be used, through α_B. This is why the source term is multiplied by α_B rather than one. In particular, if $B = 0$, then $\alpha_B = 0$ and the global transformation is irrelevant because no target anchor is observed; if $B \to \infty$, then $v_B \to 0$ and the source transformation error is fully exposed.

The target dynamics are known and deterministic. This isolates the value of source task persistence from the separate problem of learning target dynamics. Extensions to stochastic process noise with known law can be handled when the noise contributes the same additive cost under every controller, but unknown target dynamics are outside the present theorem.

P Adaptive Persistence Separation

This appendix isolates the value of conditioning the persistence decision on information revealed by the current task. Adaptivity gaps are classical in stochastic optimization: an adaptive policy may condition future probes on realized outcomes, whereas a nonadaptive policy commits before those outcomes are observed (Dean et al., 2008; Gupta et al., 2016). Our separation has a different statistical objective. A first source episode reveals whether a task is globally informative; continuing an informative task unlocks a high-order statistic about a shared parameter, while continuing an uninformative task is wasted. We prove exact acquisition costs and matching fixed-confidence constants for adaptive and precommitted persistence.

Throughout, fix an integer $L \geq 2$, an informative-task probability $p \in (0, 1]$, a parity signal $\rho \in (0, 1)$, and a reset cost $\kappa \geq 0$. The global parameter is $\sigma \in \{-1, +1\}$. We write

$$\chi(x) := \prod_{j=1}^{L} x_j, \qquad \ell_\rho := \log \frac{1+\rho}{1-\rho}, \qquad d_{\mathrm{par}}(\rho) := \rho \ell_\rho. \tag{P.1}$$

P.1 A screened parity meta-MDP

Definition P.1 (Screened parity task law)**.** A fresh task has latent

$$Z = (T, X) \in \{0, 1\} \times \{-1, +1\}^L,$$

where $T \sim \mathrm{Bernoulli}(p)$ is independent across resets. Under global model σ, define

$$\mu_\sigma(t, x) := \begin{cases} (1-p) 2^{-L}, & t = 0, \\ p\, 2^{-L} \big(1 + \rho \sigma \chi(x)\big), & t = 1. \end{cases} \tag{P.2}$$

Tasks with $T = 1$ are *informative*; tasks with $T = 0$ are *null*.

The mass function in equation P.2 is strictly positive and sums to one. Conditional on $T = 0$, the coordinates are independent fair signs. Conditional on $T = 1$, the law is the parity-tilted family of Definition C.1, with $r = L$ and bias ρ.

Definition P.2 (Screened parity meta-MDP)**.** For each (σ, t, x), let $M^{\mathrm{scr}}_{\sigma,t,x}$ be the horizon-one MDP with states

$$\mathcal{S}_{\mathrm{scr}} = \{s_0, s_1, s_\dagger\}$$

and actions

$$\mathcal{A}_{\mathrm{scr}} = \{0, 1, \ldots, L\}.$$

The initial state is s_t, so the task type $T = t$ is observed before the episode action is selected. Every action transitions to $s_\dagger$. Action 0 is safe and gives reward $1/2$; coordinate action $j \in [L]$ gives reward

$$R(j; x) = \frac{1 + x_j}{2} \in \{0, 1\}. \tag{P.3}$$

A reset samples a fresh $(T, X) \sim \mu_\sigma$; continuation preserves both T and X.

Thus the first episode on a fresh task simultaneously reveals its type and, if a coordinate action is selected, one coordinate of X. Every additional episode can reveal at most one further coordinate. A task is called *complete* if all L coordinates have been revealed before the next reset or source stopping time. It is called *informatively complete* if it is complete and $T = 1$.

Lemma P.3 (Source information appears only on informative completion)**.** *For either task type, every strict coordinate marginal is independent of σ. If $W \coloneqq \chi(X)$, then*

$$\mathbb{P}_\sigma(W = w \mid T = 1) = \frac{1}{2}(1 + \rho\sigma w), \tag{P.4}$$

$$\mathbb{P}_\sigma(W = w \mid T = 0) = \frac{1}{2}. \tag{P.5}$$

Consequently, one informatively complete task contributes exactly $d_{\mathrm{par}}(\rho)$ nats of KL information between the global models; an incomplete task or a complete null task contributes zero.

Proof. Conditional on $T = 0$, the claim follows from uniformity. Conditional on $T = 1$, proper-marginal uniformity is Lemma C.2 with $r = L$ and $\varrho = \rho$. Summing equation P.2 over the 2^{L-1} vectors with parity w gives equation P.4. Its two global-model laws are Bernoulli sign laws with directional KL

$$\frac{1 + \rho}{2}\ell_\rho + \frac{1 - \rho}{2}(-\ell_\rho) = \rho\ell_\rho.$$

The null parity law is model independent. Finally, before all L distinct coordinates of an informative task have been revealed, every observed coordinate set is a proper marginal and hence model independent. □

P.2 Fresh-target decision geometry

Set the target adaptation budget to

$$B = L - 1.$$

The target interacts with the same MDP in Definition P.2; adaptation rewards are observed but excluded from deployment evaluation, as in Appendix A.

Proposition P.4 (Target-relevant global sign)**.** *For each $\sigma \in \{-1, +1\}$, the optimal known-model target loss is*

$$\mathcal{L}^\star_{L-1}(\sigma) = \frac{1 - p\rho}{2^{L+1}}. \tag{P.6}$$

Moreover, every target strategy g satisfies

$$\frac{1}{2}\sum_{\sigma \in \{-1,+1\}} \left[\mathcal{L}_{L-1}(\sigma, g) - \mathcal{L}^\star_{L-1}(\sigma)\right] \geq \frac{p\rho}{2^{L+1}}. \tag{P.7}$$

Hence, for every

$$0 < \varepsilon < \frac{p\rho}{2^{L+1}}, \tag{P.8}$$

the two ε-optimal target-strategy sets are disjoint. There are canonical optimal strategies $g^\star_+, g^\star_-$ satisfying

$$\mathcal{L}_{L-1}(\sigma, g^\star_{-\sigma}) - \mathcal{L}^\star_{L-1}(\sigma) = \frac{p\rho}{2^L}. \tag{P.9}$$

Proof. An optimal strategy queries $L-1$ distinct coordinates unless a revealed coordinate already has value $+1$, in which case deployment can attain the oracle value one. Consider the remaining history on which all queried coordinates equal -1.

If $T = 0$, the final unobserved coordinate is fair. The safe action and the unobserved coordinate each have expected reward $1/2$, while the conditional oracle value is $3/4$. Since the all-negative adaptation history has probability $2^{-(L-1)}$, the conditional-null contribution to target loss is $2^{-(L+1)}$.

If $T = 1$ and σ is known, the final coordinate has posterior probability $(1 \pm \rho)/2$ of being $+1$, with the sign determined by σ and the observed parity. The optimal deployment action is therefore the unobserved coordinate when this posterior exceeds $1/2$, and the safe action otherwise. Exactly as in Lemma C.8, the conditional-informative contribution is $(1-\rho)2^{-(L+1)}$. Weighting the two task types gives equation P.6.

Under the uniform prior on σ, the informative parity tilt cancels, so X is uniform conditional on either value of T. The smallest possible mixture target loss is therefore $2^{-(L+1)}$ for both task types. Subtracting equation P.6 proves equation P.7. Disjointness follows exactly as in Lemma C.9. Finally, the two canonical strategies differ only on an informative task after the all-negative adaptation history. On that event, choosing the action preferred by the wrong sign loses $\rho/2$ in expected reward. Multiplying by $p2^{-(L-1)}$ proves equation P.9. □

P.3 Adaptive and precommitted persistence classes

Index source tasks by reset order, and let T_i be the type of the ith fresh task. Let N_{task} be the number of tasks initiated before source stopping, and let

$$N^{\mathrm{full}}_{\mathrm{inf}}$$

be the number of informatively complete tasks.

An *adaptive-persistence learner* is any admissible source learner from Appendix A; in particular, after the first episode on a task, its continuation decision may depend on the observed type.

Definition P.5 (Precommitted-persistence learner)**.** A source learner is *precommitted* if, immediately before initiating task i and hence before observing T_i, it chooses a block length $M_i \in \mathbb{N}$. The variable M_i may depend on the complete transcript of tasks $1, \dots, i-1$ and on private randomization, but not on any observation from task i. The learner must reset or stop after exactly M_i episodes on that task. Its within-task actions may still depend arbitrarily on the observed type and coordinate history.

Thus precommitted learners may adapt across completed tasks and may adapt their actions within the current task; only the number of episodes invested in the current task is chosen before its type is revealed. This is the precise feedback removed by the comparator class.

Proposition P.6 (Exact source information accounting)**.** *For each $\sigma \in \{-1, +1\}$, let an adaptive-persistence source learner have a stopping time finite $\mathbb{P}^{\mathcal{A}}_\sigma$-almost surely and satisfy*

$$\mathbb{E}^{\mathcal{A}}_\sigma \left[N^{\mathrm{full}}_{\mathrm{inf}} \right] < \infty.$$

Then

$$D_{\mathrm{KL}} \left(\mathbb{P}^{\mathcal{A}}_\sigma \middle\| \mathbb{P}^{\mathcal{A}}_{-\sigma} \right) = d_{\mathrm{par}}(\rho) \mathbb{E}^{\mathcal{A}}_\sigma \left[N^{\mathrm{full}}_{\mathrm{inf}} \right]. \tag{P.10}$$

Finite expected source cost implies the displayed integrability because every informatively complete task costs at least $L + \kappa$.

Proof. Fix $\sigma \in \{-1, +1\}$ and apply the conditional KL chain rule under $\mathbb{P}^{\mathcal{A}}_\sigma$ episode by episode, including the observed initial task type in each trajectory. The type law is the same under both global models. Conditional on $T = 0$, every coordinate law is model independent. Conditional on $T = 1$, safe actions, repeated coordinates, and every new coordinate before the final missing coordinate have zero conditional KL by proper-marginal uniformity. Revealing the final missing coordinate

contributes exactly $d_{\mathrm{par}}(\rho)$, independently of the realized preceding coordinate values. Persistence and action-selection kernels are fixed functions of the observed history and contribute zero KL. The random-stopping argument is the same cemetery-padding and monotone-limit argument used in Lemma C.7. Summing the only nonzero increments proves equation P.10. □

We next relate completed informative tasks to source cost. Define

$$c_{\mathrm{ad}} := \frac{\kappa+1}{p} + L - 1, \qquad c_{\mathrm{pre}} := \frac{\kappa+L}{p}. \tag{P.11}$$

Proposition P.7 (Cost per informative completion)**.** *Every adaptive-persistence learner with finite expected source cost satisfies*

$$\mathbb{E}_\sigma[\mathsf{C}_\mathsf{T}] \geq c_{\mathrm{ad}} \mathbb{E}_\sigma\left[N_{\mathrm{inf}}^{\mathrm{full}}\right]. \tag{P.12}$$

Every precommitted-persistence learner satisfies the stronger bound

$$\mathbb{E}_\sigma[\mathsf{C}_\mathsf{T}] \geq c_{\mathrm{pre}} \mathbb{E}_\sigma\left[N_{\mathrm{inf}}^{\mathrm{full}}\right]. \tag{P.13}$$

Both constants are attainable.

Proof. Every initiated task costs one episode and one reset charge. Every informatively complete task requires at least $L-1$ further episodes. Hence, pathwise,

$$\mathsf{C}_\mathsf{T} \geq (\kappa+1) N_{\mathrm{task}} + (L-1) N_{\mathrm{inf}}^{\mathrm{full}}. \tag{P.14}$$

Let $I_i = \mathbf{1}\{T_i = 1\}$. Because the event $\{N_{\mathrm{task}} \geq i\}$ is measurable before task i is drawn, Tonelli's theorem and independence give

$$\mathbb{E}_\sigma\left[\sum_{i=1}^{N_{\mathrm{task}}} I_i\right] - p\,\mathbb{E}_\sigma[N_{\mathrm{task}}] = \sum_{i\geq 1} \mathbb{E}\left[\mathbf{1}\{N_{\mathrm{task}} \geq i\}(I_i - p)\right] = 0.$$

Since an informatively complete task must be informative,

$$N_{\mathrm{inf}}^{\mathrm{full}} \leq \sum_{i=1}^{N_{\mathrm{task}}} I_i.$$

Taking expectations in equation P.14 proves equation P.12.

For a precommitted learner, let $J_i = \mathbf{1}\{M_i \geq L\}$. A task cannot be complete unless $J_i = 1$, so

$$N_{\mathrm{inf}}^{\mathrm{full}} \leq \sum_i I_i J_i.$$

Because J_i is chosen before T_i is observed,

$$\mathbb{E}[I_i J_i] = p\,\mathbb{E}[J_i].$$

Every task with $J_i = 1$ incurs at least L episodes and one reset charge; therefore

$$\mathsf{C}_\mathsf{T} \geq (\kappa + L) \sum_i J_i.$$

Taking expectations proves equation P.13.

For attainability of c_{ad}, screen every fresh task through its first episode, reset immediately on a null task, and complete every informative task. The number of screened tasks per informative completion is geometric with mean $1/p$, so the expected cost is c_{ad}. For attainability of c_{pre}, precommit to L episodes on every task and discard the parity of null tasks; the expected cost per informative completion is c_{pre}. □

P.4 Exact acquisition adaptivity gap

For an integer $K \geq 1$, let $\mathcal{V}_K^{\mathrm{ad}}$ and $\mathcal{V}_K^{\mathrm{pre}}$ denote the minimum expected source cost required to stop with at least K informatively complete tasks almost surely, over the adaptive and precommitted classes, respectively.

Theorem P.8 (Exact adaptive/precommitted separation in the screened-parity family)**.** *For every $K \geq 1$,*

$$\mathcal{V}_K^{\mathrm{ad}} = K c_{\mathrm{ad}} = K\left(\frac{\kappa+1}{p} + L - 1\right), \tag{P.15}$$

whereas

$$\mathcal{V}_K^{\mathrm{pre}} = K c_{\mathrm{pre}} = K\frac{\kappa+L}{p}. \tag{P.16}$$

Consequently, the exact acquisition adaptivity gap is

$$\mathsf{Gap}(p, L, \kappa) \coloneqq \frac{\mathcal{V}_K^{\mathrm{pre}}}{\mathcal{V}_K^{\mathrm{ad}}} = \frac{\kappa+L}{\kappa+1+p(L-1)}. \tag{P.17}$$

Writing $x = (L-1)/(\kappa+1)$,

$$\frac{1}{2}\min\left\{1+x, \frac{1}{p}\right\} \leq \mathsf{Gap}(p, L, \kappa) \leq \min\left\{1+x, \frac{1}{p}\right\}. \tag{P.18}$$

In particular, if $L - 1 \geq (\kappa+1)/p$, then $\mathsf{Gap}(p, L, \kappa) \geq 1/(2p)$.

Proof. The lower bounds in equation P.15 and equation P.16 follow immediately from Proposition P.7. The attaining procedures are the ones described at the end of its proof, stopped after the Kth informative completion. This proves the exact identities and equation P.17.

Because

$$\mathsf{Gap}(p, L, \kappa) = \frac{1+x}{1+px},$$

if $x \leq 1/p$, then $1 + px \leq 2$ and the ratio is at least $(1+x)/2$; if $x \geq 1/p$, then $1 + px \leq 2px$ and the ratio is at least $1/(2p)$. The two upper bounds follow from $1 + px \geq 1$ and $p(1+x) \leq 1 + px$. This proves equation P.18. □

The theorem isolates the acquisition effect without invoking a particular confidence level: feedback after one episode allows the learner to spend the remaining $L - 1$ episodes only on the p fraction of tasks for which continuation can reveal global information.

P.5 Matching fixed-confidence costs

Fix ε in equation P.8 and $\delta \in (0, 1/2)$. Let

$$\mathfrak{C}_\delta^{\mathrm{ad}}(\sigma)$$

and

$$\mathfrak{C}_\delta^{\mathrm{pre}}(\sigma)$$

denote the infimum expected source cost, at global model σ, over globally $(L-1, \varepsilon, \delta)$-correct adaptive and precommitted learners, respectively.

We first record the elementary sequential test used for the upper bound.

Lemma P.9 (Sequential parity test)**.** *Let $W_1, W_2, \ldots$ be i.i.d. with*

$$\mathbb{P}_\sigma(W_j = w) = \frac{1}{2}(1 + \rho\sigma w).$$

Set

$$S_m = \ell_\rho \sum_{j=1}^{m} W_j, \qquad \tau_\delta = \inf\left\{m \geq 1 : |S_m| \geq \log(1/\delta)\right\},$$

and output $\widehat{\sigma} = \operatorname{sign}(S_{\tau_\delta})$. *Then*

$$\mathbb{P}_\sigma(\widehat{\sigma} \neq \sigma) \leq \delta, \tag{P.19}$$

and

$$\frac{\operatorname{kl}(1-\delta,\delta)}{d_{\mathrm{par}}(\rho)} \leq \inf_{\widetilde{\tau},\widetilde{\sigma}} \mathbb{E}_\sigma[\widetilde{\tau}] \leq \mathbb{E}_\sigma[\tau_\delta] \leq \frac{\log(1/\delta)+\ell_\rho}{d_{\mathrm{par}}(\rho)}, \tag{P.20}$$

where the infimum is over all sequential tests having error at most δ *under both signs.*

Proof. The process $\exp(-S_m)$ is the likelihood-ratio martingale of model $-$ against model $+$ under model $+$. On the event of a wrong decision under model $+$, $S_{\tau_\delta} \leq -\log(1/\delta)$, so Ville's inequality gives probability at most δ; the other sign is symmetric. The lower bound in equation P.20 is binary data processing combined with the stopped i.i.d. KL identity

$$D_{\mathrm{KL}}\left(\mathbb{P}_+^{\widetilde{\tau}} \middle\| \mathbb{P}_-^{\widetilde{\tau}}\right) = d_{\mathrm{par}}(\rho)\mathbb{E}_+[\widetilde{\tau}].$$

For the upper bound, the increments of S_m are bounded by ℓ_ρ and have mean $d_{\mathrm{par}}(\rho)$ under model $+$. A Hoeffding bound shows that τ_δ has an exponentially decaying tail and therefore finite expectation. Wald's identity gives

$$d_{\mathrm{par}}(\rho)\mathbb{E}_+[\tau_\delta] = \mathbb{E}_+[S_{\tau_\delta}] \leq \log(1/\delta)+\ell_\rho,$$

because the overshoot is at most one increment. The argument under model $-$ is identical. This is the classical sequential probability-ratio construction (Wald, 1945; Wald & Wolfowitz, 1948). □

Theorem P.10 (Fixed-confidence adaptive/precommitted separation)**.** *For each* $\sigma \in \{-1,+1\}$*,*

$$c_{\mathrm{ad}}\frac{\operatorname{kl}(1-\delta,\delta)}{d_{\mathrm{par}}(\rho)} \leq \mathfrak{C}_\delta^{\mathrm{ad}}(\sigma) \leq c_{\mathrm{ad}}\frac{\log(1/\delta)+\ell_\rho}{d_{\mathrm{par}}(\rho)}, \tag{P.21}$$

$$c_{\mathrm{pre}}\frac{\operatorname{kl}(1-\delta,\delta)}{d_{\mathrm{par}}(\rho)} \leq \mathfrak{C}_\delta^{\mathrm{pre}}(\sigma) \leq c_{\mathrm{pre}}\frac{\log(1/\delta)+\ell_\rho}{d_{\mathrm{par}}(\rho)}. \tag{P.22}$$

Consequently,

$$\lim_{\delta\downarrow 0}\frac{\mathfrak{C}_\delta^{\mathrm{ad}}(\sigma)}{\log(1/\delta)} = \frac{c_{\mathrm{ad}}}{d_{\mathrm{par}}(\rho)}, \tag{P.23}$$

$$\lim_{\delta\downarrow 0}\frac{\mathfrak{C}_\delta^{\mathrm{pre}}(\sigma)}{\log(1/\delta)} = \frac{c_{\mathrm{pre}}}{d_{\mathrm{par}}(\rho)}, \tag{P.24}$$

and hence

$$\lim_{\delta\downarrow 0}\frac{\mathfrak{C}_\delta^{\mathrm{pre}}(\sigma)}{\mathfrak{C}_\delta^{\mathrm{ad}}(\sigma)} = \mathsf{Gap}(p,L,\kappa). \tag{P.25}$$

Proof. Because the target-strategy sets are disjoint by Proposition P.4, every globally correct source learner induces a binary test of σ with error at most δ. Binary data processing and Proposition P.6 imply

$$d_{\mathrm{par}}(\rho)\mathbb{E}_\sigma[N_{\mathrm{inf}}^{\mathrm{full}}] \geq \operatorname{kl}(1-\delta,\delta).$$

Combining this inequality with equation P.12 or equation P.13 proves the two lower bounds.

For the adaptive upper bound, screen fresh tasks sequentially. On a null task, reset after its first episode. On an informative task, query all remaining coordinates, compute the parity W_j, and feed it to the test in Lemma P.9. Let G_j be the number of freshly screened tasks from informative parity $j-1$ to informative parity j. Then G_j are i.i.d. geometric(p) random variables, independent of the parities, and the cost associated with parity j is

$$C_j^{\mathrm{ad}} = (\kappa+1)G_j + L - 1.$$

Because the parity stopping time depends only on (W_j),

$$\mathbb{E}_\sigma\left[\sum_{j=1}^{\tau_\delta} C_j^{\mathrm{ad}}\right] = c_{\mathrm{ad}}\mathbb{E}_\sigma[\tau_\delta].$$

Output the canonical target strategy $g^\star_{\widehat{\sigma}}$. Equation (P.19) gives global correctness, and equation P.20 gives the adaptive upper bound.

For the precommitted upper bound, commit to all L episodes on every fresh task, discard null-task parities, and run the same sequential test on the informative parities. The cost associated with one informative parity is

$$C_j^{\mathrm{pre}} = (\kappa + L)G_j,$$

so its mean is c_{pre}. This proves the second upper bound. The limits follow from

$$\frac{\mathrm{kl}(1-\delta, \delta)}{\log(1/\delta)} \to 1$$

and the fact that ℓ_ρ is fixed. □

P.6 Interpretation and scope

The separation is caused by one bit of feedback: the first episode reveals whether the current task can ever produce global information. Adaptive persistence spends the remaining $L - 1$ episodes only on informative tasks; precommitted persistence must decide whether to pay for those episodes before the type is observed. When informative tasks are rare and continuation is expensive, the gap approaches $1/p$.

This phenomenon is an adaptivity gap in the broad sense used in stochastic probing and stochastic packing (Dean et al., 2008; Gupta et al., 2016), but the objective and experiment differ: source tasks are not selected for their own reward, and the informative outcome is used only to recommend a strategy for an independent future task.

The theorem compares unrestricted observation-adaptive persistence with the precisely defined precommitted class of Definition P.5. It does not claim a separation against protocols allowed to condition the current task's block length on a noisy or partial type estimate; such intermediate feedback models lead to a separate round-adaptivity problem. Nor does it require computational assumptions: the separation is information theoretic and already holds in a finite horizon-one meta-MDP.

Q Assumption Audit and Counterexamples

This appendix audits the assumptions used by the main results. The purpose is not to defend every assumption as intrinsically necessary. Rather, we separate three logically different roles:

1. a *structural assumption* is necessary for the stated phenomenon, in the sense that removing it makes a principal conclusion false;
2. a *benchmark-defining assumption* specifies which aspects of a global model are target relevant and therefore which quotient parameter should be learned; and
3. a *proof-scoped assumption* is sufficient for the particular theorem or algorithmic proof given here but is not claimed to be necessary for learnability.

The distinction is important in a decision problem: two source experiments should be compared through the decisions they support, rather than through parameter recovery alone, in the sense of Blackwell's comparison of experiments (Blackwell, 1953). All counterexamples below are finite or finite dimensional, and every claimed failure can be verified exactly.

Q.1 Task persistence and the exclusivity of controlled views

The paper's basic statistical resource is that continuation preserves a single latent task while reset samples an independent latent. If this coupling is removed, the non-tensorization phenomenon disappears.

Proposition Q.1 (Resampling under continuation eliminates persistence)**.** *Modify the source protocol of Section 2 so that, conditional on the fixed global parameter θ, every source episode draws $Z_t \sim \mu_\theta$ independently, regardless of whether $C_t = \mathsf{reset}$ or $C_t = \mathsf{continue}$. Then for every admissible source*

learner and every episode t, conditional on the observed past and on the selected within-episode policy π_t, the trajectory law is

$$\overline{P}_{\theta}^{\pi_t}(\cdot) \coloneqq \int P_{\theta,z}^{\pi_t}(\cdot)\,\mu_\theta(\mathrm{d}z), \tag{Q.1}$$

and is independent of C_t. In particular, if $\overline{P}_{\theta}^{\pi} = \overline{P}_{\theta'}^{\pi}$ for every admissible one-episode policy π, then the complete stopped source-transcript laws are equal under θ and θ' for every source learner.

Proof. Under the modified protocol, Z_t is conditionally independent of the past and has law μ_θ. Integrating the conditional trajectory kernel over Z_t gives equation Q.1. The persistence action is chosen by the learner and therefore contributes the same history-dependent kernel under every global model. Induction over the observed source filtration shows that equality of all one-episode mixture kernels implies equality of every finite transcript law. Applying the same argument on the disjoint union of stopped histories gives equality at an arbitrary observed-history stopping time. □

Thus, in this modified model, continuation may alter accounting cost but cannot create cross-episode moments. The latent marginalization and the trajectory product genuinely fail to commute only when a task latent is shared across the continued episodes.

A second structural requirement is that one physical episode cannot reveal all complementary views needed to identify the task-level statistic.

Proposition Q.2 (A one-episode full-view action collapses the hierarchy)**.** *Consider the parity family of Theorem 3.1, but add a source action a_{all} whose one-episode observation is the complete latent vector $X \in \{-1,+1\}^r$. Then the pairwise persistence order of the two global models is one, and the one-episode divergence is*

$$D_{\mathrm{KL}}\left(\mu_{+}^{(r,\varrho)} \middle\| \mu_{-}^{(r,\varrho)}\right) = \varrho \log \frac{1+\varrho}{1-\varrho} = d_{\mathrm{par}}(\varrho). \tag{Q.2}$$

Proof. The action reveals X in one episode, so its observation law is exactly $\mu_{\sigma}^{(r,\varrho)}$. The two laws differ whenever $\varrho > 0$. Their likelihood ratio takes the values $(1+\varrho)/(1-\varrho)$ and its reciprocal according to the parity of X, and direct summation gives equation Q.2. Hence a one-episode experiment separates the models. □

The arbitrary-order theorem therefore expresses a controlled-access restriction, not a property of the latent distribution alone.

Q.2 Target relevance and recommendation geometry

Source identifiability does not by itself imply that source information has value for a fresh target.

Example Q.3 (A globally identifiable but target-irrelevant parameter)**.** Let $\Theta = \{0,1\}$, and suppose one source block deterministically reveals $D = \theta$. Let the target MDP be identical under the two models and have a single action with zero loss. Then the global parameter is exactly identifiable from one source block, but

$$\mathcal{L}_B(0,g) = \mathcal{L}_B(1,g) = 0 \qquad \text{for every target strategy } g.$$

Consequently, $\mathrm{Alt}_{B,\varepsilon}(\theta) = \varnothing$ and the optimal source cost is zero. Any complexity measure based on recovering θ rather than on target-incompatible decisions would assign spurious value to source data.

The converse issue is subtler: pairwise overlap of acceptable recommendation sets does not imply that one recommendation is simultaneously acceptable for a confidence set containing more than two models.

Proposition Q.4 (Pairwise overlap need not yield a common recommendation)**.** *Let $\Theta = \{1,2,3\}$ and let the target-strategy set be $\{a,b,c\}$. Define target losses by*

	a	b	c
$\theta = 1$	0	0	1
$\theta = 2$	1	0	0
$\theta = 3$	0	1	0

(Q.3)

Fix any $\varepsilon \in (0,1)$. Then every two ε-optimal strategy sets intersect, so $\mathrm{Alt}_{B,\varepsilon}(\theta) = \varnothing$ for every θ, but their total intersection is empty. If the source transcript is model independent, no randomized learner can be globally (B,ε,δ)-correct for $\delta < 1/3$.

Proof. The acceptable sets are

$$\mathcal{G}_{B,\varepsilon}(1) = \{a,b\}, \qquad \mathcal{G}_{B,\varepsilon}(2) = \{b,c\}, \qquad \mathcal{G}_{B,\varepsilon}(3) = \{a,c\}.$$

They intersect pairwise and have empty triple intersection. Since the source experiment is model independent, the output distribution is one fixed vector (p_a, p_b, p_c). Global correctness requires

$$p_c \le \delta, \qquad p_a \le \delta, \qquad p_b \le \delta.$$

Summing gives $1 \le 3\delta$. Hence no globally correct learner exists when $\delta < 1/3$. □

This is why the lower bound in Section H is stated for pairwise target-relevant alternatives, whereas the matching upper bound in Section I adds a finite decision partition with one canonical recommendation per cell. General multiple-answer identification requires answer-specific games and a sticky selection mechanism (Degenne & Koolen, 2019).

Q.3 Invariant anchors, quotient parameters, and task breadth

The reset-only nonidentifiability theorem for group actions relies on invariance of the anchor distribution. Without invariance, a marginal view may already reveal the global transformation.

Proposition Q.5 (Anchor invariance is necessary for reset-only zero information)**.** *Let $G = \mathrm{O}(1) = \{-1,+1\}$ act on $\mathbb{R}$ by multiplication, and let the source anchor be deterministic: $X = 1$ almost surely. A type-1 source view returns $Y = gX = g$. Thus one reset-only source episode identifies g exactly.*

Proof. The two possible observation laws are the point masses δ_{+1} and δ_{-1}. They are mutually singular. The source anchor law δ_1 is not invariant under the sign action, which is precisely the hypothesis missing from Theorem 5.1. □

Even with invariant anchors, the full group element need not be identifiable. The support kernel in Section J is therefore part of the estimand, not a technical artifact.

Proposition Q.6 (The support quotient is necessary, and target shift can reactivate an invisible symmetry)**.** *Let*

$$G = \{\mathrm{diag}(s_1, s_2) : s_1, s_2 \in \{-1,+1\}\}$$

act on $\mathbb{R}^2$, and let the source anchor be uniform on $\{e_1, -e_1\}$. Then

$$K_\mu = \{\mathrm{diag}(1, s_2) : s_2 \in \{-1,+1\}\}.$$

All source experiments identify at most the coset gK_μ. If the fresh target anchor is drawn from the same source law, the second sign s_2 is decision irrelevant. If instead the target anchor is e_2, the two parameters

$$g_0 = I_2, \qquad g_1 = \mathrm{diag}(1,-1)$$

have identical source laws but every predictor $\widehat{y}$ satisfies

$$\max_{j\in\{0,1\}} \mathbb{E}_{g_j} \|\widehat{y} - g_j e_2\|_2^2 \ge 1. \tag{Q.4}$$

Proof. Every source anchor lies in $\mathrm{span}(e_1)$, so changing s_2 has no effect on any source view. This gives the stated support kernel. For the shifted target, the two correct outputs are e_2 and $-e_2$. For every (possibly random) prediction $\widehat{y}$,

$$\frac{1}{2}\mathbb{E}\left[\|\widehat{y} - e_2\|_2^2 + \|\widehat{y} + e_2\|_2^2\right] = \mathbb{E}\|\widehat{y}\|_2^2 + 1 \ge 1.$$

The maximum risk is at least the average risk, proving equation Q.4. □

Thus the quotient G/K_μ is sufficient only relative to the source and target task family used to define relevance. If the target distribution moves outside the source support, a source-invisible symmetry can become control relevant.

Finally, continuation cannot replace independent anchor breadth.

Proposition Q.7 (Unlimited depth cannot replace missing anchor directions)**.** *In the noiseless orthogonal model, condition on source anchors whose span is a proper subspace $S \subsetneq \mathbb{R}^d$. Even infinitely many repeated observations of those tasks identify R only on S. For an independent $X_\star \sim \mathcal{N}(0, I_d)$,*

$$\inf_{\widehat{Y}} \sup_{R \in \mathrm{O}(d)} \mathbb{E}_R \|\widehat{Y} - RX_\star\|_2^2 \geq \dim(S^\perp). \tag{Q.5}$$

Proof. Let H be the identity on S and minus the identity on $S^\perp$. The parameters R and RH produce identical observations on all source anchors, regardless of the number of repeats. Their target outputs differ by $2RP_{S^\perp}X_\star$. The squared-loss midpoint inequality gives an average risk of at least $\mathbb{E}\|P_{S^\perp}X_\star\|_2^2 = \dim(S^\perp)$. □

Q.4 Regular noise and the singular noiseless limit

The LAN theorem in Section N assumes fixed positive effective noise. This assumption cannot be removed by taking a formal limit in the Fisher information.

Proposition Q.8 (The noiseless paired experiment is mutually singular)**.** *Let P_R^0 be the noiseless paired-view law of (X, RX) with $X \sim \mathcal{N}(0, I_d)$. For every distinct $R, S \in \mathrm{O}(d)$,*

$$P_R^0 \perp P_S^0, \qquad D_{\mathrm{KL}}(P_R^0 \| P_S^0) = +\infty. \tag{Q.6}$$

Consequently, the positive-noise DQM and finite-information LAN expansion do not extend to $\Delta_\tau = 0$.

Proof. The support of P_R^0 is the graph $\mathcal{G}_R = \{(x, Rx) : x \in \mathbb{R}^d\}$. Its intersection with $\mathcal{G}_S$ corresponds to $\ker(R - S)$, which is a proper linear subspace because $R \neq S$. A nondegenerate Gaussian vector assigns zero probability to every proper linear subspace. Hence $P_R^0(\mathcal{G}_R \setminus \mathcal{G}_S) = 1$ and $P_S^0(\mathcal{G}_R) = 0$. □

The divergence of the regular Fisher information as $\Delta_\tau \downarrow 0$ therefore signals a change of experiment, not ordinary increasingly precise LAN. Exact breadth and stabilizer arguments, not regular asymptotic theory, govern the noiseless model.

Q.5 Scope of the finite-block and bounded-stopping characteristic-time theorems

Several assumptions of Sections H, I and T are proof-scoped rather than structurally necessary. The following examples delimit what has and has not been proved.

Finite KL excludes exact one-block identification. Let $\Theta = \{0, 1\}$ and let one block return $D = \theta$ deterministically. Then the block laws are mutually singular and one block identifies the model with zero error. The fixed-confidence cost is one for every δ, rather than a nonzero constant times $\log(1/\delta)$. Thus the finite-KL assumption selects the nondegenerate logarithmic characteristic-time regime; without it, the convention $1/(+\infty) = 0$ captures only the leading $\log(1/\delta)$ coefficient, not the finite additive cost.

The observational quotient is the decision-relevant parameter. If two original models induce the same law under every depth-L experiment tree, no source learner can distinguish them. The flagship theorem therefore operates directly on the observational quotient and requires each quotient class to lie in one disjoint target-answer cell. Global identification of original representatives is neither assumed nor charged. The precursor fixed-block results in Section I retain a stronger globally identifiable formulation only for historical comparison.

Bounded log-likelihood ratios are a concentration device. In the Gaussian paired-view experiment,

$$\log \frac{p_R(y_0, y_1)}{p_S(y_0, y_1)} = \frac{1}{\Delta_\tau} y_1^\top (R - S) y_0, \tag{Q.7}$$

which is unbounded whenever $R \neq S$. Nevertheless, Sections L to N prove finite-sample minimaxity and asymptotic efficiency. Hence Assumption I.4 is not asserted to be necessary for an optimal persistence algorithm; it enables the direct Azuma–Hoeffding and stretched-exponential expected-stopping proof. Time-uniform mixture martingales provide a route to unbounded exponential-family likelihoods (Kaufmann & Koolen, 2021).

Finiteness is a theorem scope, not a learnability boundary. The orthogonal model has a continuous parameter space and is learnable at the rate established in Theorem 5.2. The finite Θ and finite $\mathcal{E}$ assumptions are used in Section 4 to obtain a finite minimax game, a finite union bound, and eventually exact MLE selection. Continuous-model characteristic times require localization, measurable design selection, and a different confidence construction.

Bounded complete trees solve the finite endogenous game. Appendix T is a precursor theorem for a supplied finite menu of bounded stopping policies. Appendices D –E remove that menu restriction under finite primitive alphabets, finite physical horizon, and maximum same-task depth L: they optimize over every deterministic depth-L experiment tree and, by randomized-to-mixed equivalence, over every randomized history-dependent bounded experiment. What remains open is the noncompact problem with unbounded task duration, continuous policy classes without discretization, or function approximation.

Q.6 Target controllability and the weight of target information

The lower transfer of matrix-estimation error into control loss requires the target system to be sensitive to every goal direction under consideration.

Proposition Q.9 (Unreachable directions can be statistically hard but control irrelevant)**.** *Let $d = 2$, $H = 1$, scalar control u, endpoint matrix $\mathsf{C}_H = e_1$, and control penalty $\Lambda = [1]$. Then*

$$W_H = \frac{1}{2} e_1 e_1^\top = \begin{pmatrix} 1/2 & 0 \\ 0 & 0 \end{pmatrix}.$$

For

$$R_0 = I_2, \qquad R_1 = \mathrm{diag}(1, -1),$$

the two transformations are distinct and can be difficult to distinguish from source data, but they induce the same oracle target control for every anchor $X_\star$. Therefore no positive lower bound proportional to the unweighted Frobenius error can hold when W_H is singular.

Proof. For $g = (g_1, g_2)$, the unique optimal control is

$$u^\star(g) = (\mathsf{C}_H^\top \mathsf{C}_H + 1)^{-1} \mathsf{C}_H^\top g = \frac{g_1}{2}.$$

The goals $R_0 X_\star$ and $R_1 X_\star$ have the same first coordinate, so their oracle controls coincide. Their difference lies in $\ker(W_H) = \mathrm{span}(e_2)$ and produces no excess control loss. □

This is why Corollary 6.2 weights the lower source term by $\lambda_{\min}(W_H)$ and assumes full row rank only when claiming a uniformly positive transfer coefficient.

The factor α_B in the source term is equally necessary. When $B = 0$,

$$\alpha_B = 0, \qquad M_B = 0,$$

so uncertainty about R has no effect on the optimal target controller: the fresh anchor itself is completely unobserved. Any formula that adds the full source matrix-estimation risk without the factor α_B is therefore false at $B = 0$.

Q.7 Screening assumptions in the adaptive-persistence gap

The adaptivity gap of Section P is created by information about the current task type arriving before the continuation decision.

Proposition Q.10 (Without early screening, task-type-directed continuation is impossible)**.** *Modify the screened-parity family so that the task type T is not revealed in the initial state. Suppose every history that has revealed fewer than all L coordinates has the same conditional law under $T = 0$ and $T = 1$. Then, for every stopping or continuation decision made before task completion, the decision is independent of T conditional on the past completed tasks. In particular, the screen-then-continue strategy used in Section P is not implementable.*

Proof. Proceed by induction over the within-task filtration. By assumption, the initial observed history has the same law under both task types. If the conditional laws agree up to a proper coordinate history, then any next action or stopping decision is generated by the same history-measurable algorithmic kernel under both types. The next incomplete observation also has the same conditional law by hypothesis. Thus every decision made before completion has the same conditional distribution under $T = 0$ and $T = 1$. □

Conversely, if the comparator is allowed to observe T before choosing its block length, it can implement the adaptive policy exactly, so the gap in equation P.17 collapses to one. The comparator definition and the exact-screening assumption are therefore part of the theorem's content, not merely expository choices.

Q.8 Assumption-to-proof-role map

Table 4: Role of the assumptions used by the flagship theorems. "Scope" means that the assumption makes the complete experiment-tree problem finite; it is not claimed to be a statistical necessity.

Assumption	Role in the proof	Status	Failure when removed
Shared latent under CONTINUE	Creates non-tensorized cross-view laws	Structural	Episodes tensorize
Fresh latent under RESET	Creates independent anchors	Structural	No new directions
Bounded task depth L	Makes the complete tree class finite	Scope	Infinite stopping problem
Finite Θ and answers	Finite GLR union and finite game	Technical	Localization required
Common domination	Likelihood-ratio factorization	Technical	Singular testing regime
Conditional exponential moments	Expected-cost concentration	Technical	Heavy-tail tools required
Matched source/target quotient	Makes source identification target sufficient	Benchmark	Invisible target directions
Target controllability	Transfers goal error to control loss	Structural	Unreachable directions have zero value

Q.9 Summary of the audit

Shared latent under continuation (structural). Continued episodes otherwise tensorize and create no cross-view information; see Proposition Q.1.

Exclusive views across episodes (structural). A one-episode full-view action collapses arbitrary persistence order to one; see Proposition Q.2.

Target-relevant alternatives (benchmark defining). A source-identifiable but target-irrelevant parameter requires no source learning.

Decision partition (upper-theorem structure). Pairwise overlap can coexist with the absence of any globally valid recommendation; see Proposition Q.4.

Invariant source anchors (structural for reset-only impossibility). Without invariance, a one-view marginal can identify the group element; see Proposition Q.5.

Table 5: Representative access models. "Endogenous persistence" means that the learner chooses whether the next source episode shares the current task latent or draws a fresh one. The entries summarize the representative works discussed below; individual papers within a broad category may impose additional assumptions.

Literature	Shared latent/group	Endogenous persistence	Controlled trajectories	Fresh target	Exact decision limit
Grouped-mixture identification	yes	no	no	no	identifiability
Task/sample meta-learning	yes	generally no	generally no	yes	task/sample rates
Multitask and transfer RL	shared structure	generally no	yes	sometimes	model specific
Latent MDPs and epistemic POMDPs	yes	no	yes	generally no	learning/control rates
Active testing and Track-and-Stop	fixed hypothesis	no task latent	experiment selection	terminal decision	characteristic time
Reservoir and infinite-arm bandits	arm latent	new/reuse arm	bandit pulls	no	source-arm identification
This work	yes	yes	yes	yes	bounded-stopping characteristic time

Support quotient and matched target family (benchmark defining). A source-invisible symmetry can become target relevant under distribution shift; see Proposition Q.6.

Independent anchor breadth (structural). Infinite continuation cannot identify unobserved directions; see Proposition Q.7.

Positive fixed observation noise (regularity for LAN). The noiseless laws are mutually singular; see Proposition Q.8.

Finite models and bounded log ratios (proof scoped). These assumptions simplify design tracking and expected-stopping analysis; they are not claimed to be general learnability boundaries.

Target controllability. This assumption is structural for isotropic transfer: unreachable goal directions have zero control value; see Proposition Q.9.

Observable current-task type (structural for the adaptive gap). Without early type information, continuation cannot be conditioned on task informativeness; see Proposition Q.10.

The audit yields two general lessons. First, the paper's impossibility results are decision theoretic: an unidentified direction matters only if it changes an admissible fresh-target decision. Second, several assumptions in the exact characteristic-time theorem delimit the present proof technique rather than the underlying phenomenon. They are stated explicitly so that future extensions—continuous model classes, unbounded likelihood ratios, multiple correct answers, and adaptive within-task stopping—do not appear as unproved consequences of the current theorem.

R Extended Related Work

Purpose of this appendix. The relevant literature spans meta-learning, multitask and transfer reinforcement learning, latent-state control, grouped-observation identifiability, active experimental design, and sequential testing. A comparison based only on topic labels is therefore misleading: several neighboring models use words such as "task," "episode," or "adaptation" for statistically different access protocols. We organize the discussion by *experiment structure*. In particular, we distinguish whether (i) several observations share a task-specific latent variable, (ii) the learner chooses whether that latent variable persists, (iii) observations are generated by controlled MDP interaction, and (iv) performance is evaluated on an independent fresh task. The comparison below is based on primary conference papers and preprints available through September 2026. It is not intended as a claim that no concurrent or differently formulated work exists.

R.1 Tasks, observations per task, and meta-learning

A large meta-learning literature separates the number of sampled tasks from the number of observations available within each task. For example, Cioba et al. (2022) optimize the allocation of a fixed passive sample budget across tasks in a mixed-linear-regression model, while Aliakbarpour et al. (2024) prove sharp few-sample-per-task phenomena for representation learning, including regimes in which changing the number of observations per task changes learnability. These works establish that task breadth and within-task depth need not collapse to total sample size. Their grouping protocol, however, is supplied by the data-generating process or selected before observing the current task. The learner does not control, after seeing trajectories from a task, whether the next controlled trajectory shares that same task latent.

The distinction is not terminological. In our source protocol, the action CONTINUE changes the conditional law of all future trajectories by preserving a latent random effect, whereas RESET draws a new random effect. Therefore the learner chooses the dependence structure of its future data. The arbitrary-order result of Theorem 3.1 shows that this difference can be qualitative: every amount of data collected with subcritical persistence can have exactly zero target-relevant information. Appendix P further compares observation-adaptive persistence with schedules that may adapt across tasks but must commit to the current task's block length before its type is observed. This feedback distinction is absent from the passive task/sample allocation models above.

R.2 MULTITASK REINFORCEMENT LEARNING, PRETRAINING, AND REPRESENTATION TRANSFER

Early multitask RL theory studies a sequence of MDPs sampled from an unknown finite family and proves that information transferred across tasks can reduce per-task exploration complexity (Brunskill & Li, 2013). Lifelong RL work also studies discovery of previously unseen MDP types across a task sequence (Brunskill & Li, 2015). In these models, the task sequence is exogenous: the learner acts within the current MDP, but does not decide whether the next episode remains in the current latent MDP or starts an independent task drawn from the population.

A more recent line studies common representations across simultaneously available or upstream tasks. Shared linear representations yield improved multitask bandit and RL rates (Hu et al., 2021); low-rank MDP theory gives provable source-to-target benefits (Cheng et al., 2022; Agarwal et al., 2023); and general function-class analyses extend the idea beyond a known linear feature map (Lu et al., 2022). Offline multitask representation learning similarly estimates a common low-rank structure from fixed upstream datasets before solving a downstream task (Ishfaq et al., 2024). Task diversity can also change the difficulty of exploration: Xu et al. (2024) show that sufficiently diverse multitask training can make otherwise-myopic exploration statistically efficient.

The closest task-selection work permits active allocation among a supplied collection of source tasks. Chen et al. (2022) and Wang et al. (2023) adaptively choose how many samples to obtain from known source tasks according to their relevance to a target. This is an important precursor to our design perspective, but the available source task identities are already instantiated. Sampling task i again does not pose a choice between retaining an unobserved latent draw and resampling a new task from a metadistribution. In our model, the task-specific anchor itself is unobserved and random; continuation creates a correspondence between controlled views of that anchor, while resetting creates a new anchor. The group-action lower bounds in Theorems 5.1 and 5.2 quantify this two-resource geometry.

RL pretraining theory asks when source interaction can help zero-shot or few-shot target performance and also identifies limitations and hardness (Ye et al., 2023). Finite-training-task meta-RL estimates a task distribution from sampled training tasks and compares the learned behavior to the Bayes-optimal policy for a fresh task (Rimon et al., 2022). These objectives are closely aligned with our fresh-target loss. The difference is the source experiment: prior theory takes the training-task sample and its within-task data budget as part of the access model, whereas our learner selects reset versus continuation online and we characterize the minimax value of that selection.

R.3 LATENT MDPS, PARTIAL OBSERVABILITY, AND RL GENERALIZATION

Latent MDPs model an unobserved MDP identity that is drawn at the beginning of an interaction epoch and remains fixed during that epoch. Recent work gives sample-efficient online guarantees without additional distributional structure (Kwon et al., 2024), while separated latent MDPs admit sharper statistical and computational analyses under separation conditions (Chen et al., 2024). These models share with ours the fact that several controlled observations may depend on one latent MDP. However, the duration of latent persistence is exogenous. Their primary objective is to learn or control the current latent-MDP process, rather than to choose when to terminate one source latent and draw another in order to perform well on an independent future task.

Generalization across unseen environments can itself induce partial observability. Ghosh et al. (2021) formalize this through an epistemic POMDP, emphasizing that a physically observed state need not reveal the unseen test condition. Structured POMDP theory, such as short-term-memory RL, provides sample-complexity guarantees when a bounded observation window is sufficient for control

(Efroni et al., 2022). These results concern the memory or belief required to infer a latent condition during interaction. By contrast, our source learner can deliberately end one latent task and request an independent one. Task persistence is therefore an acquisition action, not only a hidden-state property to be filtered.

R.4 Grouped observations, multiview learning, and identifiability

The statistical mechanism behind continuation has a classical analogue in grouped-observation mixture models. When several observations are known to share a latent component, their joint law can identify mixtures that cannot be identified from the one-observation marginal; Vandermeulen & Scott (2019) give sharp operator-theoretic identifiability results in a nonparametric setting. Our parity constructions deliberately use the same high-level mechanism. The additional content is that each view is selected by a controlled policy, the learner chooses whether observations share a task latent, and the loss is a target-control decision rather than mixture reconstruction. The filtration-level statements in Appendices C and P prove that the cancellation survives adaptive policy and reset choices.

Representation identifiability in RL is also an active topic. Reizinger et al. (2026) show that policy diversity and a contrastive successor-feature parameterization can identify environment features up to a linear transformation. Concurrent work on controlled world models derives policy-dependent conditions under which latent states and controlled dynamics are identified up to symmetry (Zhang et al., 2026). These results analyze whether a representation or transition model is identifiable under a specified distribution of policies. Our group-action analysis instead makes the sampling protocol itself the object of study: invariant one-view marginals contain zero information, paired same-task views identify an action modulo the support kernel, and independent resets shrink the residual stabilizer. The exact KL identity equation 5.1 and Theorem 5.2 quantify the statistical cost of those two acquisition resources.

R.5 Sequential experiment design, characteristic times, and interactive complexity

Chernoff's sequential design of experiments and active sequential hypothesis testing choose observation modes adaptively to maximize discrimination among hypotheses (Chernoff, 1959; Naghshvar & Javidi, 2013). Fixed-confidence pure exploration similarly derives instance-dependent information programs; Track-and-Stop attains the corresponding characteristic time for best-arm identification (Kaufmann et al., 2016; Garivier & Kaufmann, 2016). Mixture martingales supply time-uniform generalized likelihood-ratio guarantees in broader adaptive tests (Kaufmann & Koolen, 2021), while multiple-answer problems require answer-specific allocation and sticky selection (Degenne & Koolen, 2019).

Section 4 deliberately builds on this program. A supplied finite experiment menu is a useful precursor abstraction. In the finite bounded source interface studied here, however, that restriction is removed: every randomized history-dependent bounded experiment is outcome-equivalent to a model-independent mixture over the complete deterministic tree class, and Bellman pricing constructs the best omitted tree endogenously. The lower and upper proofs then use familiar change-of-measure and Track-and-Stop ideas on the induced stopped-tree laws. Recent cost-aware sequential-testing work studies exogenously supplied sensing actions with heterogeneous or random acquisition costs and identifies the ratio of expected information to expected cost as the governing design principle (Vershinin et al., 2025). We do not claim novelty for that normalization or for the generic sequential-testing proof tools. Here one statistical action is instead an internally controlled, observation-stopped multi-episode experiment whose trajectories share a sampled task latent, and RESET replaces that latent. The new object is the stopped tree law after marginalizing the shared latent. The reset/continue decision determines which block laws are available and can make their information jump from exactly zero to positive at an arbitrary persistence order. The cost normalization also charges both the episodes within a task and the reset that creates the next independent task.

Blackwell's comparison theory orders experiments by the decision problems they support (Blackwell, 1953). A recent preprint studies orders for repeatedly sampled experiments under small observation costs and large stopping budgets, characterized through pairwise KL rates (Li & Liu, 2026). Those results compare exogenously specified, repeatable experiments. Our analysis instead constructs the

experiment endogenously: continuing preserves a nuisance latent, resetting replaces it, and the target objective is defined on a later independently drawn task. After this non-tensorized experiment class is made explicit, the Persistence–Chernoff program gives its target-decision characteristic time.

The Decision–Estimation Coefficient (DEC) provides a broad necessary and sufficient complexity framework for interactive decision making (Foster et al., 2021). The persistent meta-RL protocol can in principle be encoded in such a general decision problem by augmenting the state with the source history and task-latent experiment. We therefore do not claim a new universal alternative to DEC. Our contribution is structural and evaluative: arbitrary persistence orders, exact task-completion information identities, quotient identifiability under group actions, matching orthogonal minimax rates, and their fresh-target control consequences.

Recent work on optimal RL data acquisition derives a large-deviations criterion and adaptive acquisition policy for selecting informative state–action data in an infinite-horizon MDP (Hu et al., 2026). Its controlled object is the occupancy of one RL environment. In contrast, our source learner additionally controls whether successive trajectories share a random task latent, and must balance repeated views of one task against fresh task breadth.

R.6 Reservoir bandits, lifelong learning, and adaptivity gaps

The protocol-level neighbor closest to reset versus continuation is the infinite-armed or reservoir bandit. There, an algorithm may draw a new arm from a reservoir or pull a previously sampled arm; fixed-confidence work characterizes the cost of identifying a near-optimal source arm (Aziz et al., 2018). The distinction is the objective. An arm is a candidate decision object, and the algorithm seeks a high-reward arm among those sampled. In our model, source tasks are never deployment candidates. They are exchangeable experiments used to learn a global mechanism, and performance is measured on an independent target task drawn after source learning. More directly, Wang et al. (2022) study online estimation of reservoir-distribution functionals when the learner chooses a new or existing arm and show that this feedback can change optimal sample complexity. Their within-arm observation is repeated reward sampling and the objective is a functional of the source-arm population; our task episodes may use different policies to reveal complementary views, and the decision is evaluated on an independent fresh task.

Lifelong RL likewise transfers information across a sequence of related MDPs. The online coupon-collector formulation of Brunskill & Li (2015), for example, controls cross-task discovery when the sequence is drawn from a finite set. Again, task arrivals are exogenous. The learner cannot preserve a current hidden task for another episode or discard it in favor of an independent metadistribution draw.

Appendix P is related in spirit to adaptivity-gap results in stochastic knapsack and stochastic probing (Dean et al., 2008; Gupta et al., 2016): an adaptive policy conditions future resource use on observed task type, whereas a precommitted policy does not. Our separation is decision-theoretic rather than reward-collection based. The screened source tasks contribute only likelihood information about a global parameter, and the lower bound is converted into the fixed-confidence complexity of a fresh-target control decision.

R.7 Invariant statistical models and orthogonal alignment

The mathematical ingredients of the structured example are classical. Statistical models with group invariance and Haar priors are treated by Eaton (1989); Diaconis (1988); stabilizers and bases are standard in group-action theory (Dixon & Mortimer, 1996); and orthogonal Procrustes estimation and polar factors originate in the work of Schönemann (1966) and Higham (1986). Statistical work on generalized orthogonal Procrustes instead estimates an underlying point cloud from many noisy copies transformed by sample-specific unknown rotations or reflections (Pumir et al., 2021). In our experiment, the unknown object is one globally shared transformation, every task draws a fresh latent anchor, the learner controls whether the two corresponding views are paired, and evaluation is on a fresh target anchor. We do not claim novelty for the group or Procrustes tools themselves. Their role here is to make the reset–continue tradeoff exactly solvable. Resetting supplies independent anchors whose stabilizer intersection shrinks with task breadth; continuation supplies the paired correspondence needed to reveal the group action at all. This leads to the exact residual-coset

posterior, the breadth term $(d-n)_+$, the noisy rate $\Delta_\tau p_d/n$, and the LQ-control consequence in Corollary 6.2.

R.8 Precise boundary of the contribution

The paper does *not* claim novelty for any of the following in isolation: allocating samples across tasks; exploiting grouped observations; learning a shared RL representation; selecting sensing actions through a Chernoff game; Track-and-Stop; general interactive complexity through DEC; compact-group identifiability; orthogonal Procrustes estimation; or adaptivity gaps in stochastic probing. The theorem package concerns their conjunction under a specific source-to-target access model. Concretely, its three headline contributions are:

1. an arbitrary hierarchy of pairwise target-relevant persistence orders valid under adaptive policies, resets, and source stopping;
2. an exact cost-normalized lower bound and matching expected-cost upper bound over the complete finite class of all depth-L experiment trees, equivalently all randomized history-dependent bounded persistence strategies, under a disjoint single-answer quotient partition and explicit regularity assumptions; and
3. reset–continue complementarity in structured group-action models, including orthogonal estimation, a multi-step routing realization, and fresh-target control consequences.

The primary-source audit above did not identify an earlier result proving this combination of endogenous task persistence, controlled within-task views, matching source information limits, and independent fresh-target control. This statement is intentionally narrower than a claim that no related or concurrent formulation exists; the comparison should be re-audited at the final submission date.

S A Finite-Horizon Routing Realization

This appendix gives an exact finite-state realization in which the diagnostic views are produced by multi-step state transitions and the fresh-target decision is a sequential routing policy. The construction shows that the reset–continue phenomenon is not tied to a one-step coordinate-query interface.

S.1 Source and target processes

Fix an integer $m \geq 2$ and let

$$q \coloneqq \max\{2, \lceil \log_2 m \rceil\}.$$

Choose an injective binary code

$$c : [m] \longrightarrow \{0,1\}^q, \qquad c(x) = (c_1(x), \ldots, c_q(x)).$$

The global parameter is a permutation $g \in \mathfrak{S}_m$, and every fresh task draws an anchor

$$X \sim \mathrm{Unif}([m]).$$

Write

$$T_0^g(x) = x, \qquad T_1^g(x) = g(x).$$

Source interface. The source process has horizon $H_{\mathrm{src}} = q$. At stage one the observed state is a root $s_\circ$, and the action $b \in \{0,1\}$ selects one of two disjoint diagnostic corridors. Conditional on the task anchor x, the next q observed states reveal, one bit at a time, the codeword $c(T_b^g(x))$. Formally, after choosing b the stage-$(h+1)$ state is

$$s_{b,h,c_{1:h}(T_b^g(x))}, \qquad h \in [q],$$

where $c_{1:h}$ denotes the length-h prefix. After the first decision only a dummy action is available, all rewards are zero, and the physical episode terminates after the complete codeword is observed. Let π_b denote the policy selecting corridor b. Since c is injective, one trajectory under π_b identifies $T_b^g(x)$ exactly. The two corridors are disjoint, so no single trajectory reveals both x and $g(x)$.

A source reset draws an independent anchor. A continuation restarts the physical state at $s_\circ$ while preserving the same anchor. Hence running π_0 and π_1 on one continued task reveals the correspondence $(x, g(x))$.

Target interface. A fresh target draws $X_\star \sim \mathrm{Unif}([m])$ and reveals $X_\star$ in its initial state. The target has horizon $H_{\mathrm{tgt}} = q$. At stage h, the learner selects a bit $A_h \in \{0, 1\}$, and the observed state records the chosen prefix $A_{1:h}$. The terminal reward is

$$R_{\mathrm{tgt}} = \mathbf{1}\{A_{1:q} = c(g(X_\star))\}.$$

Thus every label $y \in [m]$ corresponds to a unique q-step route, and the oracle target value is one. Learning $g(X_\star)$ changes the entire target action sequence.

S.2 Exact information and target value

Suppose the learner observes n paired source tasks, meaning that both π_0 and π_1 are executed on each task before reset. Write

$$\mathsf{D}_n = \{(X_i, g(X_i)) : i \in [n]\}, \qquad \mathcal{U}_n = \{X_i : i \in [n]\}, \qquad K_n = |\mathcal{U}_n|.$$

For a realized data set d_n, let

$$\mathcal{C}(d_n) = \{h \in \mathfrak{S}_m : h(x) = g(x) \text{ for every } x \in \mathcal{U}_n\}$$

be its consistency set. The notation on the right uses the observed paired values in d_n, so $\mathcal{C}(d_n)$ is data measurable.

Theorem S.1 (Sequential routing realization of reset–continue complementarity)**.** *The finite-horizon source and target processes above satisfy the following.*

1. *For every possibly randomized source learner that resets before every episode, including one with adaptive corridor choices and a random stopping time, the complete stopped transcript law is independent of $g \in \mathfrak{S}_m$.*
2. *Two continued trajectories on the same task under π_0 and π_1 reveal $(X, g(X))$ exactly. Consequently, after n paired tasks the global parameter is known on $\mathcal{U}_n$, and every bijection from $[m] \setminus \mathcal{U}_n$ to $[m] \setminus g(\mathcal{U}_n)$ extends to a permutation in $\mathcal{C}(\mathsf{D}_n)$.*
3. *Conditional on the paired source data, the minimax expected target regret is exactly*

$$\inf_{\pi^{\mathrm{tgt}}} \sup_{h \in \mathcal{C}(\mathsf{D}_n)} \mathbb{E}_h\left[1 - R_{\mathrm{tgt}}^{\pi^{\mathrm{tgt}}} \middle| \mathsf{D}_n\right] = \frac{(m - K_n - 1)_+}{m}. \tag{S.1}$$

The infimum ranges over all randomized target policies measurable with respect to $(\mathsf{D}_n, X_\star)$ and an independent private random seed.

Proof. For part 1, condition on the observed history before a reset-only episode and on the learner's selected corridor b. The new anchor is uniform on $[m]$. If $b = 0$, the revealed label is X; if $b = 1$, it is $g(X)$. Both are uniform on $[m]$, and the entire q-step state trajectory is a deterministic function of that label. Its conditional law is therefore independent of g. Induction over episodes, including the learner's private randomization and stopping rule, proves equality of all stopped transcript laws.

Part 2 follows from injectivity of c: the two continued trajectories reveal x and $g(x)$. Distinct observed anchors have distinct images because g is a permutation. On the unobserved domain and unused codomain, every bijection can be combined with the observed pairs to form a permutation of $[m]$.

For part 3, fix a realized source data set and let $k = K_n$. If the fresh anchor $X_\star$ belongs to $\mathcal{U}_n$, its image is known and the learner can follow the unique route $c(g(X_\star))$, obtaining reward one. Let $U = [m] \setminus \mathcal{U}_n$ and $V = [m] \setminus g(\mathcal{U}_n)$, so $|U| = |V| = r = m - k$. A randomized target policy induces, for each $x \in U$, probabilities $p_{x,y}$ over $y \in V$. Every bijection $\sigma : U \to V$ extends to a consistent global permutation. If Σ is a uniformly random bijection, then

$$\mathbb{E}_\Sigma\left[\frac{1}{r}\sum_{x \in U} p_{x,\Sigma(x)}\right] = \frac{1}{r^2}\sum_{x \in U}\sum_{y \in V} p_{x,y} = \frac{1}{r}.$$

Hence some consistent permutation has average success at most $1/r$ over an unseen uniform target anchor. Conversely, choosing uniformly from V for every unseen x attains success exactly $1/r$ for

every consistent permutation. Thus the conditional minimax regret on the unseen portion is $1 - 1/r$. Since $X_\star$ is uniform and independent of the source data,

$$\mathbb{P}(X_\star \notin \mathcal{U}_n \mid \mathsf{D}_n) = \frac{m-k}{m}.$$

Multiplying gives

$$\frac{m-k}{m}\left(1 - \frac{1}{m-k}\right) = \frac{m-k-1}{m}$$

when $k < m$; when $k = m$, the regret is zero. This is exactly equation S.1. □

Corollary S.2 (Expected target regret after random resets)**.** *For n independent paired source tasks, the minimax expected target regret is*

$$\frac{1}{m}\mathbb{E}\left[(m - K_n - 1)_+\right],$$

where K_n is the number of distinct values among n independent uniform samples from $[m]$. Reset-only source access leaves the target minimax regret at $(m-1)/m$ for every source budget.

Proof. Take expectations in equation S.1. Under reset-only access, part 1 of Theorem S.1 implies that the source transcript gives no information about g. Equivalently, the consistency set remains all of $\mathfrak{S}_m$, which is the case $K_n = 0$ in equation S.1. □

The construction has source and target horizons at least two. Information is carried by policy-dependent state trajectories, the two diagnostic routes are mutually exclusive within one episode, continuation is required to align the routes for one latent anchor, and resets are required to cover new anchors.

T Bounded Observation-Adaptive Persistent Experiments

This appendix is a reusable precursor to the complete-tree theorem. It extends the deterministic-block theory of Sections H and I to a supplied finite menu of *bounded observation-adaptive stopping experiments*. Each experiment begins with a reset, chooses episodic policies adaptively inside one task, and terminates at a bounded observation-dependent stopping time, producing a random but bounded cost. Appendices D –E instantiate this result with the complete finite class of all depth-L deterministic experiment trees and prove equivalence to every randomized history-dependent experiment that stops by depth L.

T.1 Finite menu of bounded stopping experiments

Let Θ be a finite global-model class and let $\mathcal{E}_{\mathrm{stop}}$ be a finite, nonempty experiment menu. An experiment $e \in \mathcal{E}_{\mathrm{stop}}$ consists of:

1. a reset to an independent source task;
2. an internally adaptive sequence of episodic policies;
3. a within-task stopping time τ_e satisfying
$$1 \le \tau_e \le L_e < \infty \qquad \text{almost surely under every } \theta \in \Theta; \tag{T.1}$$
4. the complete stopped transcript D_e, including all policies, trajectories, and the stopping time;
5. the random source cost
$$C_e := \kappa + \tau_e. \tag{T.2}$$

We include C_e as a coordinate of the block outcome, although it is already measurable from D_e. Under model θ, write

$$Q_\theta^e := \mathcal{L}_\theta(D_e, C_e), \qquad c_\theta(e) := \mathbb{E}_\theta[C_e]. \tag{T.3}$$

Let

$$c_{\min} := 1 + \kappa, \qquad c_{\max} := \kappa + \max_e L_e. \tag{T.4}$$

Then $c_{\min} \le C_e \le c_{\max}$ almost surely and $c_{\min} \le c_\theta(e) \le c_{\max}$.

Assumption T.1 (Bounded-stopping statistical regularity)**.** For every $e \in \mathcal{E}_{\mathrm{stop}}$, the family $\{Q_\theta^e : \theta \in \Theta\}$ is dominated by a common sigma-finite measure ν_e with density q_θ^e. Every ordered divergence

$$d_e(\theta, \theta') \coloneqq D_{\mathrm{KL}}\left(Q_\theta^e \| Q_{\theta'}^e\right) \tag{T.5}$$

is finite. For the upper theorem, we additionally assume a finite decision partition as in Assumption I.1, global separation of every distinct model pair by at least one experiment, and a constant $L < \infty$ such that

$$\left|\log \frac{q_\theta^e(x)}{q_{\theta'}^e(x)}\right| \le L \tag{T.6}$$

for all ordered model pairs, experiments, and ν_e-almost every outcome. Exact likelihood, global-MLE, and design-oracle access are assumed.

A blockwise learner selects $E_k \in \mathcal{E}_{\mathrm{stop}}$ predictably after the first $k-1$ completed blocks, receives $X_k = (D_k, C_k) \sim Q_\theta^{E_k}$ conditionally on the past, and may stop at a block stopping time T. Its realized total source cost is

$$\mathsf{C}_T^{\mathrm{stop}} \coloneqq \sum_{k=1}^{T} C_k. \tag{T.7}$$

T.2 Information per expected source cost

For $p \in \Delta(\mathcal{E}_{\mathrm{stop}})$ define

$$I_\theta(p, \theta') \coloneqq \sum_e p_e d_e(\theta, \theta'), \tag{T.8}$$

$$\bar{c}_\theta(p) \coloneqq \sum_e p_e c_\theta(e). \tag{T.9}$$

The bounded-stopping Persistence–Chernoff information is

$$\Gamma^\star_{\mathrm{stop}}(\theta) \coloneqq \sup_{p \in \Delta(\mathcal{E}_{\mathrm{stop}})} \inf_{\theta' \in \mathrm{Alt}_{B,\varepsilon}(\theta)} \frac{I_\theta(p, \theta')}{\bar{c}_\theta(p)}. \tag{T.10}$$

Equivalently, letting

$$\mathcal{W}_\theta \coloneqq \left\{ w \in \mathbb{R}_+^{\mathcal{E}_{\mathrm{stop}}} : \sum_e w_e c_\theta(e) = 1 \right\}, \tag{T.11}$$

we have

$$\Gamma^\star_{\mathrm{stop}}(\theta) = \sup_{w \in \mathcal{W}_\theta} \inf_{\theta' \in \mathrm{Alt}_{B,\varepsilon}(\theta)} \sum_e w_e d_e(\theta, \theta'). \tag{T.12}$$

The two forms are related by $w_e = p_e / \bar{c}_\theta(p)$.

Lemma T.2 (Expected-cost decomposition)**.** *Let* $N_e(T) = \sum_{k=1}^{T} \mathbf{1}\{E_k = e\}$. *If* $\mathbb{E}_\theta[\mathsf{C}_T^{\mathrm{stop}}] < \infty$, *then*

$$\mathbb{E}_\theta\left[\mathsf{C}_T^{\mathrm{stop}}\right] = \sum_e \mathbb{E}_\theta[N_e(T)] c_\theta(e). \tag{T.13}$$

Proof. Because $C_k \ge c_{\min}$, finite expected total cost implies $\mathbb{E}_\theta[T] < \infty$. Predictability of E_k and the fresh-task block law give

$$\mathbb{E}_\theta[C_k \mid \mathcal{B}_{k-1}, E_k] = c_\theta(E_k).$$

Since all summands are nonnegative, Tonelli's theorem yields

$$\begin{aligned} \mathbb{E}_\theta\left[\sum_{k=1}^{T} C_k\right] &= \sum_{k \ge 1} \mathbb{E}_\theta\left[\mathbf{1}\{k \le T\} c_\theta(E_k)\right] \\ &= \sum_e c_\theta(e) \mathbb{E}_\theta[N_e(T)]. \end{aligned}$$

□

Lemma T.3 (Stopped information decomposition)**.** *Under Assumption T.1, if* $\mathbb{E}_\theta[\mathsf{C}_T^{\mathrm{stop}}] < \infty$*, then for every* $\theta' \in \Theta$*,*

$$D_{\mathrm{KL}}\left(\mathbb{P}_\theta^{\mathcal{A}} \middle\| \mathbb{P}_{\theta'}^{\mathcal{A}}\right) = \sum_e \mathbb{E}_\theta[N_e(T)] d_e(\theta, \theta'). \tag{T.14}$$

Proof. The proof is the block-level likelihood factorization from Theorem H.5. The within-task stopping rule is already part of the experiment kernel Q_θ^e and therefore produces no additional model-dependent selection term. Predictable experiment-selection, source-stopping, and recommendation kernels are identical under both models and cancel from the Radon–Nikodym derivative. Since the experiment menu is finite, all block log-likelihood ratios are integrable under the stated finite-KL assumptions; finite expected cost implies finite expected block count. Fubini's theorem therefore gives

$$\mathbb{E}_\theta\left[\sum_{k=1}^{T} \log \frac{q_\theta^{E_k}(X_k)}{q_{\theta'}^{E_k}(X_k)}\right] = \sum_e \mathbb{E}_\theta[N_e(T)] d_e(\theta, \theta').$$

□

Theorem T.4 (Bounded-stopping target-decision lower bound)**.** *Let* Θ *and* $\mathcal{E}_{\mathrm{stop}}$ *be finite, and suppose Assumption T.1 holds without requiring the upper-theorem clauses. Every globally* (B, ε, δ)*-correct learner selecting experiments from* $\mathcal{E}_{\mathrm{stop}}$ *satisfies, for every* θ *with nonempty pairwise target-incompatible alternative set,*

$$\mathbb{E}_\theta\left[\mathsf{C}_T^{\mathrm{stop}}\right] \geq \frac{\mathrm{kl}(1-\delta, \delta)}{\Gamma_{\mathrm{stop}}^\star(\theta)}. \tag{T.15}$$

The convention $1/0 = +\infty$ *applies.*

Proof. For every $\theta' \in \mathrm{Alt}_{B,\varepsilon}(\theta)$, target-decision data processing and global correctness give

$$D_{\mathrm{KL}}\left(\mathbb{P}_\theta^{\mathcal{A}} \middle\| \mathbb{P}_{\theta'}^{\mathcal{A}}\right) \geq \mathrm{kl}(1-\delta, \delta).$$

Let $C_\theta = \mathbb{E}_\theta[\mathsf{C}_T^{\mathrm{stop}}]$ and define $w_e = \mathbb{E}_\theta[N_e(T)]/C_\theta$. By Lemma T.2, $w \in \mathcal{W}_\theta$. Using Lemma T.3,

$$\mathrm{kl}(1-\delta, \delta) \leq C_\theta \inf_{\theta' \in \mathrm{Alt}_{B,\varepsilon}(\theta)} \sum_e w_e d_e(\theta, \theta') \leq C_\theta \Gamma_{\mathrm{stop}}^\star(\theta).$$

Rearranging proves the claim. □

T.3 Bounded-Stopping Track-and-Stop

For each θ, fix a deterministic optimizer

$$p^\star(\theta) \in \operatorname*{arg\,max}_{p \in \Delta(\mathcal{E}_{\mathrm{stop}})} \inf_{\theta' \in \mathrm{Alt}_{B,\varepsilon}(\theta)} \frac{I_\theta(p, \theta')}{\bar{c}_\theta(p)}. \tag{T.16}$$

Write

$$\bar{c}_\theta^\star \coloneqq \bar{c}_\theta(p^\star(\theta)), \tag{T.17}$$

$$I_\theta^\star \coloneqq \inf_{\theta' \in \mathrm{Alt}_{B,\varepsilon}(\theta)} I_\theta(p^\star(\theta), \theta'). \tag{T.18}$$

Then

$$\Gamma_{\mathrm{stop}}^\star(\theta) = \frac{I_\theta^\star}{\bar{c}_\theta^\star}. \tag{T.19}$$

The algorithm runs one infinite block sequence independent of δ. It uses the same forced-exploration schedule $F(k) = \lfloor k^\gamma \rfloor$ with $\gamma \in (1/2, 1)$ as in Section I. On nonforced blocks it samples $E_{k+1} \sim p^\star(\widehat{\theta}_k)$, where $\widehat{\theta}_k$ is the blockwise MLE. It uses the same decision-cell GLR

$$Z_k(i) = \max_{\vartheta \in \Theta_i} \mathsf{L}_k(\vartheta) - \max_{\lambda \notin \Theta_i} \mathsf{L}_k(\lambda)$$

and threshold

$$\beta_\delta = \log\left(\frac{|\Theta| - 1}{\delta}\right).$$

Lemma T.5 (Allocation and random-cost convergence)**.** *Under the upper clauses of Assumption T.1, for every true model θ,*

$$\frac{N_e(k)}{k} \longrightarrow p_e^\star(\theta), \tag{T.20}$$

$$\frac{1}{k}\sum_{s=1}^{k} C_s \longrightarrow \bar{c}_\theta^\star \tag{T.21}$$

almost surely and in L^1. Moreover, for every $\eta > 0$ there are constants $A_\eta, b_\eta > 0$ such that the probability that either deviation exceeds η is at most $A_\eta \exp(-b_\eta k^{2\gamma-1})$.

Proof. We verify model identification, allocation tracking, and realized-cost tracking for the stopped outcomes.

Conditional block law. By construction, E_k is measurable with respect to the completed-block history $\mathcal{B}_{k-1}$. The selected experiment begins with an independent reset, and its internal stopping rule is already part of its kernel. Hence

$$\mathcal{L}_\theta(X_k \mid \mathcal{B}_{k-1}, E_k = e) = Q_\theta^e.$$

In particular, for every ordered pair (θ, λ),

$$\mathbb{E}_\theta\left[\log \frac{q_\theta^{E_k}(X_k)}{q_\lambda^{E_k}(X_k)}\middle| \mathcal{B}_{k-1}, E_k\right] = d_{E_k}(\theta, \lambda).$$

The experiment-selection kernel is algorithmic and model independent, so it contributes no likelihood factor.

MLE identification. For a wrong model $\lambda \neq \theta$, set

$$G_k(\lambda) = \mathsf{L}_k(\theta) - \mathsf{L}_k(\lambda), \qquad A_k(\lambda) = \sum_{s=1}^{k} d_{E_s}(\theta, \lambda).$$

Then $G_k(\lambda) - A_k(\lambda)$ is a martingale with increments in an interval of length at most $2L$. Cyclic forced exploration executes every experiment at least $\lfloor F(k)/|\mathcal{E}_{\mathrm{stop}}| \rfloor - 1$ times by block k. Global separation therefore gives constants $a_{\theta,\lambda} > 0$ and k_0 such that $A_k(\lambda) \geq a_{\theta,\lambda} k^\gamma$ for $k \geq k_0$. Azuma–Hoeffding yields

$$\mathbb{P}_\theta\big(G_k(\lambda) \leq 0\big) \leq A \exp(-b k^{2\gamma-1}).$$

A union bound over the finite model class proves the same tail for $\widehat{\theta}_k \neq \theta$. Since $2\gamma - 1 > 0$, the tail is summable; there are almost surely only finitely many MLE errors and their expected total number is finite.

Experiment frequencies. Let $J_s^e = \mathbf{1}\{E_s = e\}$ and let $r_s^e = \mathbb{E}_\theta[J_s^e \mid \mathcal{B}_{s-1}]$. The differences $J_s^e - r_s^e$ form bounded martingale differences. After the final MLE error, $r_s^e = p_e^\star(\theta)$ on nonforced blocks; forced blocks and erroneous-MLE blocks are $o(k)$ almost surely, with stretched-exponential deviation tails. Azuma–Hoeffding therefore gives, for every $\eta > 0$,

$$\mathbb{P}_\theta\left(\max_e \left|\frac{N_e(k)}{k} - p_e^\star(\theta)\right| > \eta\right) \leq A_\eta e^{-b_\eta k^{2\gamma-1}},$$

and proves equation T.20 almost surely.

Realized random cost. Decompose

$$\sum_{s=1}^{k} C_s = \sum_{s=1}^{k} c_\theta(E_s) + \sum_{s=1}^{k} \left(C_s - c_\theta(E_s)\right). \tag{T.22}$$

The second sum is a martingale whose increments lie in an interval of length at most $c_{\max} - c_{\min}$; Azuma–Hoeffding gives an e^{-ck} tail for a fixed normalized deviation. By the frequency result,

$$\frac{1}{k}\sum_{s=1}^{k} c_\theta(E_s) \longrightarrow \sum_e p_e^\star(\theta) c_\theta(e) = \bar{c}_\theta^\star.$$

Combining the two parts gives equation T.21 and the stated stretched-exponential tail. Since the average cost is bounded by $c_{\max}$, almost-sure convergence also implies L^1 convergence. □

Lemma T.6 (Evidence growth per realized source cost)**.** *Under the same assumptions,*

$$\frac{Z_k(a(\theta))}{k} \longrightarrow I^\star_\theta, \qquad \frac{Z_k(a(\theta))}{\sum_{s=1}^k C_s} \longrightarrow \Gamma^\star_{\mathrm{stop}}(\theta) \tag{T.23}$$

almost surely. For each wrong cell $i \neq a(\theta)$,

$$\limsup_{k\to\infty} \frac{Z_k(i)}{\sum_{s=1}^k C_s} < 0 \qquad \textit{almost surely}. \tag{T.24}$$

The true-cell lower-deviation probability has a $A\exp(-bk^{2\gamma-1})$ bound.

Proof. For each model λ, write

$$G_k(\lambda) = \mathsf{L}_k(\theta) - \mathsf{L}_k(\lambda) = \sum_{s=1}^k d_{E_s}(\theta,\lambda) + M_k(\lambda),$$

where $M_k(\lambda)$ is a bounded-increment martingale. The frequency convergence in Lemma T.5 and the vanishing forced fraction imply, simultaneously for the finite model class,

$$\frac{G_k(\lambda)}{k} \longrightarrow I_\theta(p^\star(\theta),\lambda) \quad \text{a.s.}$$

with stretched-exponential lower-deviation probability. Eventual exactness of the MLE implies that, for all sufficiently large k, the maximum likelihood within the true decision cell is attained at θ. Hence

$$Z_k(a(\theta)) = \min_{\lambda\notin\Theta_{a(\theta)}} G_k(\lambda),$$

for all large k. Taking the minimum over the finite alternative set gives

$$\frac{Z_k(a(\theta))}{k} \longrightarrow \inf_{\lambda\in \mathrm{Alt}_{B,\varepsilon}(\theta)} I_\theta(p^\star(\theta),\lambda) = I^\star_\theta.$$

A finite union bound preserves the stretched-exponential deviation tail. Dividing by the realized-cost limit $k^{-1}\sum_{s\le k} C_s \to \bar{c}^\star_\theta$ proves equation T.23.

For a wrong cell $i \neq a(\theta)$, the true model belongs to the complement of Θ_i, and therefore

$$Z_k(i) \le \max_{\lambda\in\Theta_i}\big[\mathsf{L}_k(\lambda) - \mathsf{L}_k(\theta)\big] = -\min_{\lambda\in\Theta_i} G_k(\lambda).$$

Global identifiability and forced exploration make every limiting block-time rate in the final minimum strictly positive. Since the cell and model class are finite and the realized cost has a positive limit per block, equation T.24 follows. In particular, every wrong cell statistic tends to $-\infty$ after a finite random prefix and has finite pathwise supremum. □

Theorem T.7 (Exact bounded-stopping characteristic time)**.** *Let Θ be finite with $|\Theta| \ge 2$ and let $\mathcal{E}_{\mathrm{stop}}$ be a finite menu of bounded observation-adaptive stopping experiments satisfying Assumption T.1, including its upper-theorem clauses. Assume the finite decision partition of Assumption I.1, deterministic tie-breaking, and exact likelihood, global-MLE, and design-oracle access. Then Bounded-Stopping Track-and-Stop is globally (B,ε,δ)-correct. For every θ with nonempty pairwise alternative set and $\Gamma^\star_{\mathrm{stop}}(\theta) > 0$,*

$$\lim_{\delta\downarrow 0} \frac{\mathbb{E}_\theta[\mathsf{C}^{\mathrm{stop}}_{T_\delta}]}{\log(1/\delta)} = \frac{1}{\Gamma^\star_{\mathrm{stop}}(\theta)}. \tag{T.25}$$

The same limit holds for the optimal expected cost over all learners using this experiment menu.

Proof. Anytime wrong-cell control follows from the likelihood-ratio martingale proof of Theorem I.7, since the random duration and cost are included in the dominated block outcome. Almost-sure stopping follows from Lemma T.6: wrong cells have finite pathwise suprema, whereas the true-cell GLR grows linearly at positive rate. The canonical strategy of the selected cell is therefore returned with probability at least $1-\delta$.

Let T_δ be the first GLR crossing on the common infinite experiment. The one-block GLR increment is bounded by L, so the pathwise finite-prefix and overshoot argument of Proposition I.13, together with equation T.23, gives

$$\frac{\mathsf{C}_{T_\delta}^{\mathrm{stop}}}{\beta_\delta} \longrightarrow \frac{1}{\Gamma^\star_{\mathrm{stop}}(\theta)} \qquad \text{almost surely.}$$

For expected cost, choose

$$k_\delta(\eta) = \left\lceil \frac{\beta_\delta}{I^\star_\theta - \eta} \right\rceil.$$

The stretched-exponential evidence tail implies

$$\Pr_\theta(T_\delta > k) \le A_\eta e^{-b_\eta k^{2\gamma-1}} \qquad k \ge k_\delta(\eta).$$

Since every extra block costs at most $c_{\max}$,

$$\mathbb{E}_\theta[\mathsf{C}_{T_\delta}^{\mathrm{stop}}] \le \mathbb{E}_\theta\left[\sum_{s=1}^{k_\delta(\eta)} C_s\right] + c_{\max} \sum_{k \ge k_\delta(\eta)} \Pr_\theta(T_\delta > k).$$

The tail sum is $o(\beta_\delta)$, while Lemma T.5 gives

$$\mathbb{E}_\theta\left[\sum_{s=1}^{k_\delta(\eta)} C_s\right] = \bar{c}^\star_\theta k_\delta(\eta) + o(\beta_\delta).$$

Hence

$$\limsup_{\delta \downarrow 0} \frac{\mathbb{E}_\theta[\mathsf{C}_{T_\delta}^{\mathrm{stop}}]}{\log(1/\delta)} \le \frac{\bar{c}^\star_\theta}{I^\star_\theta - \eta}.$$

Letting $\eta \downarrow 0$ yields the upper constant $1/\Gamma^\star_{\mathrm{stop}}(\theta)$. The matching lower constant is Theorem T.4, since $\mathrm{kl}(1-\delta, \delta)/\log(1/\delta) \to 1$. □

Remark T.8 (Relation to random-cost sequential testing). The information-per-expected-cost principle is consistent with recent cost-aware sequential-testing analyses with heterogeneous or random action costs (Vershinin et al., 2025). The present experiment differs in that one action is itself an internally controlled, stopped multi-episode task experiment, and its random duration is generated by observations sharing a latent task.

U Noisy Screening and the Value of Observation-Adaptive Persistence

Appendix P studies perfect revelation of a task type before the learner decides whether to continue. This appendix replaces perfect revelation by an arbitrary finite noisy screening channel and derives the exact optimal continuation rule, information rate, and fixed-confidence characteristic time. Perfect and uninformative screening appear as endpoint cases.

U.1 Screened parity family with a finite signal alphabet

Fix $L \ge 2$, $p \in (0,1]$, $\rho \in (0,1)$, and a finite signal alphabet $\mathcal{S}$. The global parameter is $\sigma \in \{-1,+1\}$. A fresh task draws a latent type $\Xi \sim \mathrm{Bernoulli}(p)$ and $X \in \{-1,+1\}^L$ according to

$$\Pr_\sigma(X = x \mid \Xi = \xi) = \begin{cases} 2^{-L}, & \xi = 0, \\ 2^{-L}\left(1 + \rho\sigma \prod_{j=1}^L x_j\right), & \xi = 1. \end{cases} \tag{U.1}$$

Before the first source action, the learner observes a signal $S \in \mathcal{S}$ drawn from a known channel $K(\cdot \mid \Xi)$, conditionally independently of X given Ξ. The first episode reveals S and one coordinate

of X. If the learner continues, the remaining $L-1$ episodes reveal the other coordinates; if it stops, the task is discarded. Write

$$\pi_s \coloneqq \Pr(S = s), \tag{U.2}$$

$$q_s \coloneqq \Pr(\Xi = 1 \mid S = s) = \frac{pK(s \mid 1)}{pK(s \mid 1) + (1-p)K(s \mid 0)} \tag{U.3}$$

for every s with $\pi_s > 0$. For notational convenience, when $\pi_s = 0$ we set $q_s = 0$; every expression below then assigns that signal zero weight. Let

$$d_{\mathrm{par}}(u) \coloneqq u \log \frac{1+u}{1-u}, \qquad u \in [0,1), \tag{U.4}$$

with $d_{\mathrm{par}}(0) = 0$, and set

$$d_s \coloneqq d_{\mathrm{par}}(q_s \rho). \tag{U.5}$$

A general within-task rule may use the signal, any proper subset of queried coordinates, and private randomization when deciding whether to continue. The next lemma shows that no proper coordinate history refines the screening signal: without loss of information or expected cost, the learner either stops after the first episode or commits immediately to revealing every remaining coordinate. Such a reduced randomized policy is represented by a vector $u = (u_s)_{s \in \mathcal{S}} \in [0,1]^{\mathcal{S}}$: after signal s, it completes the task with probability u_s, using fresh private randomization. Let $J \in \{0,1\}$ denote the completion decision. If $J = 1$, let $W = \prod_{j=1}^{L} X_j$ be the observed parity. The stopped block outcome is (S, J, JW), and its cost is

$$C(u) = \kappa + 1 + (L-1)J. \tag{U.6}$$

Lemma U.1 (Reduction to signal-only completion)**.** *Fix any bounded observation-adaptive within-task policy. There is a vector $u \in [0,1]^{\mathcal{S}}$ and a reduced policy that, after the first episode, either stops immediately or reveals all remaining coordinates, such that:*

1. *under each global sign, the reduced policy has the same joint law of (S, J, JW) as the original policy, where J records whether all L distinct coordinates are eventually revealed; and*
2. *its expected source cost is no larger under either sign.*

Consequently, optimizing information per expected cost over all bounded within-task policies reduces exactly to optimizing over the vectors u above.

Proof. For either type $\Xi \in \{0,1\}$, every strict coordinate marginal of X is uniform. More strongly, if $A \subsetneq [L]$ and x_A is a realized proper subvector, then the likelihood of x_A is the same for both task types and both global signs. Hence

$$\Pr(\Xi = 1 \mid S = s, X_A = x_A) = q_s.$$

If $j \notin A$ and at least one coordinate remains unqueried after j, then X_j is conditionally fair. If j is the final missing coordinate, the resulting parity satisfies

$$\Pr_{\sigma}(W = w \mid S = s, X_A = x_A) = \frac{1}{2}(1 + q_s \rho \sigma w),$$

which is independent of the particular proper history x_A. Repeated coordinates and safe actions are deterministic functions of the existing history and reveal nothing further.

Let u_s be the original policy's conditional probability, given $S = s$, of ever revealing all L distinct coordinates. The preceding conditional invariance implies that, given $S = s$ and completion, the parity law depends only on s, not on the path or time at which completion was chosen. Thus a policy that flips one Bernoulli(u_s) coin immediately after the first episode and, on success, queries every remaining coordinate has the same law of (S, J, JW). On failure it stops at once; on success it uses exactly L episodes. Every original path with $J = 0$ used at least one episode and every path with $J = 1$ used at least L episodes. The reduced policy therefore has no larger realized cost conditional on (S, J) and hence no larger expected cost under either sign. □

Lemma U.2 (Full-transcript likelihood-ratio sufficiency)**.** *Let $\mathcal{T}$ be the complete stopped transcript of any admissible bounded within-task policy, including every selected coordinate, observed value, private decision, and stopping time. Let J indicate whether all L distinct coordinates are revealed, and define $W = \prod_{j=1}^{L} X_j$ on $\{J = 1\}$. Then the likelihood ratio under $\sigma = +1$ relative to $\sigma = -1$ is measurable with respect to (S, J, JW) and satisfies*

$$\frac{dQ_+}{dQ_-}(\mathcal{T}) = \begin{cases} 1, & J = 0, \\ \dfrac{1 + q_S \rho W}{1 - q_S \rho W}, & J = 1. \end{cases} \tag{U.7}$$

Consequently,

$$D_{\mathrm{KL}}(Q_+ \| Q_-) = \sum_s \pi_s u_s d_{\mathrm{par}}(q_s \rho), \tag{U.8}$$

where $u_s = \Pr(J = 1 \mid S = s)$.

Proof. Fix a realized stopped transcript. Every task-type, coordinate-selection, private-randomization, and proper-coordinate factor has the same conditional law under the two global signs. If $J = 0$, at least one coordinate remains unobserved, so summing that coordinate cancels the parity tilt and the entire transcript probabilities agree. If $J = 1$, all coordinates are known and the only sign-dependent factor is the parity term. Conditional on $S = s$, mixing the null type with posterior probability $1 - q_s$ and the informative type with probability q_s gives

$$\Pr_\sigma(W = w \mid S = s, J = 1) = \frac{1}{2}(1 + q_s \rho \sigma w).$$

All policy and stopping factors cancel in the Radon–Nikodym derivative, which proves equation U.7. Taking the Q_+ expectation and conditioning on S and J gives equation U.8. □

Lemma U.3 (Signal-conditioned parity law)**.** *For every signal s with $\pi_s > 0$,*

$$\Pr_\sigma(W = w \mid S = s, J = 1) = \frac{1}{2}\left(1 + q_s \rho \sigma w\right), \qquad w \in \{-1, +1\}. \tag{U.9}$$

Consequently, the one-block KL information and expected cost of policy u are

$$I(u) \coloneqq D_{\mathrm{KL}}(Q_+^u \| Q_-^u) = \sum_{s \in \mathcal{S}} \pi_s u_s d_s, \tag{U.10}$$

$$c(u) \coloneqq \mathbb{E}[C(u)] = \kappa + 1 + (L - 1) \sum_{s \in \mathcal{S}} \pi_s u_s. \tag{U.11}$$

The reverse KL has the same value.

Proof. Conditional on $S = s$, the type is Bernoulli with success probability q_s. Under $\Xi = 0$, W is fair; under $\Xi = 1$, $\Pr_\sigma(W = w \mid \Xi = 1) = \frac{1}{2}(1 + \rho\sigma w)$ by summing equation U.1 over each parity class. Mixture over Ξ gives equation U.9. The signal law and the continuation coin are independent of σ. Conditional KL therefore contributes d_s exactly on the event that signal s occurs and the task is completed, proving equation U.10. The cost identity is immediate from equation U.6. □

U.2 Exact posterior-threshold continuation rule

Define the optimal information per expected source cost

$$\Gamma^\star_{\mathrm{scr}} \coloneqq \max_{u \in [0,1]^{\mathcal{S}}} \frac{I(u)}{c(u)}. \tag{U.12}$$

Because $d_{\mathrm{par}}(q\rho)$ is strictly increasing in $q \in [0, 1]$, the result below is equivalently a threshold rule in the posterior probability q_s.

Corollary U.4 (Posterior threshold under screened parity)**.** *Suppose at least one positive-mass signal has $d_s > 0$. A policy $u^\star$ is optimal in equation U.12 if and only if, for every $s \in \mathcal{S}$ with $\pi_s > 0$,*

$$u_s^\star = \begin{cases} 1, & d_s > (L-1)\Gamma_{\mathrm{scr}}^\star, \\ 0, & d_s < (L-1)\Gamma_{\mathrm{scr}}^\star, \\ \textit{arbitrary in } [0,1], & d_s = (L-1)\Gamma_{\mathrm{scr}}^\star. \end{cases} \tag{U.13}$$

Coordinates with $\pi_s = 0$ are immaterial and may be assigned arbitrarily. In particular, order the positive-mass signals so that $d_{(1)} \geq \cdots \geq d_{(m)}$. Then

$$\Gamma_{\mathrm{scr}}^\star = \max_{1 \leq j \leq m} \frac{\sum_{i=1}^{j} \pi_{(i)} d_{(i)}}{\kappa + 1 + (L-1)\sum_{i=1}^{j} \pi_{(i)}}. \tag{U.14}$$

Thus an optimal deterministic policy continues exactly on a posterior upper level set; randomization is needed only at a threshold tie.

Proof. This is the screened-parity specialization of the Bellman–Chernoff recursion in Theorem E.4. After signal s, every incomplete continuation has zero information, while completion requires $L-1$ additional episodes and yields d_s information. Hence the Bellman continuation advantage is $d_s - (L-1)\Gamma_{\mathrm{scr}}^\star$. For completeness, the same threshold follows directly as follows. Write $R(u) = I(u)/c(u)$. Holding every coordinate except u_s fixed, an increase by $h > 0$ changes the ratio from I/c to

$$\frac{I + h\pi_s d_s}{c + h\pi_s(L-1)}.$$

This is larger than I/c exactly when $d_s > (L-1)I/c$. The analogous statement holds for a decrease, proving the necessity of equation U.13. For sufficiency, let u satisfy that rule with $\gamma = I(u)/c(u)$. For any $v \in [0,1]^{\mathcal{S}}$,

$$I(v) - \gamma c(v) = \sum_s \pi_s (v_s - u_s)\big[d_s - (L-1)\gamma\big] \leq 0,$$

because each factor has the sign prescribed by the threshold rule and $I(u) - \gamma c(u) = 0$. Hence $I(v)/c(v) \leq \gamma$, so u is globally optimal. A linear-fractional function on the compact hypercube attains a maximum at a vertex unless a threshold coefficient vanishes, in which case either endpoint is also optimal. Sorting the d_s therefore reduces the search to deterministic upper level sets, precisely the prefixes in equation U.14. Finally,

$$\frac{d}{dq} d_{\mathrm{par}}(q\rho) = \rho \log \frac{1+q\rho}{1-q\rho} + \frac{2q\rho^2}{1-q^2\rho^2} > 0$$

for $q > 0$, proving the posterior-threshold interpretation. □

U.3 Precommitted persistence and exact characteristic times

A precommitted policy chooses, before observing S, whether the current task will receive one episode or all L episodes. It may randomize this choice and may depend on all previously completed tasks. For one fresh task, let $r \in [0,1]$ be the conditional probability of full completion. Define

$$\bar{d} := \sum_{s \in \mathcal{S}} \pi_s d_s. \tag{U.15}$$

Then its information per expected cost is

$$\frac{r\bar{d}}{\kappa + 1 + r(L-1)}. \tag{U.16}$$

This ratio is strictly increasing in r whenever $\bar{d} > 0$. Hence the optimal precommitted policy completes every task and has rate

$$\Gamma_{\mathrm{pre}}^\star = \frac{\bar{d}}{\kappa + L}. \tag{U.17}$$

For target relevance, use the target interface of Appendix P: it reveals the type and permits $L-1$ coordinate queries before deployment. The two signs have disjoint ε-optimal target-strategy sets whenever

$$0 < \varepsilon < \frac{p\rho}{2^{L+1}}. \tag{U.18}$$

Let $\mathfrak{C}_\delta^{\mathrm{scr}}(\sigma)$ and $\mathfrak{C}_\delta^{\mathrm{pre}}(\sigma)$ denote the optimal expected source costs among globally correct learners using, respectively, the finite menu of deterministic signal-contingent continuation rules and the precommitted menu.

Theorem U.5 (Exact noisy-screening characteristic times)**.** *Under equation U.18, for each $\sigma \in \{-1,+1\}$,*

$$\lim_{\delta\downarrow 0} \frac{\mathfrak{C}_\delta^{\mathrm{scr}}(\sigma)}{\log(1/\delta)} = \frac{1}{\Gamma^\star_{\mathrm{scr}}}, \tag{U.19}$$

$$\lim_{\delta\downarrow 0} \frac{\mathfrak{C}_\delta^{\mathrm{pre}}(\sigma)}{\log(1/\delta)} = \frac{1}{\Gamma^\star_{\mathrm{pre}}}. \tag{U.20}$$

Consequently, the asymptotic precommitted-to-adaptive cost ratio is

$$\frac{\Gamma^\star_{\mathrm{scr}}}{\Gamma^\star_{\mathrm{pre}}} \geq 1. \tag{U.21}$$

Proof. Every deterministic continuation set is a bounded observation-adaptive stopping experiment of Appendix T; the finite signal alphabet gives a finite menu. Its KL information and expected cost are equation U.10 and equation U.11. The binary model has one decision cell per sign, all log-likelihood ratios are bounded, and each nontrivial experiment separates the signs. Therefore Theorem T.7 applies, and optimizing over the adaptive menu gives equation U.19. The same theorem applied to the two precommitted block choices, together with equation U.17, gives equation U.20. Since full continuation is available to the adaptive learner, $\Gamma^\star_{\mathrm{scr}} \geq \Gamma^\star_{\mathrm{pre}}$. □

U.4 Binary screening channel and endpoint regimes

Let $\mathcal{S} = \{0,1\}$ and

$$\Pr(S=1 \mid \Xi=1) = a, \qquad \Pr(S=1 \mid \Xi=0) = b, \qquad 0 \leq b \leq a \leq 1. \tag{U.22}$$

Then

$$\pi_1 = pa + (1-p)b, \qquad q_1 = \frac{pa}{pa+(1-p)b}, \tag{U.23}$$

$$\pi_0 = p(1-a) + (1-p)(1-b), \qquad q_0 = \frac{p(1-a)}{p(1-a)+(1-p)(1-b)}. \tag{U.24}$$

When both signal values have positive mass, $a \geq b$ implies $q_1 \geq q_0$ and hence $d_1 \geq d_0$. Define

$$\Gamma_1 \coloneqq \frac{\pi_1 d_1}{\kappa + 1 + (L-1)\pi_1}, \tag{U.25}$$

$$\Gamma_{\mathrm{all}} \coloneqq \frac{\pi_1 d_1 + \pi_0 d_0}{\kappa + L}. \tag{U.26}$$

Corollary U.6 (Exact binary-channel interpolation)**.** *For the channel equation U.22,*

$$\Gamma^\star_{\mathrm{scr}} = \max\{\Gamma_1, \Gamma_{\mathrm{all}}\}, \qquad \Gamma^\star_{\mathrm{pre}} = \Gamma_{\mathrm{all}}. \tag{U.27}$$

The adaptive policy continues only after signal 1 *exactly when*

$$\frac{d_0}{L-1} \leq \Gamma_1; \tag{U.28}$$

otherwise it continues after both signals. At equality, every continuation probability after signal 0 *is optimal.*

For perfect screening $(a,b) = (1,0)$,

$$\Gamma^\star_{\mathrm{scr}} = \frac{p\, d_{\mathrm{par}}(\rho)}{\kappa + 1 + p(L-1)}, \qquad \Gamma^\star_{\mathrm{pre}} = \frac{p\, d_{\mathrm{par}}(\rho)}{\kappa + L}, \tag{U.29}$$

recovering Appendix P. For uninformative screening $a = b$, the posterior is $q_0 = q_1 = p$, full continuation is optimal, and

$$\Gamma^\star_{\mathrm{scr}} = \Gamma^\star_{\mathrm{pre}} = \frac{d_{\mathrm{par}}(p\rho)}{\kappa + L}. \tag{U.30}$$

Proof. The prefix formula in Corollary U.4 has only two nonempty prefixes, giving equation U.27. Adding signal 0 to the high-signal-only policy improves the ratio precisely when $d_0/(L-1) > \Gamma_1$, which proves equation U.28. Under perfect screening, $q_1 = 1$, $q_0 = 0$, and $\pi_1 = p$. Under uninformative screening, both posteriors equal p and the ratio $xd_{\mathrm{par}}(p\rho)/(\kappa + 1 + (L-1)x)$ is increasing in the total completion probability x, so completing every task is optimal. □

The theorem gives a reusable interpolation rather than only a perfect-screening separation: continuation is selected by a posterior threshold, the binary channel has a closed-form solution, and the adaptive advantage is exactly one at the uninformative endpoint.

V Upper Bounds in the Near-Square Orthogonal Regime

The intrinsic-rate upper bounds in Appendix L use finite inverse-Wishart moments and are stated most cleanly for $n \geq 2d + 2$. This appendix gives explicit finite-risk estimators for every near-square regime $n \geq d$, including $n = d$ and $n = d + 1$, where the unregularized inverse-design moments diverge. The bounds are upper bounds, not a claim that the singular transition rates below are minimax optimal.

Throughout, use the conditional-regression representation

$$Y_0 = \sqrt{a}\, Z, \qquad R_\star^\top Y_1 Y_0^\top = ZZ^\top + \sqrt{\Delta_\tau}\, GZ^\top, \tag{V.1}$$

where $Z, G \in \mathbb{R}^{d\times n}$ are independent standard Gaussian matrices. Let

$$S \coloneqq ZZ^\top, \qquad p \coloneqq n - d \geq 0. \tag{V.2}$$

V.1 Ridge conditional least squares

For $\lambda > 0$, define the ambient matrix-valued ridge estimator

$$\widetilde{R}_{n,\lambda} \coloneqq aY_1 Y_0^\top \left(Y_0 Y_0^\top + a\lambda I_d\right)^{-1}. \tag{V.3}$$

The parameter belongs to $\mathrm{O}(d)$, but the estimator is allowed to take values in $\mathbb{R}^{d\times d}$, consistently with the minimax decision problem in equation M.1.

Lemma V.1 (Exact ridge-risk decomposition)**.** *For every $n \geq d$ and $\lambda > 0$,*

$$\begin{aligned}
&\sup_{R_\star \in \mathrm{O}(d)} \mathbb{E}_{R_\star}\left[\left\|\widetilde{R}_{n,\lambda} - R_\star\right\|_{\mathrm{F}}^2\right] \\
&\quad = \mathbb{E}\,\mathrm{Tr}\left[\lambda^2 (S + \lambda I_d)^{-2}\right] + d\Delta_\tau \mathbb{E}\,\mathrm{Tr}\left[S(S + \lambda I_d)^{-2}\right] &\text{(V.4)}\\
&\quad \leq (\lambda + d\Delta_\tau)\mathbb{E}\,\mathrm{Tr}(S + \lambda I_d)^{-1}. &\text{(V.5)}
\end{aligned}$$

Proof. Substituting equation V.1 into equation V.3 gives

$$R_\star^\top \widetilde{R}_{n,\lambda} - I_d = -\lambda(S + \lambda I_d)^{-1} + \sqrt{\Delta_\tau} GZ^\top (S + \lambda I_d)^{-1}. \tag{V.6}$$

Conditionally on Z, the second term is centered, so the cross term has zero expectation. For a deterministic matrix B, standard Gaussian rows give $\mathbb{E}\|GB\|_{\mathrm{F}}^2 = d\|B\|_{\mathrm{F}}^2$. Therefore

$$\mathbb{E}\left[\left\|GZ^\top (S + \lambda I_d)^{-1}\right\|_{\mathrm{F}}^2 \middle| Z\right] = d\,\mathrm{Tr}\left[S(S + \lambda I_d)^{-2}\right],$$

which proves equation V.4. For every eigenvalue $x \geq 0$,

$$\frac{\lambda^2}{(x+\lambda)^2} \leq \frac{\lambda}{x+\lambda}, \qquad \frac{x}{(x+\lambda)^2} \leq \frac{1}{x+\lambda}.$$

Summing these scalar inequalities proves equation V.5. □

V.2 A RESOLVENT BOUND FROM A GAUSSIAN SCHUR COMPLEMENT

Lemma V.2 (Wishart resolvent bound)**.** *Let $Z \in \mathbb{R}^{d\times n}$ be standard Gaussian with $n \geq d$, let $S = ZZ^\top$, and let $p = n - d$. For every $\lambda > 0$,*

$$\mathbb{E}\,\mathrm{Tr}(S + \lambda I_d)^{-1} \leq d\,\mathbb{E}\left[\frac{1}{X_{p+1} + \lambda}\right], \tag{V.7}$$

where $X_k \sim \chi^2_k$. Moreover, for $0 < \lambda \leq 1$,

$$\mathbb{E}\left[\frac{1}{X_{p+1} + \lambda}\right] \leq \begin{cases} 2\lambda^{-1/2}, & p = 0, \\ 2\log(e + \lambda^{-1}), & p = 1, \\ (p-1)^{-1}, & p \geq 2. \end{cases} \tag{V.8}$$

Proof. Fix a row index i, write $z_i \in \mathbb{R}^n$ for row i of Z, and let Z_{-i} contain the remaining $d-1$ rows. The Schur-complement formula gives

$$\left[(S + \lambda I_d)^{-1}\right]_{ii} = \frac{1}{\lambda + z_i A_{-i,\lambda} z_i^\top}, \tag{V.9}$$

where

$$A_{-i,\lambda} = I_n - Z_{-i}^\top (Z_{-i} Z_{-i}^\top + \lambda I_{d-1})^{-1} Z_{-i}. \tag{V.10}$$

Almost surely Z_{-i} has rank $d-1$. On $\ker(Z_{-i})$, which has dimension $n - d + 1 = p + 1$, the matrix $A_{-i,\lambda}$ equals the identity. On the row space of Z_{-i}, its eigenvalues are $\lambda/(s_j^2 + \lambda) \in (0, 1)$. Hence

$$A_{-i,\lambda} \succeq P_{\ker(Z_{-i})}.$$

Conditionally on Z_{-i}, the row z_i is standard Gaussian and independent of this rank-$(p+1)$ projection. Therefore

$$z_i P_{\ker(Z_{-i})} z_i^\top \sim \chi^2_{p+1},$$

and equation V.9 is bounded above in conditional expectation by $\mathbb{E}[(X_{p+1} + \lambda)^{-1}]$. Summing the diagonal entries proves equation V.7.

For $p = 0$, the χ^2_1 density and $e^{-x/2} \leq 1$ give

$$\mathbb{E}\frac{1}{X_1 + \lambda} \leq \frac{1}{\sqrt{2\pi}} \int_0^\infty \frac{x^{-1/2}}{x + \lambda}\, dx = \sqrt{\frac{\pi}{2}}\, \lambda^{-1/2} \leq 2\lambda^{-1/2}.$$

For $p = 1$, the χ^2_2 density is $\frac{1}{2}e^{-x/2}$. Splitting the integral at one yields

$$\begin{aligned} \mathbb{E}\frac{1}{X_2 + \lambda} &\leq \frac{1}{2}\int_0^1 \frac{dx}{x + \lambda} + \frac{1}{2}\int_1^\infty e^{-x/2}\, dx \\ &\leq 2\log(e + \lambda^{-1}). \end{aligned}$$

Finally, for $p \geq 2$,

$$\mathbb{E}\frac{1}{X_{p+1} + \lambda} \leq \mathbb{E}\frac{1}{X_{p+1}} = \frac{1}{p-1},$$

using the elementary inverse moment of a chi-square variable with more than two degrees of freedom. □

V.3 COMPLETE UPPER BOUNDS IN ALL NEAR-SQUARE REGIMES

Define

$$x \coloneqq d\Delta_\tau. \tag{V.11}$$

Theorem V.3 (Near-square orthogonal upper bounds)**.** *For every $d \geq 2$, $n \geq d$, and $\Delta_\tau > 0$, the ambient-valued minimax risk satisfies*

$$\mathfrak{R}^{\mathrm{orth}}_{n,d}(\Delta_\tau) \leq \begin{cases} d\min\{1, 4\sqrt{x}\}, & n = d, \\ d\min\{1, 4x\log(e + x^{-1})\}, & n = d+1, \\ d\min\left\{1, \dfrac{x}{n-d-1}\right\}, & n \geq d+2. \end{cases} \tag{V.12}$$

For $\Delta_\tau = 0$ and $n \geq d$, the risk is zero. Hence every transition regime $d \leq n < 2d + 2$ has an explicit finite upper bound; for $n \geq 2d + 2$, the final line reduces to the intrinsic order $\Delta_\tau p_d/n$ up to a universal constant.

Proof. The constant zero estimator has risk $\sup_R \|R\|_{\mathrm{F}}^2 = d$, so it suffices to prove the nontrivial parts.

If $n = d$, let $x = d\Delta_\tau \leq 1$ and choose $\lambda = x$. Combining Lemmas V.1 and V.2 gives

$$\sup_R \mathbb{E}_R \left\| \widetilde{R}_{d,x} - R \right\|_{\mathrm{F}}^2 \leq d(x + x)\frac{2}{\sqrt{x}} = 4d\sqrt{x}.$$

If $x > 1$, use the zero estimator.

If $n = d + 1$, again take $\lambda = x$ when $x \leq 1$. Then

$$\sup_R \mathbb{E}_R \left\| \widetilde{R}_{d+1,x} - R \right\|_{\mathrm{F}}^2 \leq d(x + x)2\log(e + x^{-1}) = 4dx\log(e + x^{-1}).$$

For $x > 1$, use the zero estimator.

If $n \geq d + 2$, the unregularized conditional least-squares estimator equation L.5 is well defined almost surely and equation L.12 gives

$$\sup_R \mathbb{E}_R \left\| \widetilde{R}_n - R \right\|_{\mathrm{F}}^2 = \frac{\Delta_\tau d^2}{n - d - 1} = d\frac{x}{n - d - 1}.$$

Taking the better of this estimator and zero proves the last line. When $\Delta_\tau = 0$ and $n \geq d$, the noiseless paired anchors span $\mathbb{R}^d$ almost surely and recover R exactly, as shown in Proposition K.7. □

Remark V.4 (The singular transition is genuinely different). At $n = d$ and $n = d + 1$, the inverse-Wishart moments used by ordinary least squares diverge. Ridge regularization yields the square-root and logarithmic upper bounds in equation V.12. The global lower bound of Appendix M remains valid, but this appendix does not claim those two singular upper rates are minimax sharp. The previously uncovered transition regime is nevertheless fully covered by explicit, finite, measurable estimators.

W Robust Persistence Phenomena

The exact parity and invariance examples isolate the mechanism cleanly, but practical relevance should not depend on knife-edge cancellation. This appendix gives three quantitative robustness results: an approximate persistence hierarchy with arbitrarily small subcritical information, a near-invariance theorem for reset-only acquisition, and an exact criterion for source-to-target transfer under support shift.

W.1 Approximate persistence profiles

For a model pair (θ, θ') and maximum same-task depth m, define its information-per-cost profile

$$I_m(\theta, \theta') := \sup_{e:\, \tau_e \leq m} \frac{D_{\mathrm{KL}}(Q_\theta^e \| Q_{\theta'}^e)}{c_\theta(e)}. \tag{W.1}$$

The supremum ranges over all bounded observation-adaptive experiments of maximum depth m. Clearly I_m is nondecreasing in m.

Theorem W.1 (Quantitative approximate persistence barrier)**.** *Let θ' be pairwise target incompatible with θ. Suppose every admissible source block has depth at most m and $I_m(\theta, \theta') \leq \eta$. Then every globally (B, ε, δ)-correct learner satisfies*

$$\mathbb{E}_\theta \mathsf{C}_T \geq \frac{\mathrm{kl}(1 - \delta, \delta)}{\eta}, \tag{W.2}$$

with the convention that the right-hand side is infinite when $\eta = 0$. Conversely, consider the binary model class $\{\theta, \theta'\}$. If a depth-r experiment e_r satisfies, in both directions,

$$\min_{(\vartheta,\lambda)\in\{(\theta,\theta'),(\theta',\theta)\}} \frac{D_{\mathrm{KL}}(Q_\vartheta^{e_r} \| Q_\lambda^{e_r})}{c_\vartheta(e_r)} \geq c > 0,$$

then, under finite-alphabet uniformly bounded log-likelihood ratios, repeated independent uses of e_r with a sequential likelihood-ratio test are globally δ-correct and have expected source cost under either model

$$O\left(\frac{\log(1/\delta)}{c}\right). \tag{W.3}$$

Proof. For every selected block e, the definition of I_m gives $d_e(\theta, \theta') \leq \eta c_\theta(e)$. The stopped block information and cost decompositions therefore imply

$$D_{\mathrm{KL}}(\mathbb{P}_\theta^{\mathcal{A}} \| \mathbb{P}_{\theta'}^{\mathcal{A}}) \leq \eta\, \mathbb{E}_\theta \mathsf{C}_T.$$

Target-decision data processing gives the lower bound $D_{\mathrm{KL}} \geq \mathrm{kl}(1-\delta, \delta)$, proving equation W.2. For the upper bound, independent repetitions of e_r produce IID block observations. The two-sided sequential likelihood-ratio test with boundaries of order $\log(1/\delta)$ has both errors at most δ by the likelihood-ratio martingale. Uniformly bounded increments give bounded overshoot, while Wald's identity and the two directional information-per-cost bounds show that its expected source cost under either model is at most a universal constant times $\log(1/\delta)/c$. This proves equation W.3. □

Theorem W.2 (Robust arbitrary-order hierarchy)**.** *Fix $r \geq 2$, reset cost $\kappa \geq 0$, parity signal $\varrho \in (0, 1)$, and any $\eta > 0$. There is a finite-state persistent meta-MDP with pairwise target-incompatible global models $\sigma \in \{-1, +1\}$ such that*

$$I_{r-1}(+,-) \leq \eta, \qquad I_r(+,-) \geq \frac{d_{\mathrm{par}}(\varrho)}{r+\kappa}. \tag{W.4}$$

Hence the ratio I_{r-1}/I_r can be made arbitrarily small without exact subcritical cancellation.

Proof. Start with the order-r parity family of Appendix C. In every source episode, in addition to the original coordinate reward, reveal an independent side-channel bit S satisfying

$$\Pr_\sigma(S = s) = \frac{1}{2}(1 + \zeta\sigma s), \qquad s \in \{-1, +1\}.$$

The side channel is independent across episodes and tasks and independent of the task latent. Its one-episode divergence is $d_{\mathrm{par}}(\zeta)$. For every block of depth $m < r$, the original coordinate transcript is model independent by the adaptive parity invariance theorem. If the experiment stops after the random number $\tau_e < r$ of episodes, the KL chain rule gives exactly $\mathbb{E}_+[\tau_e] d_{\mathrm{par}}(\zeta)$ side-channel divergence, while $c_+(e) = \kappa + \mathbb{E}_+[\tau_e] \geq \mathbb{E}_+[\tau_e]$. Therefore

$$I_{r-1}(+,-) \leq d_{\mathrm{par}}(\zeta).$$

Choose $\zeta > 0$ so small that $d_{\mathrm{par}}(\zeta) \leq \eta$.

The complete r-view experiment observes the original parity, whose KL is $d_{\mathrm{par}}(\varrho)$, in addition to the side channels. Discarding the side channels cannot increase difficulty, so by data processing its KL is at least $d_{\mathrm{par}}(\varrho)$. Its cost is $r + \kappa$, yielding the second inequality in equation W.4. The target task is unchanged, so pairwise target incompatibility remains valid. □

W.2 Near invariance of reset-only marginals

Let a group G act measurably on the anchor space and let P_g^v be the one-episode marginal under view v and global action g. Define

$$\eta_{\mathrm{marg}} := \sup_{g,g'\in G} \sup_v D_{\mathrm{KL}}(P_g^v \| P_{g'}^v). \tag{W.5}$$

Exact invariance corresponds to $\eta_{\mathrm{marg}} = 0$.

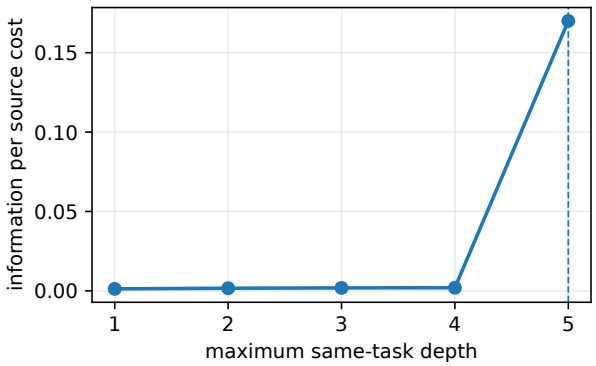


Figure 4: A robust order-r persistence profile. A weak independent side channel makes every subcritical depth slightly informative, while completing the final view unlocks the order-one parity signal.

Theorem W.3 (Reset-only information under approximate symmetry)**.** *For every adaptive reset-only learner with stopping time T satisfying $\mathbb{E}_g T < \infty$,*

$$D_{\mathrm{KL}}\left(\mathbb{P}_g^{\mathcal{A}} \middle\| \mathbb{P}_{g'}^{\mathcal{A}}\right) \le \eta_{\mathrm{marg}}\, \mathbb{E}_g[T] \qquad \forall g, g' \in G. \tag{W.6}$$

Consequently, if g and g' require incompatible target decisions, every δ-correct reset-only learner obeys

$$\mathbb{E}_g T \ge \frac{\mathrm{kl}(1-\delta, \delta)}{\eta_{\mathrm{marg}}}. \tag{W.7}$$

Proof. At every reset-only episode, the learner's selected view is predictable from the observed past, and the fresh anchor is independent of that past. The conditional trajectory KL between g and g' is therefore at most η_{marg}. The adaptive stopped KL chain rule gives

$$D_{\mathrm{KL}}(\mathbb{P}_g^{\mathcal{A}} \| \mathbb{P}_{g'}^{\mathcal{A}}) = \mathbb{E}_g \sum_{t=1}^{T} D_{\mathrm{KL}}(P_g^{V_t} \| P_{g'}^{V_t}) \le \eta_{\mathrm{marg}} \mathbb{E}_g T.$$

The target-decision event and binary data processing give equation W.7. □

Proposition W.4 (Arbitrarily near-invariant marginals with constant paired information)**.** *Let $G = \{-1, +1\}$ act on $\{-1, +1\}$ by multiplication. Let $\Pr(X = 1) = 1/2 + \epsilon$ and observe either view 0, which returns X, or view 1, which returns*

$$Y = gX\Xi, \qquad \Pr(\Xi = -1) = q < 1/2.$$

Then

$$\eta_{\mathrm{marg}} = O(\epsilon^2) \quad \text{as } \epsilon \to 0, \tag{W.8}$$

whereas the same-task paired observation (X, Y) satisfies

$$D_{\mathrm{KL}}\left(P_{g=+1}^{\mathrm{pair}} \middle\| P_{g=-1}^{\mathrm{pair}}\right) = (1-2q) \log \frac{1-q}{q}, \tag{W.9}$$

independently of ϵ.

Proof. View 0 is independent of g. For view 1,

$$\Pr_g(Y = 1) = \frac{1}{2} + g(1-2q)\epsilon.$$

The KL divergence between Bernoulli parameters $1/2 \pm (1-2q)\epsilon$ is $O(\epsilon^2)$ by Taylor expansion, proving equation W.8. In a paired task, the statistic $XY = g\Xi$. Under $g = +1$ it equals $+1$ with probability $1 - q$, and under $g = -1$ it equals $+1$ with probability q. Its binary KL is equation W.9, and XY is sufficient for g given the pair. □

Thus reset-only acquisition can be arbitrarily inefficient even when exact invariance is slightly broken, while continuation retains order-one correspondence information.

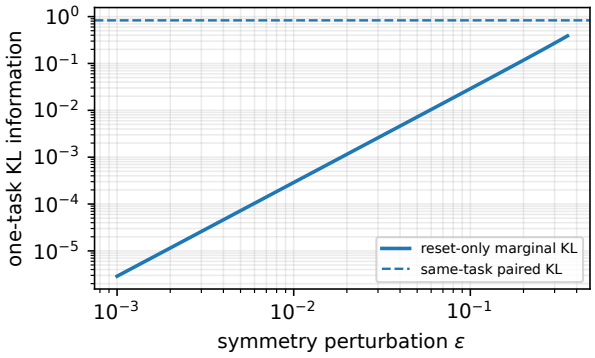


Figure 5: Approximate symmetry does not remove the value of persistence. In the binary group example, reset-only marginal information vanishes quadratically with the symmetry perturbation, while same-task paired information remains constant.

W.3 Source-to-target support shift

Let a compact group G act continuously by isometries on a metric space $(\mathcal{X}, d)$. Let μ and ν be source and target anchor laws, with support kernels

$$K_\mu = \bigcap_{x \in \mathrm{supp}(\mu)} \mathrm{Stab}_G(x), \qquad K_\nu = \bigcap_{x \in \mathrm{supp}(\nu)} \mathrm{Stab}_G(x). \tag{W.10}$$

Theorem W.5 (Exact support-shift criterion)**.** *The target map*

$$(gK_\mu, x) \longmapsto g \cdot x, \qquad x \in \mathrm{supp}(\nu),$$

is well defined if and only if

$$K_\mu \subseteq K_\nu. \tag{W.11}$$

If the inclusion fails, then no source procedure whose experiment identifies at most the quotient gK_μ can achieve zero worst-case squared target prediction risk. More precisely,

$$\inf_{\widehat{Y}} \sup_{g \in G} \mathbb{E}_g d^2(\widehat{Y}, gX_\star) \geq \frac{1}{4} \sup_{k \in K_\mu} \mathbb{E}_{X_\star \sim \nu} d^2(X_\star, kX_\star). \tag{W.12}$$

The right-hand side is strictly positive whenever $K_\mu \not\subseteq K_\nu$.

Proof. If $K_\mu \subseteq K_\nu$ and $gK_\mu = g'K_\mu$, then $g' = gk$ for some $k \in K_\mu$. Every such k fixes every point in $\mathrm{supp}(\nu)$, so $g'x = gx$ for all target anchors; the map is well defined. Conversely, if the map is well defined, compare the cosets K_μ and $kK_\mu = K_\mu$ for any $k \in K_\mu$. Well-definedness implies $kx = x$ for every target support point, so $k \in K_\nu$.

For the lower bound, fix $k \in K_\mu$ and compare parameters g and gk. They induce identical source experiments because k acts trivially on the source support. For any target predictor, averaging the squared losses under the two indistinguishable parameters and using $d^2(a,u) + d^2(a,v) \geq \frac{1}{2} d^2(u,v)$ gives

$$\max\{R_g, R_{gk}\} \geq \frac{1}{4} \mathbb{E} d^2(gX_\star, gkX_\star) = \frac{1}{4} \mathbb{E} d^2(X_\star, kX_\star).$$

Take the supremum over $k \in K_\mu$. If inclusion fails, some $k \in K_\mu$ moves a point in $\mathrm{supp}(\nu)$. Continuity gives a neighborhood on which the displacement is bounded away from zero; the support definition gives that neighborhood positive ν-mass, so the expectation is positive. □

These results show that persistence hierarchies, reset-only inefficiency, and quotient transfer survive quantitative perturbations and source-target shift.