DECISION-SUFFICIENT REPRESENTATIONS FOR PARTIALLY OBSERVED CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

A latent state for control must decide what to forget. Reconstruction spends capacity on every background detail; passive prediction discards any cue it cannot forecast, even one that changes the best action. A latent state should instead retain the distinctions that alter controlled reward consequences, and only those. We make this precise with the *decision Hankel matrix*, whose rows are histories and whose columns are reward-rooted, observation-contingent policy trees. Its rank is the minimal dimension of any exact linear representation of these consequences; a basis of its columns yields a decision code from which every trusted tree value, the optimal action values, and the complete optimal-action set are exact functions; and equality of its rows is the coarsest quotient that preserves every trusted return. At finite capacity, a taskwise decision rate–distortion objective admits a finite-sample oracle inequality and a deployment-regret guarantee, and its zero-regret rate equals the log-size of a minimum reachable action-profile cover, attained by an encoder that ignores nuisance entirely. In exact computations on finite POMDPs, the decision rank stays at 4 while a decision-null nuisance inflates the belief span 512-fold, rises by exactly one for an action-changing cue that observation prediction aliases, and, among 4.2 million encoders, regret first vanishes at four codes, reached by a single nuisance-free encoder. Compress what is decision-null, not what is hard to predict.

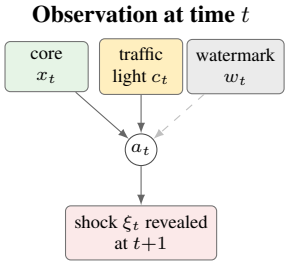
1 INTRODUCTION

Consider two bits observed by a control agent: a traffic-light cue and a video watermark. Both are fair coins, independent of everything the agent has seen, and therefore equally impossible to predict. Yet the traffic light decides whether the agent should stop, whereas the watermark changes no reward and no action. A third bit, a road shock that the agent feels only after it has acted, changes the reward but cannot inform the decision at all. A latent state for control should keep the first bit, drop the second, and never pretend to know the third.

The dominant representation objectives each get one of these bits wrong (Figure 1). Latent world models trained to reconstruct observations (Ha & Schmidhuber, 2018; Hafner et al., 2020; 2025) must keep the watermark, and a high-entropy distractor can absorb their capacity (Fu et al., 2021; Wang et al., 2022). Predictive state representations (Littman et al., 2001) keep exactly what forecasts future observations, so they alias the traffic light, which forecasts nothing; an observation-complete PSR need not even represent reward (Baisero & Amato, 2021), and state-only or inverse mutual-information objectives can be optimal while merging states that require different actions (Rakelly et al., 2021). Bisimulation (Ferns et al., 2011; Zhang et al., 2021) compares hidden states, so it demands the shock, which no history can reveal. The information bottleneck (Tishby et al., 1999) leaves the relevance variable open. None of them is rooted in what the agent's actions can achieve.

We ask *what a latent state must remember to support a trusted family of decisions*, and answer by rooting prediction in controlled reward consequences. In a finite POMDP, observation-contingent policy trees generate value vectors by alternating immediate rewards with action–observation Bellman backups. Evaluating these vectors at every history produces the *decision Hankel matrix*. This single object determines the structure of decision-sufficient memory. Its rank d is the smallest dimension of any exact linear representation of controlled reward tests. A basis of d actual tree columns gives a decision code that determines every trusted tree value, the optimal action values, and the complete set

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.



	watermark decision-null	**traffic light** observed, changes $a^\star$	**shock** revealed after acting
reconstruction belief span	**keeps it** $32 \to 4096$ as k: $8 \to 10$	keeps it	not encoded
observation prediction predictive state (PSR)	drops it rank 4	**aliases it** rank stays 4	not encoded
state similarity bisimulation	drops it 4 classes	keeps it $4 \to 8$ classes	**demands it** $4 \to 8$ classes
decision Hankel (ours) $d = \operatorname{rank} D_\Pi$	**drops it** $d = 4$ for every k	**keeps it** d: $4 \to 5$	**not encoded** $d = 4$

Figure 1: **What a latent state should remember.** Three independent fair coins reach a control agent: a decision-null watermark, an observed traffic light that changes the optimal action, and a shock revealed only after acting. Rows report what each objective does with each coin in the exact computations of Section 5 (k: nuisance size). Green marks what a controller needs; only the decision Hankel rank gets all three right.

of optimal actions. Equal rows are precisely the histories that can be merged without changing any trusted return. An observed cue that changes a decision forces a new coordinate, while randomness revealed only after acting enters through its conditional reward consequences.

Exact sufficiency is the zero-distortion endpoint of a broader principle. Real encoders have finite codebooks and learn from finite data, so we score a shared encoder by the action regret of its best task-specific head, a *taskwise decision rate–distortion* objective. With independent target and evaluation samples it admits a finite-sample oracle inequality whose price for a class of millions of encoders is only logarithmic, and under occupancy coverage its action deficiency lifts to sequential deployment regret. Under a recursive core–nuisance factorization, the minimum rate of a zero-regret encoder equals the log-size of a minimum reachable action-profile cover, and an encoder that ignores the nuisance entirely attains it.

On finite POMDPs every quantity in the theory can be computed exactly, so we can display the decision code itself. Our contributions are:

1. **Decision Hankel theory.** The rank of the decision Hankel matrix is the minimal dimension of an exact linear control-sufficient history representation; its code determines optimal planning, its rows give the coarsest safe quotient, and it separates observed cues from unrevealed randomness (Section 3).
2. **Finite capacity and finite data.** A taskwise decision rate–distortion objective with a finite-sample oracle inequality and a deployment-regret bound, each with explicit sample, target-error, class-complexity, and coverage dependence (Section 4).
3. **The exact price of relevance.** The minimum zero-regret rate equals $\log m_{\mathrm{dec}}^{\mathrm{reach}}$, the log-size of a minimum reachable action-profile cover, and a core-only encoder attains it (Section 4).
4. **The decision code, computed exactly.** On finite POMDPs the decision code is 1024 times smaller than the reconstruction target, keeps the cue that observation prediction aliases, and among 4.2 million encoders reaches zero regret with four nuisance-free codes (Section 5).

2 Related Work

Predictive states and control sufficiency. Predictive state representations replace a latent generative state by predictions of action-conditioned future observations (Littman et al., 2001), and recent frameworks connect such features to planning under partial observability (Zhang et al., 2023; 2024). Observation prediction does not by itself guarantee reward prediction: Baisero & Amato (2021) give a necessary-and-sufficient reward column-space condition and introduce reward-predictive PSRs, and mutual-information analyses separate objectives sufficient for state prediction from objectives sufficient for control (Rakelly et al., 2021). We change the representation object itself. The columns of our Hankel matrix are reward-rooted controlled policy-tree consequences, so control sufficiency,

minimality, and cue retention follow from the constructed span rather than from an auxiliary decoder condition.

Value-directed abstraction. Classical MDP abstraction studies state equivalence, model minimization, and planning-preserving quotients (Givan et al., 2003; Li et al., 2006), and belief-state planning in POMDPs is built on the same value backups we use (Kaelbling et al., 1998). Value-directed POMDP compression closes a linear subspace under reward and action–observation operators (Poupart & Boutilier, 2002). Bisimulation metrics and DeepMDP control value error through reward and transition discrepancies (Ferns et al., 2011; Gelada et al., 2019; Zhang et al., 2021), and approximate information states provide a general reward and next-information-state recursion (Subramanian et al., 2022). Our theorem works on histories and a declared policy family; it gives a matching rank lower bound and separates observed action-changing cues from unrevealed future randomness, which state-level bisimulation cannot distinguish.

Decision-aware models and capacity. Value equivalence and proper value equivalence characterize the model details that a family of policies and value functions can ignore (Grimm et al., 2020; 2021), and MuZero-style agents learn models that predict only rewards, values, and policies (Schrittwieser et al., 2020). Value-equivalent sampling casts model compression as a rate–distortion problem with regret consequences (Arumugam & Van Roy, 2022), and the comparison of experiments supplies the broader decision-theoretic view (Blackwell, 1953; Le Cam, 1964). For history representations we add the exact minimal dimension, a finite-class learner with explicit target errors and a deployment bound, and the exact zero-regret rate.

Latent world models and distractors. Latent world models learn compact predictive dynamics for imagination-based control (Ha & Schmidhuber, 2018; Hafner et al., 2020; Hansen et al., 2022; Hafner et al., 2025). Task-informed and denoised world models introduce reward, controllability, or factorization structure to separate task dynamics from distractors (Fu et al., 2021; Wang et al., 2022); posterior-predictability regularization makes the same tension explicit (Zhu et al., 2023); and analyses of reconstruction and latent self-prediction show how auxiliary objectives interact with distractions (Voelcker et al., 2024). Our theory gives the criterion these methods aim at: a distractor can be removed when it leaves every trusted controlled reward consequence unchanged, whatever its complexity or predictability, and an observed cue must be kept when it changes one.

3 Decision-Rooted Representations

3.1 Setting

We consider a finite POMDP with state set $\mathcal{S}$, action set $\mathcal{A}$, observation set $\mathcal{O}$, bounded mean-reward vectors $r_a \in \mathbb{R}^{|\mathcal{S}|}$, action–observation operators $[G_{a,o}]_{s,s'} = \Pr(S_{t+1} = s', O_{t+1} = o \mid S_t = s, A_t = a)$, and discount $\gamma \in (0, 1)$. Decisions are made from pre-action histories h in a covered set $\mathcal{H}_{\mathrm{cov}}$, and each history induces a belief $b_h \in \mathbb{R}^{|\mathcal{S}|}$. A nonempty set of trusted actions generates the *trusted policy trees*: the zero terminal $\perp$, and every tree $\sigma = (a, (\sigma_o)_{o\in\mathcal{O}})$ that takes a trusted action a and continues with subtree σ_o after observation o. Starting from $\alpha_\perp = 0$, each tree has a value vector

$$\alpha_\sigma = r_a + \gamma \sum_{o\in\mathcal{O}} G_{a,o}\alpha_{\sigma_o}. \tag{1}$$

The entry $b_h^\top \alpha_\sigma$ is the expected discounted reward of executing σ from history h, the value of a counterfactual plan. Let Σ_{fin} contain every finite trusted tree and $\mathcal{W}_\Pi = \mathrm{span}\{\alpha_\sigma : \sigma \in \Sigma_{\mathrm{fin}}\}$.

Definition 1 (Decision Hankel matrix and decision code)**.** *The decision Hankel matrix is $D_\Pi : \mathcal{H}_{\mathrm{cov}} \times \Sigma_{\mathrm{fin}} \to \mathbb{R}$, $D_\Pi(h, \sigma) = b_h^\top \alpha_\sigma$. Let $d = \mathrm{rank}(D_\Pi)$ and select d actual tree columns $\sigma_1, \ldots, \sigma_d$ that form a column basis. The* decision code *is $z_D(h) = (D_\Pi(h, \sigma_1), \ldots, D_\Pi(h, \sigma_d)) \in \mathbb{R}^d$.*

Reconstruction asks a representation to recover b_h; observation prediction asks it to recover the predictions of future observation sequences. The decision Hankel matrix asks only for the reward consequences of the agent's own options. Because $D_\Pi = [b_h^\top]_h\,[\alpha_\sigma]_\sigma$ factors through beliefs and through $\mathcal{W}_\Pi$,

$$d \;\le\; \min\big\{\dim \mathrm{span}\{b_h : h \in \mathcal{H}_{\mathrm{cov}}\},\; \dim \mathcal{W}_\Pi\big\} \;\le\; |\mathcal{S}|, \tag{2}$$

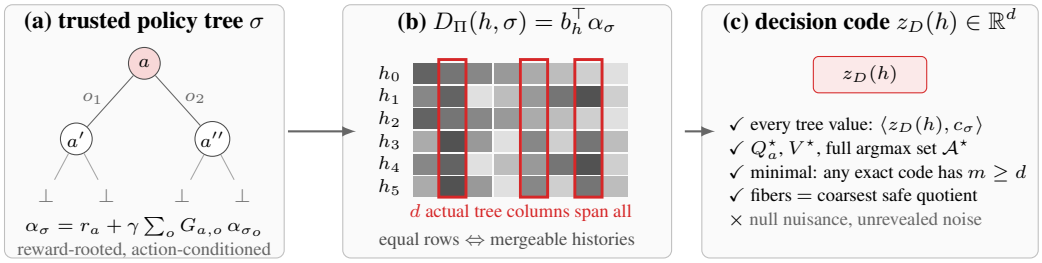


Figure 2: **The decision Hankel construction.** (a) A trusted policy tree is evaluated by alternating immediate rewards with action–observation backups. (b) Evaluating every tree at every history gives D_Π; d actual tree columns span all columns, and histories with equal rows (same shading) can be merged safely. (c) The resulting d-dimensional decision code determines every trusted tree value and the complete optimal-action set, and no exact linear code is smaller.

and the first gap can be arbitrarily large: in Section 5 the belief span is 4096 while $d = 4$. Figure 2 summarizes the construction.

3.2 Exact control sufficiency and minimality

Theorem 1 (Decision Hankel characterization)**.** *Assume finite POMDP primitives, bounded discounted rewards, trusted-continuation completeness, target-action coverage, and population-exact controlled tests (Assumptions 6–11 in Appendix B.1). Then:*

(i) ***Closure.*** *$r_a \in \mathcal{W}_\Pi$ and $G_{a,o}\mathcal{W}_\Pi \subseteq \mathcal{W}_\Pi$ for every trusted a and every o; hence $d \leq \dim \mathcal{W}_\Pi \leq |\mathcal{S}|$.*

(ii) ***Exact decoding.*** *Every finite trusted tree has a linear decoder, $D_\Pi(h, \sigma) = \langle z_D(h), c_\sigma \rangle$, and every fixed root-independent infinite continuation plan π has one, $V_\pi(h) = \langle z_D(h), c_\pi \rangle$, with truncation error at most $R_{\max}\gamma^L/(1-\gamma)$ after L reward stages.*

(iii) ***Planning.*** *On the reachable code image, the forced-action values $Q_a^\star$, the optimal value $V^\star$, and the complete optimal-action set $\mathcal{A}^\star$ are exact functions of z_D, with no action-gap or tie-breaking assumption.*

(iv) ***Minimality.*** *If $\phi : \mathcal{H}_{\text{cov}} \to \mathbb{R}^m$ linearly decodes every finite controlled reward test, then $m \geq d$ and $z_D = B\phi$ for some matrix B. The bound is attained by z_D.*

The proof has four ingredients. Policy-tree induction identifies equation 1 with the controlled discounted return, including at roots where the behavior policy is silent. A one-child graft compares two trees that differ in a single branch and exposes exactly $\gamma G_{a,o}\alpha_\sigma$, which yields the invariance in (i). The evaluation map $T\alpha = (b_h^\top \alpha)_h$ then gives $d = \dim T(\mathcal{W}_\Pi)$ and a basis of actual columns. Finally, a uniform geometric tail places every infinite continuation in the same closed span, and (iii) follows memberwise through suprema and argmax sets. Appendix B.5 gives the complete proofs.

Corollary 2 (Coarsest safe quotient and the information boundary)**.** *$z_D(h) = z_D(h')$ if and only if $D_\Pi(h, \cdot) = D_\Pi(h', \cdot)$, and this row equivalence is the coarsest quotient of histories that preserves every trusted return. Any observed cue that changes a trusted fixed-continuation value or the optimal-action set forces distinct codes. A future variable with no action-time history-measurable version cannot be recovered from z_D; only its conditional reward consequences are represented.*

Corollary 2 gives nuisance an operational definition. High entropy, unpredictability, or exogeneity neither makes a distinction removable nor makes it essential. A distinction is removable exactly when it lies in the annihilator of every trusted controlled reward consequence, and it is essential exactly when some trusted option can tell the difference.

Algorithm 1 Decision core of a finite POMDP

Require: reward vectors $\{r_a\}$, operators $\{G_{a,o}\}$ for trusted a, reachable beliefs $\{b_h\}_{h\in\mathcal{H}_{\mathrm{cov}}}$
$W \leftarrow \mathrm{orth}\{r_a\}$ ▷ reward-rooted seed
repeat
  $W \leftarrow \mathrm{orth}\big(W \cup \{G_{a,o}w : w \in W,\ a, o\}\big)$ ▷ Krylov closure, Eq. equation 3
until $\dim W$ does not change
$d \leftarrow \mathrm{rank}\big([b_h^\top]_h\, W\big)$ ▷ decision rank
select d trees of depth $\le \dim W$ whose Hankel columns are independent (pivoted QR)
return $z_D(h) = (b_h^\top \alpha_{\sigma_1}, \ldots, b_h^\top \alpha_{\sigma_d})$

3.3 Computing the decision core

Theorem 1(i) makes the decision core computable by a Krylov closure. Let $\mathcal{W}_L$ be the span of the value vectors of trusted trees of depth at most L. By equation 1 and the one-child graft,

$$\mathcal{W}_0 = \{0\}, \qquad \mathcal{W}_{L+1} = \mathrm{span}\{r_a\}_a + \textstyle\sum_{a,o} G_{a,o}\mathcal{W}_L. \tag{3}$$

An induction shows $\mathcal{W}_L \subseteq \mathcal{W}_{L+1}$, and once two consecutive spaces coincide the sequence is constant. The dimension therefore stabilizes at $\mathcal{W}_\Pi$ after at most $\dim \mathcal{W}_\Pi \le |\mathcal{S}|$ steps.

Proposition 3 (Finite depth suffices)**.** *The Hankel matrix restricted to trusted trees of depth at most $L^\star = \dim \mathcal{W}_\Pi$ has rank d, and d of its columns form a decision core.*

Algorithm 1 states the resulting procedure; every step is exact linear algebra.

4 Finite-Capacity Decision Rate–Distortion

The decision code is exact, but a learned encoder has a finite codebook and sees finite data. We therefore measure an encoder by the decisions it supports rather than by what it reconstructs. Let $\mathcal{T}$ be a finite trusted task family that shares one controlled observation system and has rewards in $[0, R_{\max}]$. Task τ has its own reachable scoring law μ_τ, n_τ independent evaluation histories, and a target estimate $\widehat{Q}_\tau$ with certificate $\|\widehat{Q}_\tau - Q^\star_\tau\|_{\infty,\mathrm{supp}(\mu_\tau)\times\mathcal{A}} \le \eta_\tau$. A finite shared encoder class Φ_M satisfies $|\,\mathrm{range}(\phi)| \le M$, and $\mathcal{G}_\phi$ is a finite class of task heads that map codes to actions. We define rate, taskwise distortion, and family distortion as

$$\begin{aligned} R(\phi) &= \log|\,\mathrm{range}(\phi)|, \qquad D(\phi) = \max_{\tau\in\mathcal{T}} D_\tau(\phi), \\ D_\tau(\phi) &= \min_{g\in\mathcal{G}_\phi} \mathbb{E}_{H\sim\mu_\tau}\Big[\max_a Q^\star_\tau(H,a) - Q^\star_\tau\big(H, g(\phi(H))\big)\Big]. \end{aligned} \tag{4}$$

The empirical selector minimizes $R(\phi) + \lambda \max_\tau \widehat{D}_\tau(\phi)$, where $\widehat{D}_\tau$ replaces μ_τ and $Q^\star_\tau$ by the evaluation sample and $\widehat{Q}_\tau$. Predictability does not enter equation 4; a code is charged only for the actions it gets wrong. Because distortion is the regret of the *best* head, the optimal action values need not factor through the code, and the family maximum makes one shared code serve every declared task. The multiplier λ prices regret in nats: the selector moves to a larger codebook only when the regret it removes exceeds the added rate divided by λ.

4.1 Finite-sample oracle inequality and deployment

Let $B = R_{\max}/(1-\gamma)$, $N_{\mathrm{cls}} = \sum_\tau \sum_{\phi\in\Phi_M} |\mathcal{G}_\phi|$, and

$$\epsilon_\tau = B\sqrt{\frac{\log(2N_{\mathrm{cls}}/\delta)}{2n_\tau}}, \qquad e_\tau = \epsilon_\tau + 2\eta_\tau, \qquad e_{\max} = \max_\tau e_\tau. \tag{5}$$

Theorem 4 (Finite-sample decision rate–distortion)**.** *Fix an independently obtained target realization satisfying the certificates. With probability at least $1-\delta$ over the evaluation histories, uniformly over encoders and heads $|\widehat{L}_\tau - L_\tau| \le e_\tau$, and*

$$R(\widehat{\phi}) + \lambda D(\widehat{\phi}) \le \inf_{\phi\in\Phi_M} \{R(\phi) + \lambda D(\phi)\} + 2\lambda e_{\max}. \tag{6}$$

Each selected task head has population action regret at most $D_\tau(\widehat{\phi}) + 2e_\tau$. If every candidate deployment occupancy satisfies $d_\tau^\pi \ll \mu_\tau$ with density at most C_τ, then

$$V_\tau^\star - V_\tau^{\widehat{\pi}_\tau} \le \frac{C_\tau}{1-\gamma}\left(D_\tau(\widehat{\phi}) + 2e_\tau\right). \tag{7}$$

The bound separates three sources of error, each with an explicit constant. Sampling enters through ϵ_τ, which grows only with $\sqrt{\log N_{\mathrm{cls}}}$: the 4.2 million encoders of Section 5 cost $\log N_{\mathrm{cls}} = 53.8$. Target error enters additively as $2\eta_\tau$ and never compounds with sampling. Distribution shift enters once, through the coverage constant C_τ, and only in the deployment statement. The proof combines one finite-class concentration event with a full-history performance-difference identity (Appendix C.5).

4.2 The exact price of relevance

What is the smallest codebook that loses no decision at all? Suppose each task-reachable history decomposes recursively as a core X and a nuisance N, and task reward and next-core kernels depend on (X, A) but not on N. At a core value x, a task *profile* $u \in \mathcal{A}^{\mathcal{T}}$ assigns one action to each task; it is *legal* at x if u_τ is optimal for every task active at x.

Definition 2 (Reachable action-profile cover)**.** *A set $\mathcal{C}$ of profiles is a reachable cover if every reachable core value has a legal profile in $\mathcal{C}$. The* decision cover number $m_{\mathrm{dec}}^{\mathrm{reach}}$ *is the minimum size of such a cover.*

The cover number counts the distinct joint decisions the task family requires. It can be far smaller than the number of core states, since states that share a legal joint action need not be told apart, and it is blind to the nuisance by construction.

Theorem 5 (Exact reachable nuisance quotient)**.** *Under the core–nuisance factorization, $Q_\tau^\star(h, a) = q_\tau^\star(X(h), a)$ on the task-reachable domain. If, at almost every core value, the nuisance laws of the tasks active there are mutually absolutely continuous, and Φ_M contains a core-only encoder realizing a minimum cover of size $m_{\mathrm{dec}}^{\mathrm{reach}} \le M$ together with its coordinate heads, then*

$$\min_{\phi \in \Phi_M:\, D(\phi)=0} R(\phi) = \log m_{\mathrm{dec}}^{\mathrm{reach}},$$

the minimum is attained by a core-only encoder, and every zero-deficiency encoder has at least $m_{\mathrm{dec}}^{\mathrm{reach}}$ active codes.

Theorem 5 turns "ignore the distractor" from a heuristic into an optimality statement: no encoder in the class, however it uses the nuisance, beats the core-only cover at zero regret. Conversely, zero regret forces every active code to carry a legal profile, so the codes of any zero-regret encoder already form a cover.

The three theorems apply one criterion at three resolutions: exact linear dimension, finite-sample rate–distortion, and the zero-regret rate. Their premises are separate; the finite-sample guarantee uses no exact controlled tests, and the nuisance quotient uses no sampling assumptions.

5 The Decision Code at Work

We now put the three coins of Figure 1 and the capacity trade-off of Section 4 into concrete finite POMDPs. Working with finite models lets us compute every quantity exactly—belief spans, predictive-state ranks, bisimulation classes, d, $m_{\mathrm{dec}}^{\mathrm{reach}}$, and population regrets over complete encoder classes—so each figure shows the objects themselves rather than estimates of them. Table 1 collects the main quantities; Appendix A gives every instance and number.

5.1 What the decision core keeps

The first family couples a controlled core (four states, three actions, $\gamma = 0.9$) with an observed nuisance of k values redrawn i.i.d. at every step. The nuisance never touches reward or the core: it is the watermark of Figure 1. Two variants add the other two coins. The *cue* is an i.i.d. bit observed before acting that changes the optimal action at one core state. The *shock* is an i.i.d. bit that changes the reward of every action and is revealed only in the next observation; knowing it in advance would change the optimal action at every core state.

Table 1: The decision code at a glance: each phenomenon, the result that describes it, and its exact size in our instances.

Phenomenon	Result	Exact quantity
decision-null nuisance	Cor. 2	$d = 4$ for all k; belief span $8 \to 4096$
observed action-changing cue	Cor. 2	$d : 4 \to 5$; observation PSR stays at 4
shock revealed after acting	Cor. 2	$d = 4$; bisimulation $4 \to 8$
decoding from d actual trees	Thm. 1	error $\leq 1.3 \times 10^{-14}$; $d-1$ columns: ≥ 0.17
depth of the decision core	Prop. 3	rank d from depth 2
zero-regret rate	Thm. 5	$M = 4$ over 4,213,660 encoders
minimum-rate encoder	Thm. 5	unique, core-only
finite-sample calibration	Thm. 4	deviation $\propto n^{-0.49}$
deployment under shift	Thm. 4	regret 0.51–0.96 for $C_\tau \leq 136$
enlarging the task family	Thm. 5	$m_{\text{dec}}^{\text{reach}} : 4 \to 6$; classes $7 \to 11$

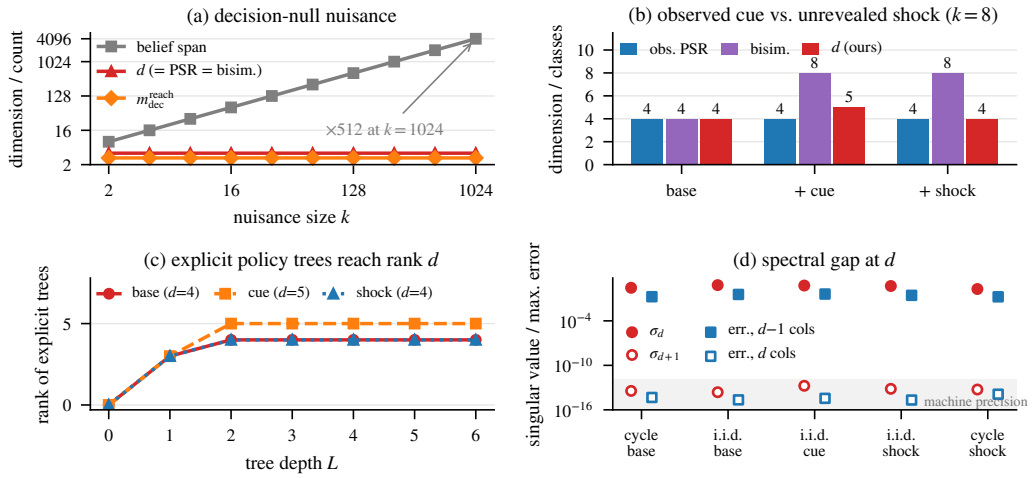


Figure 3: **What the decision code keeps** (Theorem 1, Corollary 2, Proposition 3). (a) Under an i.i.d. decision-null nuisance of size k, the belief span grows 512-fold while d (equal here to the observation-PSR rank and the bisimulation count) and $m_{\text{dec}}^{\text{reach}}$ stay constant. (b) At $k = 8$, the observed cue raises d by one and is aliased by the observation PSR; the unrevealed shock leaves d unchanged, while bisimulation splits on both. (c) Explicit policy-tree Hankel matrices (300 random trusted trees per depth) reach rank d at depth two. (d) Across five instances, σ_{d+1} and the decoding error of d tree columns sit at machine precision; $d-1$ columns fail.

Decision-null nuisance is free. As k grows from 2 to 1024, the reachable belief span, which an exact reconstruction must preserve, grows from 8 to 4096, whereas d stays at 4 and $m_{\text{dec}}^{\text{reach}}$ at 3 (Figure 3a; Table 2 in Appendix A lists every representation). The decision code is 1024 times smaller than the reconstruction target at $k = 1024$, and the ratio grows without bound.

Action-changing cues are kept; unrevealed shocks are not. The cue raises d from 4 to exactly 5 and separates the two cue values. The observation PSR stays at 4 and aliases them, because the cue predicts no future observation; reward-predictive PSRs recover the cue, which shows that the reward root is what carries it. The shock leaves d at 4: no history available at decision time reveals it, so the decision code has nothing to store. Bisimulation, which compares hidden states, doubles to 8 classes in both variants and so demands information that no controller can use (Figure 3b).

The decision core is realized by actual trees. Explicit policy-tree Hankel matrices reach rank d by depth two, far below the worst case of Proposition 3. The spectrum has a clean gap: $\sigma_{d+1} \leq 1.8 \times 10^{-13}$ against $\sigma_d \geq 1.97$. Selecting d actual tree columns by pivoted QR decodes every tree

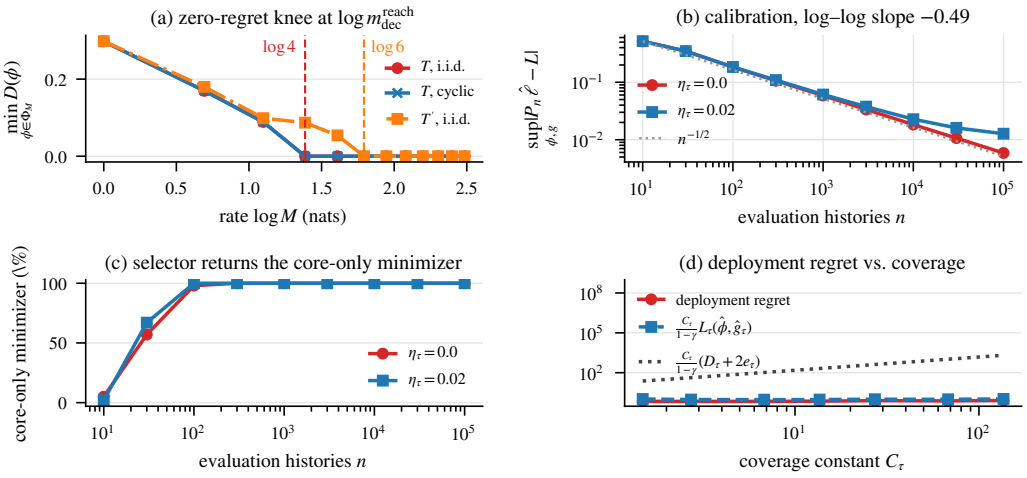


Figure 4: **Finite capacity** (Theorems 4 and 5) over 4,213,660 encoders. (a) Minimum family regret against rate reaches zero exactly at $\log m_{\text{dec}}^{\text{reach}}$ for both nuisance laws; adding the nuisance-dependent task τ_4 moves the knee from $\log 4$ to $\log 6$. (b) Uniform calibration deviation for exact ($\eta_\tau = 0$) and perturbed ($\eta_\tau = 0.02$) targets. (c) How often the empirical selector returns the core-only minimum-rate encoder ($\lambda = 10$, 100 draws per point). (d) Deployment regret, averaged over tasks, below both sides of equation 7.

value to within 1.3×10^{-14}, whereas the best $d-1$ columns leave errors of at least 0.17 (Figure 3c,d): d coordinates are enough, and no fewer will do. For all 36 instances with $k \leq 16$, brute force on the full product space reproduces every quantity.

5.2 Finite capacity and the nuisance quotient

The second family has a fully observed core $x \in \{0, \ldots, 5\}$, three actions, $\gamma = 0.9$, three tasks, and an observed nuisance $n \in \{0, \ldots, 7\}$ that affects neither reward nor the core; here $m_{\text{dec}}^{\text{reach}} = 4$. The encoder class contains all 4,213,597 partitions of the cells $(x, n \bmod 2)$, which include every core-only encoder, together with 63 encoders that use n beyond its parity, for 4,213,660 encoders in total; heads range over all code-to-action maps, so every task distortion is computed exactly.

Four codes, and none of them for the nuisance. The minimum family regret over Φ_M falls from 0.298 at $M = 1$ to 0.089 at $M = 3$ and reaches zero exactly at $M = m_{\text{dec}}^{\text{reach}} = 4$, for both an i.i.d. and a deterministic cyclic nuisance (Figure 4a). Of the 901 zero-deficiency encoders, none has fewer than four codes, and the unique one with four codes is core-only. Figure 5a shows why: the six core states collapse to four joint decisions, and the code is constant along the nuisance axis.

Finite samples recover the minimizer. The uniform calibration deviation of Theorem 4 decays with log–log slope -0.49 and levels off at the target-error floor when $\eta_\tau = 0.02$; over 3,600 evaluation draws it stays inside the radius e_τ (Figure 4b). The empirical selector returns the core-only minimum-rate encoder in 98% of draws at $n = 100$ and in every draw from $n = 300$ on (Figure 4c). As the coverage constant grows from 1.5 to 136, deployment regret stays between 0.51 and 0.96, beneath the coverage bound equation 7 (Figure 4d).

5.3 The trusted family decides what is nuisance

Adding a task τ_4 whose reward depends on $n \bmod 2$ in two core groups turns part of the nuisance into signal. The cover number rises from 4 to 6 and the knee moves with it (Figures 4a and 5b). The previous minimum-rate encoder keeps zero regret on the original tasks and incurs regret 0.189 on τ_4, entirely within the two core groups where the bit matters (16 of 48 states, Figure 5c); reselection

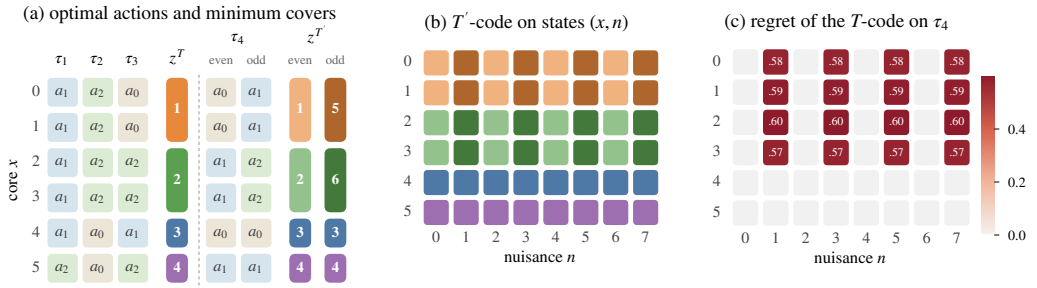


Figure 5: **Anatomy of the zero-regret code** (i.i.d. nuisance). (a) Optimal actions per core state and minimum covers: four codes for T; τ_4 splits two of them by the parity of n. (b) The T'-code on all 48 states, shaded by parent T-code. (c) The T-code loses regret on τ_4 only where parity matters.

over the same class recovers zero regret with six codes that depend on n only through its parity. The decision Hankel quotient refines from 7 to 11 classes: nuisance is defined by the declared decisions, not by the signal.

6 Implications for Latent Prediction

Predict consequences, not observations. The decision Hankel matrix names the prediction targets that suffice for control: reward-rooted, action-conditioned, observation-contingent returns. Finitely many of them span all (Proposition 3) and d of them are necessary (Theorem 1), which gives value-equivalent and reward-predictive world models (Grimm et al., 2020; Schrittwieser et al., 2020) a precise target dimension.

The trusted family is the design variable. Relevance is relative to the declared tasks and policies, and this relativity is what makes selective compression possible: a nuisance may be arbitrarily complex as long as every declared reward consequence ignores it, and enlarging the family refines the quotient at an exactly predictable rate cost.

Observability draws the line. A cue revealed before acting must be retained even when independent of the past; a shock revealed afterwards enters only through its conditional reward consequences. In the shock instance of Section 5, advance knowledge would change the optimal action at every core state, yet the decision code stores nothing for it while bisimulation spends four extra classes. Data should therefore cover observable precursors and counterfactual actions.

From exact codes to learned ones. The objective equation 4 is stated for observed histories, target estimates, and finite encoder classes, so it scores a neural encoder as it stands. The decision Hankel matrix also supplies the training signal: returns of sampled observation-contingent policy trees, with the spectrum of the resulting history-by-tree matrix indicating how many latent coordinates the declared decisions require.

7 Conclusion

A latent state for control should remember exactly what changes the consequences of its options, and the decision Hankel matrix makes this precise: its rank is the minimal exact linear memory, its basis columns determine every trusted value and the complete optimal-action set, and its rows define the coarsest safe quotient. At finite capacity, a nuisance-free encoder attains the zero-regret rate. Compress what is decision-null, not what is merely hard to predict.

AI USE STATEMENT

Generative AI tools assisted literature triage, formalization, proof drafting and checking, implementation of the exact-computation code, and manuscript drafting and editing. The authors checked cited sources against primary proceedings or official metadata and reviewed every AI-assisted proof and claim. The authors take responsibility for the final content, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The work is theoretical and computational and involves no human participants, personal data, or deployment on physical systems.

REPRODUCIBILITY STATEMENT

Appendix A specifies every instance, encoder class, and tolerance of Section 5, and Appendices B.1–C.5 state every assumption and give complete proofs of Theorems 1–5 and Corollary 2. The supplementary material contains code that regenerates every number, table, and figure of Section 5 and Appendix A on a CPU in about twenty minutes.

REFERENCES

Dilip Arumugam and Benjamin Van Roy. Deciding what to model: Value-equivalent sampling for reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2206.02072.

Andrea Baisero and Christopher Amato. Reconciling rewards with predictive state representations. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 2170–2176, 2021. doi: 10.24963/ijcai.2021/299.

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

Norm Ferns, Prakash Panangaden, and Doina Precup. Bisimulation metrics for continuous markov decision processes. *SIAM Journal on Computing*, 40(6):1662–1714, 2011. doi: 10.1137/10080484X.

Xiang Fu, Ge Yang, Pulkit Agrawal, and Tommi Jaakkola. Learning task informed abstractions. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 2021. URL https://arxiv.org/abs/2106.15612.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. DeepMDP: Learning continuous latent space models for representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2019. URL https://arxiv.org/abs/1906.02736.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in markov decision processes. *Artificial Intelligence*, 147(1–2):163–223, 2003. doi: 10.1016/S0004-3702(02)00376-4.

Christopher Grimm, André Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2011.03506.

Christopher Grimm, André Barreto, Gregory Farquhar, David Silver, and Satinder Singh. Proper value equivalence. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2106.10316.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1912.01603.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025.

Nicklas Hansen, Hao Su, and Xiaolong Wang. Temporal difference learning for model predictive control. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 2022. URL https://arxiv.org/abs/2203.04955.

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

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.

Lucien Le Cam. Sufficiency and approximate sufficiency. *The Annals of Mathematical Statistics*, 35 (4):1419–1455, 1964. doi: 10.1214/aoms/1177700372.

Lihong Li, Thomas J. Walsh, and Michael L. Littman. Towards a unified theory of state abstraction for MDPs. In *International Symposium on Artificial Intelligence and Mathematics*, 2006. URL https://dblp.org/rec/conf/isaim/LiWL06.

Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. In *Advances in Neural Information Processing Systems*, volume 14, 2001. URL https://papers.nips.cc/paper/2001/hash/1e4d36177d71bbb3558e43af9577d70e-Abstract.html.

Pascal Poupart and Craig Boutilier. Value-directed compression of POMDPs. In *Advances in Neural Information Processing Systems*, volume 15, 2002. URL https://proceedings.neurips.cc/paper/2002/hash/14ea0d5b0cf49525d1866cb1e95ada5d-Abstract.html.

Kate Rakelly, Abhishek Gupta, Carlos Florensa, and Sergey Levine. Which mutual-information representation learning objectives are sufficient for control? In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2106.07278.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588 (7839):604–609, 2020.

Jayakumar Subramanian, Amit Sinha, Raihan Seraj, and Aditya Mahajan. Approximate information state for approximate planning and reinforcement learning in partially observed systems. *Journal of Machine Learning Research*, 23, 2022. URL https://arxiv.org/abs/2010.08843.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pp. 368–377, 1999. URL https://mlanthology.org/misc/1999/tishby1999misc-information/.

Claas A. Voelcker, Tyler Kastner, Igor Gilitschenski, and Amir-massoud Farahmand. When does self-prediction help? understanding auxiliary tasks in reinforcement learning. *Reinforcement Learning Journal*, 4:1567–1597, 2024. URL https://rlj.cs.umass.edu/2024/papers/Paper197.html.

Tongzhou Wang, Simon S. Du, Antonio Torralba, Phillip Isola, Amy Zhang, and Yuandong Tian. Denoised MDPs: Learning world models better than the world itself. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 2022. URL https://arxiv.org/abs/2206.15477.

Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021.

Hongming Zhang, Tongzheng Ren, Chenjun Xiao, Dale Schuurmans, and Bo Dai. Provable representation with efficient planning for partially observable reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59759–59782, 2024. URL https://proceedings.mlr.press/v235/zhang24bd.html.

Tianjun Zhang, Tongzheng Ren, Chenjun Xiao, Wenli Xiao, Joseph E. Gonzalez, Dale Schuurmans, and Bo Dai. Energy-based predictive representations for partially observed reinforcement learning. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2477–2487, 2023. URL https://proceedings.mlr.press/v216/zhang23b.html.

Chuning Zhu, Max Simchowitz, Siri Gadipudi, and Abhishek Gupta. RePo: Resilient model-based reinforcement learning by regularizing posterior predictability. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2309.00082.

A ADDITIONAL RESULTS: EXACT INSTANCES

All computations use seed 42, exact linear algebra with numerical rank tolerance 10^{-8} relative to the largest singular value, value iteration to sup-norm tolerance 10^{-13}, and exhaustive enumeration of encoder classes. The supplementary material regenerates every number and figure below.

A.1 RANK HIERARCHY

Instances. The core has $X = 4$ states and $A = 3$ actions with $\gamma = 0.9$; core transitions are Dirichlet($\mathbf{1}$) rows and mean rewards are uniform on $[0, 1]$. The nuisance has k values, is observed, and evolves independently of the core and of actions; it never enters reward or core transitions. We use three nuisance laws: i.i.d. uniform, a deterministic cycle, and a sticky chain $0.1\, I + 0.9\, k^{-1}\mathbf{1}\mathbf{1}^{\top}$. The *cue* variant adds an i.i.d. fair bit that is observed together with the core state before acting and adds a bonus of 0.6 to one of two distinct actions, depending on the bit. The *shock* variant adds an i.i.d. fair bit that shifts the reward of every action by 0.6 in an action-dependent direction and is revealed only in the next observation.

Quantities. For every instance we compute the span of reachable beliefs, the rank of observation-test predictions (observation PSR), the rank of observation-and-reward test predictions (reward PSR), the number of state-level bisimulation classes, the decision-Hankel rank d as the dimension of $T(\mathcal{W}_{\Pi})$ obtained by Krylov closure of $\{r_a\}$ under $\{G_{a,o}\}$, the number of distinct decision rows (the size of the coarsest exact return-safe quotient), and $m^{\mathrm{reach}}_{\mathrm{dec}}$ on the decision-time planning problem. Values for $k \le 16$ are computed twice, by brute force on the full product space and by a factored formula that uses the tensor structure of $\mathcal{W}_{\Pi}$; the two agree on all seven quantities for all 36 instances, and the factored formula supplies $k > 16$.

Explicit policy trees. For five instances we construct 300 random trusted trees at each depth $L = 1, \ldots, 6$, drawing each child subtree from the trees already built at smaller depth, form the explicit history-by-tree matrix, and compare its rank with the Krylov value d. The rank reaches d by depth 2 in every case, $\sigma_{d+1} \le 1.8 \times 10^{-13}$ while $\sigma_d \ge 1.97$, d actual tree columns (selected by pivoted QR) decode all tree values with error at most 1.3×10^{-14}, and $d - 1$ columns leave errors between 0.17 and 0.42. Monte-Carlo returns of three depth-3 trees from two start states each (20,000 rollouts per pair) agree with their Hankel entries, with largest standardized deviation 1.91.

Predictable nuisance. Corollary 2 forces distinct codes for any distinction that changes a trusted fixed-continuation value. When the nuisance is predictable (cycle or sticky law) and the trusted trees may branch on nuisance observations, a tree that conditions its later actions on the upcoming nuisance value has a value that depends on the current nuisance state. The decision core therefore separates nuisance states, and d grows with k together with the observation PSR (8 at $k = 2$ to 4096 at $k = 1024$). Two routes remove this dependence. A trusted family whose trees branch only on core observations keeps $d = 4$ for every k and every nuisance law, and the control-level quotient $m^{\mathrm{reach}}_{\mathrm{dec}}$ of Theorem 5 stays at 3 for the unrestricted family (Figure 6a and Table 2).

Table 2: Rank hierarchy at nuisance size $k = 8$. "belief" is the reachable belief span; "core-br. d" is the decision-Hankel rank for trusted trees that branch only on core observations.

nuisance	variant	belief	obs. PSR	rew. PSR	d	core-br. d	bisim.	quotient	$m^{\mathrm{reach}}_{\mathrm{dec}}$
i.i.d.	base	32	4	4	4	4	4	5	3
i.i.d.	cue	64	4	5	5	5	8	9	3
i.i.d.	shock	32	4	4	4	4	8	5	3
cycle	base	32	32	32	32	4	32	33	3
cycle	cue	64	32	33	33	5	64	65	3
cycle	shock	32	32	32	32	4	64	33	3

A.2 FINITE CAPACITY

Instance and class. The core $x \in \{0, \ldots, 5\}$ is fully observed, $A = 3$, $\gamma = 0.9$, and $R_{\max} = 1$. Core transitions mix a shared Dirichlet kernel (0.7) with an action-specific one (0.3). The three

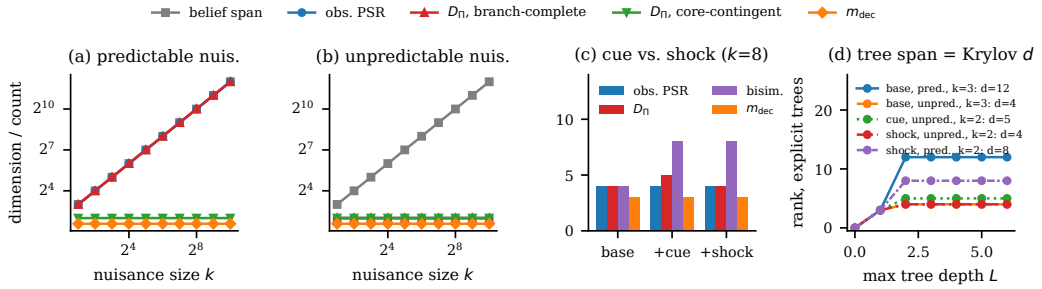


Figure 6: Complete rank hierarchy. (a) Predictable (cycle) and (b) unpredictable (i.i.d.) nuisance of size k. (c) Cue and shock variants at $k = 8$. (d) Explicit-tree rank against depth for five instances.

tasks reward, in each of four core groups $\{0,1\},\{2,3\},\{4\},\{5\}$, a group-specific best action with 0.8 and the others with 0.2, plus uniform noise of amplitude 0.1. The nuisance $n \in \{0,\dots,7\}$ is observed and follows either an i.i.d. law or a deterministic cycle. Because the system is Markov in (x,n), histories are represented by the current observation, and each scoring law μ_τ is the discounted occupancy of the uniform behavior policy. The class Φ consists of all $4{,}213{,}597$ set partitions of the 12 cells $(x, n \bmod 2)$ and 63 encoders that use n beyond its parity ($n \bmod M$, $(x+n) \bmod M$, block-plus-nuisance codes, and the identity), for $|\Phi| = 4{,}213{,}660$ and $\log N_{\text{cls}} = 53.8$. Heads range over all maps from codes to actions, so $D_\tau(\phi)$ is computed in closed form.

Capacity sweep. For both nuisance laws the zero-regret knee sits at $M = m^{\text{reach}}_{\text{dec}} = 4$ and the unique minimum-rate zero-deficiency encoder is core-only. Below the knee the minimum regret is 0.298, 0.170, and 0.089 for $M = 1,2,3$. At $M = 2$ the population-optimal encoders use the nuisance bit: under the task-maximum distortion, a decision-null bit can serve as a shared coin that balances two tasks, lowering the maximum regret from 0.184 (best core-only encoder) to 0.170. At the zero-deficiency endpoint of Theorem 5 the gap is zero, as it is at $M = 1$, $M = 3$, and every $M \geq 4$.

Sample sweep. Independent target realizations satisfy $\|\widehat{Q}_\tau - Q^\star_\tau\|_\infty \leq \eta_\tau$ with $\eta_\tau \in \{0, 0.02\}$; each setting uses 100 independent evaluation draws, for $3{,}600$ draws over $\lambda \in \{5, 10\}$, two values of η_τ, and nine sample sizes from 10 to 10^5. The uniform calibration deviation stayed inside e_τ in every draw. At $n = 10^4$ the certified radius is 0.536 and the observed deviation 0.018; the oracle margin $2\lambda e_{\max}$ exceeds the observed excess objective throughout. With $\lambda = 5$, chosen near the population switch point $\lambda^\star = 4.65$ to create a near tie, the excess objective decreases from 0.405 at $n = 10$ to 0.001 at $n = 1000$ and is zero from $n = 3000$ on.

Coverage. At $M = 3$ and $n = 5000$ (50 draws per setting), the scoring law mixes the uniform law with the uniform law on core states $\{0,1\}$, with weight $w \in [0, 0.99]$; this varies the coverage constant C_τ from 1.5 to 136. We take C_τ as the largest ratio between the maximal discounted state visitation and μ_τ, an upper bound on the candidate-uniform constant. The performance-difference identity holds to 8.0×10^{-13}. Mean deployment regret (0.51–0.96) stays below $(C_\tau/(1-\gamma))\, L_\tau(\widehat{\phi}, \widehat{g}_\tau)$ (for example 0.80 against 1.27 at $C_\tau = 1.5$) and below the full bound $(C_\tau/(1-\gamma))(D_\tau(\widehat{\phi}) + 2e_\tau)$ in every draw.

A.3 Enlarging the trusted task family

The added task τ_4 rewards, in core groups $\{0,1\}$ and $\{2,3\}$, a best action that depends on $\beta = n \bmod 2$, and agrees with a single best action in the other groups. For both nuisance laws, $m^{\text{reach}}_{\text{dec}}$ rises from 4 to 6, the knee moves from $M = 4$ to $M = 6$, the previous encoder keeps zero regret on τ_1, τ_2, τ_3 with regret 0.189 on τ_4 supported on 16 states, and the unique minimum-rate zero-deficiency encoder for T' uses six codes and depends on n only through β. With the i.i.d. nuisance, the decision-Hankel quotient refines from 7 to 11 classes and d from 6 to 8. With the cycle nuisance,

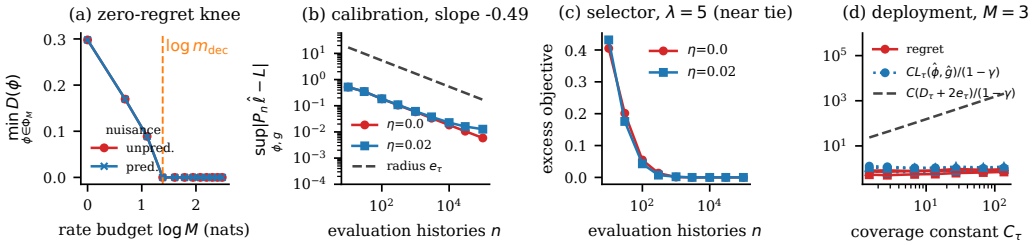


Figure 7: Complete finite-capacity results. (a) Capacity sweep for both nuisance laws. (b) Calibration deviation with the certified radius e_τ. (c) Excess objective at $\lambda = 5$. (d) Deployment regret against the coverage constant with both bounds.

the branch-complete quotient is already maximal on this instance (49 classes for both families), for the reason given in Appendix A.1.

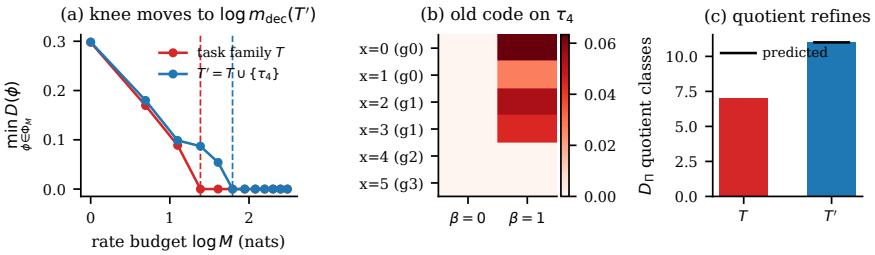


Figure 8: Enlarging the task family (i.i.d. nuisance). (a) The knee moves to $\log m_{\mathrm{dec}}^{\mathrm{reach}}(T')$. (b) Regret of the previous encoder on τ_4 by cell. (c) Decision-Hankel quotient sizes, with the size obtained by splitting each parity-dependent core value.

B EXACT DECISION-HANKEL THEORY

B.1 THEORETICAL SETUP

B.1.1 PROBLEM SETTING

Let

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, P, \mathsf{O}, r, \gamma)$$

be a finite partially observed Markov decision process (POMDP), and let the nonempty set $\mathcal{A}_{\mathrm{tr}} \subseteq \mathcal{A}$ contain the trusted actions. Decisions are made from pre-action histories h in a declared covered set $\mathcal{H}_{\mathrm{cov}}$. We write

$$b_h(s) = \Pr(S_t = s \mid H_t = h), \qquad r_a(s) = \mathbb{E}[R_t \mid S_t = s, A_t = a],$$

with beliefs represented as column vectors. For a trusted action a and observation o, the unnormalized action–observation operator is

$$[G_{a,o}]_{s,s'} = \Pr(S_{t+1} = s', O_{t+1} = o \mid S_t = s, A_t = a).$$

Thus $G_{a,o}\alpha$ is a current-state continuation-value vector for the next-observation branch o.

The deterministic finite trusted continuation trees form the support-independent, branch-complete grammar

$$\Sigma_{\mathrm{fin}} ::= \bot \quad \Big| \quad (a, (\sigma_o)_{o\in\mathcal{O}}), \qquad a \in \mathcal{A}_{\mathrm{tr}}, \quad \sigma_o \in \Sigma_{\mathrm{fin}}. \tag{T}$$

The symbol $\bot$ is a zero terminal. Define

$$\alpha_\bot = 0, \qquad \alpha_{(a,(\sigma_o)_o)} = r_a + \gamma \sum_{o\in\mathcal{O}} G_{a,o}\alpha_{\sigma_o}, \tag{R}$$

and the reward-rooted algebraic span

$$\mathcal{W}_\Pi := \mathrm{span}\{\alpha_\sigma : \sigma \in \Sigma_{\mathrm{fin}}\} \subseteq \mathbb{R}^{|\mathcal{S}|}. \tag{W}$$

Its invariance and its relation to infinite continuations are conclusions, not assumptions.

The decision Hankel matrix is the possibly infinite matrix

$$D_\Pi : \mathcal{H}_{\mathrm{cov}} \times \Sigma_{\mathrm{fin}} \to \mathbb{R}, \qquad D_\Pi(h, \sigma) = b_h^\top \alpha_\sigma. \tag{H}$$

Its algebraic rank d is the dimension of the span of its column functions on $\mathcal{H}_{\mathrm{cov}}$. A decision core is a collection of d column indices $\sigma_1, \ldots, \sigma_d$ whose column functions form a basis of that span; the empty core is used when $d = 0$. Its coordinate map is

$$z_D(h) = \big(D_\Pi(h, \sigma_1), \ldots, D_\Pi(h, \sigma_d)\big) \in \mathbb{R}^d. \tag{Z}$$

Existence of such an actual-column core is one of the conclusions below.

Let Π_∞ contain all non-anticipatory, observation-contingent infinite continuation plans supported on $\mathcal{A}_{\mathrm{tr}}$, including randomized plans, and let $\pi^{(L)}$ be the zero-terminal truncation after L reward stages. Whenever the discounted limit exists, set

$$V_\pi(h) = \lim_{L\to\infty} b_h^\top \alpha_{\pi^{(L)}}.$$

For $a \in \mathcal{A}_{\mathrm{tr}}$, let $\Pi_\infty(a)$ be the plans whose first action is forced to a, and define

$$Q_a^*(h) = \sup_{\pi\in\Pi_\infty(a)} V_\pi(h), \qquad V^*(h) = \max_{a\in\mathcal{A}_{\mathrm{tr}}} Q_a^*(h), \qquad \mathcal{A}^*(h) = \arg\max_{a\in\mathcal{A}_{\mathrm{tr}}} Q_a^*(h). \tag{P}$$

An m-dimensional exact linear controlled-test representation is a map $\phi : \mathcal{H}_{\mathrm{cov}} \to \mathbb{R}^m$ for which, for every $\sigma \in \Sigma_{\mathrm{fin}}$, a vector $c_\sigma^\phi \in \mathbb{R}^m$ satisfies

$$D_\Pi(h, \sigma) = \langle \phi(h), c_\sigma^\phi \rangle \quad \text{for every } h \in \mathcal{H}_{\mathrm{cov}}. \tag{L}$$

A quotient map η is exact return-safe when every finite controlled-test column is constant on each fiber of η. Finally,

$$h \equiv_D h' \quad \Longleftrightarrow \quad D_\Pi(h, \sigma) = D_\Pi(h', \sigma) \text{ for every } \sigma \in \Sigma_{\mathrm{fin}} \tag{E}$$

defines decision equivalence.

B.1.2 ASSUMPTIONS

Assumption 6 (Finite POMDP primitives)**.** *The sets $\mathcal{S}, \mathcal{A}, \mathcal{O}$ are finite; P and O define the joint operators $G_{a,o}$ under the convention above; every $h \in \mathcal{H}_{\mathrm{cov}}$ has the displayed posterior b_h; and rewards have the displayed state–action conditional means r_a.*

Assumption 7 (Bounded discounted rewards)**.** *The discount satisfies $0 < \gamma < 1$, and a finite $R_{\max}$ obeys*

$$|r_a(s)| \le R_{\max} \quad \textit{for all } s \in \mathcal{S},\ a \in \mathcal{A}_{\mathrm{tr}}.$$

The bound applies to conditional reward means; realized rewards may be unbounded.

Assumption 8 (Trusted continuation completeness)**.** *The target comparison contains every non-anticipatory observation-contingent continuation plan supported on $\mathcal{A}_{\mathrm{tr}}$, is closed under taking continuations after target-relevant positive-probability observation branches, and contains every forced trusted root action followed by any such continuation.*

Assumption 9 (Target-policy action coverage)**.** *The behavior policy is non-anticipatory, chooses actions from the observed history, and leaves the POMDP reward and transition–observation kernels unchanged. Every trusted action has positive behavior probability at every target-reachable covered history, and every positive-probability finite trusted action–observation extension remains in $\mathcal{H}_{\mathrm{cov}}$. No uniform lower bound is imposed. Environment-zero-probability branches and target branches outside this support are excluded from the statistical comparison.*

Assumption 10 (Exact population regime)**.** *The behavior-law conditional reward moments, observation probabilities, and next-history conditional kernels required to evaluate finite controlled reward tests are available as exact POMDP-consistent conditional versions at every covered root. At a behavior-positive root they are ordinary observational conditional quantities. At a behavior-null covered root they are supplied oracle-relative versions, which action positivity alone does not identify from the behavior joint law. All results in this section are exact population statements.*

Assumption 11 (Declared minimality class)**.** *Dimension minimality is stated over finite-dimensional history representations having exact linear decoders for every finite controlled reward test as in equation L.*

B.2 PRELIMINARIES

Write $\mu(a \mid g)$ for the behavior action kernel at a pre-action history g. The conditional versions supplied by Assumption 10 are denoted

$$m_a(g) := \mathbb{E}_\mu[R_k \mid H_k = g, A_k = a], \qquad p_a(o \mid g) := \Pr_\mu(O_{k+1} = o \mid H_k = g, A_k = a). \tag{C}$$

When $p_a(o \mid g) > 0$, let $g \oplus (a, o)$ denote the next covered pre-action history, including every policy-visible signal prescribed by the history convention. Define the finite conditional functional recursively by

$$\mathcal{J}_\perp(g) = 0, \qquad \mathcal{J}_{(a,(\sigma_o)_o)}(g) = m_a(g) + \gamma \sum_{o : p_a(o|g) > 0} p_a(o \mid g) \mathcal{J}_{\sigma_o}(g \oplus (a, o)). \tag{G}$$

The restriction to positive-probability observations prevents extrapolation of an undefined child conditional. Lemma 13 proves that a missing observation branch nevertheless contributes exactly zero to the corresponding decision-Hankel entry.

A *common randomized continuation plan* uses a root-independent observation-tree rule and an exogenous random tape Ξ_π whose law is independent of the root history, the current hidden state, and future POMDP noise. Freezing the tape fixes all observation-contingent choices simultaneously. This permits arbitrary correlation among nodewise random choices, while the plan itself is fixed independently of the root history. The finite common mixture generated by a depth-L restriction will be derived in Proposition 16.

Once the actual decision core is selected, its reachable image is

$$\mathcal{Z} := z_D(\mathcal{H}_{\mathrm{cov}}) = \{z_D(h) : h \in \mathcal{H}_{\mathrm{cov}}\} \subseteq \mathbb{R}^d. \tag{ZI}$$

The planning maps are defined on $\mathcal{Z}$.

For the final information boundary, fix any behavior, target, or coupled POMDP law $\mathbb{P}$ for which $C = \{H_t \in \mathcal{H}_{\mathrm{cov}}\}$ has positive probability, and write $\mathbb{P}_C = \mathbb{P}(\cdot \mid C)$. For a named future random element $U = U_{t+k}$, $k \geq 1$, taking values in a standard Borel space U, exact recovery from the decision core means that a Borel map $\rho : \mathbb{R}^d \to \mathsf{U}$ satisfies

$$\mathbb{P}_C\{U = \rho(z_D(H_t))\} = 1. \tag{REC}$$

This exact realization criterion is distinct from representing conditional controlled reward consequences.

B.3 Main Theorem

Theorem 12 (Decision-Hankel core characterization)**.** *Under Assumptions 6–11, the following statements hold.*

(i) ***Population semantics and algebraic closure.*** *For every supplied covered root h and every deterministic finite trusted tree σ,*

$$\mathcal{J}_\sigma(h) = b_h^\top \alpha_\sigma = D_\Pi(h, \sigma). \tag{MT1}$$

At a behavior-positive root this is an ordinary observational conditional functional; at a behavior-null covered root it is the supplied POMDP-consistent oracle-relative evaluation.

For every common randomized plan π and finite depth L, there are a finite set $\Sigma_L \subset \Sigma_{\mathrm{fin}}$ and weights $w_{\pi,L}(\sigma) \geq 0$, independent of the root history and current hidden state, with $\sum_{\sigma \in \Sigma_L} w_{\pi,L}(\sigma) = 1$, such that

$$\alpha_{\pi^{(L)}} = \sum_{\sigma \in \Sigma_L} w_{\pi,L}(\sigma)\alpha_\sigma, \qquad b_h^\top \alpha_{\pi^{(L)}} = \sum_{\sigma \in \Sigma_L} w_{\pi,L}(\sigma) D_\Pi(h, \sigma). \tag{MT2}$$

Moreover,

$$r_a \in \mathcal{W}_\Pi, \qquad G_{a,o}\mathcal{W}_\Pi \subseteq \mathcal{W}_\Pi \quad (a \in \mathcal{A}_{\mathrm{tr}},\ o \in \mathcal{O}), \tag{MT3}$$

including syntactic zero-probability branches, and every finite trusted operator composition preserves $\mathcal{W}_\Pi$ with zero defect.

(ii) ***Rank, actual core, and finite decoders.*** *The algebraic decision rank satisfies*

$$d = \mathrm{rank}(D_\Pi) \leq \dim \mathcal{W}_\Pi \leq |\mathcal{S}|. \tag{MT4}$$

There are d actual deterministic-tree columns forming a decision core, and for every $\sigma \in \Sigma_{\mathrm{fin}}$ a vector $c_\sigma \in \mathbb{R}^d$ satisfies

$$D_\Pi(h, \sigma) = \langle z_D(h), c_\sigma \rangle \quad \textit{for every } h \in \mathcal{H}_{\mathrm{cov}}. \tag{MT5}$$

When $d = 0$, the core and all coefficients are the unique empty vectors, every finite column function is zero, and the empty inner product gives the exact decoder.

(iii) ***Infinite common continuations and the exact truncation rate.*** *For every fixed root-independent common $\pi \in \Pi_\infty$, the truncation vectors converge in hidden-state ℓ_∞ norm to a vector $\alpha_\pi \in \mathcal{W}_\Pi$. There is an image coefficient $c_\pi \in \mathbb{R}^d$, independent of the covered root, such that*

$$V_\pi(h) = b_h^\top \alpha_\pi = \langle z_D(h), c_\pi \rangle. \tag{MT6}$$

For every integer $L \geq 0$,

$$\sup_{\substack{h \in \mathcal{H}_{\mathrm{cov}} \\ \pi \in \Pi_\infty}} \left| V_\pi(h) - b_h^\top \alpha_{\pi^{(L)}} \right| \leq \frac{R_{\max}\gamma^L}{1-\gamma}. \tag{MT7}$$

If π has no active reward node after stage $L-1$, the actual vector and scalar residuals are exactly zero.

(iv) ***Planning envelopes and complete ties.*** *On the reachable image $\mathcal{Z}$, there are finite maps*

$$q_a(x) = \sup_{\pi \in \Pi_\infty(a)} \langle x, c_\pi \rangle, \qquad v(x) = \max_{a \in \mathcal{A}_{\mathrm{tr}}} q_a(x), \qquad A(x) = \arg\max_{a \in \mathcal{A}_{\mathrm{tr}}} q_a(x), \tag{MT8}$$

satisfying

$$q_a(z_D(h)) = Q_a^*(h), \qquad v(z_D(h)) = V^*(h), \qquad A(z_D(h)) = \mathcal{A}^*(h). \tag{MT9}$$

The inner supremum need not be attained, and A retains the complete tie set. Each fixed-plan member is linear; a rule that chooses a different common plan from the full root history receives a single coefficient under an additional simultaneous compatibility condition.

(v) ***Minimality in the declared exact-linear class.*** *For every m-dimensional representation ϕ satisfying equation L, with scalar coordinate functions $\phi_1, \dots, \phi_m$,*

$$d \le \dim \operatorname{span}\{\phi_1, \dots, \phi_m\} \le m, \tag{MT10}$$

and a matrix $B \in \mathbb{R}^{d\times m}$ satisfies $z_D(h) = B\phi(h)$ for every covered history. The core map itself is an exact d-dimensional controlled-test representation, so the lower bound is attained. If $d = |\mathcal{S}|$, then every representation in the declared exact-linear class has $m \ge |\mathcal{S}|$; this is a dimension bound, and decision-equivalent histories may still be merged.

(vi) ***Coarsest safe quotient, cue retention, and the information boundary.*** *For all covered histories,*

$$z_D(h) = z_D(h') \quad \Longleftrightarrow \quad h \equiv_D h'. \tag{MT11}$$

The quotient by $\equiv_D$ is exact return-safe, every exact return-safe map has each fiber contained in one $\equiv_D$-class, and z_D realizes exactly these coarsest safe fibers. Thus decision-null nuisance variation may be removed.

If one fixed common trusted continuation has $V_\pi(h) \ne V_\pi(h')$, then some finite deterministic controlled-test column differs at the two histories and hence $z_D(h) \ne z_D(h')$; a forced root action is preserved by this witness. Likewise,

$$\mathcal{A}^*(h) \ne \mathcal{A}^*(h') \quad \Longrightarrow \quad z_D(h) \ne z_D(h'), \tag{MT12}$$

with no action-gap or tie-breaking assumption.

Finally, fix any behavior, target, or coupled POMDP law $\mathbb{P}$ for which $C = \{H_t \in \mathcal{H}_{\mathrm{cov}}\}$ has positive probability, set $\mathbb{P}_C = \mathbb{P}(\cdot \mid C)$, and let $U = U_{t+k}$, $k \ge 1$, take values in a standard Borel space U. If U has no $\sigma(H_t)$-measurable version under $\mathbb{P}_C$, no Borel $\rho : \mathbb{R}^d \to \mathsf{U}$ can satisfy $\mathbb{P}_C\{U = \rho(z_D(H_t))\} = 1$. Nevertheless the fixed-plan values, forced-action envelopes, optimal value, and complete optimal-action set in equation MT6 and equation MT9 remain exact functions of $z_D(H_t)$. The code therefore represents posterior effects on controlled reward consequences, not an unrevealed future realization or its full posterior law.

B.4 Proof Sketch

The proof begins with the population semantics of finite controlled trees. Lemma 13 fixes the operator orientation, posterior update, and zero-probability branch identity. Proposition 14 then applies well-founded induction to identify the conditional functional with each decision-Hankel entry. Proposition 15 supplies the equivalent stopped sequential-reweighting formula, while Proposition 16 freezes a root-independent random tape to represent every finite randomized continuation as a common finite mixture of deterministic trees.

The structural span is handled independently of observational support. Lemma 17 puts every trusted reward vector in $\mathcal{W}_\Pi$. Lemma 18 compares an all-terminal tree with a one-child graft, exposing exactly $\gamma G_{a,o}\alpha_\sigma$. Since $\gamma > 0$, Proposition 19 gives operator invariance, and Proposition 20 gives exact finite-word closure.

For the finite core, Lemma 21 maps $\mathcal{W}_\Pi$ onto the algebraic decision-column span. Proposition 22 bounds the image dimension without requiring injectivity. Lemma 23 selects a basis from actual deterministic columns, and Proposition 24 produces exact image coordinates, including the empty-core case.

The common finite mixture places every randomized truncation in $\mathcal{W}_\Pi$ by Proposition 25. Lemma 26 couples all truncations through one plan and tape and exposes the geometric tail. Proposition 27 proves that the limit stays inside the finite-dimensional span. Propositions 28 and 29 then give the exact fixed-plan decoder and the uniform scalar certificate.

Planning is performed only after every fixed common plan has its coefficient. Lemma 30 proves nonemptiness and boundedness on the reachable image. Proposition 31 transfers memberwise equality through a possibly unattained supremum, and Proposition 32 transfers it through a finite maximum and the complete set-valued argmax. Proposition 33 isolates the additional compatibility needed by a root-history-dependent selector.

Minimality follows in the scalar function space on covered histories. Lemma 34 places every Hankel column in the comparator coordinate span. Proposition 35 stacks the core-column decoder rows to obtain $z_D = B\phi$, and Proposition 36 combines necessity with the exact core decoder. Proposition 37 treats the full-rank case.

Finally, Lemma 38 identifies the fibers of z_D with complete finite decision-row equality, and Proposition 39 establishes the partition order. Proposition 41 turns a nonzero fixed infinite consequence gap into a positive-weight finite deterministic witness using two exact geometric tails. Proposition 42 preserves the entire tie set, and Proposition 43 separates impossible future-realization recovery from exact representation of conditional decision consequences.

B.5 Proof Details

B.5.1 Population Identification and Randomized Finite Continuations

Lemma 13 (POMDP-consistent conditional versions at covered roots)**.** *Under Assumptions 6, 9, and 10, fix a supplied covered pre-action root $g \in \mathcal{H}_{\mathrm{cov}}$, a trusted action $a \in \mathcal{A}_{\mathrm{tr}}$, and $o \in \mathcal{O}$. Then the supplied POMDP-consistent conditional versions satisfy*

$$m_a(g) = b_g^\top r_a, \qquad p_a(o \mid g) = b_g^\top G_{a,o}\mathbf{1}. \tag{A1.1}$$

If $p_a(o \mid g) > 0$, the posterior at the declared next covered pre-action history has the column-vector orientation

$$b_{g\oplus(a,o)} = \frac{G_{a,o}^\top b_g}{p_a(o \mid g)}, \tag{A1.2}$$

and hence, for every $x \in \mathbb{R}^{|\mathcal{S}|}$,

$$p_a(o \mid g) b_{g\oplus(a,o)}^\top x = b_g^\top G_{a,o} x. \tag{A1.3}$$

If $p_a(o \mid g) = 0$, then

$$b_g^\top G_{a,o} = 0^\top, \qquad b_g^\top G_{a,o} x = 0 \quad \textit{for every } x \in \mathbb{R}^{|\mathcal{S}|}. \tag{A1.4}$$

When $\Pr_\mu(H_t = g) > 0$, these are ordinary observational conditional identities. When the root event is behavior-null, equation A1.1–equation A1.4 describe the supplied POMDP-consistent oracle version.

Proof. Conditioning first on the current state and using the fact that the behavior action selection leaves the POMDP kernels unchanged gives

$$\begin{aligned} m_a(g) &= \sum_{s\in\mathcal{S}} b_g(s)\mathbb{E}[R_t \mid S_t = s, A_t = a] = \sum_s b_g(s) r_a(s) = b_g^\top r_a, \\ p_a(o \mid g) &= \sum_{s,s'} b_g(s) \Pr(S_{t+1} = s', O_{t+1} = o \mid S_t = s, A_t = a) \\ &= \sum_{s,s'} b_g(s)[G_{a,o}]_{s,s'} = b_g^\top G_{a,o}\mathbf{1}. \end{aligned}$$

For positive $p_a(o \mid g)$, finite-state Bayes conditioning gives, coordinatewise,

$$b_{g\oplus(a,o)}(s') = \frac{\sum_s b_g(s)[G_{a,o}]_{s,s'}}{p_a(o \mid g)},$$

which is equation A1.2. Transposing equation A1.2, multiplying by $p_a(o \mid g)$, and applying the resulting row vector to x gives equation A1.3.

If $p_a(o \mid g) = 0$, every summand $b_g(s)[G_{a,o}]_{s,s'}$ in the nonnegative double sum above is zero. Thus for each s', $\sum_s b_g(s)[G_{a,o}]_{s,s'} = 0$, proving equation A1.4. This argument is important because a child vector x may have either sign; zero total observation probability still annihilates the entire row before it is applied to x.

At a behavior-positive root, all displayed quantities are determined by ordinary conditional distributions. On a behavior-null event, conditional versions are not fixed by the behavior joint law. Assumption 10 supplies the designated version, and the lemma verifies its POMDP consistency without appeal to sequential positivity. □

Proposition 14 (Finite controlled-tree semantics and the conditional g-functional)**.** *Under Assumptions 6, 8, 9, and 10, and with the identities of Lemma 13, let $\sigma \in \Sigma_{\mathrm{fin}}$. For every current hidden state s, the expected discounted reward of executing σ from s, with zero reward after reaching $\perp$, equals $\alpha_\sigma(s)$. For every supplied covered pre-action root g, its controlled finite-tree return therefore satisfies*

$$\mathcal{V}_\sigma(g) = \mathcal{J}_\sigma(g) = b_g^\top \alpha_\sigma = D_\Pi(g, \sigma), \tag{A1.5}$$

where

$$\mathcal{J}_\perp(g) = 0, \qquad \mathcal{J}_{(a,(\sigma_o)_o)}(g) = m_a(g) + \gamma \sum_{o: p_a(o|g)>0} p_a(o \mid g) \mathcal{J}_{\sigma_o}(g \oplus (a, o)). \tag{A1.G}$$

Thus equation A1.G recovers every deterministic finite controlled test at every supplied covered root. At a behavior-positive root the right side is an ordinary observational g-functional; at a behavior-null root it is the supplied oracle-relative functional.

Proof. Give the well-founded tree grammar the height

$$\mathrm{ht}(\perp) = 0, \qquad \mathrm{ht}(a, (\sigma_o)_o) = 1 + \max_o \mathrm{ht}(\sigma_o).$$

We first prove the statewise claim by induction on this height. For $\perp$, both the controlled reward and $\alpha_\perp$ are zero. Suppose $\sigma = (a, (\sigma_o)_{o \in \mathcal{O}})$ and the claim holds for every child. Conditional on $S_t = s$, the root action is a, so additivity of expected reward and the POMDP Markov kernel give

$$\begin{aligned} \mathbb{E}_s^\sigma[\text{discounted finite reward}] &= r_a(s) + \gamma \sum_{o \in \mathcal{O}} \sum_{s' \in \mathcal{S}} [G_{a,o}]_{s,s'} \alpha_{\sigma_o}(s') \\ &= \left[r_a + \gamma \sum_o G_{a,o} \alpha_{\sigma_o} \right](s) = \alpha_\sigma(s). \end{aligned} \tag{A1.6}$$

This proves that the primitive vector recursion is exactly the finite-horizon policy-tree semantics, including variable terminal depths.

We next prove the covered-root identities in equation A1.5, again by tree height. The base tree gives $\mathcal{V}_\perp(g) = \mathcal{J}_\perp(g) = b_g^\top \alpha_\perp = 0$. For the nonterminal tree above, ordinary conditional decomposition of its controlled return gives

$$\mathcal{V}_\sigma(g) = m_a(g) + \gamma \sum_{o: p_a(o|g)>0} p_a(o \mid g) \mathcal{V}_{\sigma_o}(g \oplus (a, o)). \tag{A1.7}$$

Positive-probability trusted extensions remain in $\mathcal{H}_{\mathrm{cov}}$, so each child functional in equation A1.7 is supplied. By the induction hypothesis, equation A1.7 equals $\mathcal{J}_\sigma(g)$. Using Lemma 13 and the child induction hypotheses once more,

$$\begin{aligned} \mathcal{J}_\sigma(g) &= b_g^\top r_a + \gamma \sum_{o: p_a(o|g)>0} p_a(o \mid g) b_{g \oplus (a,o)}^\top \alpha_{\sigma_o} \\ &= b_g^\top r_a + \gamma \sum_{o: p_a(o|g)>0} b_g^\top G_{a,o} \alpha_{\sigma_o} \\ &= b_g^\top r_a + \gamma \sum_{o \in \mathcal{O}} b_g^\top G_{a,o} \alpha_{\sigma_o} \\ &= b_g^\top \alpha_\sigma. \end{aligned} \tag{A1.8}$$

The third line is valid because equation A1.4 makes every omitted environment-zero branch exactly zero without evaluating an undefined child conditional. The definition of the decision Hankel matrix gives the last equality in equation A1.5.

For the all-terminal one-stage tree, equation A1.8 reduces to $\mathcal{J}_\sigma(g) = m_a(g) = b_g^\top r_a$. Thus the terminal and one-stage baselines, as well as every variable-depth finite tree, are recovered exactly. The positive-root versus null-root interpretation follows directly from the final paragraph of Lemma 13; the algebraic equality itself is unchanged. □

Proposition 15 (Stopped-prefix sequential importance identity)**.** *Under Assumptions 6, 8, 9, and 10, fix $h \in \mathcal{H}_{\mathrm{cov}}$ and $\sigma \in \Sigma_{\mathrm{fin}}$, and let $L(\sigma)$ be its maximum reward depth. Introduce a dead symbol $\dagger$, distinct from the legal terminal tree $\perp$, and define the node, activity, target action, and stopped weights before forming any expanded product as follows.*

Set $N_0 = \sigma$ and $W_{-1} = 1$. For $0 \le k < L(\sigma)$, proceed recursively.

1. If $N_k \in \{\perp, \dagger\}$ or $W_{k-1} = 0$, set

$$I_k^\sigma = 0, \qquad W_k = 0, \qquad N_{k+1} = \dagger, \tag{A1.9a}$$

without reading a target action or a behavior propensity. 2. Otherwise there is a unique active node

$$N_k = (a_k^\sigma, (\tau_o)_{o \in \mathcal{O}}).$$

Set $I_k^\sigma = 1$ and

$$W_k = W_{k-1} \frac{\mathbf{1}\{A_{t+k} = a_k^\sigma\}}{\mu(a_k^\sigma \mid H_{t+k})}. \tag{A1.9b}$$

The supported-prefix invariant proved below ensures that this is the only clause in which a propensity is read and that its denominator is positive. 3. If $A_{t+k} \neq a_k^\sigma$, set $N_{k+1} = \dagger$. If $A_{t+k} = a_k^\sigma$, put $o_k = O_{t+k+1}$ and inspect the current POMDP-consistent conditional mass $p_{a_k^\sigma}(o_k \mid H_{t+k})$. Set

$$N_{k+1} = \begin{cases} \tau_{o_k}, & p_{a_k^\sigma}(o_k \mid H_{t+k}) > 0, \\ \dagger, & p_{a_k^\sigma}(o_k \mid H_{t+k}) = 0. \end{cases} \tag{A1.9c}$$

Thus an environment-zero observation kills the node immediately after the current reward stage; equation A1.9a makes every later weight zero before any later action or propensity is read.

Define the stopped expanded-product notation by

$$I_k^\sigma \left[\prod_{j=0}^{k} \frac{\mathbf{1}\{A_{t+j} = a_j^\sigma\}}{\mu(a_j^\sigma \mid H_{t+j})} \right]_{\mathrm{stop}} := W_k. \tag{A1.9d}$$

On a live supported prefix the bracket is the ordinary finite product. If termination or action death occurred before stage k, or an environment-zero observation occurred after an earlier reward stage, equation A1.9d is shorthand for the recursively assigned value $W_k = 0$. At the reward stage immediately preceding an environment-zero observation, the current W_k remains the valid current-action weight, while equation A1.9c guarantees $W_\ell = 0$ for every later $\ell > k$.

With these definitions,

$$\mathcal{J}_\sigma(h) = \mathbb{E}_\mu^h \left[\sum_{k=0}^{L(\sigma)-1} \gamma^k W_k R_{t+k} \right] = \mathbb{E}_\mu^h \left[\sum_{k=0}^{L(\sigma)-1} \gamma^k I_k^\sigma R_{t+k} \left[\prod_{j=0}^{k} \frac{\mathbf{1}\{A_{t+j} = a_j^\sigma\}}{\mu(a_j^\sigma \mid H_{t+j})} \right]_{\mathrm{stop}} \right] = D_\Pi(h, \sigma). \tag{A1.IPW}$$

If $\Pr_\mu(H_t = h) > 0$, the expectation is an ordinary conditional behavior-law functional. If the root is behavior-null, it denotes the supplied POMDP-consistent oracle conditional version.

Proof. We first verify that equation A1.9a–equation A1.9c are everywhere well-defined. Inductively, whenever clause 2 is entered, $W_{k-1} > 0$, every preceding target action has matched, every preceding realized observation branch has had strictly positive conditional mass, and H_{t+k} is the corresponding target-reachable covered history. At $k = 0$, this holds because h is a supplied covered target root and the root action of σ is trusted. If clause 2 is entered, Assumption 9 therefore gives $\mu(a_k^\sigma \mid H_{t+k}) > 0$, so equation A1.9b is defined. An action mismatch enters $\dagger$. After a matched

action, equation A1.9c enters a legal child only when the realized observation has positive current conditional mass; Assumption 9 then keeps that extension covered. An observation with zero current conditional mass instead enters †, including on its null sample-point fiber. At the next stage, equation A1.9a sets the weight to zero without consulting a target action or propensity. This proves the invariant and, in particular, rules out a later $0/0$ expression at an environment-zero extension.

Define the finite behavior functional

$$\mathcal{K}_\sigma(g) := \mathbb{E}^g_\mu\left[\sum_{k=0}^{L(\sigma)-1} \gamma^k W_k R_{t+k}\right]. \tag{A1.10}$$

This functional is absolutely integrable. Indeed, the supplied finite conditional reward moments make every state-action reward integrable, so finiteness of $\mathcal{S}$ and $\mathcal{A}_{\mathrm{tr}}$ gives

$$\max_{\substack{s\in\mathcal{S}\\ a\in\mathcal{A}_{\mathrm{tr}}}} \mathbb{E}[|R_t| \mid S_t = s, A_t = a] < \infty. \tag{A1.10a}$$

Apply the live-prefix propensity cancellation successively to the nonnegative quantity $|R_{t+k}|$: condition at each live history, cancel $\mathbb{E}_\mu[\mathbf{1}\{A = a\}/\mu(a \mid H) \mid H] = 1$, and then use the unchanged POMDP kernel for the next transition. The stopping rule removes action-dead prefixes and gives zero to environment-zero extensions, exactly as their zero target mass requires. Hence the resulting expectation is the absolute reward moment under the deterministic target prefix:

$$\mathbb{E}^g_\mu[W_k|R_{t+k}|] \le \max_{\substack{s\in\mathcal{S}\\ a\in\mathcal{A}_{\mathrm{tr}}}} \mathbb{E}[|R_t| \mid S_t = s, A_t = a]. \tag{A1.10b}$$

Consequently,

$$\mathbb{E}^g_\mu\left[\sum_{k=0}^{L(\sigma)-1} |\gamma|^k W_k|R_{t+k}|\right] \le \left(\sum_{k=0}^{L(\sigma)-1} |\gamma|^k\right) \max_{\substack{s\in\mathcal{S}\\ a\in\mathcal{A}_{\mathrm{tr}}}} \mathbb{E}[|R_t| \mid S_t = s, A_t = a] < \infty. \tag{A1.10c}$$

Thus all finite tower operations below are justified. No uniform lower bound on μ is required.

We prove by tree height that $\mathcal{K}_\sigma = \mathcal{J}_\sigma$. For $\perp$, the sum in equation A1.10 is empty. Let $\sigma = (a, (\sigma_o)_o)$. At the root, sequential positivity and conditioning on A_t give

$$\mathbb{E}^g_\mu\left[\frac{\mathbf{1}\{A_t = a\}}{\mu(a \mid g)} R_t\right] = \mathbb{E}_\mu[R_t \mid H_t = g, A_t = a] = m_a(g). \tag{A1.11}$$

On the event $A_t = a, O_{t+1} = o$ with $p_a(o \mid g) > 0$, every later live weight factors into the root ratio in equation A1.11 times the stopped child weight started from $g \oplus (a, o)$. On $A_t \neq a$, or after a matched action followed by an observation with $p_a(o \mid g) = 0$, every later weight is zero by equation A1.9c–equation A1.9a. Applying the finite conditional tower, using that the behavior action choice leaves the environment kernel unchanged, and then applying the induction hypothesis only to positive-mass covered children yields

$$\begin{aligned}
\mathcal{K}_\sigma(g) &= m_a(g) + \gamma \sum_{o: p_a(o|g)>0} p_a(o \mid g)\mathcal{K}_{\sigma_o}(g \oplus (a, o)) \\
&= m_a(g) + \gamma \sum_{o: p_a(o|g)>0} p_a(o \mid g)\mathcal{J}_{\sigma_o}(g \oplus (a, o)) \\
&= \mathcal{J}_\sigma(g).
\end{aligned} \tag{A1.12}$$

An observation with $p_a(o \mid g) = 0$ contributes zero, enters † immediately, and never requires either a child conditional or a later propensity. Equation equation A1.12 also covers variable-depth branches because a child equal to $\perp$ has an empty remaining sum.

By definition equation A1.9d, the expanded product in equation A1.IPW is shorthand for the stopped recursive weight on live supported prefixes. Substituting equation A1.9d into equation A1.10, and then using Proposition 14, proves equation A1.IPW. Arbitrarily small but positive propensities cancel exactly in equation A1.11–equation A1.12. At a behavior-null root, the same finite recursion is evaluated under the supplied oracle conditional version. □

Proposition 16 (Finite common mixture for randomized continuations)**.** *Under Assumptions 6, 8, 9, and 10, let π be a root-independent common randomized continuation represented by an exogenous tape Ξ_π whose law is independent of the root history, current hidden state, and future POMDP noise. For each finite depth L, freezing the tape induces a deterministic zero-terminal tree $\sigma_{\pi,L}(\Xi_\pi)$. There is a finite set $\Sigma_L \subset \Sigma_{\mathrm{fin}}$ and weights*

$$w_{\pi,L}(\sigma) := \Pr\{\sigma_{\pi,L}(\Xi_\pi) = \sigma\}, \qquad \sigma \in \Sigma_L, \tag{A1.13}$$

such that

$$w_{\pi,L}(\sigma) \geq 0, \qquad \sum_{\sigma\in\Sigma_L} w_{\pi,L}(\sigma) = 1, \tag{A1.M1}$$

and the weights are independent of h and s. Moreover,

$$\alpha_{\pi^{(L)}} = \sum_{\sigma\in\Sigma_L} w_{\pi,L}(\sigma)\alpha_\sigma, \qquad V_{\pi^{(L)}}(h) = \sum_{\sigma\in\Sigma_L} w_{\pi,L}(\sigma) D_\Pi(h,\sigma) \tag{A1.M2}$$

for every covered h.

Proof. For fixed L, a deterministic depth-L tree assigns one trusted action to each of finitely many observation prefixes of length less than L, with $\perp$ allowed at any node. Because $\mathcal{A}_{\mathrm{tr}}$, $\mathcal{O}$, and L are finite, only finitely many such trees exist. Therefore the image of the map $\xi \mapsto \sigma_{\pi,L}(\xi)$ is a finite subset Σ_L of Σ_{fin}. Defining the weight of a tree as the probability of its tape fiber gives equation A1.13 and immediately proves nonnegativity and unit total mass. Different tape realizations that induce the same deterministic tree are simply aggregated into the same fiber.

Freezing the entire tape fixes the choices at all observation-contingent nodes simultaneously, so arbitrary correlations among the plan's nodewise random choices are allowed. Independence of the tape law from h and s means that the probabilities of these fibers are the same at every root history and current hidden state. Independence from future POMDP noise permits conditioning on the tape before taking the controlled environment expectation.

For every $s \in \mathcal{S}$, finite total expectation over the tape fibers and Proposition 14 give

$$\begin{aligned}\alpha_{\pi^{(L)}}(s) &= \mathbb{E}_{\Xi_\pi}[\mathbb{E}[\text{depth-}L \text{ discounted reward} \mid S_t = s, \Xi_\pi]] \\ &= \sum_{\sigma\in\Sigma_L} w_{\pi,L}(\sigma)\alpha_\sigma(s).\end{aligned} \tag{A1.14}$$

Since equation A1.14 holds coordinatewise, it proves the vector equality in equation A1.M2. Evaluating that equality by the probability belief b_h and using the definition of D_Π gives

$$\begin{aligned}V_{\pi^{(L)}}(h) = b_h^\top \alpha_{\pi^{(L)}} &= \sum_{\sigma\in\Sigma_L} w_{\pi,L}(\sigma) b_h^\top \alpha_\sigma \\ &= \sum_{\sigma\in\Sigma_L} w_{\pi,L}(\sigma) D_\Pi(h,\sigma),\end{aligned}$$

which is the scalar equality in equation A1.M2. By Proposition 14, each scalar summand is the recovered deterministic finite return. Hence the mixture also has the ordinary observational interpretation at behavior-positive roots and the oracle-relative interpretation at behavior-null supplied roots.

When $L = 0$, the sole induced tree is $\perp$, with weight one and zero vector. A deterministic plan is the one-atom case for every L. A root-history-dependent selector would in general change the tape pushforward weights with h; the common-plan statement therefore concerns root-independent plans. □

B.5.2 Reward-Rooted Operator Invariance

Lemma 17 (All-terminal reward inclusion)**.** *Under Assumptions 6 and 8 and the support-independent grammar equation T, for every $a \in \mathcal{A}_{\mathrm{tr}}$, the all-terminal tree*

$$\tau_a^0 := \big(a, (\perp)_{o'\in\mathcal{O}}\big)$$

belongs to Σ_{fin} and satisfies

$$\alpha_{\tau_a^0} = r_a \in \mathcal{W}_\Pi.$$

Proof. Every child of τ_a^0 is the legal terminal symbol $\perp$, so grammar equation T makes τ_a^0 a legal finite tree for every trusted a. Applying recursion equation R and $\alpha_\perp = 0$ gives

$$\alpha_{\tau_a^0} = r_a + \gamma \sum_{o' \in \mathcal{O}} G_{a,o'} \alpha_\perp = r_a.$$

By definition, $\mathcal{W}_\Pi$ is the algebraic span of all α_σ with $\sigma \in \Sigma_{\text{fin}}$. Hence the generator $\alpha_{\tau_a^0} = r_a$ lies in $\mathcal{W}_\Pi$. If $r_a = 0$, the same equality simply yields the zero vector, which belongs to every linear subspace. □

Lemma 18 (Exact one-child tree difference)**.** *Under Assumptions 6, 7, and 8, for every $a \in \mathcal{A}_{\text{tr}}$, $o \in \mathcal{O}$, and $\sigma \in \Sigma_{\text{fin}}$, define*

$$\tau_{a,o,\sigma} := \big(a, (\sigma_{o'})_{o' \in \mathcal{O}}\big), \qquad \sigma_o = \sigma, \quad \sigma_{o'} = \perp \text{ for } o' \neq o.$$

Then $\tau_{a,o,\sigma} \in \Sigma_{\text{fin}}$, regardless of the probability of observing o from any covered root, and

$$\alpha_{\tau_{a,o,\sigma}} - \alpha_{\tau_a^0} = \gamma G_{a,o} \alpha_\sigma \in \mathcal{W}_\Pi.$$

Proof. Grammar equation T quantifies over every syntactic observation child and imposes no behavior-law, covered-root, or positive-probability test on a tree index. Since σ and $\perp$ are legal finite children, $\tau_{a,o,\sigma}$ is legal even when branch o has probability zero from some or all covered histories.

The primitive recursion, together with $\alpha_\perp = 0$, yields

$$\begin{aligned}\alpha_{\tau_{a,o,\sigma}} &= r_a + \gamma G_{a,o} \alpha_\sigma + \gamma \sum_{o' \neq o} G_{a,o'} \alpha_\perp \\ &= r_a + \gamma G_{a,o} \alpha_\sigma.\end{aligned}$$

Lemma 17 gives $\alpha_{\tau_a^0} = r_a$, so subtraction proves the displayed identity. Both tree vectors on the left are generators of $\mathcal{W}_\Pi$; because $\mathcal{W}_\Pi$ is a linear subspace, their difference $\gamma G_{a,o} \alpha_\sigma$ also lies in $\mathcal{W}_\Pi$. □

Proposition 19 (Invariant reward-rooted span)**.** *Under Assumptions 6, 7, and 8 and Lemma 18, for every $a \in \mathcal{A}_{\text{tr}}$ and $o \in \mathcal{O}$,*

$$G_{a,o} \mathcal{W}_\Pi \subseteq \mathcal{W}_\Pi.$$

Proof. Fix $a \in \mathcal{A}_{\text{tr}}$, $o \in \mathcal{O}$, and a generator α_σ of $\mathcal{W}_\Pi$. Lemma 18 gives

$$\gamma G_{a,o} \alpha_\sigma \in \mathcal{W}_\Pi.$$

Assumption 7 gives $\gamma > 0$. Closure of a linear subspace under multiplication by γ^{-1} therefore gives

$$G_{a,o} \alpha_\sigma = \gamma^{-1} \big(\alpha_{\tau_{a,o,\sigma}} - \alpha_{\tau_a^0}\big) \in \mathcal{W}_\Pi. \tag{A2.1}$$

Now take an arbitrary $w \in \mathcal{W}_\Pi$. By the algebraic definition of span, there are a finite $N \geq 0$, scalars $\lambda_1, \ldots, \lambda_N$, and trees $\sigma_1, \ldots, \sigma_N$ such that

$$w = \sum_{i=1}^{N} \lambda_i \alpha_{\sigma_i}.$$

Matrix linearity and equation A2.1 imply

$$G_{a,o} w = \sum_{i=1}^{N} \lambda_i G_{a,o} \alpha_{\sigma_i} \in \mathcal{W}_\Pi.$$

Thus $G_{a,o} \mathcal{W}_\Pi \subseteq \mathcal{W}_\Pi$.

This argument includes the degenerate cases. If $G_{a,o} = 0$, its image is $\{0\} \subseteq \mathcal{W}_\Pi$. If $\mathcal{W}_\Pi = \{0\}$, the only input is zero and matrix linearity gives $G_{a,o} 0 = 0$.

The positivity of γ is essential to this inference. At the excluded boundary $\gamma = 0$, the tree difference in Lemma 18 collapses to zero and need not expose the operator image. For example, take two hidden states, one trusted action, one deterministic observation, deterministic state swapping

$$G_{a,o} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad r_a = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

When $\gamma = 0$, every nonterminal tree vector is r_a, so $\mathcal{W}_\Pi = \operatorname{span}\{(1,0)^\top\}$, whereas $G_{a,o} r_a = (0,1)^\top \notin \mathcal{W}_\Pi$. Assumption 7 excludes exactly this failure mode. □

Proposition 20 (Finite operator-word closure)**.** *Under Assumptions 6, 7, and 8 and Proposition 19, let $n \geq 0$ and let*

$$\big((a_1, o_1), \ldots, (a_n, o_n)\big) \in (\mathcal{A}_{\mathrm{tr}} \times \mathcal{O})^n.$$

With

$$\mathsf{G}_{1:n} := \begin{cases} I, & n = 0, \\ G_{a_n,o_n} \cdots G_{a_1,o_1}, & n \geq 1, \end{cases}$$

one has

$$\mathsf{G}_{1:n} \mathcal{W}_\Pi \subseteq \mathcal{W}_\Pi.$$

The inclusion is exact.

Proof. Proceed by induction on n. When $n = 0$, $\mathsf{G}_{1:0} = I$, so $I\mathcal{W}_\Pi = \mathcal{W}_\Pi$. Suppose the assertion holds for a word of length n. For any $w \in \mathcal{W}_\Pi$, the induction hypothesis gives $\mathsf{G}_{1:n} w \in \mathcal{W}_\Pi$. Proposition 19, applied to the next trusted pair (a_{n+1}, o_{n+1}), then gives

$$G_{a_{n+1},o_{n+1}} \mathsf{G}_{1:n} w \in \mathcal{W}_\Pi.$$

This vector is $\mathsf{G}_{1:n+1} w$, proving the induction step. Every stage is an exact subspace inclusion, so finite composition accumulates zero defect. If any operator in the word is zero, the full composition from that point is zero and the same conclusion holds; if $\mathcal{W}_\Pi = \{0\}$, the induction remains valid from its identity base case. □

B.5.3 Decision-Hankel Rank and Exact Core Coordinates

Lemma 21 (Decision-column image identity)**.** *Under Assumption 6 and definitions equation W–equation H, let*

$$\mathbb{R}^{\mathcal{H}_{\mathrm{cov}}} := \{f : \mathcal{H}_{\mathrm{cov}} \to \mathbb{R}\}$$

be the algebraic vector space of real-valued functions on the covered-history index set, and define

$$T : \mathcal{W}_\Pi \longrightarrow \mathbb{R}^{\mathcal{H}_{\mathrm{cov}}}, \qquad (T\alpha)(h) := b_h^\top \alpha.$$

Then T is a well-defined linear map and

$$T(\mathcal{W}_\Pi) = \operatorname{span}\{D_\Pi(\cdot, \sigma) : \sigma \in \Sigma_{\mathrm{fin}}\} = \operatorname{colspan}(D_\Pi). \tag{A3.1}$$

All spans in equation A3.1 are algebraic spans, so their elements are finite linear combinations even when $\mathcal{H}_{\mathrm{cov}}$ or Σ_{fin} is infinite.

Proof. For every $\alpha \in \mathcal{W}_\Pi \subseteq \mathbb{R}^{|\mathcal{S}|}$ and covered h, $b_h^\top \alpha$ is a real scalar, so $T\alpha$ is a well-defined function on $\mathcal{H}_{\mathrm{cov}}$. For $\alpha, \beta \in \mathcal{W}_\Pi$ and scalars λ, μ,

$$(T(\lambda\alpha + \mu\beta))(h) = b_h^\top(\lambda\alpha + \mu\beta) = \lambda(T\alpha)(h) + \mu(T\beta)(h)$$

for every covered h, proving linearity as equality of functions.

For the first inclusion, take any $\alpha \in \mathcal{W}_\Pi$. By the algebraic definition of $\mathcal{W}_\Pi$, there are a finite $N \geq 0$, scalars $\lambda_1, \ldots, \lambda_N$, and deterministic finite trees $\sigma_1, \ldots, \sigma_N$ such that

$$\alpha = \sum_{j=1}^{N} \lambda_j \alpha_{\sigma_j}.$$

Linearity and the definition of D_Π give, as functions on all covered histories,

$$T\alpha = \sum_{j=1}^{N} \lambda_j T\alpha_{\sigma_j} = \sum_{j=1}^{N} \lambda_j D_\Pi(\cdot, \sigma_j).$$

Thus $T(\mathcal{W}_\Pi) \subseteq \operatorname{colspan}(D_\Pi)$.

Conversely, let $f \in \operatorname{colspan}(D_\Pi)$. By algebraic-span membership, for some finite $N \geq 0$,

$$f = \sum_{j=1}^{N} \lambda_j D_\Pi(\cdot, \sigma_j).$$

Set $\alpha = \sum_{j=1}^{N} \lambda_j \alpha_{\sigma_j}$, which belongs to $\mathcal{W}_\Pi$. Then, for every covered h,

$$(T\alpha)(h) = \sum_{j=1}^{N} \lambda_j b_h^\top \alpha_{\sigma_j} = \sum_{j=1}^{N} \lambda_j D_\Pi(h, \sigma_j) = f(h).$$

Hence $f = T\alpha \in T(\mathcal{W}_\Pi)$, proving the reverse inclusion and equation A3.1. □

Proposition 22 (Finite decision-image rank bounds)**.** *Under Assumption 6 and Lemma 21,*

$$d = \dim T(\mathcal{W}_\Pi) \leq \dim \mathcal{W}_\Pi \leq |\mathcal{S}|. \tag{A3.2}$$

The first inequality in equation A3.2 holds for noninjective T and may be strict when $\ker T \neq \{0\}$.

Proof. By definition, d is the algebraic dimension of $\operatorname{colspan}(D_\Pi)$. Lemma 21 identifies that column span exactly with $T(\mathcal{W}_\Pi)$, so

$$d = \dim T(\mathcal{W}_\Pi). \tag{A3.3}$$

Because $\mathcal{W}_\Pi$ is a subspace of the finite-dimensional space $\mathbb{R}^{|\mathcal{S}|}$, it has some finite basis $w_1, \ldots, w_q$, where $q = \dim \mathcal{W}_\Pi$, and

$$q \leq |\mathcal{S}|. \tag{A3.4}$$

For any $\alpha \in \mathcal{W}_\Pi$, write $\alpha = \sum_{j=1}^{q} \lambda_j w_j$. Linearity gives

$$T\alpha = \sum_{j=1}^{q} \lambda_j T w_j.$$

Therefore the at most q functions $Tw_1, \ldots, Tw_q$ span $T(\mathcal{W}_\Pi)$, even if some are zero or linearly dependent. Consequently

$$\dim T(\mathcal{W}_\Pi) \leq q = \dim \mathcal{W}_\Pi. \tag{A3.5}$$

Combining equation A3.3–equation A3.5 proves equation A3.2.

Preimages under T need not be unique. If $u \in \ker T$, then

$$(Tu)(h) = b_h^\top u = 0 \quad \text{for every } h \in \mathcal{H}_{\text{cov}},$$

but this does not imply $u = 0$. Likewise, $T\alpha = T\alpha'$ implies only $\alpha - \alpha' \in \ker T$, not $\alpha = \alpha'$. Such kernel directions are invisible to all covered decision columns and create no residual in the image identity. □

Lemma 23 (Actual deterministic-column core)**.** *Under Assumption 6, Lemma 21, and Proposition 22, there exist deterministic finite-tree indices*

$$\sigma_1, \ldots, \sigma_d \in \Sigma_{\text{fin}}$$

such that the actual column functions

$$D_\Pi(\cdot, \sigma_1), \ldots, D_\Pi(\cdot, \sigma_d) \tag{A3.6}$$

form a basis of $T(\mathcal{W}_\Pi) = \operatorname{colspan}(D_\Pi)$. When $d = 0$, equation A3.6 is the empty family, every deterministic column function is zero, and the empty family is the basis of the zero image space.

Proof. Write

$$\mathcal{V}_D := T(\mathcal{W}_\Pi) \quad \text{and} \quad f_\sigma := D_\Pi(\cdot, \sigma).$$

Lemma 21 gives

$$\mathcal{V}_D = \operatorname{span}\{f_\sigma : \sigma \in \Sigma_{\text{fin}}\}, \qquad \dim \mathcal{V}_D = d. \tag{A3.7}$$

First suppose $d = 0$. Then $\mathcal{V}_D = \{0\}$. Since each $f_\sigma \in \mathcal{V}_D$, every deterministic column function is identically zero. The empty family is linearly independent, its span is $\{0\}$, and hence it is a basis of $\mathcal{V}_D$. This is exactly the empty-core convention.

Now suppose $d > 0$. We select columns by a finite greedy construction. At stage $k = 0$, no column has been selected and its span is $\{0\}$, a proper subspace of $\mathcal{V}_D$. More generally, suppose $0 \le k < d$ and actual columns

$$f_{\sigma_1}, \dots, f_{\sigma_k}$$

have been selected and are linearly independent. Their span has dimension $k < d$, so it is a proper subspace of $\mathcal{V}_D$. If every generator f_σ belonged to that proper subspace, then equation A3.7 would imply that all of $\mathcal{V}_D$ belonged to it, a contradiction. Hence some deterministic index $\sigma_{k+1} \in \Sigma_{\text{fin}}$ satisfies

$$f_{\sigma_{k+1}} \notin \operatorname{span}\{f_{\sigma_1}, \dots, f_{\sigma_k}\}.$$

Adding this actual column preserves linear independence.

Repeating this argument for the finite stages $k = 0, \dots, d-1$ produces d linearly independent actual deterministic columns. A linearly independent family of d vectors in the d-dimensional space $\mathcal{V}_D$ is a basis, proving equation A3.6. The construction uses finitely many selections and finite linear combinations. Duplicate and zero columns cannot be selected after they lie in the current span, so noninjectivity of the index-to-column map causes no difficulty.

The selected indices are therefore a decision core as defined in the setting, and its coordinate map is

$$z_D(h) = \big(D_\Pi(h, \sigma_1), \dots, D_\Pi(h, \sigma_d)\big) \in \mathbb{R}^d. \tag{A3.8}$$

For $d = 0$, equation A3.8 is the unique empty vector in $\mathbb{R}^0$. □

Proposition 24 (Exact core decoder)**.** *Under Assumption 6 and Lemma 23, let*

$$\mathcal{C} := (\sigma_1, \dots, \sigma_d)$$

be the actual deterministic-column core. Every $f \in T(\mathcal{W}_\Pi)$ has a unique vector

$$\operatorname{coord}_{\mathcal{C}}(f) = \big(c_1(f), \dots, c_d(f)\big) \in \mathbb{R}^d$$

such that

$$f = \sum_{i=1}^{d} c_i(f) D_\Pi(\cdot, \sigma_i). \tag{A3.9}$$

Consequently, for every $\alpha \in \mathcal{W}_\Pi$, the image coordinate

$$c[\alpha] := \operatorname{coord}_{\mathcal{C}}(T\alpha) \tag{A3.10}$$

is well defined and depends only on $T\alpha$. In particular, for every $\sigma \in \Sigma_{\text{fin}}$, the coefficient

$$c_\sigma := c[\alpha_\sigma] \in \mathbb{R}^d \tag{A3.11}$$

is unique and satisfies

$$D_\Pi(h, \sigma) = \langle z_D(h), c_\sigma \rangle \quad \textit{for every } h \in \mathcal{H}_{\text{cov}}. \tag{A3.12}$$

When $d = 0$, all statements use the unique empty vector in $\mathbb{R}^0$, the empty sum, and the empty inner product, each equal to the unique zero image function or the scalar zero as appropriate.

Proof. Lemma 23 states that

$$D_\Pi(\cdot, \sigma_1), \dots, D_\Pi(\cdot, \sigma_d)$$

is a basis of $T(\mathcal{W}_\Pi)$. By the defining spanning and independence properties of a basis, every f in that image has coefficients satisfying equation A3.9, and those coefficients are unique. This defines $\operatorname{coord}_{\mathcal{C}}(f)$ entirely in the image function space.

For any $\alpha \in \mathcal{W}_\Pi$, $T\alpha \in T(\mathcal{W}_\Pi)$, so equation A3.10 is defined. If $\alpha, \alpha' \in \mathcal{W}_\Pi$ satisfy $T\alpha = T\alpha'$, uniqueness of the coordinates of that common image function gives

$$c[\alpha] = c[\alpha']. \tag{A3.13}$$

Equivalently, adding any $u \in \ker T$ to an admissible representative leaves the coordinate unchanged:

$$c[\alpha + u] = c[\alpha] \quad \text{whenever } \alpha, \alpha + u \in \mathcal{W}_\Pi. \tag{A3.14}$$

Equations equation A3.13–equation A3.14 are representative-independence statements in the image; they do not assert $\alpha = \alpha'$ or $u = 0$.

For a deterministic tree σ, the defining column identity gives

$$T\alpha_\sigma = D_\Pi(\cdot, \sigma),$$

so equation A3.9–equation A3.11 yield equality of functions

$$D_\Pi(\cdot, \sigma) = \sum_{i=1}^{d} (c_\sigma)_i D_\Pi(\cdot, \sigma_i). \tag{A3.15}$$

Evaluating equation A3.15 at an arbitrary covered history h and using equation A3.8 gives

$$\begin{aligned} D_\Pi(h, \sigma) &= \sum_{i=1}^{d} (c_\sigma)_i D_\Pi(h, \sigma_i) \\ &= \big\langle \big(D_\Pi(h, \sigma_1), \ldots, D_\Pi(h, \sigma_d)\big), c_\sigma \big\rangle \\ &= \langle z_D(h), c_\sigma \rangle, \end{aligned}$$

which is equation A3.12 on every covered history.

Finally, if $d = 0$, Lemma 23 gives $T(\mathcal{W}_\Pi) = \{0\}$. Thus $T\alpha = 0$ for every $\alpha \in \mathcal{W}_\Pi$, although α itself need not be zero. The space $\mathbb{R}^0$ contains the unique empty vector $()$, which is the unique coordinate of the zero image function. For every finite tree, $c_\sigma = ()$, $z_D(h) = ()$, and

$$\langle z_D(h), c_\sigma \rangle = \langle (), () \rangle = 0 = D_\Pi(h, \sigma).$$

Therefore the empty core decodes every finite column exactly and does not require $\mathcal{W}_\Pi = \{0\}$ or injectivity of T. □

B.5.4 Infinite Continuation Decoders and Uniform Truncation

Proposition 25 (Exact randomized-truncation membership)**.** *Under Assumptions 6 and 8, Proposition 16, and the invariant-span results of Propositions 19 and 20, let $\pi \in \Pi_\infty$ be a root-independent common randomized plan under the common-plan convention. For every integer $L \geq 0$,*

$$\alpha_{\pi^{(L)}} = \sum_{\sigma \in \Sigma_L} w_{\pi,L}(\sigma) \alpha_\sigma \in \mathcal{W}_\Pi. \tag{A4.1}$$

The equality and membership concern the same randomized truncation in the same hidden-state coordinates and hold exactly. When $L = 0$, $\alpha_{\pi^{(0)}} = \alpha_\perp = 0$; a deterministic plan is the one-atom special case of equation A4.1.

Proof. The common-mixture proposition gives a finite set $\Sigma_L \subset \Sigma_{\text{fin}}$, nonnegative weights of total mass one, and the exact first equality in equation A4.1. By definition,

$$\mathcal{W}_\Pi = \text{span}\{\alpha_\sigma : \sigma \in \Sigma_{\text{fin}}\}$$

is an algebraic linear span. Therefore every finite linear combination on the right side of equation A4.1 lies in $\mathcal{W}_\Pi$. Propositions 19 and 20 show that this is the same reward-rooted, trusted-operator-invariant span.

At $L = 0$, the finite mixture has the sole tree $\perp$ with weight one, so its vector is zero, which lies in every linear subspace. For a deterministic plan, the frozen-tape image has one tree of weight one, and equation A4.1 reduces to generator membership. Proposition 16 makes the common weights independent of the root history and current hidden state, so equation A4.1 defines one vector rather than a family of root-labeled mixtures. □

Lemma 26 (Same-plan discounted tail decomposition)**.** *Under Assumptions 6, 7, and 8 and Propositions 14 and 16, fix $\pi \in \Pi_\infty$. Couple all zero-terminal truncations by executing the same root-independent plan with the same exogenous tape and the same future POMDP noise. For stage $j \geq 0$, let $I_{\pi,j}$ indicate that the common plan has an active reward node at stage j, and define the auxiliary stage-mean vector $u_{\pi,j} \in \mathbb{R}^{|\mathcal{S}|}$ by*

$$u_{\pi,j}(s) := \mathbb{E}_s^\pi \left[I_{\pi,j}\, r_{A_{t+j}^\pi}(S_{t+j}^\pi) \right]. \tag{A4.2}$$

On an inactive branch the integrand in equation A4.2 is defined to be zero before any action is read. Then, for every $L \geq 0$,

$$\alpha_{\pi^{(L)}} = \sum_{j=0}^{L-1} \gamma^j u_{\pi,j}, \qquad \|u_{\pi,j}\|_\infty \leq R_{\max}, \tag{A4.3}$$

and for all integers $M \geq L \geq 0$,

$$\alpha_{\pi^{(M)}} - \alpha_{\pi^{(L)}} = \sum_{j=L}^{M-1} \gamma^j u_{\pi,j}, \tag{A4.4}$$

so

$$\begin{aligned} \|\alpha_{\pi^{(M)}} - \alpha_{\pi^{(L)}}\|_\infty &\leq R_{\max} \sum_{j=L}^{M-1} \gamma^j \\ &= \frac{R_{\max}\gamma^L(1-\gamma^{M-L})}{1-\gamma} \leq \frac{R_{\max}\gamma^L}{1-\gamma}. \end{aligned} \tag{A4.5}$$

Proof. One infinite common plan and one tape realization determine a single observation-contingent tree. Its depth-L restriction and depth-M restriction agree at every reward stage before L. Thus the random state, active-node indicator, and trusted action used at any fixed stage $j < L$ are the same objects for every later truncation. This prefix consistency is why one vector $u_{\pi,j}$, independent of the chosen truncation depth, is well defined.

Fix an initial hidden state s. Conditional on a tape realization, Proposition 14 identifies the corresponding deterministic depth-L tree vector with its statewise expected finite discounted reward. At an active stage, the tower property and the setting definition of the reward mean give

$$\mathbb{E}[R_{t+j} \mid S_{t+j}^\pi, A_{t+j}^\pi] = r_{A_{t+j}^\pi}(S_{t+j}^\pi).$$

After a terminal node, both the zero-terminal tree reward and the integrand in equation A4.2 are zero. Averaging the deterministic-tree identity over the common tape, equivalently applying Proposition 16 coordinatewise, therefore yields

$$\alpha_{\pi^{(L)}}(s) = \sum_{j=0}^{L-1} \gamma^j \mathbb{E}_s^\pi \left[I_{\pi,j} r_{A_{t+j}^\pi}(S_{t+j}^\pi) \right],$$

which is the first identity in equation A4.3.

Whenever $I_{\pi,j} = 1$, the trusted-policy assumption gives $A_{t+j}^\pi \in \mathcal{A}_{\mathrm{tr}}$. Hence Assumption 7 gives, for every initial state,

$$\begin{aligned} |u_{\pi,j}(s)| &\leq \mathbb{E}_s^\pi \left[I_{\pi,j} \left| r_{A_{t+j}^\pi}(S_{t+j}^\pi) \right| \right] \\ &\leq R_{\max} \mathbb{E}_s^\pi [I_{\pi,j}] \leq R_{\max}. \end{aligned} \tag{A4.6}$$

This calculation uses the bound on conditional means from Assumption 7.

Subtracting the two finite expansions in equation A4.3 proves the exact same-plan identity equation A4.4. The triangle inequality, equation A4.6, and $0 < \gamma < 1$ then give the first line of equation A4.5. Evaluating the finite geometric sum gives the equality in its second line; $0 \leq 1-\gamma^{M-L} \leq 1$ gives the final inequality. If $M = L$, both the sum in equation A4.4 and the difference are zero, and the same formulas remain valid. □

Proposition 27 (Closed-span infinite common-plan vector)**.** *Under Assumptions 6 and 7, Proposition 25, and Lemma 26, for every common $\pi \in \Pi_\infty$ there is a unique vector*

$$\alpha_\pi := \lim_{L\to\infty} \alpha_{\pi^{(L)}} \in \mathcal{W}_\Pi \tag{A4.7}$$

in hidden-state sup norm. Moreover, for every $L \geq 0$,

$$\|\alpha_\pi - \alpha_{\pi^{(L)}}\|_\infty \leq \frac{R_{\max}\gamma^L}{1-\gamma}. \tag{A4.8}$$

Proof. If $R_{\max} = 0$, equation A4.5 makes all truncation vectors equal, so the sequence is immediately Cauchy. If $R_{\max} > 0$, let $\varepsilon > 0$. Since $0 < \gamma < 1$, choose N with

$$\frac{R_{\max}\gamma^N}{1-\gamma} < \varepsilon.$$

For all $M \geq L \geq N$, Lemma 26 gives

$$\|\alpha_{\pi^{(M)}} - \alpha_{\pi^{(L)}}\|_\infty \leq \frac{R_{\max}\gamma^L}{1-\gamma} \leq \frac{R_{\max}\gamma^N}{1-\gamma} < \varepsilon.$$

Thus the sequence is Cauchy in $\mathbb{R}^{|\mathcal{S}|}$ with the sup norm. Coordinatewise completeness of $\mathbb{R}$, together with the finite number of state coordinates, gives a unique limit $\alpha_\pi \in \mathbb{R}^{|\mathcal{S}|}$.

It remains to prove that this limit stays in the target span. The subspace $\mathcal{W}_\Pi \subseteq \mathbb{R}^{|\mathcal{S}|}$ is finite dimensional. If it is the zero subspace, it is closed. Otherwise choose a basis $w_1, \ldots, w_q$ and form the $|\mathcal{S}| \times q$ matrix $B = (w_1 \ \cdots \ w_q)$. Its columns are independent, so $B^\top B$ is invertible. If a sequence $x_n = B\theta_n \in \mathcal{W}_\Pi$ converges coordinatewise to x, then uniqueness of the basis coordinates gives

$$\theta_n = (B^\top B)^{-1} B^\top x_n \longrightarrow \theta := (B^\top B)^{-1} B^\top x.$$

Indeed, if $B^\top Bv = 0$, then $0 = v^\top B^\top Bv = \|Bv\|_2^2$, so column independence forces $v = 0$; this verifies invertibility. Consequently $x_n = B\theta_n \to B\theta$. Since coordinatewise limits in $\mathbb{R}^{|\mathcal{S}|}$ are unique, $x = B\theta \in \mathcal{W}_\Pi$. This proves that $\mathcal{W}_\Pi$ is closed in the present finite-dimensional coordinates.

Proposition 25 gives $\alpha_{\pi^{(L)}} \in \mathcal{W}_\Pi$ for every L, so closedness implies the membership in equation A4.7. Finally fix L and let $M \to \infty$ in equation A4.5. Continuity of the finite-dimensional sup norm gives

$$\|\alpha_\pi - \alpha_{\pi^{(L)}}\|_\infty = \lim_{M\to\infty} \|\alpha_{\pi^{(M)}} - \alpha_{\pi^{(L)}}\|_\infty \leq \frac{R_{\max}\gamma^L}{1-\gamma},$$

which is equation A4.8. □

Proposition 28 (Exact fixed-plan image coefficient and decoder)**.** *Under Assumption 6, Proposition 27, and Proposition 24, let $(\sigma_1, \ldots, \sigma_d)$ be the actual-column core. For every common $\pi \in \Pi_\infty$, define its image coefficient by*

$$c_\pi := c[\alpha_\pi] = \mathrm{coord}_{\mathcal{C}}(T\alpha_\pi) \in \mathbb{R}^d. \tag{A4.9}$$

Then the discounted limit in the definition of $V_\pi(h)$ exists at every covered history and

$$V_\pi(h) = b_h^\top \alpha_\pi = \langle z_D(h), c_\pi \rangle \quad \textit{for every } h \in \mathcal{H}_{\mathrm{cov}}. \tag{A4.10}$$

The coefficient in equation A4.9 is unique as a coordinate of the image function $T\alpha_\pi$, is independent of h, and is defined without finite-truncation coefficients. If $d = 0$, c_π and $z_D(h)$ are the unique empty vectors and every quantity in equation A4.10 is zero.

Proof. Proposition 27 proves $\alpha_\pi \in \mathcal{W}_\Pi$. The exact-core proposition therefore applies to $T\alpha_\pi$, defines the unique image coordinate equation A4.9, and gives

$$b_h^\top \alpha_\pi = (T\alpha_\pi)(h) = \sum_{i=1}^{d} (c_\pi)_i D_\Pi(h, \sigma_i) = \langle z_D(h), c_\pi \rangle \tag{A4.11}$$

for every covered h. This is equality in the exact decision-column image; the vector identity $\alpha_\pi = \sum_i (c_\pi)_i \alpha_{\sigma_i}$ can fail when T has a nontrivial kernel.

Because b_h is a probability vector, every $x \in \mathbb{R}^{|\mathcal{S}|}$ satisfies

$$|b_h^\top x| \le \sum_s b_h(s)|x(s)| \le \|x\|_\infty. \tag{A4.12}$$

Apply equation A4.12 to $x = \alpha_{\pi^{(L)}} - \alpha_\pi$. Vector convergence in equation A4.7 gives

$$\lim_{L\to\infty} b_h^\top \alpha_{\pi^{(L)}} = b_h^\top \alpha_\pi. \tag{A4.13}$$

The left side is precisely the setting definition of $V_\pi(h)$, so equation A4.11–equation A4.13 prove equation A4.10.

If $d = 0$, Proposition 24 gives $T(\mathcal{W}_\Pi) = \{0\}$; hence $T\alpha_\pi = 0$, its unique coordinate is the empty vector, and equation A4.10 holds with the empty inner product. For a root-history-dependent selector that chooses different common plans at different roots, equation A4.9 applies to each fixed common member separately. □

Proposition 29 (Uniform truncation certificate and exact baselines)**.** *Under Assumption 7 and Propositions 27 and 28, for every integer $L \ge 0$,*

$$\sup_{\substack{h\in\mathcal{H}_{\mathrm{cov}}\\ \pi\in\Pi_\infty}} \left|V_\pi(h) - b_h^\top \alpha_{\pi^{(L)}}\right| \le \frac{R_{\max}\gamma^L}{1-\gamma}. \tag{A4.14}$$

The supremum is over the declared root-independent common-plan family, and the bound is deterministic with an explicit constant. If a common plan has no active reward node after stage $L-1$ for almost every tape realization and every observation branch, then

$$\alpha_\pi = \alpha_{\pi^{(L)}} \quad \textit{and} \quad V_\pi(h) = b_h^\top \alpha_{\pi^{(L)}} \quad \textit{for every covered } h, \tag{A4.15}$$

so its actual residual is exactly zero.

Proof. For arbitrary covered h, common π, and $L \ge 0$, Propositions 27 and 28, followed by equation A4.12, give

$$\begin{aligned} \left|V_\pi(h) - b_h^\top \alpha_{\pi^{(L)}}\right| &= \left|b_h^\top(\alpha_\pi - \alpha_{\pi^{(L)}})\right| \\ &\le \|\alpha_\pi - \alpha_{\pi^{(L)}}\|_\infty \\ &\le \frac{R_{\max}\gamma^L}{1-\gamma}. \end{aligned} \tag{A4.16}$$

The common-mixture semantics also gives $b_h^\top \alpha_{\pi^{(L)}} = V_{\pi^{(L)}}(h)$, so equation A4.16 compares the infinite plan with its actual same-plan zero-terminal truncation. The right side is independent of h, π, the tape weights, and the hidden state. Taking the displayed supremum proves equation A4.14.

If the common plan is terminal by depth L in the stated syntactic, common-tape sense, then $I_{\pi,j} = 0$ almost surely for every $j \ge L$, so equation A4.2 gives $u_{\pi,j} = 0$ for those stages. Equation equation A4.4 therefore gives $\alpha_{\pi^{(M)}} = \alpha_{\pi^{(L)}}$ for every $M \ge L$. Passing to the vector limit proves the first equality in equation A4.15, and belief evaluation proves the second.

At $L = 0$, $\alpha_{\pi^{(0)}} = 0$ and equation A4.14 gives the full-return bound $R_{\max}/(1-\gamma)$. If $R_{\max} = 0$, equation A4.6 makes every stage-mean vector zero, so every truncation, the limit, and every scalar residual are exactly zero. A deterministic plan is the one-atom common-mixture case and obeys the same proof. As $\gamma \uparrow 1$, the denominator in equation A4.14 records the loss explicitly; the boundary $\gamma = 1$ lies outside Assumption 7. Finally, equation A4.14 tends to zero for fixed $R_{\max}$ and $0 < \gamma < 1$, proving the exact-limit baseline. □

B.5.5 Planning Envelopes and Complete Action Ties

Lemma 30 (Bounded nonempty common-plan families)**.** *Under Assumptions 8 and 7, the setting-defined finite nonempty trusted action set, and Propositions 28 and 29, define the reachable core image*

$$\mathcal{Z} := \{z_D(h) : h \in \mathcal{H}_{\mathrm{cov}}\} \subseteq \mathbb{R}^d. \tag{A5.1}$$

For every $a \in \mathcal{A}_{\mathrm{tr}}$, the forced-root family $\Pi_\infty(a)$ is nonempty. Moreover, for every $x \in \mathcal{Z}$,

$$\varnothing \neq \{\langle x, c_\pi \rangle : \pi \in \Pi_\infty(a)\} \subseteq \left[-\frac{R_{\max}}{1-\gamma}, \frac{R_{\max}}{1-\gamma}\right]. \tag{A5.2}$$

The bound concerns decoded evaluations at reachable codes rather than coefficient norms.

Proof. Because $\mathcal{A}_{\mathrm{tr}}$ is nonempty, fix once and for all some $\bar{a} \in \mathcal{A}_{\mathrm{tr}}$. Given any $a \in \mathcal{A}_{\mathrm{tr}}$, consider the observation-contingent plan that takes a at the forced root and takes $\bar{a}$ after every later observation history. It is non-anticipatory, supported on trusted actions, and root-independent as a continuation rule. Assumption 8 contains every forced trusted root followed by such a trusted continuation, so this plan belongs to $\Pi_\infty(a)$. Hence each forced-root family is nonempty.

Now fix $x \in \mathcal{Z}$. By equation A5.1, at least one covered history h satisfies $x = z_D(h)$. For every $\pi \in \Pi_\infty(a) \subseteq \Pi_\infty$, Propositions 28 and 29, the latter evaluated at $L = 0$, give

$$|\langle x, c_\pi \rangle| = |\langle z_D(h), c_\pi \rangle| = |V_\pi(h)| \leq \frac{R_{\max}}{1-\gamma}. \tag{A5.3}$$

The value of the inner product depends only on x, not on which representative h was used to verify the bound. Combining nonemptiness with equation A5.3 proves equation A5.2. The proof controls exactly the evaluations needed by planning. □

Proposition 31 (Exact forced-action envelope on the reachable image)**.** *Under Assumptions 8 and 7, Proposition 28, and Lemma 30, for $a \in \mathcal{A}_{\mathrm{tr}}$ and $x \in \mathcal{Z}$, define*

$$q_a(x) := \sup_{\pi \in \Pi_\infty(a)} \langle x, c_\pi \rangle. \tag{A5.4}$$

Then $q_a(x)$ is a finite real number with

$$-\frac{R_{\max}}{1-\gamma} \leq q_a(x) \leq \frac{R_{\max}}{1-\gamma}, \tag{A5.5}$$

and, for every covered history h,

$$q_a(z_D(h)) = Q_a^*(h). \tag{A5.6}$$

The supremum in equation A5.4 need not be attained. Each fixed member $x \mapsto \langle x, c_\pi \rangle$ has the linear decoder, and equation A5.4 is their pointwise supremum on $\mathcal{Z}$.

Proof. For fixed a and $x \in \mathcal{Z}$, Lemma 30 says that the set in equation A5.4 is a nonempty subset of the closed interval displayed in equation A5.2. Completeness of $\mathbb{R}$ therefore gives a finite supremum, and the same interval bounds that supremum, proving equation A5.5.

Fix $h \in \mathcal{H}_{\mathrm{cov}}$. The index set $\Pi_\infty(a)$ in equation A5.4 is the identical root-independent forced-action family that appears in the setting definition of $Q_a^*(h)$. For each one of its members, the fixed-plan decoder gives the exact memberwise equality

$$\langle z_D(h), c_\pi \rangle = V_\pi(h). \tag{A5.7}$$

Taking the supremum of the two equal scalar families indexed by the same $\pi \in \Pi_\infty(a)$ yields

$$q_a(z_D(h)) = \sup_{\pi \in \Pi_\infty(a)} \langle z_D(h), c_\pi \rangle = \sup_{\pi \in \Pi_\infty(a)} V_\pi(h) = Q_a^*(h),$$

which is equation A5.6.

The proof involves no interchange of a limit and a supremum: Proposition 28 first constructs V_π and c_π for each fixed common π; only afterward does equation A5.4 take a pointwise supremum over the completed fixed-plan equalities. □

Proposition 32 (Optimal value and complete maximizing-action set)**.** *Under the setting-defined finite nonempty trusted action set and Proposition 31, define, for every $x \in \mathcal{Z}$,*

$$v(x) := \max_{a \in \mathcal{A}_{\mathrm{tr}}} q_a(x), \qquad A(x) := \arg\max_{a \in \mathcal{A}_{\mathrm{tr}}} q_a(x) = \{a \in \mathcal{A}_{\mathrm{tr}} : q_a(x) = v(x)\}. \tag{A5.8}$$

Then $v(x)$ is finite, $A(x)$ is a nonempty subset of $\mathcal{A}_{\mathrm{tr}}$, and

$$-\frac{R_{\max}}{1-\gamma} \le v(x) \le \frac{R_{\max}}{1-\gamma}. \tag{A5.9}$$

For every covered history h,

$$v(z_D(h)) = V^*(h), \qquad A(z_D(h)) = \mathcal{A}^*(h). \tag{A5.10}$$

The second equality preserves the entire tie set, with no action-gap or uniqueness assumption.

Proof. For fixed $x \in \mathcal{Z}$, Proposition 31 places each $q_a(x)$ in the finite interval equation A5.5. Because $\mathcal{A}_{\mathrm{tr}}$ is finite and nonempty, the finite list of real numbers $\{q_a(x) : a \in \mathcal{A}_{\mathrm{tr}}\}$ has at least one largest element. Consequently $v(x)$ is a finite real number, the set of all actions attaining that outer finite maximum is nonempty, and equation A5.9 follows from the common bounds on the list.

Now fix a covered history h. Applying equation A5.6 separately for every trusted action gives equality of the complete action-indexed vectors

$$\big(q_a(z_D(h))\big)_{a\in\mathcal{A}_{\mathrm{tr}}} = \big(Q_a^*(h)\big)_{a\in\mathcal{A}_{\mathrm{tr}}}. \tag{A5.11}$$

Taking the finite maximum of the two identical vectors and using the setting definition of V^* yields

$$v(z_D(h)) = \max_{a\in\mathcal{A}_{\mathrm{tr}}} q_a(z_D(h)) = \max_{a\in\mathcal{A}_{\mathrm{tr}}} Q_a^*(h) = V^*(h).$$

Moreover,

$$\begin{aligned} A(z_D(h)) &= \{a : q_a(z_D(h)) = v(z_D(h))\} \\ &= \{a : Q_a^*(h) = V^*(h)\} = \mathcal{A}^*(h), \end{aligned}$$

which proves equation A5.10 as equality of full sets.

The outer maximum over actions is attained because the action set is finite; the inner suprema defining q_a need not be. Because the proof compares exact real numbers, zero action gaps, all-action ties, and partially overlapping tie sets are covered directly. □

Proposition 33 (Common-family envelopes versus a root-labeled selector)**.** *Under Proposition 28, let*

$$s : \mathcal{H}_{\mathrm{cov}} \to \Pi_\infty, \qquad h \longmapsto s(h) =: \pi_h,$$

be any rule that selects a possibly different root-independent common plan after observing the full root history, and define its pointwise selected return by

$$F_s(h) := V_{\pi_h}(h). \tag{A5.12}$$

Then

$$F_s(h) = \langle z_D(h), c_{\pi_h}\rangle \quad \textit{for each covered } h. \tag{A5.13}$$

The fixed-plan decoder supplies a single coefficient for F_s only if the additional simultaneous compatibility condition

$$\textit{there exists } c_s \in \mathbb{R}^d \textit{ such that } \langle z_D(h), c_s\rangle = \langle z_D(h), c_{\pi_h}\rangle \quad \textit{for every covered } h \tag{A5.14}$$

holds. More generally, F_s factors through z_D as an arbitrary set map on $\mathcal{Z}$ if and only if it is constant on decision-core fibers:

$$z_D(h) = z_D(h') \quad \Longrightarrow \quad F_s(h) = F_s(h'). \tag{A5.15}$$

Neither equation A5.14 nor equation A5.15 follows from fixed-common-plan decoding when the selector may use root-history distinctions. Propositions 31 and 32, are consistent with this boundary, because their envelopes use the same complete common-plan index families at every $x \in \mathcal{Z}$.

Proof. For each fixed covered h, $\pi_h = s(h)$ is a member of the common-plan family. Applying Proposition 28 to that fixed member at that fixed history proves equation A5.13. But the right side of equation A5.13 contains c_{π_h}, whose index can change with h. Replacing it by one vector c_s is valid exactly when all the equations in equation A5.14 hold simultaneously. Proposition 28 proves those equations when the same plan π is fixed for all histories; for a history-varying collection of coefficient vectors, equation A5.14 is a separate condition.

For arbitrary code factorization, suppose first that $F_s = f_s \circ z_D$ for some map $f_s : \mathcal{Z} \to \mathbb{R}$. Equal codes then give equal f_s-values, proving the necessity of equation A5.15. Conversely, if equation A5.15 holds, define $f_s(x)$ to be $F_s(h)$ for any covered h with $z_D(h) = x$. Condition equation A5.15 makes this definition independent of the representative, so $F_s = f_s \circ z_D$. This proves the equivalence.

A selector may choose different common members at two histories with the same code; the individual decoded values at that common code need not be equal across different members, so fiber consistency is not automatic. By contrast, $q_a(x)$, $v(x)$, and $A(x)$ are formed from index sets fixed independently of a representative history, hence are functions of x by construction. If a special selector happens to be fiber-compatible, or happens to attain an already-defined envelope when an optimizer exists, its selected value may factor through the corresponding map; even then, a single linear coefficient requires equation A5.14. □

B.5.6 Minimal Exact-Linear Dimension and the Full-Rank Boundary

Lemma 34 (Exact comparator coordinate-span inclusion)**.** *Under Assumption 11, let*

$$\phi = (\phi_1, \dots, \phi_m) : \mathcal{H}_{\mathrm{cov}} \to \mathbb{R}^m, \qquad m \in \mathbb{N}_0,$$

be any exact linear controlled-test representation, where $\phi_j(h)$ denotes the j-th coordinate of $\phi(h)$. Then, in the algebraic function space

$$\mathbb{R}^{\mathcal{H}_{\mathrm{cov}}} = \{f : \mathcal{H}_{\mathrm{cov}} \to \mathbb{R}\},$$
$$\mathrm{colspan}(D_\Pi) \subseteq \mathrm{span}\{\phi_1, \dots, \phi_m\}. \tag{A6.1}$$

Consequently,

$$d \le \dim \mathrm{span}\{\phi_1, \dots, \phi_m\} \le m. \tag{A6.2}$$

For $m = 0$, the coordinate family is empty, its span is $\{0\}$, and equation A6.1–equation A6.2 imply $d = 0$.

Proof. Fix $\sigma \in \Sigma_{\mathrm{fin}}$. Exact linear decoding supplies at least one vector

$$c_\sigma^\phi = \big((c_\sigma^\phi)_1, \dots, (c_\sigma^\phi)_m\big) \in \mathbb{R}^m$$

such that, for every covered history h,

$$\begin{aligned} D_\Pi(h, \sigma) &= \langle \phi(h), c_\sigma^\phi \rangle \\ &= \sum_{j=1}^{m} (c_\sigma^\phi)_j \phi_j(h). \end{aligned} \tag{A6.3}$$

Thus, as equality of functions on the entire covered-history set,

$$D_\Pi(\cdot, \sigma) = \sum_{j=1}^{m} (c_\sigma^\phi)_j \phi_j(\cdot) \in \mathrm{span}\{\phi_1, \dots, \phi_m\}. \tag{A6.4}$$

Every element of $\mathrm{colspan}(D_\Pi)$ is a finite algebraic linear combination of such column functions. Taking the same finite combination of the right-hand sides of equation A6.4 proves equation A6.1. The argument holds for infinite $\mathcal{H}_{\mathrm{cov}}$ and Σ_{fin}.

By definition, the algebraic dimension of $\mathrm{colspan}(D_\Pi)$ is d. Inclusion equation A6.1 therefore gives the first inequality in equation A6.2. For the second, consider the linear map

$$\mathbb{R}^m \longrightarrow \mathbb{R}^{\mathcal{H}_{\mathrm{cov}}}, \qquad a \longmapsto \big[h \longmapsto \langle \phi(h), a \rangle\big]. \tag{A6.5}$$

Its image is exactly $\mathrm{span}\{\phi_1, \dots, \phi_m\}$, because the images of the standard coordinate vectors are precisely the functions $\phi_1, \dots, \phi_m$. Hence that image is spanned by at most m functions, which proves

$$\dim \mathrm{span}\{\phi_1, \dots, \phi_m\} \le m.$$

If the coordinate functions are redundant, the last inequality can be strict. Equivalently, the map in equation A6.5 may have a nontrivial kernel, and a column may admit multiple decoder vectors. Neither phenomenon affects equation A6.1: exact equality places the column function in the coordinate-function span regardless of decoder uniqueness. When $m = 0$, the empty inner product in equation A6.3 is zero, so every deterministic column is zero, the column span is $\{0\}$, and $d = 0$. The same conclusion holds if the covered-history set is empty. □

Proposition 35 (Linear factorization through any exact comparator)**.** *Under Assumption 11, Lemma 34, and Lemma 23, let*

$$\sigma_1, \dots, \sigma_d$$

be the actual deterministic-tree core and let

$$z_D(h) = \big(D_\Pi(h, \sigma_1), \dots, D_\Pi(h, \sigma_d)\big).$$

For every exact linear controlled-test representation $\phi : \mathcal{H}_{\mathrm{cov}} \to \mathbb{R}^m$, there exists a matrix

$$B \in \mathbb{R}^{d \times m}$$

such that

$$z_D(h) = B\phi(h) \quad \textit{for every } h \in \mathcal{H}_{\mathrm{cov}}. \tag{A6.6}$$

If $d = 0$, B is the unique $0 \times m$ matrix and equation A6.6 is equality of the unique empty vectors in $\mathbb{R}^0$.

Proof. For each of the finitely many core indices σ_i, exact linear decoding supplies at least one vector

$$c^\phi_{\sigma_i} \in \mathbb{R}^m$$

such that

$$D_\Pi(h, \sigma_i) = \langle \phi(h), c^\phi_{\sigma_i} \rangle \quad \text{for every covered } h. \tag{A6.7}$$

Choose one such vector for each $i = 1, \dots, d$, and define

$$B := \begin{bmatrix} (c^\phi_{\sigma_1})^\top \\ \vdots \\ (c^\phi_{\sigma_d})^\top \end{bmatrix} \in \mathbb{R}^{d \times m}. \tag{A6.8}$$

For any covered h and each $i = 1, \dots, d$, equations equation A6.7–equation A6.8 give

$$[B\phi(h)]_i = (c^\phi_{\sigma_i})^\top \phi(h) = D_\Pi(h, \sigma_i) = [z_D(h)]_i. \tag{A6.9}$$

Equality of every coordinate proves equation A6.6.

If some comparator coordinates are redundant, the decoder vectors in equation A6.7, and hence B, need not be unique. The proposition asserts existence of B. In particular, B may have a nontrivial kernel, and $\phi(h)$ need not be recoverable from $z_D(h)$.

When $d = 0$, there are no core decoder rows to choose. The empty stack in equation A6.8 is the unique $0 \times m$ matrix, and both $z_D(h)$ and $B\phi(h)$ are the unique vector $() \in \mathbb{R}^0$. When $m = 0$, Lemma 34 already forces $d = 0$, so the only possible case is the same 0×0 empty factorization. □

Proposition 36 (Exact-linear minimum)**.** *Under Assumption 11, Lemma 34, Lemma 23, and Proposition 24, the decision-core map*

$$z_D : \mathcal{H}_{\mathrm{cov}} \to \mathbb{R}^d$$

is an exact linear controlled-test representation. Every exact linear controlled-test representation $\phi : \mathcal{H}_{\mathrm{cov}} \to \mathbb{R}^m$ satisfies $m \geq d$. Therefore dimension d is the minimum in the declared exact-linear comparison class and z_D attains it. This remains exact when $d = 0$.

Proof. Proposition 24 states that, for every deterministic finite tree σ, there is a vector $c_\sigma \in \mathbb{R}^d$ satisfying

$$D_\Pi(h, \sigma) = \langle z_D(h), c_\sigma \rangle \quad \text{for every } h \in \mathcal{H}_{\mathrm{cov}}. \tag{A6.10}$$

Equation equation A6.10 is exactly the definition of a d-dimensional exact linear controlled-test representation. Thus z_D belongs to the declared comparison class.

Now let $\phi : \mathcal{H}_{\mathrm{cov}} \to \mathbb{R}^m$ be any other member of that class. Lemma 34 gives

$$d \leq \dim \operatorname{span}\{\phi_1, \dots, \phi_m\} \leq m, \tag{A6.11}$$

and therefore $m \geq d$. Since equation A6.10 supplies an exact representation whose codomain dimension is exactly d, the lower bound is attained.

The actual core also shows that the d coordinate functions of z_D are the actual basis columns

$$D_\Pi(\cdot, \sigma_1), \dots, D_\Pi(\cdot, \sigma_d);$$

for $d > 0$ they are linearly independent, and for $d = 0$ the family is empty. In the latter case, Proposition 24 proves that every deterministic column is zero, $z_D(h) = ()$, $c_\sigma = ()$, and the empty inner product in equation A6.10 is zero. Hence the zero-dimensional core is an exact attainment. □

Proposition 37 (Full algebraic rank in the declared class)**.** *Under Assumption 11 and Proposition 36, suppose, for this specialization, that*

$$d = |\mathcal{S}|. \tag{A6.12}$$

Then every declared exact linear controlled-test representation $\phi : \mathcal{H}_{\mathrm{cov}} \to \mathbb{R}^m$ satisfies

$$m \geq |\mathcal{S}|, \tag{A6.13}$$

and the core representation z_D attains dimension $|\mathcal{S}|$. Statement equation A6.13 is a dimension lower bound in the declared exact-linear class, and it leaves decision-equivalent histories free to merge. In particular,

$$h \equiv_D h' \quad \Longrightarrow \quad z_D(h) = z_D(h') \tag{A6.14}$$

remains true under equation A6.12, so decision-equivalent histories may still be merged.

Proof. For an arbitrary declared exact-linear comparator, Proposition 36 gives $m \geq d$. Substituting the local full-rank condition equation A6.12 proves equation A6.13. The same proposition shows that z_D is exact and has codomain dimension

$$d = |\mathcal{S}|,$$

so the lower bound is attained within the class.

To verify the merger boundary directly, suppose $h \equiv_D h'$. By definition equation E,

$$D_\Pi(h, \sigma) = D_\Pi(h', \sigma) \quad \text{for every } \sigma \in \Sigma_{\mathrm{fin}}.$$

This equality holds in particular for each selected core index σ_i. Therefore

$$\begin{aligned} z_D(h) &= \big(D_\Pi(h, \sigma_1), \ldots, D_\Pi(h, \sigma_d)\big) \\ &= \big(D_\Pi(h', \sigma_1), \ldots, D_\Pi(h', \sigma_d)\big) = z_D(h'), \end{aligned}$$

which proves equation A6.14; the converse is established in the quotient step. Full algebraic column rank is therefore compatible with a non-injective history-to-core map.

Finally, the proof of equation A6.13 uses the exact equality equation A6.3 to place every original decision column in a finite coordinate-function span; this exact span inclusion is what defines the declared exact-linear class. □

B.5.7 Safe Quotients, Cue Retention, and the Information Boundary

Lemma 38 (Decision-core fibers equal finite decision rows)**.** *Under Assumption 6 and Proposition 24, for all $h, h' \in \mathcal{H}_{\mathrm{cov}}$,*

$$z_D(h) = z_D(h') \quad \Longleftrightarrow \quad D_\Pi(h, \sigma) = D_\Pi(h', \sigma) \text{ for every } \sigma \in \Sigma_{\mathrm{fin}} \quad \Longleftrightarrow \quad h \equiv_D h'. \tag{A7.1}$$

When $d = 0$, $z_D(h)$ is the unique empty vector for every covered history, every deterministic finite column function is zero, and all covered histories lie in one decision-equivalence class.

Proof. Suppose first that $z_D(h) = z_D(h')$. For an arbitrary $\sigma \in \Sigma_{\mathrm{fin}}$, Proposition 24 gives

$$D_\Pi(h, \sigma) = \langle z_D(h), c_\sigma \rangle = \langle z_D(h'), c_\sigma \rangle = D_\Pi(h', \sigma). \tag{A7.2}$$

Thus every deterministic finite column agrees, which is exactly $h \equiv_D h'$.

Conversely, suppose $h \equiv_D h'$. The selected core indices $\sigma_1, \ldots, \sigma_d$ are themselves members of Σ_{fin}, so row equality gives

$$D_\Pi(h, \sigma_i) = D_\Pi(h', \sigma_i) \quad \text{for every } i = 1, \ldots, d.$$

The definition of z_D as the tuple of these entries therefore gives $z_D(h) = z_D(h')$.

If $d = 0$, Proposition 24 gives $c_\sigma = ()$ and $D_\Pi(h, \sigma) = \langle (), () \rangle = 0$ for every h, σ. Hence both sides of equation A7.1 hold for every pair. □

Proposition 39 (Coarsest exact return-safe quotient)**.** *Under Assumption 6 and Lemma 38, let*

$$q_D : \mathcal{H}_{\mathrm{cov}} \longrightarrow \mathcal{H}_{\mathrm{cov}}/\equiv_D, \qquad q_D(h) := [h]_D, \tag{A7.3}$$

be the canonical decision-equivalence quotient. Then:

1. q_D is exact return-safe. 2. If $\eta : \mathcal{H}_{\mathrm{cov}} \to \mathcal{Y}$ is any exact return-safe map, then

$$\eta(h) = \eta(h') \Longrightarrow h \equiv_D h'. \tag{A7.4}$$

Consequently there is a unique map on the attained image,

$$\kappa_\eta : \eta(\mathcal{H}_{\mathrm{cov}}) \longrightarrow \mathcal{H}_{\mathrm{cov}}/\equiv_D, \qquad \kappa_\eta(\eta(h)) := [h]_D, \tag{A7.5}$$

satisfying $q_D = \kappa_\eta \circ \eta$. 3. z_D is exact return-safe and has exactly the same fibers as q_D.

Thus every exact return-safe partition refines the decision-equivalence partition, while the decision-equivalence quotient has the largest safe fibers and is coarsest in partition order. The code z_D realizes that coarsest quotient exactly.

Proof. Equality of complete deterministic-column rows is reflexive, symmetric, and transitive, so $\equiv_D$ is an equivalence relation and equation A7.3 is well defined. If $q_D(h) = q_D(h')$, then $h \equiv_D h'$; by definition, every finite deterministic controlled-test column has equal values at the two histories. Hence each such column is constant on every q_D-fiber, proving exact return safety.

Now let η be exact return-safe and suppose $\eta(h) = \eta(h')$. Return safety says that every finite controlled-test column is constant on that fiber, so

$$D_\Pi(h, \sigma) = D_\Pi(h', \sigma) \quad \text{for all } \sigma \in \Sigma_{\mathrm{fin}}.$$

This is $h \equiv_D h'$, proving equation A7.4. Therefore the definition in equation A7.5 is independent of the representative h: two representatives of one attained η-value are decision equivalent. Equation $q_D = \kappa_\eta \circ \eta$ follows immediately. This direction is essential: the coarser decision quotient factors through every safer, possibly finer, map η; an arbitrary finer η need not factor through q_D, because it may retain decision-null nuisance information.

Finally, Lemma 38 says precisely

$$z_D(h) = z_D(h') \Longleftrightarrow q_D(h) = q_D(h').$$

Equal z_D-values therefore force equality of every finite column, proving that z_D is safe, and the biconditional proves equality of the two fiber partitions. At $d = 0$, the one empty-code fiber is safe because every finite column is zero, so the same conclusion holds without a nonempty core. □

Proposition 40 (Full algebraic rank as a class boundary)**.** *Under Propositions 36 and 37 and Lemma 38, suppose locally that*

$$d = |\mathcal{S}|. \tag{A7.6}$$

Then every m-dimensional exact linear controlled-test representation in the declared comparator class satisfies

$$m \geq |\mathcal{S}|, \tag{A7.7}$$

and the core attains dimension $|\mathcal{S}|$. Nevertheless,

$$h \equiv_D h' \quad \Longleftrightarrow \quad z_D(h) = z_D(h') \tag{A7.8}$$

still holds. Hence any decision-equivalent histories that occur may still be merged by the coarsest exact return-safe quotient.

Proof. Proposition 36 gives $m \geq d$ for every declared exact-linear comparator and gives an exact core representation of dimension d. Substitution of equation A7.6 proves equation A7.7 and attainment. Independently, equation A7.8 is Lemma 38; its proof applies unchanged when d equals the ambient hidden-state dimension.

Thus full algebraic decision rank prevents reduction below the state count inside the class whose coordinates linearly decode every finite controlled test. Whether two distinct histories share a complete decision row is a separate question; when they do, Proposition 39 permits their merger. □

Proposition 41 (A fixed infinite consequence has a finite deterministic cue witness)**.** *Under Assumption 8, Propositions 16 and 29, and Lemma 38, fix $h, h' \in \mathcal{H}_{\mathrm{cov}}$ and one root-independent common trusted plan $\pi \in \Pi_\infty$. If*

$$\delta := V_\pi(h) - V_\pi(h') \neq 0, \tag{A7.9}$$

then there exist an integer $L \geq 1$ and a deterministic tree $\sigma \in \Sigma_L \subset \Sigma_{\mathrm{fin}}$ with $w_{\pi,L}(\sigma) > 0$ such that

$$D_\Pi(h, \sigma) \neq D_\Pi(h', \sigma), \qquad z_D(h) \neq z_D(h'). \tag{A7.10}$$

If $\pi \in \Pi_\infty(a)$ for a trusted forced root action a, the witnessing positive-weight tree in equation A7.10 also has root action a. The depth is pair-and-plan specific and may be chosen so that

$$\frac{2R_{\max}\gamma^L}{1-\gamma} < |\delta|.$$

A deterministic plan is the one-atom case, and a finite continuation is the exact zero-tail specialization.

Proof. Define the auxiliary inherited tail scale

$$\varepsilon_L := \frac{R_{\max}\gamma^L}{1-\gamma}. \tag{A7.11}$$

The tail certificate includes $0 < \gamma < 1$ and finite $R_{\max}$, so $\varepsilon_L \to 0$. Hence, for the pair-and-plan-specific nonzero δ, choose an integer $L \geq 1$ such that

$$2\varepsilon_L < |\delta|. \tag{A7.12}$$

This remains valid when $R_{\max} = 0$, in which case $\varepsilon_L = 0$; the fixed-plan bound would in fact make the premise $\delta \neq 0$ impossible, so no inconsistency arises.

Write

$$\Delta_L := b_h^\top \alpha_{\pi^{(L)}} - b_{h'}^\top \alpha_{\pi^{(L)}}. \tag{A7.13}$$

Apply Proposition 29 at h and h' for this same plan and depth. The triangle inequality gives the exact two-tail comparison

$$\begin{aligned} |\Delta_L - \delta| &= \left|\left(b_h^\top \alpha_{\pi^{(L)}} - V_\pi(h)\right) - \left(b_{h'}^\top \alpha_{\pi^{(L)}} - V_\pi(h')\right)\right| \\ &\leq \left|b_h^\top \alpha_{\pi^{(L)}} - V_\pi(h)\right| + \left|b_{h'}^\top \alpha_{\pi^{(L)}} - V_\pi(h')\right| \\ &\leq 2\varepsilon_L < |\delta|. \end{aligned} \tag{A7.14}$$

If $\Delta_L = 0$, equation A7.14 would read $|\delta| < |\delta|$, a contradiction. Thus $\Delta_L \neq 0$.

Proposition 16 uses the same weights at both roots, so

$$\Delta_L = \sum_{\sigma \in \Sigma_L} w_{\pi,L}(\sigma)\left[D_\Pi(h, \sigma) - D_\Pi(h', \sigma)\right]. \tag{A7.15}$$

The sum is finite. If every term with positive weight had zero bracket, then equation A7.15 would be zero. Therefore some $\sigma \in \Sigma_L$ satisfies

$$w_{\pi,L}(\sigma) > 0, \qquad D_\Pi(h, \sigma) - D_\Pi(h', \sigma) \neq 0.$$

Lemma 38 then gives $z_D(h) \neq z_D(h')$.

When $\pi \in \Pi_\infty(a)$, freezing its exogenous tape cannot change the forced first action. Because $L \geq 1$, every deterministic restriction in the support has root action a, including the witness above. If the common plan is already terminal by a finite depth, Proposition 29 makes both actual tails zero beyond that depth, so the same argument reduces to an exact finite mixture; a deterministic finite tree reduces further to direct row separation. The conclusion is pair specific and uses no uniform margin. □

Proposition 42 (Different complete optimal-action sets force cue separation)**.** *Under Propositions 31 and 32, for all $h, h' \in \mathcal{H}_{\mathrm{cov}}$,*

$$\mathcal{A}^*(h) \neq \mathcal{A}^*(h') \quad \Longrightarrow \quad z_D(h) \neq z_D(h'). \tag{A7.16}$$

The sets in equation A7.16 are the complete sets of maximizing trusted actions. No positive action gap, unique maximizer, optimizer-attaining plan, or selected tie-break representative is required.

Proof. We prove the contrapositive. Suppose $z_D(h) = z_D(h') =: x \in \mathcal{Z}$. Proposition 31 gives, for every trusted action a,

$$Q_a^*(h) = q_a(x) = Q_a^*(h'). \tag{A7.17}$$

Thus the entire finite action-indexed value vectors coincide. Applying the full-argmax map to the same point x gives

$$\mathcal{A}^*(h) = A(x) = \mathcal{A}^*(h'). \tag{A7.18}$$

Equations equation A7.17–equation A7.18 remain valid when the inner plan suprema are unattained, when all actions tie, or when tie sets have more than one element. Therefore different complete sets imply different codes, proving equation A7.16. If $d = 0$, every covered history has the same empty code, so equation A7.18 correctly shows that the premise of equation A7.16 cannot occur. □

Proposition 43 (Exact future-realization nonrecovery and exact decision-consequence maps)**.** *Under Assumption 6, fix a law $\mathbb{P}$, the positive-probability covered event $C = \{H_t \in \mathcal{H}_{\mathrm{cov}}\}$, and the conditional law $\mathbb{P}_C = \mathbb{P}(\cdot \mid C)$. Let $U = U_{t+k}$, $k \geq 1$, take values in a standard Borel space U, and suppose U has no $\sigma(H_t)$-measurable version under $\mathbb{P}_C$. Then there is no Borel map*

$$\rho : \mathbb{R}^d \longrightarrow \mathsf{U}$$

such that

$$\mathbb{P}_C\{U = \rho(z_D(H_t))\} = 1. \tag{A7.19}$$

Separately, Proposition 28 and Propositions 31 and 32 give the following exact factorizations pointwise for every covered history, and therefore $\mathbb{P}_C$-almost surely when evaluated at H_t:

$$\begin{aligned} &V_\pi(H_t) = \ell_\pi(z_D(H_t)), \quad \ell_\pi(x) := \langle x, c_\pi \rangle \quad \text{for every fixed common } \pi, \\ &Q_a^*(H_t) = q_a(z_D(H_t)) \qquad\qquad\qquad\quad \text{for every trusted } a, \\ &V^*(H_t) = v(z_D(H_t)), \\ &\mathcal{A}^*(H_t) = A(z_D(H_t)). \end{aligned} \tag{A7.20}$$

Here q_a, v, A are exact maps on the reachable image $\mathcal{Z}$, and A is set-valued with values in the finite power set of trusted actions. Each ℓ_π is linear, and equation A7.20 concerns conditional controlled reward consequences rather than the realization U or its full posterior law.

Proof. For each finite state s, the posterior coordinate $b_{H_t}(s)$ is a version of a conditional probability given H_t and is therefore $\sigma(H_t)$-measurable. Conditioning further on the event $C \in \sigma(H_t)$ does not change this measurability. Each core coordinate is the finite sum

$$D_\Pi(H_t, \sigma_i) = b_{H_t}^\top \alpha_{\sigma_i} = \sum_{s \in \mathcal{S}} b_{H_t}(s) \alpha_{\sigma_i}(s), \tag{A7.21}$$

so it is $\sigma(H_t)$-measurable. Since d is finite, the vector $z_D(H_t)$ is $\sigma(H_t)$-measurable and

$$\sigma(z_D(H_t)) \subseteq \sigma(H_t) \quad \text{under } \mathbb{P}_C. \tag{A7.22}$$

If a Borel ρ satisfied equation A7.19, then $\widetilde{U} := \rho(z_D(H_t))$ would be a $\sigma(H_t)$-measurable U-valued random element by measurable composition. Equation equation A7.19 would make $\widetilde{U}$ a $\sigma(H_t)$-measurable version of U under $\mathbb{P}_C$, contradicting the premise. This proves exact nonrecovery. History measurability of U is therefore necessary for decoding from the generally coarser z_D.

For the positive statements, on C the realized history belongs to $\mathcal{H}_{\mathrm{cov}}$. Proposition 28, evaluated at H_t, gives the first line of equation A7.20 for each fixed common plan. Propositions 31 and 32, evaluated at the same reachable code, give the remaining lines. These are set-theoretic exact factorization statements on $\mathcal{Z}$. They represent the conditional controlled reward consequences after future uncertainty has been integrated out. □

B.5.8 PROOF OF THE MAIN THEOREM

Proof of Theorem 12. For part (i), Lemma 13 gives the POMDP-consistent reward, observation, posterior, and zero-branch identities. Proposition 14 applies them by finite-tree induction to prove equation MT1. Proposition 15 gives the equivalent stopped-prefix population functional, with the positive-root and oracle-relative null-root interpretations stated above. Proposition 16 proves the common finite weights and both identities in equation MT2. Lemmas 17 and 18 prove reward inclusion and the exact tree-difference identity, respectively; Propositions 19 and 20 yield equation MT3 and its zero-defect finite-composition extension.

For part (ii), Lemma 21 identifies the decision-column span with the image of $\mathcal{W}_{\Pi}$. Proposition 22 gives equation MT4. Lemma 23 selects a basis of actual deterministic columns, including the empty basis when $d = 0$, and Proposition 24 gives equation MT5 together with representative independence modulo the kernel of the evaluation map.

For part (iii), Proposition 25 places every common randomized truncation in $\mathcal{W}_{\Pi}$. Lemma 26 proves the exact same-plan stage expansion and geometric Cauchy estimate. Proposition 27 constructs $\alpha_{\pi} \in \mathcal{W}_{\Pi}$, Proposition 28 supplies c_{π} and equation MT6, and Proposition 29 proves equation MT7 with the stated modes, dependence, and exact finite-termination baseline.

For part (iv), Lemma 30 proves that each forced-action value family is nonempty and bounded on $\mathcal{Z}$. Proposition 31 gives the first identity in equation MT9, and Proposition 32 gives the other two while retaining every tie. Proposition 33 proves the final selector qualification.

For part (v), Lemma 34 proves equation MT10. Proposition 35 constructs the stated $d \times m$ matrix B. Proposition 36 proves that the core belongs to the declared class and attains dimension d, and Proposition 37 gives the class-restricted full-rank statement.

For part (vi), Lemma 38 proves equation MT11. Proposition 39 proves safety, maximal safe fibers, and the correct quotient-factorization direction; Proposition 40 shows that this remains compatible with full algebraic decision rank. Proposition 41 proves the fixed-consequence cue implication while preserving any forced root action. Proposition 42 proves equation MT12 for complete action sets. Proposition 43 proves the exact Borel nonrecovery statement and, separately, the positive fixed-plan and planning-consequence factorizations. Together these give the six parts of Theorem 12. □

C FINITE-CAPACITY AND FINITE-SAMPLE THEORY

C.1 THEORETICAL SETUP

C.1.1 PROBLEM SETTING

Fix a finite discounted partially observed control system with finite state, observation, and action sets $\mathcal{S}, \mathcal{O}, \mathcal{A}$, common initial law ρ, discount $\gamma \in [0,1)$, and a finite trusted task family $\mathcal{T}$. The tasks share the controlled state/observation system and differ through bounded reward kernels. For task τ, let $\mathcal{R}_\tau$ be the set of pre-decision histories reachable from ρ under arbitrary history policies, and let $\mathcal{H}_\tau \subseteq \mathcal{R}_\tau$ be a fixed task-reachable essential support with $\mu_\tau(\mathcal{H}_\tau) = 1$. For $h \in \mathcal{R}_\tau$, define

$$Q^\star_\tau(h,a) := \sup_\pi \mathbb{E}^\pi_\tau\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \,\middle|\, H_t = h, A_t = a\right], \qquad V^\star_\tau(h) := \max_{a\in\mathcal{A}} Q^\star_\tau(h,a).$$

Let V^π_τ denote the scalar discounted value from ρ and $V^\star_\tau := \sup_\pi V^\pi_\tau$. Put

$$B := \frac{R_{\max}}{1-\gamma},$$

so $0 \le Q^\star_\tau \le B$ on $\mathcal{R}_\tau$.

A shared, task-blind encoder ϕ receives the history h but no separate task label. Let Φ_M be a finite class of such encoders, with

$$\mathcal{Z}_\phi := \mathrm{range}(\phi), \qquad |\mathcal{Z}_\phi| \le M, \qquad R(\phi) := \log|\mathcal{Z}_\phi|.$$

Each task uses a head from the finite class $\mathcal{G}_\phi \subseteq \mathcal{A}^{\mathcal{Z}_\phi}$. For $g \in \mathcal{G}_\phi$, set $\pi_{\tau,\phi,g}(h) := g(\phi(h))$ on $\mathcal{R}_\tau$ and, on the task's scoring domain, define

$$\ell_{\tau,\phi,g}(h) := V^\star_\tau(h) - Q^\star_\tau\big(h, g(\phi(h))\big) \in [0,B], \qquad L_\tau(\phi,g) := \mathbb{E}_{H\sim\mu_\tau}\ell_{\tau,\phi,g}(H),$$
$$D_\tau(\phi) := \min_{g\in\mathcal{G}_\phi} L_\tau(\phi,g), \qquad D(\phi) := \max_{\tau\in\mathcal{T}} D_\tau(\phi).$$

Thus D_τ scores histories in task τ's own domain, whereas D requires one shared codebook to serve the union of all task-reachable scoring domains.

For every task, an independent target-estimation split returns $\widehat{Q}_\tau : \mathcal{H}_\tau \times \mathcal{A} \to \mathbb{R}$, and a separate evaluation collection contains $H_{\tau,1}, \dots, H_{\tau,n_\tau}$. Write

$$P_{n_\tau} f := \frac{1}{n_\tau}\sum_{i=1}^{n_\tau} f(H_{\tau,i})$$

and define

$$\widehat{\ell}_{\tau,\phi,g}(h) := \max_{a\in\mathcal{A}} \widehat{Q}_\tau(h,a) - \widehat{Q}_\tau\big(h, g(\phi(h))\big), \qquad \widehat{D}_\tau(\phi) := \min_{g\in\mathcal{G}_\phi} P_{n_\tau}\widehat{\ell}_{\tau,\phi,g}.$$

Choose $\widehat{g}_{\tau,\phi} \in \arg\min_{g\in\mathcal{G}_\phi} P_{n_\tau}\widehat{\ell}_{\tau,\phi,g}$, fix $\lambda > 0$, and set

$$\widehat{\phi} \in \arg\min_{\phi\in\Phi_M}\left\{R(\phi) + \lambda \max_{\tau\in\mathcal{T}} \widehat{D}_\tau(\phi)\right\}, \qquad \widehat{g}_\tau := \widehat{g}_{\tau,\widehat{\phi}}, \qquad \widehat{\pi}_\tau := \pi_{\tau,\widehat{\phi},\widehat{g}_\tau}.$$

The exact indexed class count is

$$N_{\mathrm{cls}} := \sum_{\tau\in\mathcal{T}}\sum_{\phi\in\Phi_M} |\mathcal{G}_\phi| \le |\mathcal{T}|\,|\Phi_M|\,|\mathcal{A}|^M.$$

For a task-τ history policy π, its normalized discounted occupancy is

$$d^\pi_\tau(E) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr^\pi_\tau(H_t \in E), \qquad E \subseteq \mathcal{R}_\tau.$$

For the structural clauses, write the observable decomposition as $(X,N) = (x(H), n_0(H))$. For each task, put $\mathcal{X}^{\mathrm{reach}}_\tau := x(\mathcal{R}_\tau)$ and define the core Bellman operator on bounded measurable core action-values by

$$(\mathcal{B}_\tau q)(x,a) := \bar{r}_\tau(x,a) + \gamma\mathbb{E}_{X'\sim P_{X,\tau}(\cdot|x,a)} \max_{a'\in\mathcal{A}} q(X',a').$$

The proof below constructs this operator's unique bounded measurable fixed point, denotes it by $q_\tau^\star$, and proves $q_\tau^\star = \mathcal{B}_\tau q_\tau^\star$ before using it in the structural conclusions. Fix weights $w_\tau > 0$ with $\sum_\tau w_\tau = 1$ and let

$$\bar{\mu} := \sum_{\tau \in \mathcal{T}} w_\tau \mu_\tau, \qquad \bar{\mu}_X := \sum_{\tau \in \mathcal{T}} w_\tau \mu_{\tau,X}, \qquad r_\tau(x) := \frac{d\mu_{\tau,X}}{d\bar{\mu}_X}(x).$$

Up to $\bar{\mu}_X$-null sets, define

$$\mathcal{X}_\tau := \{x : r_\tau(x) > 0\}, \qquad \mathcal{T}(x) := \{\tau \in \mathcal{T} : r_\tau(x) > 0\},$$

$$\mathcal{U}_{\mathrm{reach}}(x) := \left\{ u \in \mathcal{A}^{\mathcal{T}} : u_\tau \in \arg\max_{a \in \mathcal{A}} q_\tau^\star(x,a) \text{ for every } \tau \in \mathcal{T}(x) \right\},$$

where coordinates outside $\mathcal{T}(x)$ are unconstrained, and

$$m_{\mathrm{dec}}^{\mathrm{reach}} := \min\left\{ |\mathcal{C}| : \mathcal{C} \subseteq \mathcal{A}^{\mathcal{T}},\ \mathcal{C} \cap \mathcal{U}_{\mathrm{reach}}(x) \neq \varnothing \text{ for } \bar{\mu}_X\text{-almost every } x \right\}.$$

This integer is finite because $\mathcal{A}$ and $\mathcal{T}$ are finite. Let $\nu_{\tau,x} := \mathcal{L}_{\mu_\tau}(N \mid X = x)$ for $\tau \in \mathcal{T}(x)$ and define the nuisance conditional under the history mixture by

$$\bar{\nu}_x := \sum_{\tau \in \mathcal{T}} w_\tau r_\tau(x) \nu_{\tau,x},$$

with zero-weight terms omitted. Since $\sum_\tau w_\tau r_\tau(x) = 1$ almost everywhere, $\bar{\nu}_x$ is a probability law. For any encoder, define its conditionally active and globally mixture-active codes by

$$\mathcal{Z}_\phi(x) := \{z \in \mathcal{Z}_\phi : \bar{\nu}_x\{n : \phi(x,n) = z\} > 0\}, \qquad \mathcal{Z}_\phi^{\bar{\mu}} := \{z \in \mathcal{Z}_\phi : \bar{\mu}(\phi(H) = z) > 0\}.$$

Every $Q_\tau^\star$ and $\widehat{Q}_\tau$ is a full-history/action target on task τ's own domain. Both may depend on the full history rather than on $Z = \phi(H)$; deployment uses only the shared encoder and the relevant task head.

C.1.2 Assumptions

Assumption 44 (Finite bounded discounted tasks)**.** *The sets $\mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{T}$ are finite, $0 \le \gamma < 1$, and each task reward kernel is supported on $[0, R_{\max}]$. All tasks use the common initial law and controlled state/observation system.*

Assumption 45 (Fixed finite shared representation classes)**.** *The integer $M \ge 1$, the class Φ_M is finite and nonempty, every $\mathcal{G}_\phi$ is finite and nonempty, $|\mathrm{range}(\phi)| \le M$, the encoder has no separate task-label input, and all classes are fixed before any target or evaluation split is observed.*

Assumption 46 (Taskwise reachable iid evaluation)**.** *For every τ, $n_\tau \ge 1$, $H_{\tau,1}, \dots, H_{\tau,n_\tau} \overset{\mathrm{iid}}{\sim} \mu_\tau$, and $\mu_\tau(\mathcal{H}_\tau) = 1$ with $\mathcal{H}_\tau \subseteq \mathcal{R}_\tau$. Each history is extracted by the same pre-specified discounted-time rule from a different task-τ episode. All taskwise evaluation collections and target-estimation splits are mutually independent.*

Assumption 47 (Independent taskwise target certificates)**.** *On the pre-evaluation certificate event*

$$\mathcal{E}_Q := \bigcap_{\tau \in \mathcal{T}} \left\{ \sup_{h \in \mathcal{H}_\tau} \sup_{a \in \mathcal{A}} |\widehat{Q}_\tau(h,a) - Q_\tau^\star(h,a)| \le \eta_\tau \right\},$$

the displayed bounds hold for known $\eta_\tau \ge 0$. We condition the result on this independently obtained event; if task τ's finite model is known, then $\eta_\tau = 0$.

Assumption 48 (Taskwise candidate-policy coverage)**.** *For every τ there is a finite $C_\tau \ge 1$ such that, for every $\phi \in \Phi_M$ and $g \in \mathcal{G}_\phi$,*

$$d_\tau^{\pi_{\tau,\phi,g}} \ll \mu_\tau, \qquad \left\| \frac{d d_\tau^{\pi_{\tau,\phi,g}}}{d\mu_\tau} \right\|_\infty \le C_\tau.$$

Assumption 49 (Task-reachable observable core–nuisance factorization)**.** *On $\bigcup_\tau \mathcal{R}_\tau$, every history is measurably identified with an observable pair $(X, N) = (x(H), n_0(H))$. Writing $H_{t+1} = (H_t, A_t, \Delta H_{t+1})$, there is a measurable F_X such that $X_{t+1} = F_X(X_t, A_t, \Delta H_{t+1})$. For every task τ, $h = (x,n) \in \mathcal{R}_\tau$, and $a \in \mathcal{A}$, there are a reward kernel K_τ^r and a controlled core kernel $P_{X,\tau}$ such that*

$$\mathcal{L}_\tau(R_t \mid H_t = h, A_t = a) = K_\tau^r(\cdot \mid x, a), \qquad \mathcal{L}_\tau(X_{t+1} \mid H_t = h, A_t = a) = P_{X,\tau}(\cdot \mid x, a),$$

with $\bar{r}_\tau(x,a) := \mathbb{E}_{K_\tau^r(\cdot|x,a)} R_t$. The nuisance dynamics may otherwise be arbitrary and unpredictable. This assumption is local to each task domain; value factorization is derived from it.

Assumption 50 (Shared-core nuisance support equivalence)**.** *For $\bar{\mu}_X$-almost every x,*

$$\nu_{\tau,x} \sim \nu_{\sigma,x} \qquad \text{for every } \tau, \sigma \in \mathcal{T}(x),$$

where $\sim$ denotes mutual absolute continuity. Thus a measurable nuisance set, and hence a deterministic encoder code at fixed x, has positive conditional probability for one task active at x if and only if it has positive conditional probability for every such task. No density-ratio bound is imposed.

Assumption 51 (Adequate shared core capacity)**.** *The inequality $m_{\mathrm{dec}}^{\mathrm{reach}} \le M$ holds, and there are a minimizing cover $\mathcal{C}_{\mathrm{dec}} = \{u_1, \dots, u_{m_{\mathrm{dec}}^{\mathrm{reach}}}\}$, a measurable selector*

$$c_{\mathrm{dec}} : \bigcup_{\tau} \mathcal{X}_{\tau}^{\mathrm{reach}} \longrightarrow \{1, \dots, m_{\mathrm{dec}}^{\mathrm{reach}}\}, \qquad u_{c_{\mathrm{dec}}(x)} \in \mathcal{U}_{\mathrm{reach}}(x) \quad \bar{\mu}_X\text{-almost everywhere},$$

and a core-only shared encoder

$$\phi_{\mathrm{dec}}(h) := c_{\mathrm{dec}}(x(h)) \in \Phi_M, \qquad |\mathcal{Z}_{\phi_{\mathrm{dec}}}| = |\mathcal{Z}_{\phi_{\mathrm{dec}}}^{\bar{\mu}}| = m_{\mathrm{dec}}^{\mathrm{reach}},$$

whose coordinate heads $g_{\mathrm{dec},\tau}(j) := (u_j)_\tau$ belong to $\mathcal{G}_{\phi_{\mathrm{dec}}}$ for every task. This is a capacity and class-inclusion condition defined through the reachable core problems; $D(\phi_{\mathrm{dec}}) = 0$ is derived from it.

C.2 Preliminaries

Throughout, $\log$ is the natural logarithm, so the fixed-length rate $R(\phi) = \log|\mathcal{Z}_\phi|$ is measured in nats. Let $\mathbb{P}_{\mathrm{eval}}$ denote the marginal law of all evaluation collections. A certified target realization is fixed first, and probability is taken over the independent evaluation collections.

For $\delta \in (0,1)$, define the taskwise sampling, calibrated, and maximal radii

$$\epsilon_\tau(\delta) := B\sqrt{\frac{\log(2N_{\mathrm{cls}}/\delta)}{2n_\tau}}, \qquad e_\tau(\delta) := \epsilon_\tau(\delta) + 2\eta_\tau, \qquad e_{\max}(\delta) := \max_{\tau \in \mathcal{T}} e_\tau(\delta).$$

They are the only statistical radii in the main theorem. The first records the exact finite-class sampling deviation, the second adds the two target errors, and the third enters the shared maximum-deficiency objective.

Every target, empirical loss, and population risk keeps its task index and its declared domain. In particular, $Q_\tau^\star$ is used only on $\mathcal{R}_\tau$, $\widehat{Q}_\tau$ only on $\mathcal{H}_\tau \times \mathcal{A}$, and L_τ only under μ_τ. The quotient conclusions are $\bar{\mu}_X$-almost-everywhere statements and constrain only coordinates $\tau \in \mathcal{T}(x)$. These are domain conventions.

C.3 Main Theorem

Theorem 52 (Taskwise decision-risk control and exact reachable-core quotients)**.** *The theorem has three conclusions, each under the premises stated in its own clause.*

(i) *Under Assumptions 44, 45, 46, and 47, fix $\delta \in (0,1)$ and any independently obtained target realization satisfying $\mathcal{E}_Q$. Under $\mathbb{P}_{\mathrm{eval}}$, one event has probability at least $1-\delta$ and, simultaneously for every $\tau \in \mathcal{T}$,*

$$\sup_{\phi \in \Phi_M,\, g \in \mathcal{G}_\phi} \left| P_{n_\tau} \widehat{\ell}_{\tau,\phi,g} - L_\tau(\phi, g) \right| \le e_\tau(\delta), \tag{8}$$

$$R(\widehat{\phi}) + \lambda D(\widehat{\phi}) \le \inf_{\phi \in \Phi_M} \{R(\phi) + \lambda D(\phi)\} + 2\lambda e_{\max}(\delta), \tag{9}$$

and

$$L_\tau(\widehat{\phi}, \widehat{g}_\tau) \le D_\tau(\widehat{\phi}) + 2e_\tau(\delta). \tag{10}$$

If, in addition, Assumption 48 holds, the same event also gives, separately for every task,

$$\begin{aligned} V_\tau^\star - V_\tau^{\widehat{\pi}_\tau} &= \frac{1}{1-\gamma} \mathbb{E}_{H \sim d_\tau^{\widehat{\pi}_\tau}} \ell_{\tau,\widehat{\phi},\widehat{g}_\tau}(H) \\ &\le \frac{C_\tau}{1-\gamma} L_\tau(\widehat{\phi}, \widehat{g}_\tau) \\ &\le \frac{C_\tau}{1-\gamma} \big(D_\tau(\widehat{\phi}) + 2e_\tau(\delta)\big). \end{aligned} \tag{11}$$

The exact indexed count and its all-maps envelope are

$$N_{\mathrm{cls}} = \sum_{\tau\in\mathcal{T}} \sum_{\phi\in\Phi_M} |\mathcal{G}_\phi| \le |\mathcal{T}|\,|\Phi_M|\,|\mathcal{A}|^M, \tag{12}$$

while $\epsilon_\tau(\delta)$ retains the sharper exact N_{cls}. When $B = 0$, every true loss and every ϵ_τ is zero and the true-loss event is deterministic.

(ii) *Under Assumptions 44 and 49 alone, the core Bellman operator has a unique bounded measurable fixed point $q_\tau^\star$, constructed as the uniform limit of its zero iterates, and deterministically*

$$Q_\tau^\star(h,a) = q_\tau^\star(x(h),a) \quad (\tau\in\mathcal{T},\ h\in\mathcal{R}_\tau,\ a\in\mathcal{A}). \tag{13}$$

The same equality therefore holds on $\mathcal{H}_\tau$. Equation equation 13 is stated for the chosen task-reachable version on $\mathcal{R}_\tau$.

(iii) *Under Assumptions 44, 45, 49, 50, and 51,*

$$D(\phi_{\mathrm{dec}}) = 0, \qquad R(\phi_{\mathrm{dec}}) = \log m_{\mathrm{dec}}^{\mathrm{reach}}. \tag{14}$$

Conversely, for every $\phi\in\Phi_M$ satisfying $D(\phi)=0$, choose one family of heads $g_\tau\in\mathcal{G}_\phi$ with $L_\tau(\phi,g_\tau)=0$ and set $u_z=(g_\tau(z))_{\tau\in\mathcal{T}}$. Then, for $\bar\mu_X$-almost every x and every $z\in\mathcal{Z}_\phi$,

$$z\in\mathcal{Z}_\phi(x) \quad\Longrightarrow\quad u_z\in\mathcal{U}_{\mathrm{reach}}(x). \tag{15}$$

For those same heads, define $\mathcal{C}_\phi := \{u_z : z\in\mathcal{Z}_\phi^{\bar\mu}\}$. This is a valid reachable-profile cover and

$$m_{\mathrm{dec}}^{\mathrm{reach}} \le |\mathcal{C}_\phi| \le |\mathcal{Z}_\phi^{\bar\mu}| \le |\mathcal{Z}_\phi|. \tag{16}$$

Consequently,

$$\min_{\phi\in\Phi_M:\, D(\phi)=0} R(\phi) = \log m_{\mathrm{dec}}^{\mathrm{reach}}, \tag{17}$$

and the minimum is attained by the core-only encoder $\phi_{\mathrm{dec}}(h) = c_{\mathrm{dec}}(x(h))$. In equation 15, only coordinates $\tau\in\mathcal{T}(x)$ are constrained.

If every $\eta_\tau = 0$, then on $\mathcal{E}_Q$, $\widehat{Q}_\tau = Q_\tau^\star$ on $\mathcal{H}_\tau\times\mathcal{A}$, $\widehat{\ell}_{\tau,\phi,g} = \ell_{\tau,\phi,g}$ on $\mathcal{H}_\tau$, and $e_\tau = \epsilon_\tau$ exactly, so only sampling error remains. Under homogeneous taskwise data and coverage parameters, the constants reduce by equality to the common-parameter form stated immediately after this result. The structural specialization equation 14–equation 17 has no remainder. Coverage is used only for equation 11, and nuisance support equivalence and realizable capacity only for clause (iii).

Corollary 53 (Homogeneous-task finite-sample guarantee)**.** *Under Assumptions 44, 45, 46, 47, and 48, suppose that, for every $\tau\in\mathcal{T}$,*

$$(\mu_\tau, n_\tau, \eta_\tau, C_\tau) = (\mu, n, \eta, C), \tag{18}$$

where $n\ge 1$, $\eta\ge 0$, and $C\ge 1$. For every $\delta\in(0,1)$ define

$$\epsilon := B\sqrt{\frac{\log(2N_{\mathrm{cls}}/\delta)}{2n}}, \qquad e := \epsilon + 2\eta. \tag{19}$$

For every fixed target realization satisfying $\mathcal{E}_Q$, one event has $\mathbb{P}_{\mathrm{eval}}$-probability at least $1-\delta$ and, simultaneously for every task,

$$\sup_{\phi\in\Phi_M,\, g\in\mathcal{G}_\phi} \left| P_{n_\tau}\widehat{\ell}_{\tau,\phi,g} - L_\tau(\phi,g)\right| \le e, \tag{20}$$

$$R(\widehat{\phi}) + \lambda D(\widehat{\phi}) \le \inf_{\phi\in\Phi_M}\{R(\phi)+\lambda D(\phi)\} + 2\lambda e, \tag{21}$$

$$L_\tau(\widehat{\phi},\widehat{g}_\tau) \le D_\tau(\widehat{\phi}) + 2e, \tag{22}$$

and

$$\begin{aligned} V_\tau^\star - V_\tau^{\widehat{\pi}_\tau} &= \frac{1}{1-\gamma}\mathbb{E}_{d_\tau^{\widehat{\pi}_\tau}}\ell_{\tau,\widehat{\phi},\widehat{g}_\tau} \\ &\le \frac{C}{1-\gamma}L_\tau(\widehat{\phi},\widehat{g}_\tau) \\ &\le \frac{C}{1-\gamma}\big(D_\tau(\widehat{\phi}) + 2e\big). \end{aligned} \tag{23}$$

Consequently,

$$\max_{\tau\in\mathcal{T}}\big(V_\tau^\star - V_\tau^{\widehat{\pi}_\tau}\big) \le \frac{C}{1-\gamma}\big(D(\widehat{\phi}) + 2e\big). \tag{24}$$

If $\eta = 0$, then $e = \epsilon$, the estimated and true losses coincide on every scoring support, and every target-estimation term vanishes exactly.

Proof. Apply Proposition 78. Its conditions are exactly equation 18–equation 19, and its conclusions are equation 20–equation 24. □

C.4 Proof Sketch

The statistical argument begins with the true losses only. For $B > 0$, Lemma 54 and Proposition 55 give, for each labeled (τ, ϕ, g),

$$2\exp\left[-\frac{2n_\tau\epsilon_\tau(\delta)^2}{B^2}\right] = \frac{\delta}{N_{\mathrm{cls}}}.$$

A single union over exactly N_{cls} triples therefore has total displayed budget δ. When $B = 0$, all true losses vanish and the event is deterministic.

On $\mathcal{E}_Q$, maximum stability and the selected-action certificate give $|\widehat{\ell}_{\tau,\phi,g} - \ell_{\tau,\phi,g}| \le 2\eta_\tau$ on $\mathcal{H}_\tau$. Proposition 57 adds this deterministic bridge to the one true-loss event. Finite minimum and maximum stability then yield the taskwise deficiency bounds and the shared $e_{\max}$ envelope, while exact empirical minimization gives the $2\lambda e_{\max}$ oracle in Proposition 59. A separate two-sided head comparison gives Proposition 60.

For deployment, Lemma 61 telescopes the full-history Bellman defects on $\mathcal{R}_\tau$. Proposition 62 controls the terminal-value and omitted-reward tails separately and identifies the normalized discounted occupancy. Candidate-uniform Radon–Nikodym domination then places the occupancy on the scoring support and gives Proposition 63. The same calibrated event is reused throughout.

The structural argument starts independently. Matching finite-horizon history and core recursions remain on the same task-reachable domains. After the induction in Proposition 65, Proposition 66 proves bounded-measurable completeness and Bellman self-mapping, constructs the uniformly Cauchy zero iterates, proves their measurable fixed-point limit and uniqueness, and then removes the two geometric tails. The realizable profile selector is therefore taskwise optimal and yields the zero-deficiency, exact-rate achiever of Propositions 68 and 69.

For the converse, zero taskwise risks are disintegrated at a fixed core point. On active task-reachable domains this gives

$$r_\tau(x)p_{\tau,z}(x)\Delta_{\tau,z}(x) = 0.$$

Mutual absolute continuity transfers positivity of the same encoder cell across all active tasks, so Proposition 73 makes every conditionally active code profile legal, universally over the finite code range on one common almost-everywhere core set. Mixture disintegration and a finite null-set removal then give the global cover and cardinality chain in Proposition 76. The exact achiever and converse meet in Propositions 80 and 81. Finally, Proposition 82 conjoins the three premise-local implications, while Proposition 78 gives the homogeneous-task corollary without dropping a term.

C.5 Proof Details

C.5.1 Joint concentration of true task losses

Lemma 54 (Two-sided Hoeffding inequality)**.** *Let $Y_1, \dots, Y_n$ be independent and identically distributed random variables with $Y_i \in [0, B]$ almost surely, where $B > 0$, and let $m = \mathbb{E}Y_i$. Then, for every $t > 0$,*

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^n Y_i - m\right| > t\right) \le 2\exp\left(-\frac{2nt^2}{B^2}\right).$$

Proof. This is the two-sided bounded-summand inequality of Hoeffding (1963), specialized to the common interval $[0, B]$. □

Proposition 55 (Joint heterogeneous true-loss concentration)**.** *Under Assumptions 44, 45, and 46, let* $\delta \in (0,1)$ *and define*

$$\epsilon_\tau(\delta) = B\sqrt{\frac{\log(2N_{\mathrm{cls}}/\delta)}{2n_\tau}}.$$

For each task set

$$\mathcal{E}_{\tau,\mathrm{true}}(\delta) := \left\{ \sup_{\phi\in\Phi_M,\, g\in\mathcal{G}_\phi} |P_{n_\tau}\ell_{\tau,\phi,g} - L_\tau(\phi,g)| \le \epsilon_\tau(\delta) \right\},$$

and let

$$\mathcal{E}_{\mathrm{true}}(\delta) := \bigcap_{\tau\in\mathcal{T}} \mathcal{E}_{\tau,\mathrm{true}}(\delta).$$

Then

$$\Pr\big(\mathcal{E}_{\mathrm{true}}(\delta)\big) \ge 1-\delta.$$

If $B > 0$*, the proof uses one finite union and its displayed Hoeffding tail budgets satisfy the exact identity*

$$\sum_{\tau\in\mathcal{T}} \sum_{\phi\in\Phi_M} \sum_{g\in\mathcal{G}_\phi} 2\exp\left[-\frac{2n_\tau\epsilon_\tau(\delta)^2}{B^2}\right] = \delta.$$

If $B = 0$*, the event* $\mathcal{E}_{\mathrm{true}}(\delta)$ *holds surely.*

Proof. First fix an indexed triple (τ,ϕ,g). On $\mathcal{H}_\tau \subseteq \mathcal{R}_\tau$, the definitions give

$$0 \le Q^\star_\tau(h,a) \le B \quad \text{and} \quad V^\star_\tau(h) = \max_{a\in\mathcal{A}} Q^\star_\tau(h,a).$$

Consequently,

$$0 \le \ell_{\tau,\phi,g}(h) = V^\star_\tau(h) - Q^\star_\tau\big(h, g(\phi(h))\big) \le B.$$

Since $H_{\tau,i} \overset{\mathrm{iid}}{\sim} \mu_\tau$ and $\mu_\tau(\mathcal{H}_\tau) = 1$, the scalar random variables

$$Y_{\tau,\phi,g,i} := \ell_{\tau,\phi,g}(H_{\tau,i}), \qquad i = 1,\dots,n_\tau,$$

are iid in $[0,B]$, with $\mathbb{E}Y_{\tau,\phi,g,i} = L_\tau(\phi,g)$. The class and hence this fixed loss function are chosen before the evaluation data are observed.

Suppose first that $B > 0$. Because $N_{\mathrm{cls}} \ge 1$ and $\delta \in (0,1)$, the logarithm in $\epsilon_\tau(\delta)$ is positive. Lemma 54 gives

$$\Pr(|P_{n_\tau}\ell_{\tau,\phi,g} - L_\tau(\phi,g)| > \epsilon_\tau(\delta)) \le 2\exp\left[-\frac{2n_\tau\epsilon_\tau(\delta)^2}{B^2}\right].$$

The task-specific radius preserves the heterogeneous n_τ, but its substitution into the exponent gives the same indexed budget for every triple:

$$\frac{2n_\tau\epsilon_\tau(\delta)^2}{B^2} = \log\left(\frac{2N_{\mathrm{cls}}}{\delta}\right),$$

and therefore

$$2\exp\left[-\frac{2n_\tau\epsilon_\tau(\delta)^2}{B^2}\right] = 2\exp\left[-\log\left(\frac{2N_{\mathrm{cls}}}{\delta}\right)\right] = \frac{\delta}{N_{\mathrm{cls}}}.$$

For later use, define the auxiliary bad event

$$A_{\tau,\phi,g} := \{|P_{n_\tau}\ell_{\tau,\phi,g} - L_\tau(\phi,g)| > \epsilon_\tau(\delta)\}.$$

The complement of the desired joint event is exactly the union of these bad events. Applying Boole's inequality once, over the full labeled triple index set, yields

$$\begin{aligned}\Pr\big(\mathcal{E}_{\mathrm{true}}(\delta)^c\big) &= \Pr\left(\bigcup_{\tau\in\mathcal{T}}\bigcup_{\phi\in\Phi_M}\bigcup_{g\in\mathcal{G}_\phi} A_{\tau,\phi,g}\right)\\ &\le \sum_{\tau\in\mathcal{T}}\sum_{\phi\in\Phi_M}\sum_{g\in\mathcal{G}_\phi}\Pr(A_{\tau,\phi,g})\\ &\le \sum_{\tau\in\mathcal{T}}\sum_{\phi\in\Phi_M}\sum_{g\in\mathcal{G}_\phi} 2\exp\left[-\frac{2n_\tau\epsilon_\tau(\delta)^2}{B^2}\right]\\ &= \left(\sum_{\tau\in\mathcal{T}}\sum_{\phi\in\Phi_M}|\mathcal{G}_\phi|\right)\frac{\delta}{N_{\mathrm{cls}}} = \delta.\end{aligned}$$

Thus $\Pr(\mathcal{E}_{\mathrm{true}}(\delta)) \ge 1-\delta$. The equality on the last line is the exact one-union *budget identity* for the Hoeffding–Boole upper bound.

It remains to handle $B = 0$ without division by B. Since $B = R_{\max}/(1-\gamma)$ and $1-\gamma > 0$, this branch has $R_{\max} = 0$. Every reward is then zero almost surely. Hence, on every task scoring domain and for every action,

$$Q^\star_\tau(h,a) = V^\star_\tau(h) = 0, \qquad \ell_{\tau,\phi,g}(h) = 0.$$

It follows that

$$P_{n_\tau}\ell_{\tau,\phi,g} = L_\tau(\phi,g) = \epsilon_\tau(\delta) = 0$$

for every indexed triple. Thus every $\mathcal{E}_{\tau,\mathrm{true}}(\delta)$, and hence their joint intersection, holds surely. This proves the proposition in both branches. □

C.5.2 Calibration by independently certified targets

Lemma 56 (Maximum stability and independent conditioning)**.** *For real vectors $(u_a)_{a\in\mathcal{A}}$ and $(v_a)_{a\in\mathcal{A}}$ on the same nonempty finite action set,*

$$\left|\max_a u_a - \max_a v_a\right| \le \max_a |u_a - v_a|.$$

If $\mathscr{F}_1$ and $\mathscr{F}_2$ are independent sigma-fields and $A \in \mathscr{F}_2$, then

$$\mathbb{E}[\mathbf{1}_A \mid \mathscr{F}_1] = \Pr(A) \quad \textit{almost surely}.$$

Proof. Writing $r = \max_a |u_a - v_a|$ gives $\max_a u_a \le \max_a v_a + r$; interchanging u and v gives the reverse one-sided inequality. For the conditioning identity, if $C \in \mathscr{F}_1$, independence yields $\mathbb{E}[\mathbf{1}_C\mathbf{1}_A] = \Pr(C)\Pr(A)$, the defining integral identity for the displayed conditional expectation. □

Proposition 57 (Conditional simultaneous taskwise target calibration)**.** *Under Assumptions 47 and 46, and using Proposition 55, fix $\delta \in (0,1)$, and define the procedure-defined radii*

$$\epsilon_\tau(\delta) = B\sqrt{\frac{\log(2N_{\mathrm{cls}}/\delta)}{2n_\tau}}, \qquad e_\tau(\delta) = \epsilon_\tau(\delta) + 2\eta_\tau.$$

For each task, define the calibrated event

$$\mathcal{E}_{\tau,\mathrm{cal}}(\delta) := \left\{\sup_{\phi\in\Phi_M,\, g\in\mathcal{G}_\phi}\left|P_{n_\tau}\widehat{\ell}_{\tau,\phi,g} - L_\tau(\phi,g)\right| \le e_\tau(\delta)\right\},$$

and define its simultaneous taskwise intersection

$$\mathcal{E}_{\mathrm{cal}}(\delta) := \bigcap_{\tau\in\mathcal{T}} \mathcal{E}_{\tau,\mathrm{cal}}(\delta).$$

Let $\mathscr{F}_Q$ be the sigma-field generated by all target-estimation splits. Then the following statements hold.

1. On $\mathcal{E}_Q$, for every task τ, history $h \in \mathcal{H}_\tau$, encoder $\phi \in \Phi_M$, and head $g \in \mathcal{G}_\phi$,

$$\left|\widehat{\ell}_{\tau,\phi,g}(h) - \ell_{\tau,\phi,g}(h)\right| \le 2\eta_\tau.$$

2. Let $\Pr_{\mathrm{eval}}$ denote the marginal law of all evaluation collections. For every fixed target-split realization $q = (q_\tau)_{\tau\in\mathcal{T}}$ satisfying the certificate bounds, let $\mathcal{E}_{\mathrm{cal}}(\delta; q)$ denote the calibrated event with q_τ substituted for $\widehat{Q}_\tau$. Then

$$\Pr_{\mathrm{eval}}\big(\mathcal{E}_{\mathrm{cal}}(\delta; q)\big) \ge 1 - \delta.$$

This fixed-realization statement holds even when $\Pr(\mathcal{E}_Q) = 0$. Its sigma-field conditional consequence is

$$\mathbf{1}_{\mathcal{E}_Q} \Pr\big(\mathcal{E}_{\mathrm{cal}}(\delta) \mid \mathscr{F}_Q\big) \ge (1-\delta)\mathbf{1}_{\mathcal{E}_Q} \quad \textit{almost surely}.$$

Consequently, whenever $\Pr(\mathcal{E}_Q) > 0$,

$$\Pr\big(\mathcal{E}_{\mathrm{cal}}(\delta) \mid \mathcal{E}_Q\big) \ge 1 - \delta.$$

3. If $\eta_\tau = 0$ for a task, then on $\mathcal{E}_Q$

$$\widehat{Q}_\tau(h, a) = Q^\star_\tau(h, a) \quad \textit{for every } (h, a) \in \mathcal{H}_\tau \times \mathcal{A},$$

and hence $\widehat{\ell}_{\tau,\phi,g} = \ell_{\tau,\phi,g}$ pointwise on $\mathcal{H}_\tau$, $e_\tau(\delta) = \epsilon_\tau(\delta)$, and $\mathcal{E}_{\tau,\mathrm{cal}}(\delta)$ agrees almost surely with $\mathcal{E}_{\tau,\mathrm{true}}(\delta)$ on the evaluation sample. If every $\eta_\tau = 0$, then

$$\mathbf{1}_{\mathcal{E}_Q}\mathbf{1}_{\mathcal{E}_{\mathrm{cal}}(\delta)} = \mathbf{1}_{\mathcal{E}_Q}\mathbf{1}_{\mathcal{E}_{\mathrm{true}}(\delta)} \quad \textit{almost surely},$$

and the entire target-estimation remainder disappears exactly.

Proof. Fix τ, $h \in \mathcal{H}_\tau$, $\phi \in \Phi_M$, and $g \in \mathcal{G}_\phi$, and write $b = g(\phi(h)) \in \mathcal{A}$. On $\mathcal{E}_Q$, Assumption 47 gives

$$\max_{a\in\mathcal{A}} |\widehat{Q}_\tau(h, a) - Q^\star_\tau(h, a)| \le \eta_\tau.$$

Lemma 56 therefore yields

$$\left|\max_{a\in\mathcal{A}} \widehat{Q}_\tau(h, a) - \max_{a\in\mathcal{A}} Q^\star_\tau(h, a)\right| \le \eta_\tau,$$

while the same certificate at the selected action gives

$$|\widehat{Q}_\tau(h, b) - Q^\star_\tau(h, b)| \le \eta_\tau.$$

Using the definitions of the estimated and true losses and applying the triangle inequality once gives the pointwise bridge

$$\begin{aligned}
|\widehat{\ell}_{\tau,\phi,g}(h) - \ell_{\tau,\phi,g}(h)| &= \left|\Big(\max_a \widehat{Q}_\tau(h, a) - \max_a Q^\star_\tau(h, a)\Big) - \Big(\widehat{Q}_\tau(h, b) - Q^\star_\tau(h, b)\Big)\right| \\
&\le \left|\max_a \widehat{Q}_\tau(h, a) - \max_a Q^\star_\tau(h, a)\right| + |\widehat{Q}_\tau(h, b) - Q^\star_\tau(h, b)| \\
&\le 2\eta_\tau.
\end{aligned}$$

This argument is deterministic on $\mathcal{E}_Q$ and simultaneous in all indices.

By Assumption 46, every $H_{\tau,i}$ lies in $\mathcal{H}_\tau$ almost surely. Because the task family and each n_τ are finite, all evaluation histories lie in their corresponding supports simultaneously with probability one. Consequently, on $\mathcal{E}_Q$, for every indexed triple,

$$P_{n_\tau}|\widehat{\ell}_{\tau,\phi,g} - \ell_{\tau,\phi,g}| \le 2\eta_\tau \quad \text{almost surely}.$$

On the event $\mathcal{E}_{\mathrm{true}}(\delta)$, the same triple also satisfies

$$|P_{n_\tau}\ell_{\tau,\phi,g} - L_\tau(\phi, g)| \le \epsilon_\tau(\delta).$$

Thus, on $\mathcal{E}_Q \cap \mathcal{E}_{\mathrm{true}}(\delta)$,

$$\begin{aligned}|P_{n_\tau}\widehat{\ell}_{\tau,\phi,g} - L_\tau(\phi,g)| &\le |P_{n_\tau}(\widehat{\ell}_{\tau,\phi,g} - \ell_{\tau,\phi,g})| + |P_{n_\tau}\ell_{\tau,\phi,g} - L_\tau(\phi,g)| \\ &\le P_{n_\tau}|\widehat{\ell}_{\tau,\phi,g} - \ell_{\tau,\phi,g}| + \epsilon_\tau(\delta) \\ &\le 2\eta_\tau + \epsilon_\tau(\delta) = e_\tau(\delta).\end{aligned}$$

The event is already simultaneous over every (τ,ϕ,g), and the pointwise certificate is uniform over the same indices. Therefore the displayed implication holds jointly, giving

$$\mathbf{1}_{\mathcal{E}_Q}\mathbf{1}_{\mathcal{E}_{\mathrm{true}}(\delta)} \le \mathbf{1}_{\mathcal{E}_Q}\mathbf{1}_{\mathcal{E}_{\mathrm{cal}}(\delta)} \quad \text{almost surely.}$$

The same inclusion can be stated without assigning a probability to the certificate event. Fix any target-split output $q = (q_\tau)_{\tau\in\mathcal{T}}$ that satisfies the displayed certificate bounds, and evaluate the estimated losses using this fixed q. The preceding pointwise derivation is deterministic in q, so under the evaluation marginal it gives

$$\mathcal{E}_{\mathrm{true}}(\delta) \subseteq \mathcal{E}_{\mathrm{cal}}(\delta; q) \quad \Pr_{\mathrm{eval}}\text{-almost surely.}$$

The evaluation marginal does not depend on q, and Proposition 55 controls that same marginal. Hence

$$\Pr_{\mathrm{eval}}\big(\mathcal{E}_{\mathrm{cal}}(\delta;q)\big) \ge \Pr_{\mathrm{eval}}\big(\mathcal{E}_{\mathrm{true}}(\delta)\big) \ge 1-\delta.$$

This proves the fixed-certified-realization form even when the random event $\mathcal{E}_Q$ itself has probability zero.

For the corresponding almost-sure sigma-field formulation, let $\mathscr{F}_{\mathrm{eval}}$ be the sigma-field generated by all evaluation collections. The event $\mathcal{E}_{\mathrm{true}}(\delta)$ is $\mathscr{F}_{\mathrm{eval}}$-measurable, and $\mathscr{F}_{\mathrm{eval}}$ is independent of $\mathscr{F}_Q$. Lemma 56 and Proposition 55 give

$$\Pr\big(\mathcal{E}_{\mathrm{true}}(\delta) \mid \mathscr{F}_Q\big) = \Pr\big(\mathcal{E}_{\mathrm{true}}(\delta)\big) \ge 1-\delta \quad \text{almost surely.}$$

Since $\mathcal{E}_Q \in \mathscr{F}_Q$, conditioning the preceding event inclusion on $\mathscr{F}_Q$ gives

$$\begin{aligned}\mathbf{1}_{\mathcal{E}_Q}\Pr\big(\mathcal{E}_{\mathrm{cal}}(\delta) \mid \mathscr{F}_Q\big) &= \mathbb{E}\big[\mathbf{1}_{\mathcal{E}_Q}\mathbf{1}_{\mathcal{E}_{\mathrm{cal}}(\delta)} \,\big|\, \mathscr{F}_Q\big] \\ &\ge \mathbb{E}\big[\mathbf{1}_{\mathcal{E}_Q}\mathbf{1}_{\mathcal{E}_{\mathrm{true}}(\delta)} \,\big|\, \mathscr{F}_Q\big] \\ &= \mathbf{1}_{\mathcal{E}_Q}\Pr\big(\mathcal{E}_{\mathrm{true}}(\delta) \mid \mathscr{F}_Q\big) \\ &\ge (1-\delta)\mathbf{1}_{\mathcal{E}_Q}.\end{aligned}$$

If $\Pr(\mathcal{E}_Q) > 0$, taking expectations and dividing by that probability yields the ordinary event-conditional statement. If it is zero, the almost-sure sigma-field inequality is vacuous on that null fiber, while the preceding fixed-realization evaluation-marginal statement remains the nonvacuous certified guarantee.

Finally, if $\eta_\tau = 0$, then on $\mathcal{E}_Q$ the nonnegative supremum in the certificate is at most zero, so $\widehat{Q}_\tau = Q^\star_\tau$ everywhere on $\mathcal{H}_\tau \times \mathcal{A}$. Substitution into the two loss definitions gives $\widehat{\ell}_{\tau,\phi,g} = \ell_{\tau,\phi,g}$ pointwise on $\mathcal{H}_\tau$. Since evaluation histories lie in that support almost surely, the corresponding empirical losses and taskwise good-event indicators agree almost surely, while $e_\tau = \epsilon_\tau$ by equality. Applying this task by task proves the all-η-zero joint-event statement. □

C.5.3 Stability of the shared penalized objective

Lemma 58 (Stability of finite extrema)**.** *Let I be a nonempty finite set and let $(a_i)_{i\in I}$ and $(b_i)_{i\in I}$ be real families. Then*

$$\left|\min_{i\in I} a_i - \min_{i\in I} b_i\right| \le \max_{i\in I}|a_i - b_i|, \qquad \left|\max_{i\in I} a_i - \max_{i\in I} b_i\right| \le \max_{i\in I}|a_i - b_i|.$$

Proof. Set $r = \max_i |a_i - b_i|$. Since $a_i \le b_i + r$ for every i, $\max_i a_i \le \max_i b_i + r$; exchange the two families for the reverse bound. Applying the maximum result to $(-a_i)_i$ and $(-b_i)_i$ proves the minimum result. No uniqueness of an extremizer is required. □

Proposition 59 (Heterogeneous shared-encoder oracle)**.** *Under Assumption 45 and Proposition 57, fix $\delta \in (0,1)$, and define*

$$e_{\max}(\delta) := \max_{\tau \in \mathcal{T}} e_\tau(\delta).$$

On $\mathcal{E}_{\mathrm{cal}}(\delta)$, the following statements hold simultaneously.

1. For every $\tau \in \mathcal{T}$ and $\phi \in \Phi_M$,

$$|\widehat{D}_\tau(\phi) - D_\tau(\phi)| \le e_\tau(\delta).$$

2. For every $\phi \in \Phi_M$,

$$\left|\max_{\tau \in \mathcal{T}} \widehat{D}_\tau(\phi) - D(\phi)\right| \le e_{\max}(\delta).$$

3. Every procedure-defined exact minimizer

$$\widehat{\phi} \in \arg\min_{\phi \in \Phi_M} \left\{ R(\phi) + \lambda \max_{\tau \in \mathcal{T}} \widehat{D}_\tau(\phi) \right\}$$

satisfies

$$R(\widehat{\phi}) + \lambda D(\widehat{\phi}) \le \inf_{\phi \in \Phi_M} \{R(\phi) + \lambda D(\phi)\} + 2\lambda e_{\max}(\delta).$$

These conclusions hold for every resolution of ties in the head minima, task maxima, and encoder minima.

Proof. On $\mathcal{E}_{\mathrm{cal}}(\delta)$, for every fixed (τ, ϕ),

$$\max_{g \in \mathcal{G}_\phi} \left| P_{n_\tau} \widehat{\ell}_{\tau,\phi,g} - L_\tau(\phi, g) \right| \le e_\tau(\delta).$$

Lemma 58 and the definitions give

$$\begin{aligned} |\widehat{D}_\tau(\phi) - D_\tau(\phi)| &= \left| \min_{g \in \mathcal{G}_\phi} P_{n_\tau} \widehat{\ell}_{\tau,\phi,g} - \min_{g \in \mathcal{G}_\phi} L_\tau(\phi, g) \right| \\ &\le \max_{g \in \mathcal{G}_\phi} \left| P_{n_\tau} \widehat{\ell}_{\tau,\phi,g} - L_\tau(\phi, g) \right| \\ &\le e_\tau(\delta). \end{aligned}$$

This compares the minimum values themselves, so it remains valid if either minimum has multiple optimizers.

For fixed ϕ, Lemma 58 and the taskwise bounds just proved yield

$$\begin{aligned} \left| \max_{\tau \in \mathcal{T}} \widehat{D}_\tau(\phi) - \max_{\tau \in \mathcal{T}} D_\tau(\phi) \right| &\le \max_{\tau \in \mathcal{T}} |\widehat{D}_\tau(\phi) - D_\tau(\phi)| \\ &\le \max_{\tau \in \mathcal{T}} e_\tau(\delta) \\ &= e_{\max}(\delta). \end{aligned}$$

Because $D(\phi) = \max_\tau D_\tau(\phi)$, this proves the second claim. The heterogeneous radii enter only through their task-uniform envelope $e_{\max}$.

Introduce the auxiliary objective abbreviations

$$\widehat{J}(\phi) := R(\phi) + \lambda \max_{\tau \in \mathcal{T}} \widehat{D}_\tau(\phi), \qquad J(\phi) := R(\phi) + \lambda D(\phi).$$

The same deterministic $R(\phi)$ occurs in both objectives, so the task-maximum bound gives, simultaneously for every $\phi \in \Phi_M$,

$$|\widehat{J}(\phi) - J(\phi)| = \lambda \left| \max_{\tau} \widehat{D}_\tau(\phi) - D(\phi) \right| \le \lambda e_{\max}(\delta).$$

Finiteness and nonemptiness of Φ_M ensure both objective minima are attained. Let $\widehat{\phi}$ be any exact empirical minimizer selected by the procedure and let $\phi^\star \in \arg\min_{\phi\in\Phi_M} J(\phi)$ be any population minimizer. Two objective transfers and exact empirical optimality give

$$\begin{aligned} J(\widehat{\phi}) &\le \widehat{J}(\widehat{\phi}) + \lambda e_{\max}(\delta) \\ &\le \widehat{J}(\phi^\star) + \lambda e_{\max}(\delta) \\ &\le J(\phi^\star) + 2\lambda e_{\max}(\delta) \\ &= \inf_{\phi\in\Phi_M} J(\phi) + 2\lambda e_{\max}(\delta). \end{aligned}$$

The middle inequality holds for every exact empirical minimizer, including all ties; any tied population minimizer may likewise be used. Expanding J gives the claimed oracle. Selected-head population risk is controlled separately in Proposition 60.

□

C.5.4 Selected heads and infinite-discounted deployment

Proposition 60 (Exact selected-head population comparison)**.** *Under Assumption 45 and Proposition 57, fix $\delta \in (0,1)$. On $\mathcal{E}_{\mathrm{cal}}(\delta)$, simultaneously for every $\tau \in \mathcal{T}$ and $\phi \in \Phi_M$,*

$$L_\tau(\phi, \widehat{g}_{\tau,\phi}) \le D_\tau(\phi) + 2e_\tau(\delta).$$

Consequently, for every realization of the target and evaluation data in this event,

$$L_\tau(\widehat{\phi}, \widehat{g}_\tau) \le D_\tau(\widehat{\phi}) + 2e_\tau(\delta) \qquad \text{for every } \tau \in \mathcal{T}.$$

If $\eta_\tau = 0$, the remainder in these inequalities is exactly $2\epsilon_\tau(\delta)$.

Proof. Fix τ and ϕ. Finiteness and nonemptiness give a population comparison head

$$g^{\star,L}_{\tau,\phi} \in \arg\min_{g\in\mathcal{G}_\phi} L_\tau(\phi, g), \qquad L_\tau(\phi, g^{\star,L}_{\tau,\phi}) = D_\tau(\phi).$$

On $\mathcal{E}_{\mathrm{cal}}(\delta)$, the preceding result applies to both the empirical minimizer and this comparison head. Exact empirical optimality then gives

$$\begin{aligned} L_\tau(\phi, \widehat{g}_{\tau,\phi}) &\le P_{n_\tau}\widehat{\ell}_{\tau,\phi,\widehat{g}_{\tau,\phi}} + e_\tau(\delta) \\ &= \widehat{D}_\tau(\phi) + e_\tau(\delta) \\ &\le P_{n_\tau}\widehat{\ell}_{\tau,\phi,g^{\star,L}_{\tau,\phi}} + e_\tau(\delta) \\ &\le L_\tau(\phi, g^{\star,L}_{\tau,\phi}) + 2e_\tau(\delta) \\ &= D_\tau(\phi) + 2e_\tau(\delta). \end{aligned}$$

Every line holds simultaneously for all ϕ, so after the data are realized it may be evaluated at $\phi = \widehat{\phi}$. By definition, $\widehat{g}_{\tau,\widehat{\phi}} = \widehat{g}_\tau$. Finally, $\eta_\tau = 0$ gives $e_\tau(\delta) = \epsilon_\tau(\delta)$ by equality.

□

Lemma 61 (Reachable-history Bellman defect telescope)**.** *Under Assumption 44, fix a task τ and any measurable deterministic history policy π whose task-τ trajectory is reachable from ρ. Define on $\mathcal{R}_\tau$ the auxiliary Bellman defect*

$$\Delta^\pi_\tau(h) := V^\star_\tau(h) - Q^\star_\tau(h, \pi(h)).$$

Then $0 \le \Delta^\pi_\tau \le B$, the reachable version of the optimal target obeys

$$Q^\star_\tau(h,a) = \mathbb{E}_\tau[R_t + \gamma V^\star_\tau(H_{t+1}) \,|\, H_t = h, A_t = a] \quad (h \in \mathcal{R}_\tau),$$

and

$$V^\star_\tau = \mathbb{E}_\tau V^\star_\tau(H_0), \qquad H_0 \text{ is distributed according to the initial-history law induced by } \rho.$$

For every integer $T \ge 0$,

$$\sum_{t=0}^{T} \gamma^t \mathbb{E}^\pi_\tau \Delta^\pi_\tau(H_t) = V^\star_\tau - \sum_{t=0}^{T} \gamma^t \mathbb{E}^\pi_\tau R_t - \gamma^{T+1}\mathbb{E}^\pi_\tau V^\star_\tau(H_{T+1}).$$

All targets and defects in this statement are evaluated only on $\mathcal{R}_\tau$.

Proof. For $m \geq 0$, define on $\mathcal{R}_\tau$

$$Q_{\tau,m}(h,a) := \sup_\sigma \mathbb{E}_\tau^\sigma \left[\sum_{k=0}^{m-1} \gamma^k R_{t+k} \,\middle|\, H_t = h, A_t = a \right], \qquad Q_{\tau,0} := 0,$$

and put $V_{\tau,m}(h) := \max_a Q_{\tau,m}(h,a)$. The finite action set makes each pointwise maximum attained; with a fixed action ordering, the first maximizer supplies a measurable finite-horizon selector in the standard finite POMDP history model. Splitting off the first reward therefore gives the exact finite-horizon recursion

$$Q_{\tau,m+1}(h,a) = \mathbb{E}_\tau[R_t + \gamma V_{\tau,m}(H_{t+1}) \,|\, H_t = h, A_t = a].$$

This is a derivation in the current history and reward-timing convention: the first included reward is R_t, and the continuation begins at the pre-decision history H_{t+1}.

Nonnegative bounded rewards give, uniformly on the reachable task domain,

$$0 \leq Q_\tau^\star(h,a) - Q_{\tau,m}(h,a) \leq \sum_{k=m}^{\infty} \gamma^k R_{\max} = B\gamma^m.$$

Taking a finite maximum over actions yields the same bound for $V_\tau^\star - V_{\tau,m}$. Hence the recursion passes uniformly to the limit and proves the displayed infinite-horizon Bellman identity without introducing a new version of $Q_\tau^\star$. The finite-horizon functions are measurable by recursion, their uniform limits are measurable, and hence the defect Δ_τ^π is measurable as well.

For completeness, let $V_{\tau,m}^{\star,\rho}$ be the optimal scalar return from ρ with exactly the first m rewards. The same finite backward choice gives

$$V_{\tau,m}^{\star,\rho} = \mathbb{E}_\tau V_{\tau,m}(H_0),$$

Extending an optimal finite-horizon policy arbitrarily and bounding the omitted nonnegative reward tail gives

$$0 \leq V_\tau^\star - V_{\tau,m}^{\star,\rho} \leq B\gamma^m.$$

The uniform convergence of $V_{\tau,m}$ now yields $V_\tau^\star = \mathbb{E}_\rho V_\tau^\star(H_0)$.

Because π is a history policy and $\mathcal{R}_\tau$ contains histories reachable under arbitrary history policies, $H_t \in \mathcal{R}_\tau$ almost surely under π. Applying the Bellman identity at $A_t = \pi(H_t)$, then taking expectations, gives

$$\mathbb{E}_\tau^\pi \Delta_\tau^\pi(H_t) = \mathbb{E}_\tau^\pi V_\tau^\star(H_t) - \mathbb{E}_\tau^\pi R_t - \gamma \mathbb{E}_\tau^\pi V_\tau^\star(H_{t+1}).$$

Multiplication by γ^t and summation from 0 through T cancels every intermediate value term. The initial term is the scalar $V_\tau^\star$ by the identity just proved, leaving exactly the claimed finite telescope. Finally, $0 \leq \Delta_\tau^\pi \leq B$ follows from $V_\tau^\star = \max_a Q_\tau^\star$ and $0 \leq Q_\tau^\star, V_\tau^\star \leq B$.

□

Proposition 62 (Exact infinite-discounted performance-difference identity)**.** *Under Assumption 44 and Lemma 61, for every task τ and every measurable deterministic candidate history policy π,*

$$V_\tau^\star - V_\tau^\pi = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_\tau^\pi \Delta_\tau^\pi(H_t) = \frac{1}{1-\gamma} \mathbb{E}_{H \sim d_\tau^\pi} \Delta_\tau^\pi(H).$$

At finite T, the terminal-value tail and omitted-reward tail are each bounded by $B\gamma^{T+1}$; consequently, the absolute difference between the finite telescope and $V_\tau^\star - V_\tau^\pi$ is at most $2B\gamma^{T+1}$.

Proof. Write

$$S_T := \sum_{t=0}^{T} \gamma^t \mathbb{E}_\tau^\pi \Delta_\tau^\pi(H_t).$$

The two truncation terms have distinct sources. First, bounded optimal values give the terminal-value bound

$$0 \leq \gamma^{T+1} \mathbb{E}_\tau^\pi V_\tau^\star(H_{T+1}) \leq B\gamma^{T+1}.$$

Second, nonnegative Tonelli and the definition of scalar policy value give $V_\tau^\pi = \sum_{t=0}^\infty \gamma^t \mathbb{E}_\tau^\pi R_t$. Bounded nonnegative rewards therefore give the omitted-reward bound

$$0 \le V_\tau^\pi - \sum_{t=0}^{T} \gamma^t \mathbb{E}_\tau^\pi R_t = \sum_{t=T+1}^{\infty} \gamma^t \mathbb{E}_\tau^\pi R_t \le B\gamma^{T+1}.$$

Subtracting the infinite scalar value gap from the finite identity in Lemma 61 gives the exact signed truncation relation

$$S_T - (V_\tau^\star - V_\tau^\pi) = \sum_{t=T+1}^{\infty} \gamma^t \mathbb{E}_\tau^\pi R_t - \gamma^{T+1} \mathbb{E}_\tau^\pi V_\tau^\star(H_{T+1}).$$

The two separate bounds above and the triangle inequality therefore yield

$$|S_T - (V_\tau^\star - V_\tau^\pi)| \le 2B\gamma^{T+1} \longrightarrow 0.$$

Because $\Delta_\tau^\pi \ge 0$, monotone convergence also gives

$$\lim_{T\to\infty} S_T = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_\tau^\pi \Delta_\tau^\pi(H_t).$$

Finally, the definition of the normalized discounted occupancy and nonnegative series integration give

$$d_\tau^\pi(\mathcal{R}_\tau) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t = 1$$

and

$$\mathbb{E}_{H\sim d_\tau^\pi} \Delta_\tau^\pi(H) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{E}_\tau^\pi \Delta_\tau^\pi(H_t).$$

Combining the last three displays proves both equalities. If $\gamma = 0$, both tails are zero, $d_\tau^\pi = \mathcal{L}(H_0)$, and the identity reduces exactly to the immediate action regret at the initial history.

□

Proposition 63 (Taskwise RN lift and pathwise selected-policy certificate)**.** *Under Assumptions 45, 44, and 48, Proposition 57, Proposition 60, and Proposition 62, fix $\delta \in (0,1)$. Then every task and candidate pair (ϕ, g) satisfy*

$$V_\tau^\star - V_\tau^{\pi_{\tau,\phi,g}} = \frac{1}{1-\gamma} \mathbb{E}_{H\sim d_\tau^{\pi_{\tau,\phi,g}}} \ell_{\tau,\phi,g}(H) \le \frac{C_\tau}{1-\gamma} L_\tau(\phi, g),$$

where the expectation of $\ell_{\tau,\phi,g}$ is taken on its scored domain $\mathcal{H}_\tau$, which has full $d_\tau^{\pi_{\tau,\phi,g}}$-measure. On $\mathcal{E}_{\mathrm{cal}}(\delta)$, the data-dependent selected pair obeys the complete chain

$$\begin{aligned} V_\tau^\star - V_\tau^{\widehat{\pi}_\tau} &= \frac{1}{1-\gamma} \mathbb{E}_{H\sim d_\tau^{\widehat{\pi}_\tau}} \ell_{\tau,\widehat{\phi},\widehat{g}_\tau}(H) \\ &\le \frac{C_\tau}{1-\gamma} L_\tau(\widehat{\phi}, \widehat{g}_\tau) \\ &\le \frac{C_\tau}{1-\gamma} \big(D_\tau(\widehat{\phi}) + 2e_\tau(\delta)\big). \end{aligned}$$

Proof. Fix τ, ϕ, g, and abbreviate $\pi = \pi_{\tau,\phi,g}$. View μ_τ as a measure on $\mathcal{R}_\tau$ supported on $\mathcal{H}_\tau$. The coverage assumption gives a nonnegative density

$$r_\tau^\pi := \frac{\mathrm{d}d_\tau^\pi}{\mathrm{d}\mu_\tau}, \qquad 0 \le r_\tau^\pi \le C_\tau \quad \mu_\tau\text{-almost surely.}$$

It also gives $d_\tau^\pi(\mathcal{R}_\tau \setminus \mathcal{H}_\tau) = 0$. On the scored domain $\mathcal{H}_\tau$,

$$\Delta_\tau^\pi(h) = V_\tau^\star(h) - Q_\tau^\star(h, g(\phi(h))) = \ell_{\tau,\phi,g}(h).$$

Thus Proposition 62 and the Radon–Nikodym formula yield

$$\begin{aligned} V_\tau^\star - V_\tau^\pi &= \frac{1}{1-\gamma}\int_{\mathcal{H}_\tau} \ell_{\tau,\phi,g}(h)\, d_\tau^\pi(\mathrm{d}h) \\ &= \frac{1}{1-\gamma}\int_{\mathcal{H}_\tau} \ell_{\tau,\phi,g}(h) r_\tau^\pi(h)\, \mu_\tau(\mathrm{d}h) \\ &\le \frac{C_\tau}{1-\gamma}\int_{\mathcal{H}_\tau} \ell_{\tau,\phi,g}(h)\, \mu_\tau(\mathrm{d}h) \\ &= \frac{C_\tau}{1-\gamma} L_\tau(\phi, g). \end{aligned}$$

The same task, encoder, head, and nonnegative action-regret object occur on both measures; the only transfer factor is C_τ.

Assumption 48 holds for every candidate pair before data-dependent selection. Therefore, after any target and evaluation realization, the realized pair $(\widehat{\phi}, \widehat{g}_\tau)$ is one of the candidates to which the last display applies. On $\mathcal{E}_{\mathrm{cal}}(\delta)$, apply Proposition 60 to the final population loss. This gives the stated complete chain on exactly the same event.

□

C.5.5 Task-reachable Bellman factorization through the core

Lemma 64 (Canonical task-domain truncations and support closure)**.** *Under Assumptions 44 and 49, fix $\tau \in \mathcal{T}$. For $h \in \mathcal{R}_\tau$, $a \in \mathcal{A}$, and an integer $k \ge 0$, define the canonical value containing exactly the next k rewards by*

$$Q_{\tau,k}^{\mathrm{can}}(h,a) := \sup_\pi \mathbb{E}_\tau^\pi \left[\sum_{j=0}^{k-1} \gamma^j R_{t+j} \,\middle|\, H_t = h, A_t = a \right], \qquad Q_{\tau,0}^{\mathrm{can}} \equiv 0, \tag{25}$$

where the empty sum is zero and the fixed first action is not reselected by π. For $x \in \mathcal{X}_\tau^{\mathrm{reach}} := x(\mathcal{R}_\tau)$, define

$$q_{\tau,0}(x,a) := 0, \qquad q_{\tau,k+1} := \mathcal{B}_\tau q_{\tau,k}. \tag{26}$$

Then, for every $x \in \mathcal{X}_\tau^{\mathrm{reach}}$ and $a \in \mathcal{A}$,

$$P_{X,\tau}(\mathcal{X}_\tau^{\mathrm{reach}} \mid x, a) = 1, \tag{27}$$

so equation 26 is well defined on its stated domain. Moreover, for every $h = (x, n) \in \mathcal{R}_\tau$,

$$Q_{\tau,k+1}^{\mathrm{can}}(h,a) = \bar{r}_\tau(x,a) + \gamma\, \mathbb{E}_\tau \left[\max_{a' \in \mathcal{A}} Q_{\tau,k}^{\mathrm{can}}(H_{t+1}, a') \,\middle|\, H_t = h, A_t = a \right], \tag{28}$$

and

$$q_{\tau,k+1}(x,a) = \bar{r}_\tau(x,a) + \gamma\, \mathbb{E}_{X' \sim P_{X,\tau}(\cdot|x,a)} \max_{a' \in \mathcal{A}} q_{\tau,k}(X', a'). \tag{29}$$

Both families obey

$$0 \le Q_{\tau,k}^{\mathrm{can}}(h,a),\ q_{\tau,k}(x,a) \le R_{\max} \sum_{j=0}^{k-1} \gamma^j \le B, \qquad B := \frac{R_{\max}}{1-\gamma}. \tag{30}$$

In particular, their first backup is the same core-only quantity:

$$Q_{\tau,1}^{\mathrm{can}}(h,a) = q_{\tau,1}(x(h),a) = \bar{r}_\tau(x(h),a). \tag{31}$$

Proof. Fix $x \in \mathcal{X}_\tau^{\mathrm{reach}}$ and choose a reachable history $h \in \mathcal{R}_\tau$ with $x(h) = x$. Because $\mathcal{R}_\tau$ is the task-τ set reachable under arbitrary history policies, a policy witnessing reachability of h can be continued by choosing any prescribed $a \in \mathcal{A}$ at h. The conditional law generated by that continuation is supported on one-step histories that are themselves task-τ reachable. Thus, on the chosen task-reachable conditional version,

$$\Pr_\tau(H_{t+1} \in \mathcal{R}_\tau \mid H_t = h, A_t = a) = 1. \tag{32}$$

The recursively observed core satisfies $X_{t+1} = x(H_{t+1})$, and Assumption 49 identifies its conditional law with $P_{X,\tau}(\cdot \mid x, a)$. Equation equation 32 therefore gives equation 27. If two reachable histories have the same core x, the assumption assigns both the same kernel $P_{X,\tau}(\cdot \mid x, a)$, so this closure statement and recursion equation 26 are independent of the representing history.

We next verify equation 28 directly. The definition in equation 25 counts R_t as reward zero, so a $(k+1)$-reward return decomposes exactly as

$$R_t + \gamma \sum_{j=0}^{k-1} \gamma^j R_{t+1+j}. \tag{33}$$

Conditional on the successor history, any continuation policy's remaining expected return is bounded above by $\max_{a'} Q^{\mathrm{can}}_{\tau,k}(H_{t+1}, a')$. Conversely, backward induction over the finite horizon, with a fixed tie-breaking rule on the finite action set, selects a measurable maximizing action at every reachable successor history and attains that maximum. Randomization cannot improve a maximum of finitely many conditional expected returns. Taking the conditional expectation of equation 33, using $\mathbb{E}_\tau[R_t \mid H_t = h, A_t = a] = \bar{r}_\tau(x, a)$, yields equation 28. Values on $\mathcal{R}_\tau$ suffice because equation 32 places the successor in $\mathcal{R}_\tau$ almost surely.

Equation equation 29 is the definition of $\mathcal{B}_\tau$, and equation 27 makes its expectation internal to $\mathcal{X}^{\mathrm{reach}}_\tau$. Nonnegative rewards give the lower bounds in equation 30. The upper bound for $Q^{\mathrm{can}}_{\tau,k}$ follows directly from equation 25, and the same bound for $q_{\tau,k}$ follows by induction from equation 26 and equation 29. Setting $k = 0$ in equation 28–equation 29 gives equation 31. This proves the lemma.

□

Proposition 65 (Exact matching-horizon factorization)**.** *Under Assumptions 44 and 49 and Lemma 64, for every task τ, integer $k \geq 0$, history $h \in \mathcal{R}_\tau$, and action $a \in \mathcal{A}$,*

$$Q^{\mathrm{can}}_{\tau,k}(h, a) = q_{\tau,k}(x(h), a). \tag{34}$$

Proof. Fix τ. The proof is by induction on the common number k of included rewards. At $k = 0$, both sides of equation 34 are zero by equation 25–equation 26.

Assume equation 34 holds for one k, simultaneously for all $h \in \mathcal{R}_\tau$ and $a \in \mathcal{A}$. Fix $h = (x, n) \in \mathcal{R}_\tau$ and $a \in \mathcal{A}$. Lemma 64 ensures that $H_{t+1} \in \mathcal{R}_\tau$ almost surely under the conditional law in the Bellman backup. Hence the induction hypothesis is legal at every successor used by the expectation, and equation 28 becomes

$$\begin{aligned} Q^{\mathrm{can}}_{\tau,k+1}(h, a) &= \bar{r}_\tau(x, a) + \gamma\, \mathbb{E}_\tau\left[\max_{a' \in \mathcal{A}} q_{\tau,k}(x(H_{t+1}), a') \,\middle|\, H_t = h, A_t = a\right] \\ &= \bar{r}_\tau(x, a) + \gamma\, \mathbb{E}_{X' \sim P_{X,\tau}(\cdot \mid x, a)} \max_{a' \in \mathcal{A}} q_{\tau,k}(X', a') \\ &= (\mathcal{B}_\tau q_{\tau,k})(x, a) = q_{\tau,k+1}(x, a). \end{aligned} \tag{35}$$

The second equality uses exactly the recursive identity $X_{t+1} = x(H_{t+1})$ and its task-local marginal $P_{X,\tau}(\cdot \mid x, a)$. The reward term and continuation term may be dependent; linearity of conditional expectation separates their expectations. The continuation is core-only by the induction hypothesis, so its expectation needs only the next-core marginal. Equation equation 35 closes the induction.

The argument fixed one task throughout. It therefore permits task-dependent K^r_τ and $P_{X,\tau}$, permits arbitrary nuisance dynamics consistent with the two stated marginals, and evaluates $Q^{\mathrm{can}}_{\tau,k}$ only on $\mathcal{R}_\tau$. This proves the proposition.

□

Proposition 66 (Canonical discounted limit and exact factorization)**.** *Under Assumptions 44 and 49, Lemma 64, and Proposition 65, fix $\tau \in \mathcal{T}$. The operator $\mathcal{B}_\tau$ is a γ-contraction self-map on the complete sup-normed space of bounded measurable functions on $\mathcal{X}^{\mathrm{reach}}_\tau \times \mathcal{A}$. The iterates $q_{\tau,k} = \mathcal{B}^k_\tau 0$ converge uniformly to a bounded measurable limit $q_{\tau,\infty}$, and this limit is the unique bounded measurable fixed point of $\mathcal{B}_\tau$; consequently it is the object denoted $q^\star_\tau$ above. Let $Q^\star_\tau$ be the task-reachable infinite-horizon value defined above. For every $k \geq 0$, with the convention $\gamma^0 = 1$,*

$$0 \leq Q^\star_\tau(h, a) - Q^{\mathrm{can}}_{\tau,k}(h, a) \leq B\gamma^k \quad (h \in \mathcal{R}_\tau, a \in \mathcal{A}), \tag{36}$$

and

$$0 \le q^\star_\tau(x,a) - q_{\tau,k}(x,a) \le B\gamma^k \quad (x \in \mathcal{X}^{\mathrm{reach}}_\tau, a \in \mathcal{A}). \tag{37}$$

Consequently,

$$Q^\star_\tau(h,a) = q^\star_\tau(x(h),a) \quad \textit{for every } h \in \mathcal{R}_\tau, a \in \mathcal{A}, \tag{38}$$

and therefore for every $h \in \mathcal{H}_\tau$.

Proof. We first identify the history-side limit directly with $Q^\star_\tau$. For any admissible task-τ continuation policy and any $k \ge 0$, nonnegative bounded rewards give the pathwise tail bound

$$0 \le \sum_{j=0}^{\infty} \gamma^j R_{t+j} - \sum_{j=0}^{k-1} \gamma^j R_{t+j} \le R_{\max} \sum_{j=k}^{\infty} \gamma^j = B\gamma^k. \tag{39}$$

The same policy class and the same chosen conditional task-reachable version occur in the finite and infinite definitions. Taking conditional expectations and then suprema over policies shows first that $Q^{\mathrm{can}}_{\tau,k} \le Q^\star_\tau$, and then that $Q^\star_\tau \le Q^{\mathrm{can}}_{\tau,k} + B\gamma^k$. This proves equation 36, uniformly over the stated reachable domain.

We now construct the core-side fixed point. Equip $\mathcal{X}^{\mathrm{reach}}_\tau$ with the inherited measurable structure and consider the real bounded measurable functions on $\mathcal{X}^{\mathrm{reach}}_\tau \times \mathcal{A}$, with the sup norm. This space is complete. Indeed, if (f_j) is sup-norm Cauchy, then $(f_j(x,a))$ is Cauchy in $\mathbb{R}$ at every (x,a), defining a pointwise limit f. Given $\varepsilon > 0$, choose J so that $\|f_j - f_\ell\|_\infty \le \varepsilon$ for $j, \ell \ge J$; letting $\ell \to \infty$ pointwise gives $\|f_j - f\|_\infty \le \varepsilon$ for every $j \ge J$. Thus convergence is uniform, and f is bounded and measurable as a pointwise limit of measurable functions.

Let q be any member of this space. Because $\mathcal{A}$ is finite, $x' \mapsto \max_{a'} q(x',a')$ is bounded and measurable. The defining measurability of the stochastic kernel $P_{X,\tau}$, the measurability of $\bar{r}_\tau$, and the forward-closure conclusion of Lemma 64 therefore imply that $\mathcal{B}_\tau q$ is measurable on the same domain and evaluates q only on $\mathcal{X}^{\mathrm{reach}}_\tau$. Moreover,

$$\|\mathcal{B}_\tau q\|_\infty \le R_{\max} + \gamma \|q\|_\infty < \infty. \tag{40}$$

Thus $\mathcal{B}_\tau$ is a self-map of this complete bounded-measurable space. For any two members q, q', the elementary finite-max inequality gives

$$\begin{aligned} \|\mathcal{B}_\tau q - \mathcal{B}_\tau q'\|_\infty &\le \gamma \sup_{x,a} \mathbb{E}_{X' \sim P_{X,\tau}(\cdot|x,a)} \left| \max_{a'} q(X',a') - \max_{a'} q'(X',a') \right| \\ &\le \gamma \|q - q'\|_\infty. \end{aligned} \tag{41}$$

It remains to construct and identify its fixed point. Lemma 64 gives $q_{\tau,0} = 0$ and the recursion $q_{\tau,k+1} = \mathcal{B}_\tau q_{\tau,k}$; because zero is bounded and measurable and the self-map property was just proved, induction shows that every iterate is bounded and measurable. Their first increment satisfies $\|q_{\tau,1} - q_{\tau,0}\|_\infty \le R_{\max}$. Repeated use of equation 41 gives, for every $j \ge 0$,

$$\|q_{\tau,j+1} - q_{\tau,j}\|_\infty \le R_{\max} \gamma^j. \tag{42}$$

Indeed, $j = 0$ is the first-increment bound, and the induction step follows from $\|q_{\tau,j+1} - q_{\tau,j}\|_\infty \le \gamma \|q_{\tau,j} - q_{\tau,j-1}\|_\infty$. Consequently, for integers $m > k \ge 0$,

$$\|q_{\tau,m} - q_{\tau,k}\|_\infty \le \sum_{j=k}^{m-1} \|q_{\tau,j+1} - q_{\tau,j}\|_\infty \le R_{\max} \sum_{j=k}^{m-1} \gamma^j \le B\gamma^k. \tag{43}$$

The sequence is uniformly Cauchy, so completeness constructs a bounded measurable limit

$$q_{\tau,\infty} := \lim_{k \to \infty} q_{\tau,k}, \qquad \|q_{\tau,\infty}\|_\infty \le B, \tag{44}$$

where the bound follows from equation 30. The self-map contraction and the recursion imply

$$\begin{aligned} \|\mathcal{B}_\tau q_{\tau,\infty} - q_{\tau,\infty}\|_\infty &\le \|\mathcal{B}_\tau q_{\tau,\infty} - \mathcal{B}_\tau q_{\tau,k}\|_\infty + \|q_{\tau,k+1} - q_{\tau,\infty}\|_\infty \\ &\le \gamma \|q_{\tau,\infty} - q_{\tau,k}\|_\infty + \|q_{\tau,k+1} - q_{\tau,\infty}\|_\infty \longrightarrow 0. \end{aligned} \tag{45}$$

Hence $q_{\tau,\infty} = \mathcal{B}_\tau q_{\tau,\infty}$. If p is any other bounded measurable fixed point, equation 41 gives

$$\|p - q_{\tau,\infty}\|_\infty = \|\mathcal{B}_\tau p - \mathcal{B}_\tau q_{\tau,\infty}\|_\infty \le \gamma \|p - q_{\tau,\infty}\|_\infty. \tag{46}$$

Because $1-\gamma>0$, this norm is zero. Thus the constructed limit is the unique bounded measurable fixed point, which justifies the notation above $q_\tau^\star := q_{\tau,\infty}$.

Finally, $\mathcal{B}_\tau$ is monotone because finite maxima, integration against a probability kernel, positive scaling by γ, and addition all preserve pointwise order. Since $q_{\tau,1}=\mathcal{B}_\tau 0\geq 0=q_{\tau,0}$, so $q_{\tau,k+1}\geq q_{\tau,k}$ pointwise. Letting $m\to\infty$ in equation 43 therefore gives exactly

$$0\leq q_\tau^\star(x,a)-q_{\tau,k}(x,a)\leq B\gamma^k, \tag{47}$$

which is equation 37.

For any $h\in\mathcal{R}_\tau$ and $a\in\mathcal{A}$, Proposition 65, equation 36, and equation 37 now yield the explicit same-target comparison

$$\begin{aligned}|Q_\tau^\star(h,a)-q_\tau^\star(x(h),a)|&\leq \left|Q_\tau^\star(h,a)-Q_{\tau,k}^{\mathrm{can}}(h,a)\right|\\&\quad+\left|Q_{\tau,k}^{\mathrm{can}}(h,a)-q_{\tau,k}(x(h),a)\right|\\&\quad+|q_{\tau,k}(x(h),a)-q_\tau^\star(x(h),a)|\\&\leq 2B\gamma^k.\end{aligned} \tag{48}$$

Since $\gamma^k\to 0$, equation 38 follows. The middle residual in equation 48 is exactly zero by Proposition 65.

The boundary cases are included. If $R_{\max}=0$, then $B=0$, equation 42 makes every core iterate zero, and the history tail makes every history value zero. If $\gamma=0$, the self-map has contraction modulus zero, $q_{\tau,1}=\bar{r}_\tau$ is already its fixed point, and the $k=0$ tails in equation 36–equation 37 are bounded by B, while every $k\geq 1$ tail vanishes. Thus one may take $k=1$ in equation 48. Finally, $\mathcal{H}_\tau\subseteq\mathcal{R}_\tau$, so equation 38 restricts pointwise to $\mathcal{H}_\tau$. This proves the proposition. □

C.5.6 A REALIZABLE CORE-PROFILE ACHIEVER

Lemma 67 (Taskwise activity of the reachable-profile selector)**.** *Under Assumption 51 and the mixture with $w_\tau>0$, for every $\tau\in\mathcal{T}$,*

$$\mu_{\tau,X}\left(\left\{x : r_\tau(x)>0,\quad u_{c_{\mathrm{dec}}(x)}\in\mathcal{U}_{\mathrm{reach}}(x),\quad (u_{c_{\mathrm{dec}}(x)})_\tau\in\arg\max_{a\in\mathcal{A}} q_\tau^\star(x,a)\right\}\right)=1. \tag{49}$$

Proof. Because $\bar{\mu}_X=\sum_{\sigma\in\mathcal{T}} w_\sigma\mu_{\sigma,X}$ and $w_\tau>0$, every $\bar{\mu}_X$-null set is $\mu_{\tau,X}$-null. Thus $\mu_{\tau,X}\ll\bar{\mu}_X$, consistently with the derivative $r_\tau=\mathrm{d}\mu_{\tau,X}/\mathrm{d}\bar{\mu}_X$. For the measurable set where this derivative vanishes, the defining Radon–Nikodym identity gives

$$\mu_{\tau,X}\big(\{x : r_\tau(x)=0\}\big)=\int_{\{r_\tau=0\}} r_\tau(x)\,\bar{\mu}_X(\mathrm{d}x)=0. \tag{50}$$

Hence $r_\tau(x)>0$, equivalently $\tau\in\mathcal{T}(x)$, for $\mu_{\tau,X}$-almost every x.

Assumption 51 states that

$$u_{c_{\mathrm{dec}}(x)}\in\mathcal{U}_{\mathrm{reach}}(x)\quad\text{for }\bar{\mu}_X\text{-almost every } x. \tag{51}$$

The domination just proved transfers the same null set in equation 51 to $\mu_{\tau,X}$. Moreover, if $H\sim\mu_\tau$, then $H\in\mathcal{H}_\tau\subseteq\mathcal{R}_\tau$ almost surely and consequently $X=x(H)\in\mathcal{X}_\tau^{\mathrm{reach}}$ almost surely, so every core value used below lies on the task's declared domain.

On the intersection of the full-measure sets from equation 50–equation 51, task τ is active. By the definition of $\mathcal{U}_{\mathrm{reach}}(x)$, membership in equation 51 then imposes exactly

$$(u_{c_{\mathrm{dec}}(x)})_\tau\in\arg\max_{a\in\mathcal{A}} q_\tau^\star(x,a). \tag{52}$$

Coordinates outside $\mathcal{T}(x)$ are unconstrained, but equation 50 shows that such an inactive τ-coordinate occurs only on a $\mu_{\tau,X}$-null set and is therefore never scored by task τ's population loss. If equation 52 contains several actions, every tied maximizer has the same optimal value, so no uniqueness or margin is needed. Equations equation 50–equation 52 prove equation 49.

□

Proposition 68 (Exact zero deficiency of the core profile achiever)**.** *Under Assumptions 45 and 51, Proposition 66, and Lemma 67, the realizable coordinate heads satisfy, for every $\tau \in \mathcal{T}$,*

$$\ell_{\tau,\phi_{\mathrm{dec}},g_{\mathrm{dec},\tau}}(H) = 0 \quad \mu_\tau\text{-almost surely}, \tag{53}$$

and therefore

$$D_\tau(\phi_{\mathrm{dec}}) = 0 \quad \text{for every } \tau, \qquad D(\phi_{\mathrm{dec}}) = 0. \tag{54}$$

Proof. Fix a task τ. Under $H \sim \mu_\tau$, put $X = x(H)$. By Lemma 67, outside a μ_τ-null set the selected profile has an active optimal coordinate. On that full-measure set, the realizability identities and equation 52 give

$$g_{\mathrm{dec},\tau}(\phi_{\mathrm{dec}}(H)) = g_{\mathrm{dec},\tau}(c_{\mathrm{dec}}(X)) = (u_{c_{\mathrm{dec}}(X)})_\tau \in \arg\max_{a\in\mathcal{A}} q^\star_\tau(X,a). \tag{55}$$

Proposition 66 applies to this same τ, this reachable history H, and every action. It therefore yields the exact, zero-residual chain

$$\begin{aligned} Q^\star_\tau(H, g_{\mathrm{dec},\tau}(\phi_{\mathrm{dec}}(H))) &= q^\star_\tau\big(X, (u_{c_{\mathrm{dec}}(X)})_\tau\big) \\ &= \max_{a\in\mathcal{A}} q^\star_\tau(X,a) \\ &= \max_{a\in\mathcal{A}} Q^\star_\tau(H,a) = V^\star_\tau(H). \end{aligned} \tag{56}$$

The first and third equalities use the preceding result on the same task domain; the middle equality is precisely equation 55. Substitution into the loss proves equation 53.

Assumption 51 places $g_{\mathrm{dec},\tau}$ in $\mathcal{G}_{\phi_{\mathrm{dec}}}$. Hence equation 53 makes it a legal zero-risk competitor:

$$0 \le D_\tau(\phi_{\mathrm{dec}}) = \min_{g\in\mathcal{G}_{\phi_{\mathrm{dec}}}} L_\tau(\phi_{\mathrm{dec}}, g) \le L_\tau(\phi_{\mathrm{dec}}, g_{\mathrm{dec},\tau}) = 0. \tag{57}$$

The first inequality follows from the pointwise nonnegativity of every regret loss. Thus each taskwise deficiency is exactly zero. Taking the maximum over the finite task family gives equation 54. Action ties do not change equation 56, and disjoint task-core supports do not create a conflict because inactive coordinates are never scored by the corresponding task, as proved in Lemma 67.

□

Proposition 69 (Exact full-range rate of the core profile achiever)**.** *Under Assumptions 45 and 51,*

$$R(\phi_{\mathrm{dec}}) = \log|\mathcal{Z}_{\phi_{\mathrm{dec}}}| = \log m^{\mathrm{reach}}_{\mathrm{dec}}. \tag{58}$$

The encoder is core-only, and its full code range has no additional mixture-inactive code. In the boundary case $m^{\mathrm{reach}}_{\mathrm{dec}} = 1$, equation 58 reads $R(\phi_{\mathrm{dec}}) = 0$.

Proof. The rate definition is the fixed-length full-range rate $R(\phi) = \log|\mathcal{Z}_\phi|$. Assumption 51 supplies the exact counts

$$|\mathcal{Z}_{\phi_{\mathrm{dec}}}| = |\mathcal{Z}^{\bar\mu}_{\phi_{\mathrm{dec}}}| = m^{\mathrm{reach}}_{\mathrm{dec}}. \tag{59}$$

Because $\bar\mu$ is a probability measure and the code range is finite, its pushforward assigns positive mass to at least one code. Thus $|\mathcal{Z}^{\bar\mu}_{\phi_{\mathrm{dec}}}| \ge 1$, and equation 59 shows that the logarithm in equation 58 is well defined. Since $\mathcal{Z}^{\bar\mu}_{\phi_{\mathrm{dec}}} \subseteq \mathcal{Z}_{\phi_{\mathrm{dec}}}$ and the two finite sets have equal cardinality, they coincide; in particular, an off-mixture range extension cannot add a rate surcharge.

Substituting the first equality in equation 59 into the rate definition proves equation 58. With the natural-log convention, this rate is measured in nats. Finally, $\phi_{\mathrm{dec}}(h) = c_{\mathrm{dec}}(x(h))$ depends only on the observable core, so this exact-rate achiever is nuisance-free by construction. If $m^{\mathrm{reach}}_{\mathrm{dec}} = 1$, then $\log 1 = 0$. □

C.5.7 Conditional support transfer for active codes

Lemma 70 (Simultaneous zero-risk head attainment)**.** *Under Assumption 45, fix $\phi \in \Phi_M$. If the condition $D(\phi) = 0$ holds, then for every $\tau \in \mathcal{T}$ there exists $g_\tau \in \mathcal{G}_\phi$ such that*

$$L_\tau(\phi, g_\tau) = D_\tau(\phi) = 0. \tag{60}$$

Consequently, for every $z \in \mathcal{Z}_\phi$, the full task profile

$$u_z := \big(g_\tau(z)\big)_{\tau\in\mathcal{T}} \in \mathcal{A}^{\mathcal{T}} \tag{61}$$

is well defined.

Proof. Every loss $\ell_{\tau,\phi,g}$ is nonnegative, so every $L_\tau(\phi, g)$ and every $D_\tau(\phi)$ is nonnegative. Since

$$0 = D(\phi) = \max_{\tau \in \mathcal{T}} D_\tau(\phi),$$

we have $D_\tau(\phi) = 0$ for every task. The finite nonempty class $\mathcal{G}_\phi$ attains the minimum in the definition of $D_\tau(\phi)$; choose any minimizer and call it g_τ. This proves equation 60. The task family is finite and each head is a total map on $\mathcal{Z}_\phi$, so the coordinate tuple in equation 61 is defined for every code.

□

Lemma 71 (Measurable active-domain disintegration identities)**.** *Under Assumption 45, Proposition 66, and Lemma 70, retain the local conditional hypothesis $D(\phi) = 0$ and the zero-risk heads g_τ. There exists a measurable set $\Omega_{\mathrm{dis},\phi}$ with $\bar{\mu}_X(\Omega_{\mathrm{dis},\phi}) = 1$ such that the following holds. For every $x \in \Omega_{\mathrm{dis},\phi}$, every active task $\tau \in \mathcal{T}(x)$, and every $z \in \mathcal{Z}_\phi$, all of the quantities below are defined on the same task-reachable core domain:*

$$E_z(x) := \{n : \phi(x, n) = z\}, \qquad p_{\tau,z}(x) := \nu_{\tau,x}(E_z(x)), \tag{62}$$

$$\Delta_{\tau,z}(x) := \max_{a \in \mathcal{A}} q^\star_\tau(x, a) - q^\star_\tau\big(x, g_\tau(z)\big) \geq 0, \tag{63}$$

and they satisfy the exact active-domain identity

$$r_\tau(x) p_{\tau,z}(x) \Delta_{\tau,z}(x) = 0. \tag{64}$$

The maps in equation 62–equation 63 are measurable on the active, task-reachable subspace on which they are used.

Proof. We first make the domain restriction explicit. Under $H \sim \mu_\tau$, the definitions give $H \in \mathcal{H}_\tau \subseteq \mathcal{R}_\tau$ almost surely; hence

$$\mu_{\tau,X}(\mathcal{X}^{\mathrm{reach}}_\tau) = 1. \tag{65}$$

Using $d\mu_{\tau,X} = r_\tau \, d\bar{\mu}_X$, equation 65 implies

$$\int_{(\mathcal{X}^{\mathrm{reach}}_\tau)^c} r_\tau(x) \, \bar{\mu}_X(dx) = 0. \tag{66}$$

Therefore, outside a $\bar{\mu}_X$-null set, the implication

$$r_\tau(x) > 0 \quad \Longrightarrow \quad x \in \mathcal{X}^{\mathrm{reach}}_\tau \tag{67}$$

holds. Thus the core value in equation 63 is evaluated only where Proposition 66 defines its same-task value.

For measurability, regard the shared deterministic code cell as the measurable set on the union-of-task-reachable observable-pair domain

$$E_z := \{(x, n) : \phi(x, n) = z\}$$

and let $E_z(x)$ be its section. For a regular conditional kernel $x \mapsto \nu_{\tau,x}$, the map $x \mapsto \nu_{\tau,x}(E_z(x))$ is measurable: this is immediate for measurable rectangles, is preserved under complements and monotone disjoint unions, and hence holds for the sigma-field generated by rectangles. We use the resulting version only at points satisfying $r_\tau(x) > 0$. The finite maximum in equation 63, the fixed action $g_\tau(z)$, and the bounded measurability of $q^\star_\tau$ constructed by Proposition 66 make $\Delta_{\tau,z}$ measurable on the same active, task-reachable subspace.

Proposition 66 is now applied only to $H \in \mathcal{H}_\tau$. For μ_τ-almost every $h = (x, n)$,

$$\begin{aligned} \ell_{\tau,\phi,g_\tau}(h) &= \max_{a \in \mathcal{A}} Q^\star_\tau(h, a) - Q^\star_\tau\big(h, g_\tau(\phi(h))\big) \\ &= \max_{a \in \mathcal{A}} q^\star_\tau(x, a) - q^\star_\tau\big(x, g_\tau(\phi(x, n))\big) \\ &= \sum_{z \in \mathcal{Z}_\phi} \mathbf{1}_{E_z(x)}(n) \Delta_{\tau,z}(x). \end{aligned} \tag{68}$$

Disintegrating the nonnegative measurable expression in equation 68 under the task's own joint law of (X, N), using equation 60, and then changing only the core measure gives

$$\begin{aligned} 0 = L_\tau(\phi, g_\tau) &= \int_{\{x: r_\tau(x)>0\}\cap \mathcal{X}_\tau^{\mathrm{reach}}} \sum_{z\in\mathcal{Z}_\phi} p_{\tau,z}(x)\Delta_{\tau,z}(x)\,\mu_{\tau,X}(dx) \\ &= \int_{\{x: r_\tau(x)>0\}\cap \mathcal{X}_\tau^{\mathrm{reach}}} r_\tau(x) \sum_{z\in\mathcal{Z}_\phi} p_{\tau,z}(x)\Delta_{\tau,z}(x)\,\bar{\mu}_X(dx). \end{aligned} \tag{69}$$

The restriction in equation 69 loses no $\mu_{\tau,X}$-mass: the complement of $\{r_\tau > 0\}$ has $\mu_{\tau,X}$-measure zero by the Radon–Nikodym identity, and equation 65 handles the reachable-core restriction.

Every summand in equation 69 is nonnegative. Hence the integrand is zero $\bar{\mu}_X$-almost everywhere on its explicitly stated domain. Since $\mathcal{Z}_\phi$ is finite, each summand is then zero there, proving equation 64 for every code. Finally, $\mathcal{T}$ and $\mathcal{Z}_\phi$ are finite, so we may intersect the full sets from equation 67, equation 69, and all task/code identities to obtain one $\Omega_{\mathrm{dis},\phi}$ of full $\bar{\mu}_X$-measure.

Crucially, equation 64 is stated only after $\tau \in \mathcal{T}(x)$ has made $r_\tau(x) > 0$ and equation 67 has made the task-reachable core value legal.

□

Lemma 72 (Mutual-absolute-continuity transfer of code-cell support)**.** *Under Assumption 50 and Lemma 71, there exists a measurable set $\Omega_{\mathrm{supp},\phi} \subseteq \Omega_{\mathrm{dis},\phi}$ with $\bar{\mu}_X(\Omega_{\mathrm{supp},\phi}) = 1$ such that, for every $x \in \Omega_{\mathrm{supp},\phi}$ and $z \in \mathcal{Z}_\phi$,*

$$z \in \mathcal{Z}_\phi(x) \quad\Longrightarrow\quad p_{\tau,z}(x) > 0 \quad \textit{for every } \tau \in \mathcal{T}(x). \tag{70}$$

Proof. Intersect $\Omega_{\mathrm{dis},\phi}$ with the $\bar{\mu}_X$-full set on which $\sum_\tau w_\tau r_\tau(x) = 1$, the conditional mixture is

$$\bar{\nu}_x = \sum_{\sigma\in\mathcal{T}(x)} w_\sigma r_\sigma(x)\nu_{\sigma,x}, \tag{71}$$

and all active-task conditional laws are mutually absolutely continuous. Only active terms appear in equation 71.

Fix x in that common full set and a code $z \in \mathcal{Z}_\phi(x)$. By the definition of conditional code activity and equation 71,

$$0 < \bar{\nu}_x(E_z(x)) = \sum_{\sigma\in\mathcal{T}(x)} w_\sigma r_\sigma(x) p_{\sigma,z}(x). \tag{72}$$

The sum is finite and nonnegative. Therefore some $\sigma \in \mathcal{T}(x)$ has $p_{\sigma,z}(x) > 0$; for that task the coefficient $w_\sigma r_\sigma(x)$ is also strictly positive. Now fix any $\tau \in \mathcal{T}(x)$. If $p_{\tau,z}(x) = 0$, then $\nu_{\tau,x}(E_z(x)) = 0$. Mutual absolute continuity, specifically $\nu_{\sigma,x} \ll \nu_{\tau,x}$, would imply $\nu_{\sigma,x}(E_z(x)) = 0$, contradicting $p_{\sigma,z}(x) > 0$. Thus $p_{\tau,z}(x) > 0$, proving equation 70.

This argument uses only null-set equivalence for the same encoder cell at the same core point, with no density-ratio lower bound. If $|\mathcal{T}(x)| = 1$, the positive term in equation 72 already proves equation 70; if task core supports are disjoint, the cross-task part of the assumption is vacuous.

□

Proposition 73 (Profile legality of every conditionally active code)**.** *Under Assumptions 45 and 50, Proposition 66, and the local conditional hypothesis $D(\phi) = 0$, choose the heads and profiles in Lemma 70. Then, for $\bar{\mu}_X$-almost every x,*

$$z \in \mathcal{Z}_\phi(x) \quad\Longrightarrow\quad u_z \in \mathcal{U}_{\mathrm{reach}}(x) \qquad (z \in \mathcal{Z}_\phi). \tag{73}$$

Proof. Fix $x \in \Omega_{\mathrm{supp},\phi}$ and $z \in \mathcal{Z}_\phi(x)$. Let $\tau \in \mathcal{T}(x)$ be arbitrary. Lemma 72 gives $p_{\tau,z}(x) > 0$, while activity gives $r_\tau(x) > 0$. The active-domain identity from Lemma 71 is therefore an equality of defined quantities with two strictly positive factors:

$$0 = r_\tau(x) p_{\tau,z}(x) \Delta_{\tau,z}(x).$$

Consequently $\Delta_{\tau,z}(x) = 0$, and the definition in equation 63 yields

$$g_\tau(z) \in \arg\max_{a \in \mathcal{A}} q^\star_\tau(x, a). \tag{74}$$

Because τ was arbitrary, equation 74 holds for every active task. The definition of $\mathcal{U}_{\mathrm{reach}}(x)$ places no constraint on coordinates $\tau \notin \mathcal{T}(x)$. Hence $u_z \in \mathcal{U}_{\mathrm{reach}}(x)$, proving equation 73. □

C.5.8 From conditional code activity to a global profile cover

Lemma 74 (Mixture activity and the conditional code partition)**.** *Under Assumption 45 and the mixture disintegration defined above, fix $\phi \in \Phi_M$. For every $z \in \mathcal{Z}_\phi$, define the measurable encoder cell and its nuisance section by*

$$E_z := \{(x, n) : \phi(x, n) = z\}, \qquad E_z(x) := \{n : (x, n) \in E_z\}, \tag{75}$$

and put

$$\bar{p}_z(x) := \bar{\nu}_x(E_z(x)), \qquad \beta_z := \bar{\mu}\{h : \phi(h) = z\}. \tag{76}$$

Then $\bar{p}_z$ is measurable, $0 \le \bar{p}_z \le 1$, and

$$\beta_z = \int \bar{p}_z(x)\, \bar{\mu}_X(dx). \tag{77}$$

Moreover, there is a measurable set $\Omega_{\mathrm{mix},\phi}$ with $\bar{\mu}_X(\Omega_{\mathrm{mix},\phi}) = 1$ such that

$$\sum_{z \in \mathcal{Z}_\phi} \bar{p}_z(x) = 1 \qquad (x \in \Omega_{\mathrm{mix},\phi}). \tag{78}$$

Proof. The deterministic encoder is measurable, hence every E_z in equation 75 is measurable. To check section measurability, let $\mathscr{D}$ be the class of measurable sets E for which $x \mapsto \bar{\nu}_x(E(x))$ is measurable. Probability-kernel measurability puts every measurable rectangle in $\mathscr{D}$; closure under complements and increasing disjoint unions follows from finite additivity and monotone convergence of the probability kernel. Thus the monotone-class construction for the product sigma-field gives $E_z \in \mathscr{D}$. This proves measurability of $\bar{p}_z$, while its range $[0, 1]$ follows because $\bar{\nu}_x$ is a probability law on the retained mixture domain.

By construction, $\bar{\nu}_x$ is the regular conditional nuisance law under the joint history mixture. Applying that disintegration to the indicator of the same cell E_z gives

$$\begin{aligned} \beta_z &= \int \mathbf{1}_{E_z}(x, n)\, \bar{\mu}(dx, dn) \\ &= \int \left[\int \mathbf{1}_{E_z}(x, n)\, \bar{\nu}_x(dn) \right] \bar{\mu}_X(dx) \\ &= \int \bar{p}_z(x)\, \bar{\mu}_X(dx), \end{aligned} \tag{79}$$

which is equation 77. Equivalently, on the active-domain version,

$$\bar{p}_z(x) = \sum_{\tau \in \mathcal{T}(x)} w_\tau r_\tau(x) \nu_{\tau,x}(E_z(x)), \tag{80}$$

where inactive zero-weight terms are omitted.

Finally, the cells $\{E_z : z \in \mathcal{Z}_\phi\}$ are a finite disjoint partition of the encoder's history domain. Conditioning that partition on $X = x$ preserves total mass for $\bar{\mu}_X$-almost every x, so

$$\sum_{z \in \mathcal{Z}_\phi} \bar{p}_z(x) = \bar{\nu}_x\left(\bigsqcup_{z \in \mathcal{Z}_\phi} E_z(x) \right) = 1. \tag{81}$$

Because the sum is finite, the equality holds simultaneously for all its terms on one common full set $\Omega_{\mathrm{mix},\phi}$.

□

Lemma 75 (Finite-null local-to-global code activity)**.** *Under Assumption 45 and Lemma 74, there exists a measurable set $\Omega_{\mathrm{act},\phi} \subseteq \Omega_{\mathrm{mix},\phi}$ with $\bar{\mu}_X(\Omega_{\mathrm{act},\phi}) = 1$ such that*

$$\mathcal{Z}_\phi(x) \subseteq \mathcal{Z}_\phi^{\bar{\mu}} \qquad (x \in \Omega_{\mathrm{act},\phi}). \tag{82}$$

Proof. By definition and equation 76,

$$z \notin \mathcal{Z}_\phi^{\bar{\mu}} \quad \Longleftrightarrow \quad \beta_z = 0. \tag{83}$$

Fix a code satisfying equation 83. Lemma 74 gives the exact nonnegative identity

$$0 = \int \bar{p}_z(x)\, \bar{\mu}_X(dx), \qquad \bar{p}_z(x) \geq 0. \tag{84}$$

For every integer $k \geq 1$,

$$0 = \int \bar{p}_z(x)\, \bar{\mu}_X(dx) \geq \int_{\{\bar{p}_z \geq 1/k\}} \bar{p}_z(x)\, \bar{\mu}_X(dx) \geq \frac{1}{k}\, \bar{\mu}_X\{x : \bar{p}_z(x) \geq 1/k\}. \tag{85}$$

Hence each set on the right is null, and therefore

$$\bar{\mu}_X\{x : \bar{p}_z(x) > 0\} = \bar{\mu}_X\left(\bigcup_{k=1}^{\infty} \{x : \bar{p}_z(x) \geq 1/k\}\right) = 0. \tag{86}$$

There are only finitely many codes in $\mathcal{Z}_\phi$. Consequently the set

$$\Omega_{\mathrm{act},\phi} := \Omega_{\mathrm{mix},\phi} \setminus \bigcup_{z \in \mathcal{Z}_\phi :\, \beta_z = 0} \{x : \bar{p}_z(x) > 0\} \tag{87}$$

has full $\bar{\mu}_X$-measure. If $x \in \Omega_{\mathrm{act},\phi}$ and $z \in \mathcal{Z}_\phi(x)$, then by definition $\bar{p}_z(x) > 0$. Equation equation 87 rules out $\beta_z = 0$, so $\beta_z > 0$ and therefore $z \in \mathcal{Z}_\phi^{\bar{\mu}}$. This proves equation 82.

The conclusion holds almost everywhere: a globally inactive code may be conditionally active at an exceptional core point, but all such points for all inactive codes are removed in the single finite union equation 87 before the cover is formed.

□

Proposition 76 (Global active-profile cover and exact code lower bound)**.** *Under Assumption 45, Proposition 73, Lemmas 74 and 75, and the local converse hypothesis $D(\phi) = 0$, define*

$$\mathcal{C}_\phi := \{u_z : z \in \mathcal{Z}_\phi^{\bar{\mu}}\} \subseteq \mathcal{A}^{\mathcal{T}}. \tag{88}$$

Then $\mathcal{C}_\phi$ is a valid $\mathcal{U}_{\mathrm{reach}}$-cover and

$$m_{\mathrm{dec}}^{\mathrm{reach}} \leq |\mathcal{C}_\phi| \leq |\mathcal{Z}_\phi^{\bar{\mu}}| \leq |\mathcal{Z}_\phi|. \tag{89}$$

Proof. Proposition 73 supplies a measurable set $\Omega_{\mathrm{leg},\phi}$ of full $\bar{\mu}_X$-measure on which every conditionally active code has a legal profile. Intersect it with the full set from Lemma 75:

$$\Omega_\phi^\star := \Omega_{\mathrm{leg},\phi} \cap \Omega_{\mathrm{act},\phi}. \tag{90}$$

Fix $x \in \Omega_\phi^\star$. Equation equation 78 is valid there because $\Omega_{\mathrm{act},\phi} \subseteq \Omega_{\mathrm{mix},\phi}$. Its finitely many summands are nonnegative and sum to one, so at least one code $z_x \in \mathcal{Z}_\phi$ satisfies $\bar{p}_{z_x}(x) > 0$. Thus

$$z_x \in \mathcal{Z}_\phi(x). \tag{91}$$

Lemma 75 upgrades equation 91 to $z_x \in \mathcal{Z}_\phi^{\bar{\mu}}$, while Proposition 73 gives

$$u_{z_x} \in \mathcal{U}_{\mathrm{reach}}(x). \tag{92}$$

By equation 88, the same profile u_{z_x} lies in $\mathcal{C}_\phi$. Therefore

$$\mathcal{C}_\phi \cap \mathcal{U}_{\mathrm{reach}}(x) \neq \varnothing \qquad (x \in \Omega_\phi^\star), \tag{93}$$

which is precisely the definition of a valid reachable-profile cover. No measurable selection of z_x is required: equation 93 is a pointwise existence statement on a common full set.

The first inequality in equation 89 now follows from the definition of $m_{\mathrm{dec}}^{\mathrm{reach}}$ as the minimum cardinality of any such cover. The map $z \mapsto u_z$ sends $\mathcal{Z}_\phi^{\bar{\mu}}$ onto $\mathcal{C}_\phi$, so elementary finite-image counting yields the second inequality even when several codes have the same profile. Finally, $\mathcal{Z}_\phi^{\bar{\mu}} \subseteq \mathcal{Z}_\phi$ by definition, which gives the third inequality.

Action ties cause no difficulty because profile legality requires membership in an argmax set, not uniqueness. If task-core supports are disjoint, $\mathcal{U}_{\mathrm{reach}}(x)$ constrains only the task coordinates active at that x, exactly as in the preceding result. Nuisance-dependent splitting may create several globally active codes with the same profile; this can only make $|\mathcal{C}_\phi| < |\mathcal{Z}_\phi^{\bar{\mu}}|$, in the safe direction for equation 89. The one-profile case $m_{\mathrm{dec}}^{\mathrm{reach}} = 1$ and codes active only on null core sets are both covered by equation 87–equation 89. □

C.5.9 The finite-sample oracle and its baseline reductions

Proposition 77 (Base finite-sample oracle and taskwise deployment guarantee)**.** *Under Assumptions 44, 45, 46, and 47, and Propositions 55, 57, 59, and 60, let $\delta \in (0,1)$, fix any target-estimation realization satisfying $\mathcal{E}_Q$, and define*

$$\epsilon_\tau(\delta) = B\sqrt{\frac{\log(2N_{\mathrm{cls}}/\delta)}{2n_\tau}}, \qquad e_\tau(\delta) = \epsilon_\tau(\delta) + 2\eta_\tau, \qquad e_{\max}(\delta) = \max_{\tau\in\mathcal{T}} e_\tau(\delta).$$

Here and below, log *is the natural logarithm, consistently with the exponential Hoeffding tail; hence $R(\phi) = \log|\mathcal{Z}_\phi|$ is measured in nats. Then, under the evaluation marginal, one event has probability at least $1-\delta$ and simultaneously satisfies, for every task,*

$$\sup_{\phi\in\Phi_M,\, g\in\mathcal{G}_\phi} \left| P_{n_\tau}\widehat{\ell}_{\tau,\phi,g} - L_\tau(\phi,g)\right| \le e_\tau(\delta),$$

$$R(\widehat{\phi}) + \lambda D(\widehat{\phi}) \le \inf_{\phi\in\Phi_M}\{R(\phi) + \lambda D(\phi)\} + 2\lambda e_{\max}(\delta),$$

and

$$L_\tau(\widehat{\phi},\widehat{g}_\tau) \le D_\tau(\widehat{\phi}) + 2e_\tau(\delta).$$

If, in addition, Assumption 48 holds, so that Proposition 63 applies, the same event also satisfies

$$\begin{aligned} V_\tau^\star - V_\tau^{\widehat{\pi}_\tau} &= \frac{1}{1-\gamma}\mathbb{E}_{H\sim d_\tau^{\widehat{\pi}_\tau}}\ell_{\tau,\widehat{\phi},\widehat{g}_\tau}(H) \\ &\le \frac{C_\tau}{1-\gamma}L_\tau(\widehat{\phi},\widehat{g}_\tau) \\ &\le \frac{C_\tau}{1-\gamma}\big(D_\tau(\widehat{\phi}) + 2e_\tau(\delta)\big) \end{aligned}$$

for every task.

The exact indexed class count obeys

$$\begin{aligned} N_{\mathrm{cls}} &= \sum_{\tau\in\mathcal{T}}\sum_{\phi\in\Phi_M}|\mathcal{G}_\phi| \\ &\le \sum_{\tau\in\mathcal{T}}\sum_{\phi\in\Phi_M}|\mathcal{A}|^{|\mathcal{Z}_\phi|} \\ &\le |\mathcal{T}|\,|\Phi_M|\,|\mathcal{A}|^M. \end{aligned}$$

Consequently, without replacing the sharper exact-count radius,

$$\epsilon_\tau(\delta) \le B\sqrt{\frac{\log\big(2|\mathcal{T}|\,|\Phi_M|\,|\mathcal{A}|^M/\delta\big)}{2n_\tau}}.$$

The following reductions are exact.

1. If every $\eta_\tau = 0$, then, on $\mathcal{E}_Q$, $\widehat{\ell}_{\tau,\phi,g} = \ell_{\tau,\phi,g}$ on $\mathcal{H}_\tau$, $\mathcal{E}_{\mathrm{cal}}(\delta)$ and $\mathcal{E}_{\mathrm{true}}(\delta)$ agree almost surely under the evaluation marginal, and $e_\tau(\delta) = \epsilon_\tau(\delta)$. Therefore the oracle, selected-head, and optional deployment conclusions above reduce respectively to

$$R(\widehat{\phi}) + \lambda D(\widehat{\phi}) \le \inf_{\phi \in \Phi_M} \{R(\phi) + \lambda D(\phi)\} + 2\lambda \max_{\tau \in \mathcal{T}} \epsilon_\tau(\delta),$$

$$L_\tau(\widehat{\phi}, \widehat{g}_\tau) \le D_\tau(\widehat{\phi}) + 2\epsilon_\tau(\delta),$$

and, when coverage is imposed,

$$V_\tau^\star - V_\tau^{\widehat{\pi}_\tau} \le \frac{C_\tau}{1-\gamma}\big(D_\tau(\widehat{\phi}) + 2\epsilon_\tau(\delta)\big).$$

2. If

$$(\mu_\tau, n_\tau, \eta_\tau, C_\tau) = (\mu, n, \eta, C) \quad \textit{for every } \tau,$$

define the specialization shorthand

$$\epsilon := B\sqrt{\frac{\log(2N_{\mathrm{cls}}/\delta)}{2n}}, \qquad e := \epsilon + 2\eta.$$

Then $\epsilon_\tau = \epsilon$, $e_\tau = e_{\max} = e$, the uniform statistical radius is e, the oracle remainder is exactly $2\lambda e$, every selected-head remainder is exactly $2e$, and, when the common coverage premise is imposed,

$$V_\tau^\star - V_\tau^{\widehat{\pi}_\tau} \le \frac{C}{1-\gamma}\big(D_\tau(\widehat{\phi}) + 2e\big).$$

Taking the maximum over the finite task family gives the exact common-parameter form

$$\begin{aligned}\max_{\tau \in \mathcal{T}}\big(V_\tau^\star - V_\tau^{\widehat{\pi}_\tau}\big) &\le \frac{C}{1-\gamma}\max_{\tau \in \mathcal{T}}\big(D_\tau(\widehat{\phi}) + 2e\big) \\ &= \frac{C}{1-\gamma}\big(D(\widehat{\phi}) + 2e\big).\end{aligned}$$

The conclusions of this proposition are independent of Assumptions 49, 50, or 51.

Proof. Fix a target-estimation realization satisfying $\mathcal{E}_Q$. Proposition 57 supplies the single event $\mathcal{E}_{\mathrm{cal}}(\delta)$, with evaluation probability at least $1 - \delta$, and supplies on it the first displayed uniform deviation. Its probability is inherited from Proposition 55.

On exactly that event, Proposition 59 gives the displayed shared-encoder oracle. Proposition 60 gives the selected-head population inequality for every task. If the taskwise coverage premise is also imposed, Proposition 63 gives the entire equality-and-two-inequalities deployment chain. These are deterministic implications on $\mathcal{E}_{\mathrm{cal}}(\delta)$, so composing them does not alter its probability. Their task indices, scoring laws, losses, and selected objects agree exactly; in particular, the deployment result uses the selected-head population loss before its final substitution.

It remains to prove the local class calculation and the two requested specializations. For each ϕ, $\mathcal{G}_\phi \subseteq \mathcal{A}^{\mathcal{Z}_\phi}$, so

$$|\mathcal{G}_\phi| \le |\mathcal{A}^{\mathcal{Z}_\phi}| = |\mathcal{A}|^{|\mathcal{Z}_\phi|}.$$

The nonempty head classes make the finite action set nonempty, and $|\mathcal{Z}_\phi| \le M$; hence

$$|\mathcal{A}|^{|\mathcal{Z}_\phi|} \le |\mathcal{A}|^M.$$

Summing this inequality over the labeled task and encoder indices gives the displayed all-maps class bound. Since $\log$ and the nonnegative square-root map are increasing on their domains and $B \ge 0$, the class bound gives the displayed radius envelope. The exact N_{cls}-based radius remains the primary radius.

If every $\eta_\tau = 0$, the exact-target part of Proposition 57 yields $\widehat{Q}_\tau = Q_\tau^\star$ on the task scoring support, pointwise equality of estimated and true losses there, equality of the calibrated and true event indicators on $\mathcal{E}_Q$, and

$$e_\tau(\delta) = \epsilon_\tau(\delta) + 2 \cdot 0 = \epsilon_\tau(\delta).$$

Thus

$$e_{\max}(\delta) = \max_{\tau \in \mathcal{T}} \epsilon_\tau(\delta).$$

Substitution into the already proved oracle, selected-head, and deployment formulas gives the first exact reduction.

Under the homogeneous specialization,

$$n_\tau = n, \qquad \eta_\tau = \eta \quad \text{for every } \tau,$$

so direct substitution gives

$$\epsilon_\tau(\delta) = B\sqrt{\frac{\log(2N_{\mathrm{cls}}/\delta)}{2n}} = \epsilon, \qquad e_\tau(\delta) = \epsilon + 2\eta = e.$$

Because every task has the same radius,

$$e_{\max}(\delta) = \max_{\tau \in \mathcal{T}} e = e.$$

The oracle and selected-head constants therefore become $2\lambda e$ and $2e$ by equality. Under the common coverage constant, the taskwise deployment formula has the common multiplier $C/(1-\gamma)$. Finally,

$$\max_{\tau \in \mathcal{T}} \big(D_\tau(\widehat{\phi}) + 2e\big) = \max_{\tau \in \mathcal{T}} D_\tau(\widehat{\phi}) + 2e = D(\widehat{\phi}) + 2e,$$

which proves the common-parameter maximum form.

The boundary $B = 0$ is already covered by the concentration and calibration propositions: $\epsilon_\tau = 0$, and only the explicit $2\eta_\tau$ target perturbation can remain. If additionally every $\eta_\tau = 0$, all true values, true losses, and deployment value gaps vanish, consistently with every displayed exact-target formula. This completes the proposition. □

Proposition 78 (Homogeneous-task rate specialization bridge)**.** *Under Assumptions 44, 45, 46, 47, and 48, suppose that for every $\tau \in \mathcal{T}$,*

$$(\mu_\tau, n_\tau, \eta_\tau, C_\tau) = (\mu, n, \eta, C), \tag{94}$$

where $n \geq 1$, $\eta \geq 0$, and $C \geq 1$. For every $\delta \in (0,1)$ define

$$\epsilon := B\sqrt{\frac{\log(2N_{\mathrm{cls}}/\delta)}{2n}}, \qquad e := \epsilon + 2\eta. \tag{95}$$

For every fixed target realization satisfying $\mathcal{E}_Q$, one event has evaluation probability at least $1-\delta$ and, simultaneously for every τ,

$$\sup_{\phi \in \Phi_M,\, g \in \mathcal{G}_\phi} \Big| P_{n_\tau} \widehat{\ell}_{\tau,\phi,g} - L_\tau(\phi, g) \Big| \leq e, \tag{96}$$

$$R(\widehat{\phi}) + \lambda D(\widehat{\phi}) \leq \inf_{\phi \in \Phi_M} \{R(\phi) + \lambda D(\phi)\} + 2\lambda e, \tag{97}$$

$$L_\tau(\widehat{\phi}, \widehat{g}_\tau) \leq D_\tau(\widehat{\phi}) + 2e, \tag{98}$$

and

$$\begin{aligned} V_\tau^\star - V_\tau^{\widehat{\pi}_\tau} &= \frac{1}{1-\gamma} \mathbb{E}_{d_\tau^{\widehat{\pi}_\tau}} \ell_{\tau,\widehat{\phi},\widehat{g}_\tau} \\ &\leq \frac{C}{1-\gamma} L_\tau(\widehat{\phi}, \widehat{g}_\tau) \\ &\leq \frac{C}{1-\gamma} \big(D_\tau(\widehat{\phi}) + 2e\big). \end{aligned} \tag{99}$$

Consequently,

$$\max_{\tau \in \mathcal{T}} \big(V_\tau^\star - V_\tau^{\widehat{\pi}_\tau}\big) \leq \frac{C}{1-\gamma} \big(D(\widehat{\phi}) + 2e\big). \tag{100}$$

If $\eta = 0$, then $e = \epsilon$, estimated and true losses agree on every task scoring support, and every target-estimation term vanishes exactly.

Proof. Condition equation 94 and the governing assumptions verify all sample-size, target-error, and coverage conditions of Proposition 77. Direct substitution in the technical radii gives

$$\epsilon_\tau = \epsilon, \qquad e_\tau = e, \qquad e_{\max} = e.$$

The exact class count and certificate event remain unchanged, and the same joint evaluation event is reused. Substituting these equalities into the uniform-deviation, oracle, and selected-head conclusions gives equation 96–equation 98. The full deployment chain from Proposition 77, with $C_\tau = C$, gives equation 99. Finally,

$$\max_\tau \big(D_\tau(\widehat{\phi}) + 2e\big) = D(\widehat{\phi}) + 2e$$

gives equation 100. Every simplification is an equality. When $\eta = 0$, Proposition 57 gives exact target/loss equality and $e = \epsilon$. □

Expanded radii. For each task, the radius expands as

$$2e_\tau(\delta) = 2B\sqrt{\frac{\log(2N_{\mathrm{cls}}/\delta)}{2n_\tau}} + 4\eta_\tau, \tag{101}$$

and

$$\begin{aligned}\frac{C_\tau}{1-\gamma}\big(D_\tau(\widehat{\phi}) + 2e_\tau(\delta)\big) = \frac{C_\tau}{1-\gamma}D_\tau(\widehat{\phi}) \\ + \frac{2C_\tau R_{\max}}{(1-\gamma)^2}\sqrt{\frac{\log(2N_{\mathrm{cls}}/\delta)}{2n_\tau}} \\ + \frac{4C_\tau\eta_\tau}{1-\gamma}.\end{aligned} \tag{102}$$

C.5.10 The minimal-assumption core implication

Proposition 79 (Minimal-assumption core factorization)**.** *Under Assumptions 44 and 49, for every task $\tau \in \mathcal{T}$, every history $h \in \mathcal{R}_\tau$, and every action $a \in \mathcal{A}$,*

$$Q^\star_\tau(h, a) = q^\star_\tau(x(h), a). \tag{103}$$

Consequently, for every $h \in \mathcal{H}_\tau$ and $a \in \mathcal{A}$,

$$Q^\star_\tau(h, a) = q^\star_\tau(x(h), a). \tag{104}$$

Both statements use the chosen task-reachable version of $Q^\star_\tau$.

Proof. Fix an arbitrary task $\tau \in \mathcal{T}$. Proposition 66 uses only Assumptions 44 and 49; it constructs the unique bounded measurable core fixed point and gives equation 103 for every $h \in \mathcal{R}_\tau$ and $a \in \mathcal{A}$. Its full-history object is exactly the task-reachable version of $Q^\star_\tau$, and its constructed core object is $q^\star_\tau$ evaluated at $x(h) \in \mathcal{X}^{\mathrm{reach}}_\tau$. Both sides therefore refer to the same task, action, horizon, and version domain, leaving the exact residual

$$|Q^\star_\tau(h, a) - q^\star_\tau(x(h), a)| = 0. \tag{105}$$

Because $\mathcal{H}_\tau \subseteq \mathcal{R}_\tau$, restricting equation 103 to $h \in \mathcal{H}_\tau$ gives equation 104. Since τ was arbitrary, the implication holds separately on every task's own domain. □

C.5.11 The exact reachable quotient and minimum rate

Proposition 80 (Reachable quotient achievability and active-code converse)**.** *Under Assumptions 44, 45, 49, 50, and 51, and Propositions 66, 68, 69, 73, and 76, the following conclusions hold.*

First, the core-only encoder satisfies

$$D(\phi_{\mathrm{dec}}) = 0, \qquad R(\phi_{\mathrm{dec}}) = \log m^{\mathrm{reach}}_{\mathrm{dec}}. \tag{106}$$

Second, for every $\phi \in \Phi_M$ satisfying $D(\phi) = 0$, one can choose heads $g_\tau \in \mathcal{G}_\phi$ with $L_\tau(\phi, g_\tau) = 0$ for every task and define $u_z = (g_\tau(z))_{\tau \in \mathcal{T}}$ so that, for $\bar{\mu}_X$-almost every x,

$$z \in \mathcal{Z}_\phi(x) \quad \Longrightarrow \quad u_z \in \mathcal{U}_{\mathrm{reach}}(x) \qquad (z \in \mathcal{Z}_\phi). \tag{107}$$

For those same heads and profiles,

$$\mathcal{C}_\phi := \{u_z : z \in \mathcal{Z}_\phi^{\bar{\mu}}\} \tag{108}$$

is a valid $\mathcal{U}_{\mathrm{reach}}$-cover and obeys

$$m_{\mathrm{dec}}^{\mathrm{reach}} \le |\mathcal{C}_\phi| \le |\mathcal{Z}_\phi^{\bar{\mu}}| \le |\mathcal{Z}_\phi|. \tag{109}$$

Proof. Proposition 66 fixes the common structural convention before the two sides are assembled. It constructs $q_\tau^\star$ as the unique bounded measurable core fixed point, making the conditional core gaps used by the converse measurable, and proves that every such core action value equals the corresponding full-history action value on the same task's reachable domain.

Under the realizability premise, Proposition 68 gives the first equality in equation 106, and Proposition 69 gives the second. The latter uses the full-range rate and the assumed exact count

$$|\mathcal{Z}_{\phi_{\mathrm{dec}}}| = |\mathcal{Z}_{\phi_{\mathrm{dec}}}^{\bar{\mu}}| = m_{\mathrm{dec}}^{\mathrm{reach}}. \tag{110}$$

Achievability is thus established independently of the converse and of support equivalence.

Now fix an arbitrary $\phi \in \Phi_M$ under the local converse antecedent $D(\phi) = 0$. Proposition 73 uses the measurable fixed point of Proposition 66, chooses the zero-risk heads, defines the profiles, and proves equation 107. Its conclusion checks only coordinates $\tau \in \mathcal{T}(x)$; coordinates outside $\mathcal{T}(x)$ remain unconstrained exactly as required by $\mathcal{U}_{\mathrm{reach}}(x)$.

Proposition 76 uses exactly those heads and profiles and the antecedent $D(\phi) = 0$. It first removes the finite union of local-only null activity sets, then proves that the image in equation 108 meets $\mathcal{U}_{\mathrm{reach}}(x)$ for almost every x, and finally gives equation 109. The intermediate image count is retained because distinct codes may induce the same action profile; no injectivity of $z \mapsto u_z$ is assumed. This completes the independent converse and the proposition.

□

Proposition 81 (Minimum zero-deficiency representation rate)**.** *Under Assumptions 44, 45, 49, 50, and 51, and Proposition 80, the zero-deficiency set is nonempty and*

$$\min_{\phi \in \Phi_M :\, D(\phi) = 0} R(\phi) = \log m_{\mathrm{dec}}^{\mathrm{reach}}. \tag{111}$$

Here log *is the natural logarithm, so equation 111 is a rate in nats. The minimum is attained by ϕ_{dec}, which can therefore be chosen core-only and nuisance-free.*

Proof. Equation equation 106 and the assumed membership $\phi_{\mathrm{dec}} \in \Phi_M$ show that the constraint set in equation 111 is nonempty and that its minimum is at most $\log m_{\mathrm{dec}}^{\mathrm{reach}}$. Moreover, equation 110 and the fact that $\bar{\mu}$ is a probability law show $|\mathcal{Z}_{\phi_{\mathrm{dec}}}^{\bar{\mu}}| \ge 1$, hence

$$1 \le m_{\mathrm{dec}}^{\mathrm{reach}} \le M. \tag{112}$$

Thus every logarithm below is applied to a positive integer.

Conversely, take any $\phi \in \Phi_M$ with $D(\phi) = 0$. Equation equation 109 gives

$$m_{\mathrm{dec}}^{\mathrm{reach}} \le |\mathcal{Z}_\phi^{\bar{\mu}}| \le |\mathcal{Z}_\phi|. \tag{113}$$

Since the natural logarithm is increasing on $(0, \infty)$, the full-range rate therefore satisfies

$$R(\phi) = \log |\mathcal{Z}_\phi| \ge \log m_{\mathrm{dec}}^{\mathrm{reach}}. \tag{114}$$

This lower bound holds for the full-range rate. Combining equation 114, which holds for every feasible encoder, with the feasible equality candidate equation 106 proves equation 111 and shows that ϕ_{dec} attains it.

Finally, Assumption 51 defines $\phi_{\mathrm{dec}}(h) = c_{\mathrm{dec}}(x(h))$. It has no nuisance input once the observable core is fixed, so this particular minimizer is core-only and nuisance-free. When $m_{\mathrm{dec}}^{\mathrm{reach}} = 1$, equations equation 111–equation 114 give the exact boundary value $R = \log 1 = 0$. □

C.5.12 ASSEMBLY OF THE THREE CLAUSES

Proposition 82 (Three-part theorem with clause-local premises)**.** *Throughout,* $\log$ *denotes the natural logarithm, so* $R(\phi) = \log|\mathcal{Z}_\phi|$ *is measured in nats. The following three conclusions hold, each under its own stated premises.*

1. Under Assumptions 44, 45, 46, and 47, for every $\delta \in (0,1)$*, define*

$$\epsilon_\tau(\delta) := B\sqrt{\frac{\log(2N_{\mathrm{cls}}/\delta)}{2n_\tau}}, \qquad e_\tau(\delta) := \epsilon_\tau(\delta) + 2\eta_\tau, \qquad e_{\max}(\delta) := \max_{\tau\in\mathcal{T}} e_\tau(\delta). \tag{115}$$

Conditional on any fixed independently obtained target realization satisfying $\mathcal{E}_Q$*, the event* $\mathcal{E}_{\mathrm{cal}}(\delta)$ *under the evaluation marginal has probability at least* $1-\delta$ *and simultaneously satisfies, for every task,*

$$\sup_{\phi\in\Phi_M,\, g\in\mathcal{G}_\phi} \left| P_{n_\tau}\widehat{\ell}_{\tau,\phi,g} - L_\tau(\phi,g)\right| \le e_\tau(\delta), \tag{116}$$

$$R(\widehat{\phi}) + \lambda D(\widehat{\phi}) \le \inf_{\phi\in\Phi_M}\{R(\phi) + \lambda D(\phi)\} + 2\lambda e_{\max}(\delta), \tag{117}$$

and

$$L_\tau(\widehat{\phi},\widehat{g}_\tau) \le D_\tau(\widehat{\phi}) + 2e_\tau(\delta). \tag{118}$$

If, in addition, Assumption 48 holds, the same event also satisfies

$$\begin{aligned} V_\tau^\star - V_\tau^{\widehat{\pi}_\tau} &= \frac{1}{1-\gamma}\mathbb{E}_{H\sim d_\tau^{\widehat{\pi}_\tau}}\ell_{\tau,\widehat{\phi},\widehat{g}_\tau}(H) \\ &\le \frac{C_\tau}{1-\gamma}L_\tau(\widehat{\phi},\widehat{g}_\tau) \\ &\le \frac{C_\tau}{1-\gamma}\big(D_\tau(\widehat{\phi}) + 2e_\tau(\delta)\big). \end{aligned} \tag{119}$$

The exact class count and its all-maps envelope are

$$N_{\mathrm{cls}} = \sum_{\tau\in\mathcal{T}}\sum_{\phi\in\Phi_M}|\mathcal{G}_\phi| \le |\mathcal{T}|\,|\Phi_M|\,|\mathcal{A}|^M, \tag{120}$$

while equation 115 retains the sharper exact N_{cls}*. In particular, without replacing that primary radius,*

$$\epsilon_\tau(\delta) \le B\sqrt{\frac{\log\big(2|\mathcal{T}|\,|\Phi_M|\,|\mathcal{A}|^M/\delta\big)}{2n_\tau}}. \tag{121}$$

When $B = 0$*, the true-loss event is deterministic and* $\epsilon_\tau = 0$*.*

2. Under Assumptions 44 and 49 alone, deterministically, $q_\tau^\star$ *is the unique bounded measurable fixed point constructed in Proposition 66, and*

$$Q_\tau^\star(h,a) = q_\tau^\star(x(h),a) \quad \text{for every } \tau\in\mathcal{T},\ h\in\mathcal{R}_\tau,\ a\in\mathcal{A}, \tag{122}$$

and therefore for every $h\in\mathcal{H}_\tau$*. Equation equation 122 is stated for each task's chosen reachable version on* $\mathcal{R}_\tau$*.*

3. Under Assumptions 44, 45, 49, 50, and 51,

$$D(\phi_{\mathrm{dec}}) = 0, \qquad R(\phi_{\mathrm{dec}}) = \log m_{\mathrm{dec}}^{\mathrm{reach}}. \tag{123}$$

For every $\phi\in\Phi_M$ *under the retained converse antecedent* $D(\phi)=0$*, one can choose heads* $g_\tau\in\mathcal{G}_\phi$ *with* $L_\tau(\phi,g_\tau)=0$*, define* $u_z = (g_\tau(z))_{\tau\in\mathcal{T}}$*, and obtain, for* $\bar{\mu}_X$*-almost every* x *and every* $z\in\mathcal{Z}_\phi$*,*

$$z\in\mathcal{Z}_\phi(x) \Longrightarrow u_z\in\mathcal{U}_{\mathrm{reach}}(x). \tag{124}$$

For those same heads and profiles,

$$\mathcal{C}_\phi := \{u_z : z\in\mathcal{Z}_\phi^{\bar{\mu}}\} \tag{125}$$

is a valid reachable-profile cover and

$$m_{\mathrm{dec}}^{\mathrm{reach}} \le |\mathcal{C}_\phi| \le |\mathcal{Z}_\phi^{\bar{\mu}}| \le |\mathcal{Z}_\phi|. \tag{126}$$

Consequently,

$$\min_{\phi\in\Phi_M:\, D(\phi)=0} R(\phi) = \log m_{\mathrm{dec}}^{\mathrm{reach}} \tag{127}$$

nats, and the minimum is attained by the core-only encoder $\phi_{\mathrm{dec}}(h) = c_{\mathrm{dec}}(x(h))$. Only active task coordinates are constrained in equation 124, and equation 127 exhibits a core-only minimizer.

The conjunction preserves the following exact baselines. If every $\eta_\tau = 0$, then on $\mathcal{E}_Q$, $\widehat{\ell}_{\tau,\phi,g} = \ell_{\tau,\phi,g}$ on $\mathcal{H}_\tau$, and $e_\tau(\delta) = \epsilon_\tau(\delta)$; moreover, $\mathcal{E}_{\mathrm{cal}}(\delta)$ and $\mathcal{E}_{\mathrm{true}}(\delta)$ agree almost surely under the evaluation marginal. The sampling radius remains. If

$$(\mu_\tau, n_\tau, \eta_\tau, C_\tau) = (\mu, n, \eta, C) \quad \textit{for every } \tau, \tag{128}$$

then, with

$$\epsilon := B\sqrt{\frac{\log(2N_{\mathrm{cls}}/\delta)}{2n}}, \qquad e := \epsilon + 2\eta, \tag{129}$$

one has $e_\tau = e_{\max} = e$, the oracle and selected-head remainders are exactly $2\lambda e$ and $2e$, and, when the common coverage premise is imposed,

$$V_\tau^\star - V_\tau^{\widehat{\pi}_\tau} \le \frac{C}{1-\gamma}\big(D_\tau(\widehat{\phi}) + 2e\big) \quad \textit{for every } \tau,$$

and therefore

$$\max_{\tau\in\mathcal{T}}\big(V_\tau^\star - V_\tau^{\widehat{\pi}_\tau}\big) \le \frac{C}{1-\gamma}\big(D(\widehat{\phi}) + 2e\big). \tag{130}$$

Finally, the full core-plus-support-plus-capacity specialization retains the exact conclusions equation 123, equation 126, and equation 127, with zero remainder. Coverage enters only equation 119, and support equivalence and quotient realizability enter only the third implication.

Proof. Fix $\delta \in (0,1)$ and a target-estimation realization satisfying $\mathcal{E}_Q$. Proposition 77 supplies one evaluation event of probability at least $1-\delta$ and proves equation 115–equation 120, including equation 119 only under the additional coverage premise. It also supplies the exact all-$\eta_\tau = 0$ and homogeneous substitutions equation 128–equation 130.

Separately, under exactly Assumptions 44 and 49, Proposition 79 uses the fixed-point construction and proves equation 122 pointwise on each $\mathcal{R}_\tau$, hence on $\mathcal{H}_\tau$. This deterministic conclusion holds independently of the statistical event.

Finally, under exactly the five premises displayed in the third implication, Proposition 80 proves equation 123–equation 126, retaining $D(\phi) = 0$ as the antecedent for each arbitrary converse encoder, the $\bar{\mu}_X$-almost-everywhere scope, the universal quantifier over $z \in \mathcal{Z}_\phi$, the active-task coordinate convention, and the distinction between mixture-active and full-range codes. Proposition 81 applies the full-range natural-log rate to equation 126, proves equation 127, and exhibits the core-only equality candidate. This structural implication is deterministic.

Each of the three implications has thus been proved by a preceding result under its own antecedents, and the baseline statements are equality substitutions into those results. This proves the proposition. □

C.5.13 Proof of the Main Theorem

Proof of Theorem 52. Part (i) is Proposition 77, including its coverage-conditional deployment chain. Part (ii) is Proposition 79, backed by the constructed unique bounded measurable fixed point. Part (iii) follows from Propositions 80 and 81. Proposition 82 states the three implications together, each with its own assumptions. □