Latent stationary overlap and minimax off-policy evaluation in finite POMDPs

CausalSmith*

September 13, 2026

Abstract

This paper characterizes the observable-data minimax risk for stationary off-policy evaluation in finite partially observed Markov decision processes. The data are one behavior-policy trajectory of length T, the trajectory horizon, with bounded rewards and binary actions; policies are known functions of the observed state; the joint observed–latent Markov state satisfies stationary start, sequential ignorability, one-step policy overlap $L = \exp(\zeta)$, where ζ is the log-overlap scale, geometric total-variation contraction $\alpha = \exp(-1/t_0)$, where t_0 is the mixing scale, and latent stationary overlap $d_e(s) \le C d_b(s)$, where C is the stationary-overlap constant and d_e and d_b are the target and behavior stationary joint-state laws. For sufficiently large T, and cardinality-uniformly over finite observed and latent alphabets, the minimax squared-error risk is comparable, uniformly over $C \ge 1$, to

$$T^{-1} + T^{-\beta} q_C^{2(1-\beta)}, \qquad q_C = \frac{C-1}{C}, \qquad \beta = \frac{2}{2 + t_0 \zeta},$$

where q_C is the overlap-distance radius, the normalized distance of the stationary density-ratio bound from unit overlap, and β is the rate exponent set by the mixing and log-overlap scales, up to constants depending only on t_0 and ζ. For every fixed $C > 1$, the surface has the Hu–Wager partial-history exponent $T^{-\beta}$; the parametric term changes the rate on shrinking overlap-distance scales $q_C \lesssim T^{-1/2}$. A clipped partial-history importance weighting (PHIW) estimator with radius-calibrated history depth attains this surface when the depth is calibrated from (t_0, ζ, C). A signed-depth finite POMDP pair supplies matching hard alternatives while satisfying the same overlap and contraction conditions. A finite controlled insulin-policy demonstration shows that bounded latent stationary occupancy and bounded action overlap can hold together in a stylized treatment model.

1 Introduction

Off-policy evaluation asks how well the value of a target policy can be learned from data generated by a different behavior policy. In longitudinal causal inference this question is tied to the g-formula, propensity-score balancing, marginal structural models, and dynamic treatment regimes [Robins, 1986, Rosenbaum and Rubin, 1983, Robins et al., 2000, Murphy, 2003]. In

*Machine-generated by the CausalSmith research pipeline. Each displayed formal statement is either mapped to a Lean declaration or marked as a presentation-level definition, and each displayed proof renders a Lean proof, as recorded in the verification appendix (scope, toolchain, commit, and any statement conditional on a published input). Cited work otherwise supplies framing and positioning.

reinforcement learning and stochastic control it is the policy-evaluation problem for Markov decision processes and partially observed Markov decision processes [Smallwood and Sondik, 1973, Kaelbling et al., 1998, Sutton and Barto, 2018]. This paper studies the stationary mean reward of a known target policy from a single stationary behavior trajectory when the Markov state has an observed coordinate and a latent coordinate.

The fully observed benchmark is by now well developed. Importance weighting, doubly robust methods, and stationary density-ratio methods use the observed Markov state to replace full trajectory likelihood ratios by more stable value-function or occupancy-ratio objects [Precup et al., 2000, Jiang and Li, 2016, Thomas and Brunskill, 2016, Liu et al., 2018, Nachum et al., 2019, Uehara et al., 2020, Kallus and Uehara, 2020]. Econometric work on adaptive and longitudinal treatment settings gives related average-reward and policy-learning theory [Liao et al., 2021, 2022]. These analyses identify the role of overlap in observed states and actions when the state available to the analyst is rich enough to be Markov.

Partial observation changes the statistical object. Hu and Wager [2023] show that, in finite-state POMDPs with known policies, bounded rewards, one-step policy overlap, and geometric mixing, partial-history importance weighting can evaluate a target policy at a polynomial mean-squared-error rate governed by the mixing and overlap constants. The estimator uses a recent observable history: a longer history reduces latent-state bias, while multiplying more policy ratios increases variance. The corresponding minimax lower bound establishes the same partial-history exponent within the stated bounded-reward, binary-action finite-state model class.

This paper's main message is that bounding the latent stationary density ratio at any fixed radius $C > 1$ preserves the partially observed Hu–Wager exponent $T^{-\beta}$. The overlap-distance radius $q_C = (C-1)/C$ changes the rate in the local layer where q_C shrinks on the $T^{-1/2}$ scale. The model class $\mathcal{M}_T(t_0, \zeta, C)$ in Definition 10 consists of finite observed and latent alphabets, known memoryless policies on $\{0,1\}$, a stationary behavior start, bounded rewards, sequential ignorability, one-step policy overlap $L = \exp(\zeta)$ with log-overlap scale ζ, uniform total-variation contraction $\alpha = \exp(-1/t_0)$ with mixing scale t_0, and latent stationary overlap $d_e(s) \le C d_b(s)$ for the target and behavior stationary joint-state laws. The estimand $\theta(K, e)$ in Definition 12 is the stationary target-policy mean reward over the joint observed–latent state, and $\mathbb{E}_M$ denotes expectation under the observed law induced by model M. The minimax comparison ranges over finite alphabets with constants depending only on fixed (t_0, ζ) after the threshold in Theorem 1; estimators observe the observed trajectory, its observed alphabet size, and the policy pair.

Assumption interpretation Latent stationary overlap is plausible when changing from the behavior policy to the target policy changes the long-run frequency of each joint observed–latent regime by a bounded factor. An analyst would justify it through design knowledge, mechanistic modeling, sensitivity analysis, or external evidence about how the target policy shifts latent regimes. Because the bound is imposed on the joint state, it is stronger than overlap for the observed state alone and carries information about unobserved occupancy. Sequential ignorability and known memoryless binary policies match randomized or logged-policy settings in which treatment probabilities are functions of the observed state. Uniform total-variation contraction is a Doeblin-type mixing condition for both policies; it supplies the stationary laws, the bias decay in the PHIW upper bound, and the signed-depth membership argument. Bounded rewards keep squared-error risk on a common scale. Cardinality-uniform finite alphabets mean that the risk constants are stable as the observed and latent alphabets vary.

Limitations Latent stationary overlap cannot be checked from the observed trajectory alone. Uniform Doeblin-type contraction can fail in slowly mixing clinical dynamics, in which case the PHIW bias term and the signed-depth lower-bound calibration require model-specific replacement. Non-stationary starts require burn-in or explicit initial-distribution terms beyond the stationary-start statements used here. The lower-bound alternatives use latent alphabet size of order $\log T$, so a sharp frontier for a fixed hidden alphabet remains an open question. The radius-calibrated depth attains the displayed frontier using (t_0, ζ, C), equivalently q_C, and data-driven selection for an unknown overlap radius, including the possible cost of logarithmic adaptation, remains open. The local path $C_T = 1 + \delta_T$ is a mathematical local-asymptotic slice of the uniform surface.

The main result is the all-radius minimax surface in Theorem 1. For fixed $t_0 > 0$ and $\zeta > 0$, define the exponent $\beta = 2/(2 + t_0\zeta)$ and the overlap-distance radius $q_C = (C-1)/C$. Uniformly over every $C \geq 1$, the cardinality-uniform observable-data minimax squared-error risk is comparable to

$$\mathfrak{r}_T(t_0, \zeta, C) = T^{-1} + T^{-\beta} q_C^{2(1-\beta)},$$

with constants and a large-sample threshold depending only on t_0 and ζ. The radius-calibrated partial-history estimator in Definition 20 and Definition 16 attains the upper bound; its depth is computed from (t_0, ζ, C), so the overlap radius C is a tuning input supplied to the estimator rather than a quantity it learns from the data. The lower bound combines the radius-explicit signed-depth construction with a uniform parametric floor.

The formula has three immediate interpretations. At the unit-overlap boundary $C = 1$, the radius is zero, the target and behavior stationary laws coincide, and Theorem 3 gives the T^{-1} minimax scale. For every fixed $C > 1$, the radius is constant and Theorem 2 gives the partially observed rate $T^{-\beta}$. Along local paths $C_T = 1 + \delta_T$ with $\delta_T \to 0$, Theorem 4 gives the transition

$$T^{-1} + T^{-\beta} \delta_T^{2(1-\beta)}$$

up to constants, with an elbow at the scale $T\delta_T^2 \asymp 1$. Below that scale, immediate weighting reaches the parametric order; above it, the radius-calibrated PHIW depth used in the upper bound grows logarithmically with the effective radius.

The lower-bound alternatives are finite signed-depth POMDPs. Their hidden state records a depth coordinate and a sign coordinate; the target policy forces the rare action, while the behavior policy takes that action with probability $1/L$. The reward signal appears at terminal depth, and the stationary target value changes sign across the two alternatives. Lemma 18 places these hard alternatives in the latent-overlap model class, with amplitude ε_C comparable to the radius q_C through $q_C/2 \leq \varepsilon_C \leq q_C$. The observable divergence bounds in Lemmas 31 and 35 pass through the signed-depth factor $\alpha^{2Q} L^{-Q}$, and Lemma 40 supplies the parametric floor. Together these statements match the bias–variance difficulty faced by PHIW at the radius-calibrated depth.

The paper also includes a finite controlled insulin-policy demonstration. The construction in Definition 24 specifies glucose, diet, activity, and lagged treatment coordinates in a 360-state controlled Markov model designed to make the assumptions checkable by finite arithmetic. The target policy treats above a glucose threshold and the behavior policy randomizes treatment. Theorem 5 verifies finite state count, one-half contraction, bounded one-step policy overlap, finite latent stationary overlap with the sufficient bound $C = 720$, and a strictly negative interval arithmetic bound for the immediate-weighting diagnostic. This value of C gives $q_C = 719/720 \approx$

0.9986, so the depth prescribed at this radius by Definition 20 is

$$\kappa_T^{\mathrm{rad}}(720) = \min\left\{\left\lfloor \frac{T}{2} \right\rfloor, \left\lfloor \frac{\log\{T(719/720)^2\}}{2\log(1/\alpha) + \log L} \right\rfloor\right\}$$

once $T(719/720)^2 > 1$, and zero on the complementary branch. Because q_C is within 0.15% of one, this depth differs from the fixed-radius balanced depth k_T of Definition 17 only by a bounded rounding adjustment at large T. Since the reward is a function of observed glucose, the diagnostic compares behavior-stationary and target-stationary reward averages in this stylized finite treatment model.

The appendix closes with notes on the formal results, presentation-level definitions, and cited external inputs.

The paper proceeds as follows. The related-work section places the result among dynamic treatment regimes, off-policy evaluation, and partially observed models. The setup section defines the finite POMDP class, the stationary target value, minimax risk, the partial-history estimator, and the signed-depth family. The main-results section states the uniform all-radius theorem first and then its fixed-overlap, unit-overlap, and shrinking-radius regimes. The insulin-policy section gives the finite treatment-policy adaptation and its interval-certified diagnostic. The discussion interprets the risk surface and records limitations. The appendix collects the auxiliary constructions, covariance bounds, signed-depth likelihood calculations, radius-sensitive risk bounds, and parametric lower-bound ingredients.

2 Related work

The paper sits at the intersection of dynamic treatment regimes, off-policy evaluation, and partially observed Markov decision processes. Sequential causal estimands trace back to the potential-outcome and g-formula traditions of Robins [1986], Rosenbaum and Rubin [1983], and Robins et al. [2000], with dynamic regimes developed in econometrics, biostatistics, and reinforcement learning by work such as Murphy [2003]. In that tradition, the central object is a target-policy value identified from data generated by a different behavior policy under sequential ignorability and overlap. The present analysis keeps that causal interpretation, while specializing to finite-state partially observed Markov decision processes with a stationary target value and a latent stationary overlap condition.

A second line of work studies off-policy evaluation in Markov decision processes. Classical dynamic-programming and POMDP foundations appear in Smallwood and Sondik [1973], Kaelbling et al. [1998], and Sutton and Barto [2018]. For fully observed MDPs, importance weighting, density-ratio estimation, and doubly robust methods provide estimators whose behavior depends on policy ratios, state-occupancy ratios, and mixing properties [Precup et al., 2000, Jiang and Li, 2016, Thomas and Brunskill, 2016, Liu et al., 2018, Nachum et al., 2019, Uehara et al., 2020, Kallus and Uehara, 2020]. Recent econometric work extends these ideas to adaptive and longitudinal settings, including Liao et al. [2021] and Liao et al. [2022]. These methods supply the fully observed benchmark: when the state observed by the analyst is the Markov state, occupancy-ratio control can stabilize long-horizon evaluation. The minimax comparisons in Theorems 1, 2 and 4 instead quantify the cost of evaluating a stationary target policy when the overlap restriction is imposed on the latent stationary law.

The closest partially observed benchmark is Hu and Wager [2023, Model 3 and Theorems 2.1, 3.1]. Their finite-state POMDP analysis uses sequential ignorability, known policies, bounded

rewards, one-step policy overlap, and geometric mixing to study partial-history estimators and the minimax partial-history rate. The new ingredient here is the latent stationary overlap radius $d_e/d_b \le C$ on the joint hidden–observed state. For each fixed $C > 1$, Theorem 2 keeps the Hu–Wager exponent $T^{-\beta}$; the fixed-overlap upper bound uses the balanced PHIW depth, and the lower bound uses signed-depth alternatives satisfying the same latent-overlap radius. A separate construction is needed for the converse, and the reason is a short calculation. In the chain built in the proof of Hu and Wager [2023, Theorem 3.1], the hidden coordinate is a depth ladder $h_1, \dots, h_Q$; one action advances the chain one rung with probability $1 - \delta$, the other resets it to h_1, the target policy always advances, and the behavior policy advances with probability $e^{-\zeta_\pi}$, where ζ_π is their one-step log-overlap scale. We record the consequence as a short calculation of our own, read off their construction rather than quoted from a numbered result: reaching the terminal rung requires $Q - 1$ consecutive advances, so the two stationary masses at that rung are

$$d^{\pi}(h_Q) = (1-\delta)^{Q-1}, \qquad d^{\pi_0}(h_Q) = (1-\delta)^{Q-1} e^{-(Q-1)\zeta_\pi},$$

and their ratio is $e^{(Q-1)\zeta_\pi}$, exponential in the depth. Because their rate-calibrated choice takes $Q \asymp \log T$, that ratio eventually exceeds any fixed C: their hard instances leave $\mathcal{M}_T(t_0, \zeta, C)$ for every fixed radius, so the fixed-radius converse here is obtained from the signed-depth family of Definition 18 instead, whose reset step is perturbed by ε_C precisely so that the stationary ratio stays inside the radius.

A closely related fully observed comparison is Mehrabi and Wager [2025]. Their state is the Markov state, and they weaken the usual requirement that the target-to-behavior stationary density ratio be uniformly bounded to a tail condition on that ratio: Mehrabi and Wager [2025, Definition 1] asks only that the $(1 + \delta)$-th power of the stationary ratio have a bounded tail under the behavior stationary law. In the square-integrable regime $\delta > 1$ their truncated doubly robust estimator is $\sqrt{T}$-consistent and asymptotically normal with the semiparametric efficiency variance [Mehrabi and Wager, 2025, Theorem 3.4]. Four differences keep that result and the present one from being two readings of one condition. Their overlap norm is a tail bound on a ratio of *observed*-state stationary laws, whereas Assumption 7 is a pointwise bound $d_e \le C d_b$ on a ratio of *joint observed–latent* stationary laws. Their estimator is given the Markov state, whereas the estimators admitted by Definition 14 never see the latent coordinate. Their state space is general and the bounded-ratio assumption they relax is credible mainly on bounded state spaces, whereas the class here is uniform over all finite observed and latent cardinalities. Finally, their inferential target is a limit distribution for one estimator, whereas the target here is a worst-case squared-error rate over the whole class. Theorem 1 shows the resulting contrast: fixed latent stationary overlap with any fixed $C > 1$ retains the partial-history rate $T^{-\beta}$, while the parametric term affects the frontier on shrinking scales $q_C \lesssim T^{-1/2}$. Theorems 3 and 4 record the endpoint and local slices of the same uniform surface.

A related POMDP literature recovers target values through proxy, bridge, coverage, or observability structure. Proxy and bridge approaches such as Tennenholtz et al. [2020], Bennett and Kallus [2021], Miao et al. [2022], Shi et al. [2022], Uehara et al. [2023], and Kuang et al. [2025] impose relationships that make observed variables informative about latent states or bridge functions. Zhang and Jiang [2024] give coverage conditions for future-dependent value functions, and Zhang and Jiang [2025] separates tractable and hard regimes for history-dependent policies. The present minimax analysis works in the finite POMDP experiment class with sequential ignorability, bounded rewards, contraction, one-step policy overlap, and latent stationary overlap, and quantifies the partial-history evaluation rate implied by those conditions.

The insulin-policy example connects the abstract finite-state model to dynamic treatment settings where the observed record is a coarse clinical history and the physiological state contains latent components. Prior work on diabetes management and treatment-policy learning, including Maahs et al. [2012], Luckett et al. [2020], and related reinforcement-learning formulations such as Nahum-Shani et al. [2018], motivates finite controlled models with clinically interpretable state variables and actions. The adaptation in Theorem 5 provides a finite controlled Markov model on which four of the paper's structural constants are certified by finite arithmetic — state count, contraction rate, one-step policy overlap, and latent stationary overlap — together with a certified nonzero immediate-weighting bias, linking the mechanism behind the theory to a stylized treatment-policy evaluation problem.

3 Setup and estimands

We observe a single trajectory from a finite partially observed Markov decision process. Throughout, T denotes the trajectory horizon, $\|\cdot\|_{\mathrm{TV}}$ is total variation distance, and all index ranges are the finite ranges displayed in the definitions below. Generic constants may depend on the primitive radius and mixing parameters explicitly named in a result, while the minimax statements are uniform over finite observed and latent cardinalities. Asymptotic notation is used only for the horizon regime governed by the fixed parameters in the displayed model classes.

The finite alphabets are the observed state space $\mathcal{X}$, the latent state space $\mathcal{H}$, and the binary action set $\mathcal{A}$. The joint state S_t has observed coordinate X_t and latent coordinate H_t; the action and reward are A_t and Y_t. The behavior policy b generates the data, the target policy e defines the counterfactual policy value, and the one-step law K governs rewards and successor states.

Definition 1. [`synth_8`] (Policy-overlap factor) *For a real parameter ζ, define the policy-overlap constant L by*

$$L(\zeta) = \exp(\zeta).$$

The positive log-overlap scale ζ parameterizes the policy-overlap constant L. Larger values allow the target policy to place relatively more mass on actions that are less common under the behavior policy.

Definition 2. [`synth_3`] (Full finite trajectory) *For natural numbers T, n_X, n_H, a full trajectory of length T is a pair*

$$\big((S_t)_{t\in\{0,\dots,T\}}, ((A_t, Y_t))_{t\in\{0,\dots,T-1\}}\big) \in \big(\{0,\dots,n_X-1\}\times\{0,\dots,n_H-1\}\big)^{\{0,\dots,T\}} \times \big(\{\text{false},\text{true}\}\times\mathbb{R}\big)^{\{0,\dots,T-1\}}.$$

Here $S_t = (X_t, H_t)$ is the observed–latent joint state at state time t, with observed coordinate $X_t \in \{0,\dots,n_X-1\}$ and latent coordinate $H_t \in \{0,\dots,n_H-1\}$. At transition time t, $A_t \in \{\text{false},\text{true}\}$ is the Boolean action and $Y_t \in \mathbb{R}$ is the reward; when actions are written numerically, false *is identified with* 0 *and* true *with* 1.

The full trajectory records both the observed and latent coordinates. The econometrician observes only the observed coordinate, action, and reward, but the structural restrictions are most compactly stated on the finite observed–latent state process.

The next group of conditions specifies the controlled Markov experiment and the way actions enter the observed-data law.

Assumption 1. [ass:pomdp-kernel] (Markov kernel) *For each* $t = 0, \dots, T-1$,

$$\mathcal{L}(Y_t, X_{t+1}, H_{t+1} \mid X_{0:t}, H_{0:t}, A_{0:t}, Y_{0:t-1}) = K(\cdot \mid X_t, H_t, A_t).$$

Assumption 1 is the time-homogeneous POMDP condition used in Model 3 of Hu and Wager [2023]. It makes the current joint state and action sufficient for the next reward and state transition, so the latent coordinate carries the unobserved predictive information.

Definition 3. [synth_1] (Memoryless binary policy carrier) *For* $n_X \in \mathbb{N}$, *a memoryless policy on an observed-state alphabet of size* n_X *is a function*

$$p : \{0, \dots, n_X - 1\} \times \{\text{false}, \text{true}\} \to \mathbb{R}.$$

When the action alphabet is written as $\mathcal{A} = \{0, 1\}$, *we identify* false *with* 0 *and* true *with* 1. *This is a carrier only: no nonnegativity and no normalization is imposed here. The probability-vector restrictions on the behavior and target carriers are imposed later, on* b *in Assumption 2 and on* e *in Assumption 5.*

Memoryless policies depend on the observed coordinate rather than on the latent state. This matches the information set of the policy comparison: both the behavior and target policies are revealed functions of observed state and action. The object just defined is only the carrier for such a function; that the behavior carrier b and the target carrier e are genuine action distributions is a restriction imposed by the assumptions below, on b in Assumption 2 and on e in Assumption 5.

Assumption 2. [ass:sequential-ignorability] (Observed-state sequential ignorability) *Identify the numeric action alphabet* $\mathcal{A} = \{0, 1\}$ *with the Boolean action set by* $0 \leftrightarrow$ false *and* $1 \leftrightarrow$ true. *The behavior carrier* b *is a probability vector at each observed state,*

$$b(a \mid x) \geq 0 \quad \text{for all } x \in \mathcal{X},\ a \in \mathcal{A}, \qquad \sum_{a \in \mathcal{A}} b(a \mid x) = 1 \quad \text{for all } x \in \mathcal{X}.$$

For every $t \in \{0, \dots, T-1\}$, *the joint law of the state-and-epoch history strictly before* t, *the current joint state* (X_t, H_t), *and the current action* A_t *factors as the law of that history and current joint state followed by the behavior kernel evaluated at the current observed state:*

$$\mathbb{P}\big\{\big((X_s, H_s)_{s<t}, (A_s, Y_s)_{s<t}, (X_t, H_t), A_t\big) \in dh\, da\big\} = \mathbb{P}\big\{\big((X_s, H_s)_{s<t}, (A_s, Y_s)_{s<t}, (X_t, H_t)\big) \in dh\big\}\ b(da \mid X_t(h)).$$

Equivalently, for every $a \in \mathcal{A}$ *and every* $t \in \{0, \dots, T-1\}$,

$$\Pr(A_t = a \mid X_{0:t}, H_{0:t}, A_{0:t-1}, Y_{0:t-1}) = b(a \mid X_t).$$

Assumption 2 is the observed-state sequential ignorability and randomization condition in Hu and Wager [2023]. Conditional on the current observed state, action assignment follows the behavior policy even when the latent coordinate affects future rewards and states.

Definition 4. [synth_4] (Policy transition operator) *For a raw experiment* M *with one-step kernel* K, *a memoryless policy* p, *and joint states* $s = (x, h)$ *and* s', *define the policy-induced transition operator* P_p *by*

$$P_p(s, s') := \sum_{a \in \{0,1\}} p(a \mid x)\, K\big(\{(r, \widetilde{s}) : \widetilde{s} = s'\} \mid s, a\big).$$

The policy-induced transition operator P_p is the reward-marginalized Markov kernel on joint states obtained after averaging actions under policy p. It is the transition law used to define stationarity and mixing under both the behavior and target policies.

Definition 5. [`synth_2`] (Joint-state alphabet) *For natural numbers n_X and n_H, the finite observed–latent joint-state alphabet is*

$$\mathcal{S}(n_X, n_H) := \{i \in \mathbb{N} : 0 \le i < n_X\} \times \{j \in \mathbb{N} : 0 \le j < n_H\}.$$

The joint alphabet $\mathcal{S}$ is finite but otherwise unrestricted. This cardinality-uniform formulation lets the risk statements range over both observed and latent state complexities.

Definition 6. [`synth_9`] (Behavior and target stationary joint-state laws) *For a memoryless policy $p \in \{b, e\}$, let P_p be the policy-induced Markov kernel on the joint state space $\mathcal{S} = \mathcal{X} \times \mathcal{H}$. Define $d_p \in \Delta(\mathcal{S})$ to be the stationary joint-state law satisfying*

$$d_p P_p = d_p.$$

We write d_b for the behavior-policy stationary joint-state law and d_e for the target-policy stationary joint-state law.

The stationary law d_p provides the population distribution attached to a fixed memoryless policy. In particular, d_b describes the stationary behavior experiment, while d_e supplies the stationary target distribution in the value estimand.

We now impose the remaining restrictions that define the population experiment analyzed below.

Assumption 3. [`ass:stationary-start`] (Stationary start) *The initial joint state S_0 has the behavior stationary law d_b:*

$$S_0 \sim d_b.$$

Assumption 3 is the behavior-stationary initialization condition in Hu and Wager [2023]. It aligns the observed trajectory with the stationary behavior distribution from the first period.

Assumption 4. [`ass:bounded-reward`] (Bounded rewards) *For each $t = 0, \ldots, T-1$,*

$$\Pr(|Y_t| \le 1) = 1.$$

Assumption 4 is the uniformly bounded rewards condition in Hu and Wager [2023]. The common $[-1, 1]$ scale fixes the squared-error normalization used in the minimax risk.

Assumption 5. [`ass:policy-overlap`] (Policy overlap) *The target carrier e is a probability vector at each observed state,*

$$e(a \mid x) \ge 0 \quad \text{for all } x \in \mathcal{X},\ a \in \mathcal{A}, \qquad \sum_{a \in \mathcal{A}} e(a \mid x) = 1 \quad \text{for all } x \in \mathcal{X},$$

and for every $(a, x) \in \mathcal{A} \times \mathcal{X}$ the policy-overlap constant L satisfies

$$e(a \mid x) \le L b(a \mid x).$$

Assumption 5 is the uniform one-step policy overlap condition in Hu and Wager [2023]. It bounds observable action likelihood ratios state by state, which is the primitive overlap control for partial-history weighting, and it also records that the target policy is a genuine action distribution at every observed state, which is what makes the one-step ratios average to one.

Definition 7. [`synth_7`] (Derived contraction factor) *For any real t_0, define the contraction parameter α by*

$$\alpha := \exp\left(-\frac{1}{t_0}\right).$$

The positive mixing scale t_0 determines the contraction parameter α. Smaller α corresponds to faster geometric forgetting in total variation distance.

Assumption 6. [`ass:uniform-contraction`] (Uniform contraction) *For every $(p, \nu, \nu') \in \{b, e\} \times \Delta(\mathcal{S})^2$, the contraction parameter α and the policy-induced transition operator P_p satisfy*

$$\|\nu P_p - \nu' P_p\|_{\mathrm{TV}} \leq \alpha \|\nu - \nu'\|_{\mathrm{TV}}.$$

Assumption 6 is the uniform geometric total-variation contraction condition in Hu and Wager [2023]. It supplies a common mixing rate for the behavior and target policy kernels on the joint state space.

Assumption 7. [`ass:latent-stationary-overlap`] (Latent stationary overlap) *For the overlap constant C and experiment M, the latent stationary overlap condition holds when $C \geq 1$ and, writing d_e and d_b for the stationary laws of the target-policy and behavior-policy joint-state transition kernels, respectively,*

$$d_e(s) \leq C\, d_b(s), \qquad s \in \mathcal{S}.$$

Assumption 7 is the strong stationary distributional overlap condition imposed on the unobserved joint state. The stationary density-ratio bound C relates the target and behavior stationary laws over the full finite state space, including the latent coordinate.

The observed-data object is a trajectory of observed coordinates, actions, and rewards. The next definitions collect the seven restrictions into the model class and define the value and risks used in the main results.

Definition 8. [`synth_12`] (Observable trajectory space) *For a horizon T and observed-state cardinality n_X, define the observable trajectory space O_T by*

$$\mathsf{O}_T = \{0, \ldots, T-1\} \to \big(\{0, \ldots, n_X - 1\} \times \{0, 1\} \times \mathbb{R}\big).$$

Thus an element of O_T records, at each epoch $t < T$, the observed coordinate X_t, the binary action A_t, and the real reward Y_t. Equivalently, it is the sequence $((X_t, A_t, Y_t))_{t=0}^{T-1}$.

The observable trajectory space O_T is the statistical sample space for estimators. It keeps the latent path out of the estimator input while retaining the observed policy information b and e.

Definition 9. [`synth_11`] (Model-class assumption predicates) *For an observable-data POMDP experiment $(\mathcal{X}, \mathcal{H}, K, b, e, \mathcal{L}(\mathsf{O}_T))$ with $\mathcal{S} = \mathcal{X} \times \mathcal{H}$, $\mathcal{A} = \{0, 1\}$, policy-induced kernels P_p, stationary laws d_p, $\alpha = \exp(-1/t_0)$, and $L = \exp(\zeta)$, write*

$$\mathrm{A}_{\mathrm{pomdp}}, \quad \mathrm{A}_{\mathrm{ign}}, \quad \mathrm{A}_{\mathrm{start}}, \quad \mathrm{A}_{\mathrm{reward}}, \quad \mathrm{A}_{\mathrm{act}}, \quad \mathrm{A}_{\mathrm{mix}}, \quad \mathrm{A}_{\mathrm{stat}}$$

for the following seven restrictions, respectively:

$$\mathcal{L}(Y_t, X_{t+1}, H_{t+1} \mid X_{0:t}, H_{0:t}, A_{0:t}, Y_{0:t-1}) = K(\cdot \mid X_t, H_t, A_t), \qquad t = 0, \ldots, T-1,$$

as in Assumption 1;

$$\Pr(A_t = a \mid X_{0:t}, H_{0:t}, A_{0:t-1}, Y_{0:t-1}) = b(a \mid X_t), \qquad (a, t) \in \mathcal{A} \times \{0, \ldots, T-1\},$$

as in Assumption 2;

$$S_0 \sim d_b,$$

as in Assumption 3;

$$\Pr(|Y_t| \leq 1) = 1, \qquad t = 0, \ldots, T-1,$$

as in Assumption 4;

$$e(a \mid x) \leq Lb(a \mid x), \qquad (a, x) \in \mathcal{A} \times \mathcal{X}, \qquad e \text{ a probability vector at each } x,$$

as in Assumption 5;

$$\|\nu P_p - \nu' P_p\|_{\mathrm{TV}} \leq \alpha \|\nu - \nu'\|_{\mathrm{TV}}, \qquad (p, \nu, \nu') \in \{b, e\} \times \Delta(\mathcal{S})^2,$$

as in Assumption 6; and

$$d_e(s) \leq C d_b(s), \qquad s \in \mathcal{S},$$

as in Assumption 7.

Definition 9 gives short predicate names for the seven restrictions. These names are bookkeeping devices for the class definition and preserve the direct link between the compact model notation and the assumptions just stated.

Definition 10. [`def:model-class`] (Latent-overlap POMDP class) *For $T \in \mathbb{N}$ and $t_0, \zeta, C \in \mathbb{R}$, $\mathcal{M}_T(t_0, \zeta, C)$ is the class of all observable-data POMDP experiments*

$$(\mathcal{X}, \mathcal{H}, K, b, e, \mathcal{L}(\mathsf{O}_T))$$

with the following defining properties:

- ***(Horizon and alphabets.)*** *The trajectory horizon satisfies $1 \leq T$, the observed and latent alphabets $\mathcal{X}$ and $\mathcal{H}$ are finite, and $|\mathcal{A}| = 2$.*
- ***(Parameter scales.)*** *The mixing and log-overlap parameters satisfy $0 < t_0$ and $0 < \zeta$.*
- ***(Raw experiment.)*** *The tuple has a reward-transition kernel K, behavior policy b, target policy e, and probability law $\mathcal{L}(\mathsf{O}_T)$ for the observed trajectory generated from the joint state coordinates (X_t, H_t).*
- ***(Seven restrictions.)*** *The seven restrictions in Assumptions 1 to 7 hold, with policy-overlap factor $L = \exp(\zeta)$ and contraction factor $\alpha = \exp(-1/t_0)$.*

Equivalently,

$$\mathcal{M}_T(t_0, \zeta, C) := \left\{ (\mathcal{X}, \mathcal{H}, K, b, e, \mathcal{L}(\mathsf{O}_T)) : \begin{array}{l} 1 \leq T, \quad |\mathcal{X}| < \infty, \quad |\mathcal{H}| < \infty, \quad |\mathcal{A}| = 2, \\ 0 < t_0, \quad 0 < \zeta, \\ \mathrm{A}_{\mathrm{pomdp}}, \mathrm{A}_{\mathrm{ign}}, \mathrm{A}_{\mathrm{start}}, \mathrm{A}_{\mathrm{reward}}, \mathrm{A}_{\mathrm{act}}, \mathrm{A}_{\mathrm{mix}}, \mathrm{A}_{\mathrm{stat}} \text{ hold} \end{array} \right\}.$$

The class ranges over all finite observed and latent cardinalities and all revealed behavior-target policy pairs satisfying these conditions.

The model class $\mathcal{M}_T(t_0, \zeta, C)$ is the population domain for the minimax analysis. Its parameters separately control the time horizon, mixing scale, one-step action overlap, and latent stationary overlap.

Definition 11. [`synth_5`] (Target reward regression) *For any raw finite-state experiment M with target policy e and one-step law K, and for every joint state $s = (x, h) \in \{0, \dots, n_X - 1\} \times \{0, \dots, n_H - 1\}$, define the target-policy reward regression g_e by*

$$g_e(s) = \sum_{a \in \{0,1\}} e(a \mid x) \int r \, K(dr, ds' \mid s, a).$$

The target-policy reward regression g_e averages the one-period reward under the target action distribution at a fixed joint state. It is the reward component of the stationary target value.

Definition 12. [`def:target-value`] (Target value $\theta(K, e)$) *For the joint state S_t and the target-policy reward regression g_e,*

$$\theta(K, e) := \sum_{s=(x,h) \in \mathcal{S}} d_e(s) g_e(s) = \sum_{s=(x,h) \in \mathcal{S}} d_e(s) \sum_{a \in \mathcal{A}} e(a \mid x) \, E_K[Y_t \mid S_t = s, A_t = a].$$

The estimand $\theta(K, e)$ is the stationary mean reward obtained by running the target policy and evaluating rewards under its stationary joint-state law. It is the population target for observable-data off-policy evaluation.

Definition 13. [`synth_13`] (Estimator-specific risk) *Let*

$$M = (\mathcal{X}, \mathcal{H}, K, b, e, \mathcal{L}(\mathsf{O}_T)) \in \mathcal{M}_T(t_0, \zeta, C)$$

be a model in the latent-overlap class of Definition 10, whose observed trajectory is

$$\mathsf{O}_T = ((X_t, A_t, Y_t))_{t=0}^{T-1},$$

and let $\mathrm{Law}_M^{\mathrm{obs}}$ denote the probability law of O_T induced by M under the behavior policy b.

For any observable-data estimator $\widehat{\theta}$, its estimator-specific worst-case squared risk over $\mathcal{M}_T(t_0, \zeta, C)$ is

$$R_T(t_0, \zeta, C; \widehat{\theta}) := \sup_{M \in \mathcal{M}_T(t_0, \zeta, C)} \mathbb{E}_{\mathrm{Law}_M^{\mathrm{obs}}} \left[\left\{ \widehat{\theta}(n_X, b, e, \mathsf{O}_T) - \theta(K, e) \right\}^2 \right],$$

where $n_X = |\mathcal{X}|$ is the revealed observed-alphabet size of M. In particular,

$$R_T(t_0, \zeta, C; \widehat{\theta}_{k_T})$$

is this quantity evaluated at the clipped partial-history importance-weighted estimator $\widehat{\theta}_{k_T}$ with balanced history depth k_T.

The estimator-specific risk in Definition 13 evaluates a fixed observable-data rule uniformly over the model class. This quantity is used for the partial-history estimator bound, while the minimax risk below optimizes over all admissible observable-data estimators.

Definition 14. [def:minimax-risk] (Observable minimax risk) *For a trajectory horizon $T \geq 1$ and parameters t_0, ζ, C, the cardinality-uniform observable-data minimax squared-error risk is*

$$R_T(t_0, \zeta, C) := \inf_{\widehat{\theta}} \sup_{M \in \mathcal{M}_T(t_0,\zeta,C)} \mathbb{E}_{\mathrm{Law}_M^{\mathrm{obs}}} \left[\left\{ \widehat{\theta}(n_X, b, e, \mathsf{O}_T) - \theta(K, e) \right\}^2 \right].$$

Here $\mathcal{M}_T(t_0, \zeta, C)$ is the finite observed- and latent-state model class in Definition 10, with unrestricted finite observed alphabet size n_X and unrestricted finite latent alphabet size, and $\theta(K, e)$ is the stationary target-policy mean reward in Definition 12. The positive-horizon requirement $1 \leq T$ is part of membership in $\mathcal{M}_T(t_0, \zeta, C)$, not a separate restriction on the supremum; at the degenerate horizon $T = 0$ the class is empty and the display is not used. The infimum ranges over all observable-data estimators $\widehat{\theta}$ that take as inputs only n_X, b, e, and the observed trajectory view O_T, are measurable in O_T for each n_X, b, e, and are uniformly valued in $[-1, 1]$.

The minimax risk $R_T(t_0, \zeta, C)$ measures the best achievable squared-error performance from the observed trajectory and the revealed policies. The uniformity over finite cardinalities makes the rate a statement about the structural restrictions rather than a fixed finite state dimension.

The final definitions in this section introduce the concrete estimator and the lower-bound family used in the main results.

Definition 15. [synth_6] (One-step policy ratio) *For behavior and target policy functions $b, e : \{0, \ldots, n_X - 1\} \times \{0, 1\} \to \mathbb{R}$, the observable one-step policy ratio is*

$$\rho_t = \begin{cases} 0, & b(X_t, A_t) = 0, \\ e(X_t, A_t)/b(X_t, A_t), & b(X_t, A_t) \neq 0. \end{cases}$$

The observable one-step policy ratio ρ_t compares target and behavior action probabilities along the observed path. The zero-cell convention is compatible with the domination condition in Assumption 5.

Definition 16. [def:phiw-estimator] (Partial-history estimator $\widehat{\theta}_k$) *For $0 \leq k < T$, define*

$$Z_t(k) := Y_t \prod_{j=t-k}^{t} \rho_j,$$

where ρ_t is the observable one-step policy ratio with the zero-cell convention. The clipped partial-history importance-weighted estimator is

$$\widehat{\theta}_k := \mathrm{clip}_{[-1,1]} \left\{ \frac{1}{T-k} \sum_{t=k}^{T-1} Z_t(k) \right\}.$$

The score $Z_t(k)$ weights the reward by the recent observable policy-ratio history, and the estimator $\widehat{\theta}_k$ averages these scores with clipping to the reward scale. The depth k is the tuning parameter that trades the contraction of latent-history bias against the variance from multiplying policy ratios.

Definition 17. [def:history-depth] (History depth k_T) *For the contraction parameter α and policy-overlap constant L, the balanced fixed-radius history depth is*

$$k_T := \min \left\{ \left\lfloor \frac{T}{2} \right\rfloor, \max \left\{ 0, \left\lfloor \frac{\log(T)}{2\log(1/\alpha) + \log L} \right\rfloor \right\} \right\}.$$

The fixed-depth rule k_T balances the two primitives α and L at horizon T. It is the estimator depth used in the fixed-radius upper bound.

Definition 18. [`def:signed-depth-family`] (Signed-depth family) *For a trajectory length T, parameters $t_0, \zeta, C \in \mathbb{R}$, a terminal depth $Q \in \mathbb{N}$, and an alternative index $v \in \{-1, +1\}$, the signed-depth family is defined under*

$$t_0 > 0, \qquad \zeta > 0, \qquad C > 1, \qquad Q \geq 1.$$

It is the raw POMDP experiment in O_T with one observed state and $2(Q+1)$ hidden states obtained by embedding the following finite-reward model. The observed coordinate X_t is constant, and the hidden coordinate is written

$$H_t = (J_t, U_t) \in \{0, \ldots, Q\} \times \{-1, +1\}.$$

Writing α for the mixing value determined by t_0, L for the policy factor determined by ζ, $\varepsilon_C = \min\{(C-1)/2, 1/2\}$, and $c_0 = (1-\alpha)/4$, the behavior and target policies are

$$b(A_t = 1) = \frac{1}{L}, \qquad b(A_t = 0) = 1 - \frac{1}{L}, \qquad e(A_t = 1) = 1, \qquad e(A_t = 0) = 0.$$

The initial hidden-state law is the probability measure obtained from the weights

$$w_j \frac{1 + u\varepsilon_C L^{-j}}{2}, \qquad w_j = \frac{(1-\alpha)\alpha^j}{1 - \alpha^{Q+1}}, \qquad j \in \{0, \ldots, Q\},\ u \in \{-1, +1\}.$$

The transition kernel is the probability measure on reward-symbol and next-hidden-state pairs whose real weight at (y, H_{t+1}), with $y \in \{-1, +1\}$, is the product of the reward weight

$$\frac{1 + yvc_0 U_t \mathbf{1}\{J_t = Q\}}{2}$$

and the hidden-transition weight specified as follows. A reset sign u' has weight

$$\frac{1 + u'\varepsilon_C}{2}.$$

If $J_t = Q$, then only transitions with $J_{t+1} = 0$ receive this reset-sign weight. If $J_t < Q$, transitions with $J_{t+1} = 0$ receive weight $(1-\alpha)(1 + U_{t+1}\varepsilon_C)/2$, while transitions with $J_{t+1} = J_t + 1$ receive weight

$$\alpha\, \mathbf{1}\{U_{t+1} = U_t\} \quad \text{when } A_t = 1, \qquad \frac{\alpha}{2} \quad \text{when } A_t = 0,$$

and all other hidden transitions receive weight zero. The finite reward symbols are mapped to real rewards by $0 \mapsto -1$ and $1 \mapsto 1$, and the trajectory law is the finite chronological law generated by this initial distribution, behavior policy, and kernel, then embedded into the unrestricted real-reward raw experiment.

The signed-depth two-point sub-experiment in Definition 18 is the finite construction used for the lower bounds. Its terminal depth Q, alternative sign v, latent depth coordinate J_t, latent sign coordinate U_t, reset-sign perturbation ε_C, reward-amplitude constant c_0, and truncated geometric mass w_j make the overlap and contraction parameters explicit inside a finite POMDP.

4 Main results

The preceding section defined the observed-data experiment class, the stationary target value, the minimax risk, and the partial-history importance-weighted estimators. This section gives the risk surface as the latent stationary overlap radius varies. The organizing parameters are the normalized distance from unit overlap and the radius-calibrated history depth, where the radius C is supplied to the estimator as a tuning input.

Definition 19. [`def:overlap-distance-radius`] (Overlap-distance radius q_C) *For $C \geq 1$, define*

$$q_C := \frac{C-1}{C}.$$

The radius q_C is zero at equality of the two stationary laws and increases with the allowed stationary density-ratio bound. It is the scale on which the hidden-state part of the risk is measured.

Definition 20. [`def:radius-adaptive-history-depth`] (Radius-calibrated history depth $\kappa_T^{\mathrm{rad}}(C)$) *For $T \in \mathbb{N}$ and $C \geq 1$, define*

$$\kappa_T^{\mathrm{rad}}(C) := \begin{cases} 0, & Tq_C^2 \leq 1, \\ \min\left\{ \left\lfloor \frac{T}{2} \right\rfloor, \left\lfloor \frac{\log(Tq_C^2)}{2\log(1/\alpha) + \log L} \right\rfloor \right\}, & Tq_C^2 > 1. \end{cases}$$

The depth $\kappa_T^{\mathrm{rad}}(C)$ shortens the history when the radius is close to unit overlap and lengthens it when the latent stationary shift is large enough to make hidden-state memory relevant. The threshold $Tq_C^2 \leq 1$ is the parametric boundary layer in the theorem below.

For the all-radius theorem it is convenient to index models directly and write their target values without carrying the kernel and policy notation in each display.

Definition 21. [`synth_16`] (Finite signed-depth observed-word law) *For $C > 1$, $Q \geq 1$, and $v \in \{-1, +1\}$, let*

$$\mathbb{P}_{Q,v}^{\mathrm{fin},T}$$

be the finite observed-word law through horizon T generated by the signed-depth construction of Definition 18. Its observed state is constant, its latent state is $H_t = (J_t, U_t) \in \{0, \ldots, Q\} \times \{-1, +1\}$, its behavior and target policies satisfy $b(1) = 1/L$ and $e(1) = 1$, and its reward alphabet is $\{-1, +1\}$ with

$$\Pr_v(Y_t = y \mid J_t, U_t, A_t) = \frac{1 + yvc_0 U_t \mathbf{1}\{J_t = Q\}}{2}.$$

The initial latent law is $d_b(j, u) = w_j(1 + u\varepsilon_C L^{-j})/2$, and the resulting finite observation consists of the action-reward word $(A_t, Y_t)_{t=0}^{T-1}$, equivalently the finite observed trajectory with constant observed-state coordinate.

The next theorem is the master characterization. Its constants depend on t_0 and ζ, while the comparison holds uniformly over every radius $C \geq 1$ and over the finite observed and hidden cardinalities allowed by Definition 10.

Definition 22. [synth_14] (Target value of an indexed model) *For an indexed model $m \in \mathcal{M}_T(t_0, \zeta, C)$, write K_m and e_m for its joint reward-transition kernel and target policy. The target value of m is*

$$\theta_m := \theta(K_m, e_m),$$

with $\theta(K, e)$ the stationary target-policy causal value from Definition 12.

Theorem 1 gives a single observable estimator *rule* — one formula, applied at every radius — together with a matching minimax surface. The rule is not radius-free: the depth it selects is computed from the supplied value of C, so different radii give different depths and different estimators. The expression for $\mathfrak{r}_T(t_0, \zeta, C)$ has a parametric term and a hidden-state term scaled by q_C. The radius-calibrated rule in Definition 20 implements this balance directly: in the boundary layer it uses depth zero, and away from that layer it increases the observable history length according to the effective radius Tq_C^2.

The lower-bound calculations use the finite observed law induced by the signed-depth construction through horizon T. Naming that law separates the observable experiment used in the converse from the latent construction that generates it.

Definition 23. [synth_10] (Unclipped partial-history average) *For the balanced history depth k_T, define the unclipped partial-history average by*

$$\widehat{\theta}^{\mathrm{raw}}_{k_T} := \frac{1}{T - k_T} \sum_{t=k_T}^{T-1} Z_t(k_T), \qquad Z_t(k_T) := Y_t \prod_{j=t-k_T}^{t} \rho_j.$$

Here $\rho_j = e(A_j \mid X_j)/b(A_j \mid X_j)$ on behavior-supported cells, with the zero-cell convention used for ρ_j.

For comparison with the clipped estimator in Definition 16, we also record the unclipped average at the fixed-radius balancing depth. This object is useful for interpreting the bias–variance calculation, while the risk bounds below are stated for the clipped estimator.

Theorem 1. [thm:uniform-overlap-frontier] (Uniform overlap frontier) *Fix $t_0 > 0$ and $\zeta > 0$. Let $R_T(t_0, \zeta, C)$ be the minimax squared-error risk in Definition 14, let $q_C = (C-1)/C$ be the overlap-distance radius in Definition 19, and set*

$$\beta = \frac{2}{2 + t_0\zeta}, \qquad \mathfrak{r}_T(t_0, \zeta, C) = T^{-1} + T^{-\beta} q_C^{2(1-\beta)}.$$

There exist constants $c_{\mathrm{lo}}, c_{\mathrm{hi}} > 0$, with $c_{\mathrm{lo}} \le c_{\mathrm{hi}}$, and an integer $T_\star \ge 1$, all depending only on t_0 and ζ, such that for every $C \ge 1$ and every $T \ge T_\star$,

$$c_{\mathrm{lo}} \mathfrak{r}_T(t_0, \zeta, C) \le R_T(t_0, \zeta, C) \le c_{\mathrm{hi}} \mathfrak{r}_T(t_0, \zeta, C),$$

and the radius-adaptive partial-history importance-weighted estimator with depth $\kappa_T^{\mathrm{rad}}(C)$ from Definition 20 satisfies

$$\sup_{m \in \mathcal{M}_T(t_0, \zeta, C)} \mathbb{E}_m\Big[\big(\widehat{\theta}_{\kappa_T^{\mathrm{rad}}(C)} - \theta_m\big)^2\Big] \le c_{\mathrm{hi}} \mathfrak{r}_T(t_0, \zeta, C).$$

Moreover, for every sequence $C_T \ge 1$, if $Tq_{C_T}^2$ is eventually bounded, then there is a constant $c_{\mathrm{imm}} > 0$, depending on the sequence and on t_0, ζ, such that eventually

$$\sup_{m \in \mathcal{M}_T(t_0, \zeta, C_T)} \mathbb{E}_m\Big[\big(\widehat{\theta}_0 - \theta_m\big)^2\Big] \le \frac{c_{\mathrm{imm}}}{T}.$$

Holding the radius at a nonunit value gives the fixed-radius slice of Theorem 1. The next theorem states that slice in the fixed-depth notation of Definition 17, together with the signed-depth alternatives from Definition 18 that witness the converse.

Theorem 2. [thm:fixed-c-minimax] (Fixed-overlap minimax rate) *Fix $t_0, \zeta, C \in \mathbb{R}$. Assume:*

- ***(Positive memory exponent.)*** $t_0 > 0$.
- ***(Positive overlap exponent.)*** $\zeta > 0$.
- ***(Fixed latent-overlap radius.)*** $C > 1$.

Let the exponent β be

$$\beta = \frac{2}{2 + t_0 \zeta}.$$

There exist constants $c_\star, c^\star \in \mathbb{R}$ and an integer $T_\star \geq 1$ such that

$$0 < c_\star \leq c^\star$$

and, for every integer $T \geq T_\star$,

$$c_\star T^{-\beta} \leq R_T(t_0, \zeta, C) \leq c^\star T^{-\beta},$$

where $R_T(t_0, \zeta, C)$ is the minimax squared-error risk in Definition 14. Moreover, the clipped partial-history importance-weighted estimator $\widehat{\theta}_{k_T}$, with k_T as in Definition 17, satisfies

$$\sup_{M \in \mathcal{M}_T(t_0, \zeta, C)} \mathbb{E}_M \left[\{\widehat{\theta}_{k_T} - \theta(M)\}^2 \right] \leq c^\star T^{-\beta},$$

with $\theta(M)$ the stationary target-policy mean reward in Definition 12.

For each integer $T \geq T_\star$, there is an integer $Q \geq 1$ such that the lower bound is witnessed by the two signed-depth alternatives $M^T_{Q,v}$, $v \in \{-1, +1\}$, generated by Definition 18 with one observed state and $2(Q+1)$ latent states: for every measurable observable-data estimator $\widehat{\theta}$, under the usual squared-error integrability conditions for both alternatives,

$$c_\star T^{-\beta} \leq \max_{v \in \{-1,+1\}} \mathbb{E}_{M^T_{Q,v}} \left[\{\widehat{\theta} - \theta(M^T_{Q,v})\}^2 \right].$$

The exponent β reflects the balance between geometric forgetting and the variance cost of multiplying observable policy ratios. Since q_C is a positive constant on this slice, the all-radius expression in Theorem 1 reduces to the power $T^{-\beta}$ up to constants. The estimator achieves that rate, and the two-point construction shows that every observable-data estimator faces the same order of risk over the class.

The unit-overlap boundary is the $C = 1$ slice of the same surface. At this endpoint, the normalized radius is $q_C = 0$, and the risk expression in Theorem 1 is on the T^{-1} scale.

Theorem 3. [prop:unit-overlap-boundary] (Unit overlap boundary) *Let $t_0 > 0$ and $\zeta > 0$. At unit latent overlap, $C = 1$, the following assertions hold.*

For every trajectory length T, every finite observed and hidden state cardinalities, and every experiment $M \in \mathcal{M}_T(t_0, \zeta, 1)$ in the model class of Definition 10, the target stationary law d_e and

the behavior stationary law d_b of the corresponding policy-induced joint-state transition kernels coincide:

$$d_e = d_b.$$

Moreover, there is a constant $c_{\mathrm{imm}} > 0$, depending only on t_0 and ζ, such that for every integer $T \geq 1$, the worst-case squared risk of the immediate observable estimator $\widehat{\theta}_0$ satisfies

$$\sup_{M \in \mathcal{M}_T(t_0,\zeta,1)} \mathbb{E}_M\left[\left(\widehat{\theta}_0 - \theta(K,e)\right)^2\right] \leq \frac{c_{\mathrm{imm}}}{T}.$$

Finally, for the minimax squared-error risk $R_T(t_0,\zeta,1)$ of Definition 14, there are constants $c_{\mathrm{lo}}, c_{\mathrm{hi}} > 0$ with $c_{\mathrm{lo}} \leq c_{\mathrm{hi}}$ and an integer $T_\star \geq 1$ such that, for every $T \geq T_\star$,

$$\frac{c_{\mathrm{lo}}}{T} \leq R_T(t_0,\zeta,1) \leq \frac{c_{\mathrm{hi}}}{T}.$$

The hidden-state exponent at these positive scales satisfies

$$\frac{2}{2+t_0\zeta} < 1.$$

The boundary result identifies the equality case as an immediate-weighting regime. The coincidence $d_e = d_b$ gives the stationary invariance needed for depth zero, and the matching minimax bounds place the endpoint on the T^{-1} scale.

The remaining regime reads Theorem 1 along a deterministic path approaching the unit-overlap boundary. Write the nonnegative excess as δ_T; then $C_T = 1 + \delta_T$ traces a local path into the nonunit region.

Theorem 4. [`thm:shrinking-overlap-frontier`] (Shrinking-overlap frontier) *The following conditions hold:*

- ***(Fixed smoothness.)*** *The constants satisfy $t_0 > 0$ and $\zeta > 0$.*
- ***(Shrinking radius.)*** *The deterministic radius excess sequence $(\delta_T)_{T\geq 1}$ satisfies $\delta_T \geq 0$ for every T.*
- ***(Local limit.)*** *The sequence satisfies $\delta_T \to 0$.*
- ***(Rate exponent.)*** *Let $\beta = 2/(2+t_0\zeta)$.*
- ***(Local rate and depth.)*** *Define*

$$r_T(\delta) = T^{-1} + T^{-\beta}\delta_T^{2(1-\beta)}$$

and

$$\kappa_T = \begin{cases} 0, & T\delta_T^2 \leq 1, \\ \min\left\{\lfloor T/2 \rfloor, \left\lfloor \dfrac{\log(T\delta_T^2)}{2\log(1/\alpha) + \log L} \right\rfloor\right\}, & T\delta_T^2 > 1. \end{cases}$$

Then there exist constants $c_{\mathrm{loc}}, C_{\mathrm{loc}}$ with $0 < c_{\mathrm{loc}} \leq C_{\mathrm{loc}}$, depending only on t_0 and ζ, such that, for all sufficiently large T,

$$c_{\mathrm{loc}}\, r_T(\delta) \leq R_T(t_0,\zeta,1+\delta_T) \leq C_{\mathrm{loc}}\, r_T(\delta),$$

where R_T is the minimax squared-error risk in Definition 14. Moreover, the observable partial-history importance-weighted estimator $\widehat{\theta}_{\kappa_T}$ of Definition 16 satisfies

$$\sup_{M \in \mathcal{M}_T(t_0,\zeta,1+\delta_T)} \mathbb{E}_M\left[\left(\widehat{\theta}_{\kappa_T} - \theta(M)\right)^2\right] \le C_{\mathrm{loc}}\, r_T(\delta)$$

for all sufficiently large T. Finally, if there is a constant $D \ge 0$ such that $T\delta_T^2 \le D$ for all sufficiently large T, then there is a constant $c_{\mathrm{imm}} > 0$ such that

$$\sup_{M \in \mathcal{M}_T(t_0,\zeta,1+\delta_T)} \mathbb{E}_M\left[\left(\widehat{\theta}_0 - \theta(M)\right)^2\right] \le \frac{c_{\mathrm{imm}}}{T}$$

for all sufficiently large T.

The local rate decomposes into the parametric term and a hidden-state term scaled by the shrinking excess. When $T\delta_T^2$ remains bounded, the immediate estimator already attains the parametric order. When the product grows, the depth in the theorem increases logarithmically with the effective radius $T\delta_T^2$, matching the same risk expression.

5 A finite insulin-policy adaptation

The minimax characterization in Theorem 1 is stated for an abstract finite partially observed model class. This section records a finite insulin-policy construction that keeps the glucose-threshold structure familiar from mobile-health and diabetes treatment applications while putting every component on a finite alphabet [Maahs et al., 2012, Liao et al., 2021, Luckett et al., 2020]. The construction, denoted $\mathcal{I}_{\mathrm{grid}}$, uses refreshed diet and activity coordinates to create latent heterogeneity, uniform mixing, and stationary overlap in a single controlled Markov model.

Definition 24. [`def:insulin-grid`] (Insulin grid $\mathcal{I}_{\mathrm{grid}}$) *The finite insulin-grid state is*

$$s = (g, d_1, d_2, x_1, x_2, a_1) \in \{50, 100, 150, 200, 250\} \times \{0, 78\}^2 \times \{0, 31, 850\}^2 \times \{0, 1\},$$

so the state space has 360 states. The diet coordinates form

$$H_t = (d_1, d_2),$$

and all other coordinates form X_t. The behavior and target policies are

$$b(1 \mid X_t) = 0.3, \qquad e(1 \mid X_t) = \mathbf{1}\{g \ge 125\}.$$

With probability $1/2$, the next full state is drawn uniformly. With the remaining probability, draw

$$d_0 \in \{0, 78\} \quad \textit{with probabilities} \quad (0.8, 0.2),$$

draw

$$x_0 \in \{0, 31, 850\} \quad \textit{with probabilities} \quad (0.4, 0.4, 0.2),$$

draw $\xi \sim N(0, 25)$, and set

$$g' = \mathrm{round}_{\{50,100,150,200,250\}}\, \mathrm{clip}_{[50,250]}\big(10 + 0.9g + 0.1d_1 + 0.1d_2 - 0.01x_1 - 0.01x_2 - 2A_t - 4a_1 + \xi\big).$$

The memory coordinates then shift to

$$(g', d_0, d_1, x_0, x_1, A_t).$$

The reward is

$$Y_t = \begin{cases} -1, & g \le 70, \\ -2/3, & g > 150, \\ -1/3, & 70 < g \le 80 \text{ or } 120 < g \le 150, \\ 0, & \text{otherwise.} \end{cases}$$

This construction defines $\mathcal{I}_{\text{grid}}$.

Two conventions in the displayed transition fix the numbers the certificate uses, so we state them explicitly. First, $N(0, 25)$ is the centered Gaussian of *variance* 25, that is standard deviation 5. Second, the composite clip-and-round operator is the nearest-grid map with cutpoints at the midpoints $75, 125, 175, 225$ of the glucose grid, the two extreme cells absorbing everything beyond them; each cell is the interval closed on the left, so a value landing exactly on a midpoint is assigned to the higher grid point, a tie that has probability zero under the continuous noise. Writing m for the deterministic part of the displayed expression and Φ for the standard normal distribution function, the next-glucose weights are therefore

$$\Pr\{g' = 50\} = \Phi\left(\frac{75 - m}{5}\right), \qquad \Pr\{g' = 250\} = 1 - \Phi\left(\frac{225 - m}{5}\right),$$

and, for $i \in \{1, 2, 3\}$,

$$\Pr\{g' = 50 + 50i\} = \Phi\left(\frac{75 + 50i - m}{5}\right) - \Phi\left(\frac{75 + 50(i-1) - m}{5}\right).$$

The observed coordinate contains the glucose level, lagged activity, and lagged treatment, while the diet coordinates are latent. The target rule is a deterministic threshold policy on the observed glucose coordinate, and the behavior rule randomizes treatment with fixed probability. The refresh step supplies a direct minorization component, so the example aligns with the contraction and overlap conditions used by Theorem 1 within a finite-state adaptation of the insulin-policy setting.

The next two facts record the mixing and stationary-law ingredients behind the certified contraction and stationary-overlap constants. They apply to any policy rule on the observed grid states, including the behavior and target policies in Definition 24.

Lemma 1. [`lem:insulin-policy-contraction`] (Uniform one-half contraction of the insulin-grid policy kernels) *Let* $T \in \mathbb{N}$ *and let* p *assign to each of the* 90 *observed states of the refreshed finite insulin adaptation* $\mathcal{I}_{\text{grid}}$ *of Definition 24 a probability law on the action set* $\{0, 1\}$. *Write* S_{grid} *for the* 360*-point joint state space,* $x(s)$ *for the observed coordinate of a grid state* s, K_T *for the transition kernel of* $\mathcal{I}_{\text{grid}}$, *and*

$$P_p^{(T)}(s, s') = \sum_{a \in \{0,1\}} p(a \mid x(s))\, K_T\{(q, s'') : \ s'' = s' \mid s, a\}, \qquad s, s' \in \mathsf{S}_{\text{grid}},$$

for the induced transition matrix. Then, for all probability vectors ν, ν' *on* S_{grid},

$$\|\nu P_p^{(T)} - \nu' P_p^{(T)}\|_{\text{TV}} \le \frac{1}{2}\|\nu - \nu'\|_{\text{TV}},$$

equivalently, in the ℓ^1 *norm* $\|\cdot\|_1 = 2\|\cdot\|_{\mathrm{TV}}$,

$$\|\nu P_p^{(T)} - \nu' P_p^{(T)}\|_1 \le \frac{1}{2}\|\nu - \nu'\|_1.$$

The proof is deferred to Section B.

Lemma 2. [`lem:insulin-stationary-law`] (The selected insulin-grid stationary law is stationary) *Let* $T \in \mathbb{N}$ *and let* p *assign to each observed state of the refreshed finite insulin adaptation* $\mathcal{I}_{\mathrm{grid}}$ *of Definition 24 a probability law on* $\{0,1\}$. *Suppose that the induced transition matrix* $P_p^{(T)}$ *on the joint state space* $\mathsf{S}_{\mathrm{grid}}$ *satisfies*

$$\|\nu P_p^{(T)} - \nu' P_p^{(T)}\|_{\mathrm{TV}} \le \frac{1}{2}\|\nu - \nu'\|_{\mathrm{TV}}$$

for all probability vectors ν, ν' *on* $\mathsf{S}_{\mathrm{grid}}$. *Then the stationary law* d_p *selected for* $P_p^{(T)}$ *is a probability vector and satisfies the stationarity equations*

$$d_p(s') = \sum_{s \in \mathsf{S}_{\mathrm{grid}}} d_p(s) P_p^{(T)}(s, s'), \qquad s' \in \mathsf{S}_{\mathrm{grid}}.$$

The proof is deferred to Section B.

Lemma 1 gives the one-half total-variation contraction supplied by the refresh component. Lemma 2 then identifies the selected stationary distribution as a genuine invariant law for each policy-induced transition matrix, providing the stationary objects needed for the target value and the bias diagnostic.

The reward in Definition 24 is a function of the current grid state. The following regressions make this explicit under target-policy averaging and under one-step target-to-behavior reweighting.

Lemma 3. [`lem:insulin-target-reward-regression`] (Target-policy reward regression on the insulin grid) *Let* $T \in \mathbb{N}$, *let* $\mathcal{I}_{\mathrm{grid}}$ *be the refreshed finite insulin adaptation of Definition 24 with transition kernel* K_T *and target policy* e, *and for a grid state* $s = (g, d_1, d_2, x_1, x_2, a_1)$ *write* $x(s) = (g, x_1, x_2, a_1)$ *for its observed coordinate and*

$$r_{\mathrm{ins}}(s) = \begin{cases} -1, & g \le 70, \\ -2/3, & g > 150, \\ -1/3, & 70 < g \le 80 \text{ or } 120 < g \le 150, \\ 0, & \text{otherwise}, \end{cases}$$

for the insulin reward vector. Writing p_1 *for the numeric reward coordinate of a reward/state pair* p *drawn by* K_T, *we have, for every* $s \in \mathsf{S}_{\mathrm{grid}}$,

$$\sum_{a \in \{0,1\}} e(a \mid x(s)) \int p_1 \, K_T(dp \mid s, a) = r_{\mathrm{ins}}(s).$$

The proof is deferred to Section B.

Lemma 4. [lem:insulin-immediate-reward-regression] (Immediate weighted reward regression on the insulin grid) *Let $T \in \mathbb{N}$, let $\mathcal{I}_{\mathrm{grid}}$ be the refreshed finite insulin adaptation of Definition 24 with transition kernel K_T, behavior policy b and target policy e, and for a grid state $s = (g, d_1, d_2, x_1, x_2, a_1)$ write $x(s) = (g, x_1, x_2, a_1)$. Let*

$$\rho(x, a) = \begin{cases} e(a \mid x)/b(a \mid x), & b(a \mid x) \neq 0, \\ 0, & \text{otherwise}, \end{cases}$$

and let r_{ins} be the insulin reward vector

$$r_{\mathrm{ins}}(s) = \begin{cases} -1, & g \leq 70, \\ -2/3, & g > 150, \\ -1/3, & 70 < g \leq 80 \text{ or } 120 < g \leq 150, \\ 0, & \text{otherwise}, \end{cases}$$

Writing p_1 for the numeric reward coordinate of a reward/state pair p drawn by K_T, we have, for every $s \in \mathsf{S}_{\mathrm{grid}}$,

$$\sum_{a \in \{0,1\}} b(a \mid x(s))\, \rho(x(s), a) \int p_1 \, K_T(dp \mid s, a) = r_{\mathrm{ins}}(s).$$

The proof is deferred to Section B.

Together, Lemmas 3 and 4 show that the immediate reward regression is the same state reward under target averaging and under one-step reweighting. The remaining discrepancy in current-action weighting therefore comes from the stationary law used to average this regression: behavior-stationary weighting versus the target-stationary value in Definition 12.

The finite diagnostic uses certified interval arithmetic for stationary reward averages. The next statements first give the general enclosure principle and then instantiate it for the behavior and target stationary reward averages on the insulin grid.

Lemma 5. [lem:certified-stationary-reward-enclosure] (Stationary reward enclosure) *Let S be a finite set, and let P be a stochastic matrix on S whose entries satisfy*

$$P(s, s') \in \mathfrak{A}(s, s')$$

for all $s, s' \in \mathsf{S}$. The intervals are rational intervals $[\ell, u]$ with $\ell \leq u$, combined by exact interval arithmetic:

$$[a, b] + [c, d] = [a + c, b + d], \qquad [a, b] \ominus [c, d] = [a - d, b - c],$$

and $[a, b] \cdot [c, d]$ is the interval whose endpoints are the smallest and largest of the four endpoint products. For an interval $\mathfrak{r}$, write $\ell(\mathfrak{r})$ and $u(\mathfrak{r})$ for its endpoints and

$$|\mathfrak{r}|_{\max} = \max\{|\ell(\mathfrak{r})|, |u(\mathfrak{r})|\}.$$

Suppose the following conditions hold.

- ***(Initial row and interval rows.)*** *There are an exact rational probability vector μ_0 on S, a number $m \in \mathbb{N}$, and interval rows $\mathfrak{T}_0, \ldots, \mathfrak{T}_{m+1}$ on S such that*

  $$\mu_0(s) \in \mathfrak{T}_0(s)$$

  for every $s \in \mathsf{S}$.

- ***(Chunked recurrence.)*** *For every $k \le m$ and every output state s', there are pairwise disjoint blocks $\mathcal{G}_1, \dots, \mathcal{G}_J \subseteq \mathsf{S}$ covering S, each with at most a fixed positive number of states; block intervals $\Xi_1(s'), \dots, \Xi_J(s')$; and partial-sum intervals $\Sigma_1, \dots, \Sigma_{J+1}$ such that*

$$\sum_{s \in \mathcal{G}_j} \mathfrak{T}_k(s) \cdot \mathfrak{A}(s, s') \subseteq \Xi_j(s')$$

for each j,

$$0 \in \Sigma_{J+1}, \qquad \Xi_j(s') + \Sigma_{j+1} \subseteq \Sigma_j$$

for $j = J, \dots, 1$, and

$$\Sigma_1 = \mathfrak{T}_{k+1}(s').$$

- ***(Reward bounds.)*** *There are a reward vector $r : \mathsf{S} \to \mathbb{R}$, rational intervals $\mathfrak{r}(s) \ni r(s)$, and a rational bound $B \ge 0$ such that*

$$|\mathfrak{r}(s)|_{\max} \le B$$

for every $s \in \mathsf{S}$.

- ***(Contraction and stationarity.)*** *There are a rational coefficient $0 \le \rho < 1$ and a probability vector π such that*

$$\|\nu P - \nu' P\|_1 \le \rho \|\nu - \nu'\|_1$$

for all probability vectors ν, ν' on S, and

$$\pi P = \pi.$$

Define

$$\delta = \sum_{s \in \mathsf{S}} \left| \mathfrak{T}_m(s) \ominus \mathfrak{T}_{m+1}(s) \right|_{\max}, \qquad \mathfrak{E} = \sum_{s \in \mathsf{S}} \mathfrak{T}_m(s) \cdot \mathfrak{r}(s),$$

and

$$\mathfrak{J} = \left[\ell(\mathfrak{E}) - \frac{B\delta}{1-\rho}, u(\mathfrak{E}) + \frac{B\delta}{1-\rho} \right].$$

Then

$$\sum_{s \in \mathsf{S}} \pi(s)\, r(s) \in \mathfrak{J}.$$

The proof is deferred to Section B.

Lemma 5 converts a finite number of interval matrix calculations into a bound on the stationary reward average. In the insulin grid, it supplies the certificate used to compare the behavior-stationary immediate-weighted average with the target-stationary value.

Lemma 6. [`lem:insulin-bias-interval-evaluation`] (Insulin bias interval bounds) *For $\star \in \{\mathrm{b}, \mathrm{e}\}$, let $\mathfrak{T}_{\star,0}$ and $\mathfrak{T}_{\star,1}$ be the terminal and successor rational interval rows on the 360-state grid $\mathsf{S}_{\mathrm{grid}}$ of Definition 24 recorded by the zero-step chunked iterate certificate, in blocks of at most 40 states, for the rational interval enclosure of the behavior-policy case $\star = \mathrm{b}$ or target-policy case $\star = \mathrm{e}$. Let $\mathfrak{r}_{\mathrm{ins}}(s) = [r_{\mathrm{ins}}(s), r_{\mathrm{ins}}(s)]$, where $r_{\mathrm{ins}}(s) \in \{-1, -2/3, -1/3, 0\}$ is the grid reward from Definition 24. Using the exact rational interval arithmetic of Lemma 5, define*

$$\delta_\star = \sum_{s \in \mathsf{S}_{\mathrm{grid}}} \left| \mathfrak{T}_{\star,0}(s) \ominus \mathfrak{T}_{\star,1}(s) \right|_{\max}, \qquad \mathfrak{E}_\star = \sum_{s \in \mathsf{S}_{\mathrm{grid}}} \mathfrak{T}_{\star,0}(s) \cdot \mathfrak{r}_{\mathrm{ins}}(s),$$

$$\mathfrak{J}_\star = \left[\ell(\mathfrak{E}_\star) - \frac{\delta_\star}{1-1/2},\, u(\mathfrak{E}_\star) + \frac{\delta_\star}{1-1/2}\right], \qquad \mathfrak{B} = \mathfrak{J}_{\mathrm{b}} \ominus \mathfrak{J}_{\mathrm{e}}.$$

Equivalently,

$$\mathfrak{B} = \left[\ell(\mathfrak{J}_{\mathrm{b}}) - u(\mathfrak{J}_{\mathrm{e}}),\, u(\mathfrak{J}_{\mathrm{b}}) - \ell(\mathfrak{J}_{\mathrm{e}})\right].$$

Then the endpoints of this rational interval satisfy

$$-\frac{6}{1000} < \ell(\mathfrak{B}) \qquad \textit{and} \qquad u(\mathfrak{B}) < -\frac{58}{10000}.$$

The proof is deferred to Section B.

Lemma 6 is the finite arithmetic step behind the diagnostic. It encloses the behavior- and target-stationary reward averages in rational intervals and shows that their interval difference lies strictly below zero.

The first diagnostic isolates the stationary bias of current-action weighting in this finite construction. It compares the one-step reweighted reward under the behavior stationary law with the stationary target value.

Lemma 7. [`lem:insulin-bias-certificate`] (Insulin-grid bias certificate) *For every natural horizon T, let $\Delta_{\mathrm{imm}}(T)$ be the stationary immediate-weighting bias for the refreshed finite insulin adaptation $\mathcal{I}_{\mathrm{grid}}$ of Definition 24,*

$$\Delta_{\mathrm{imm}}(T) = \sum_s d_b(s) \sum_{a\in\{0,1\}} b(s_1,a)\,\rho(s_1,a) \int p_1\, K_T(dp \mid s,a) - \theta(K_T,e),$$

where d_b is the stationary law of the behavior-policy transition kernel on the insulin grid, K_T is the refreshed insulin-grid transition kernel, b and e are its behavior and target policies, $\rho(s_1,a)$ is the one-step target-to-behavior policy ratio, and $\theta(K_T,e)$ is the target value in Definition 12. Then

$$-\frac{6}{1000} < \Delta_{\mathrm{imm}}(T) < -\frac{58}{10000}.$$

The proof is deferred to Section B.

The interval in Lemma 7 is a finite-state certificate for the immediate-weighting bias diagnostic Δ_{imm}. The theorem below records the model-class properties of the same construction and restates the diagnostic in the notation of the insulin-grid transition law.

Theorem 5. [`prop:finite-insulin-demonstration`] (Finite insulin witness) *For every horizon T, the refreshed insulin experiment $\mathcal{I}_{\mathrm{grid}}$ of Definition 24 satisfies:*

- ***(State count.)*** *Its observed–latent joint-state alphabet has cardinality $90 \cdot 4 = 360$.*
- ***(Uniform contraction.)*** *It satisfies Assumption 6 with contraction parameter $\alpha = 1/2$.*
- ***(Policy overlap.)*** *It satisfies Assumption 5 with one-step overlap constant $L = 10/3$.*
- ***(Stationary overlap.)*** *It satisfies Assumption 7 with stationary-overlap constant $C = 720$.*

Define the immediate-weighting bias diagnostic by

$$\Delta_{\mathrm{imm}}(T) = \sum_{s} d_b(s) \sum_{a\in\{0,1\}} b(a \mid s_X)\, \frac{e(a \mid s_X)}{b(a \mid s_X)} \int r\, K(ds', dr \mid s, a) - \theta(K, e),$$

where K, b, e are the transition kernel, behavior policy, and target policy of $\mathcal{I}_{\mathrm{grid}}$, s_X is the observed coordinate of s, and d_b is the stationary law induced by b. This diagnostic is interval-certified as strictly negative:

$$-0.006 < \Delta_{\mathrm{imm}}(T) < -0.0058.$$

Theorem 5 certifies five things about the grid, and exactly these five. Its joint observed–latent alphabet has 360 states. It contracts in total variation at rate $\alpha = 1/2$, so it meets Assumption 6. Its policy pair has one-step overlap constant $L = 10/3$, so it meets Assumption 5. Its two stationary joint-state laws satisfy the latent stationary overlap bound of Assumption 7 at radius $C = 720$. And its immediate-weighting diagnostic Δ_{imm} lies in the interval $(-0.006, -0.0058)$, certified by interval arithmetic.

The last of these is the substantive one, and it is what the construction was built to isolate. Immediate weighting reweights the observed reward by the current-action ratio alone, so it averages the observed-glucose reward under behavior-stationary occupancy; the target value averages the same reward under target-stationary occupancy. The certified interval quantifies the gap between these two occupancies in a model whose latent stationary overlap is finite and whose mixing is fast. Bounded latent occupancy and fast forgetting are therefore compatible with a stationary bias for current-action weighting, and the source of that bias is occupancy rather than slow mixing or a divergent action ratio. This is the mechanism the depth parameter of Definition 16 exists to control.

6 Discussion and limitations

Theorems 1, 2 and 4 describe a risk surface indexed by the stationary latent-overlap radius. For each fixed $C > 1$, the minimax risk $R_T(t_0, \zeta, C)$ has rate $T^{-\beta}$, while at $C = 1$ it has the parametric T^{-1} scale. The radius-sensitive statements refine this picture by expressing how the risk changes as C approaches one. In that local regime, the normalized radius q_C from Definition 19 is the relevant coordinate: it measures how far the target stationary occupancy may move from the behavior stationary occupancy while retaining the finite-state POMDP restrictions of Definition 10.

The elbow $q_C \asymp T^{-1/2}$ has a direct statistical interpretation. Below this scale, the local overlap distance is small enough that the parametric floor governs the risk, as recorded by Theorem 3 and the small-radius branch of Theorem 4. Above this scale, the hidden-state occupancy discrepancy becomes visible in the observable-data experiment through longer histories, and the larger-radius branch of Theorem 4 gives the corresponding degradation. The radius-calibrated history depth $\kappa_T^{\mathrm{rad}}(C)$ in Definition 20 operationalizes this tradeoff: it selects depth zero in the parametric neighborhood and increases the observable history length when the overlap radius makes latent imbalance statistically consequential.

The role of latent stationary overlap is therefore distinct from ordinary one-step action overlap. The one-step condition controls the variance of products of observable policy ratios, while the stationary density-ratio condition controls how target and behavior policies populate hidden

states. Theorem 1 is the master rate statement: Theorems 2 to 4 are the fixed-overlap, unit-overlap, and local shrinking-overlap slices. The proof balances the PHIW bias scale $q_C\alpha^k$, coming from contraction after conditioning on a length-k observed history, against the variance scale L^k/T, coming from products of one-step policy ratios. The signed-depth lower bound chooses a terminal depth Q so that the observable KL is controlled by a constant times $T\varepsilon_C^2\alpha^{2Q}L^{-Q}$, while the target value gap is calibrated to the upper-bound bias scale. The radius-calibrated depth in Definitions 16 and 20 attains the displayed frontier when calibrated with (t_0, ζ, C), and a known shrinking law for q_C gives the corresponding changing depth.

The finite insulin demonstration in Theorem 5 is a controlled finite-state policy example. The construction supplies a refreshed Markov model with finite contraction, one-step policy overlap, and the sufficient stationary density-ratio bound $C = 720$, hence $q_C = 719/720 \approx 0.9986$. Its immediate-weighting diagnostic remains separated from zero under the displayed interval arithmetic because current-action weighting averages the observed-glucose reward under behavior-stationary occupancy rather than target-stationary occupancy. The example therefore exhibits, in a stylized treatment-policy model, the occupancy mechanism that the depth parameter of the partial-history estimator is designed to control. Its certified content is the four structural constants and the bias interval; establishing the full model-class predicate for this grid — in particular fixing an initial law and verifying the stationary-start, Markov-kernel, sequential-ignorability, and bounded-reward conditions at $t_0 = 1/\log 2$ and $\zeta = \log(10/3)$ — is a separate verification we do not carry out here, so the frontier theorem is not being applied to this grid.

Limitations. The signed-depth converse in Definition 18 uses a latent alphabet whose terminal depth grows with T, so the lower-bound construction is calibrated to a sequence of finite experiments. Confidence intervals and policy-learning regret are separate research directions built on top of the value-estimation problem studied here. The finite insulin adaptation in Definition 24 is a controlled finite-state demonstration designed to instantiate the paper's overlap and contraction constants within an interpretable treatment-policy example. Three limits of that demonstration are worth stating. Its certificate covers the four structural constants and the bias interval, not the full model-class predicate, so no rate guarantee of this paper is asserted for it. It is not a finite-sample comparison of estimators, and it does not show that the depth calibrated from the sufficient bound $C = 720$ is the practically preferable depth on this grid. And a data-driven choice of depth for an unknown radius remains open here as elsewhere in the paper.

Appendices

A Proofs and auxiliary constructions

This appendix collects the auxiliary statements used to support the main rate calculations. The first group records the observable filtration and the finite-prefix quantities that enter the signed-depth likelihood comparison.

Definition 25. [`def:filter-quantities`] (Filter quantities $\mathcal{F}_{t-1}^{\mathrm{obs}}, \ell_t, q_t^v, a_t^{(Q)}$) *Define*

$$\mathcal{F}_{t-1}^{\mathrm{obs}} := \sigma(A_0, Y_0, \ldots, A_{t-1}, Y_{t-1}),$$

$$\ell_t := \min\{Q, t\},$$

$$q_t^v := \Pr_v(J_t = Q \mid \mathcal{F}_{t-1}^{\mathrm{obs}}),$$

and

$$a_t^{(Q)} := L^{-(Q-\ell_t)} \prod_{j=t-\ell_t}^{t-1} \mathbf{1}\{A_j = 1\},$$

where the empty product is equal to one.

The quantities in Definition 25 separate two sources of information in the observed history. The posterior q_t^v captures the chance that the hidden depth has reached its terminal value, while the factor $a_t^{(Q)}$ records the visible action pattern needed for that event under the rare-action construction. This separation is the device that makes the observed likelihood calculation finite-prefix and observable.

The upper-bound argument begins with local second-moment and covariance controls for the PHIW scores. These statements isolate, respectively, the contribution from a single weighted window and the dependence between overlapping windows.

Lemma 8. [`lem:phiw-window-second-moment`] (Window second moment) *Let $T, n_X, n_H \in \mathbb{N}$, let $k \in \mathbb{N}$, let $L \in \mathbb{R}$, and let M be a raw POMDP experiment on finite observed and hidden alphabets. Suppose that:*

- ***(Sequential ignorability.)*** *M satisfies Assumption 2.*
- ***(Policy overlap.)*** *M satisfies Assumption 5 with constant L.*
- ***(Overlap level.)*** *$L \geq 1$.*
- ***(Kernel law.)*** *M satisfies Assumption 1.*
- ***(Bounded rewards.)*** *M satisfies Assumption 4.*
- ***(Window position.)*** *$t \in \{0, \ldots, T-1\}$ and $k \leq t$.*

For the depth-k score $Z_t(k)$ of Definition 16, under the observed trajectory law of M,

$$\mathbb{E}_{\mathrm{obs}}\left[Z_t(k)^2\right] \leq L^{k+1}.$$

Proof of Lemma 8. 1. Write an observed trajectory as

$$o = ((x_j, a_j, y_j))_{j=0}^{T-1} \in \mathsf{O}_T,$$

and set

$$\mathcal{W}_{t,k} := \{j \in \{0, \ldots, T-1\} : j \leq t,\ t - j \leq k\}.$$

For every j, define the observable zero-cell policy ratio

$$\rho_j(o) := \begin{cases} 0, & b(a_j \mid x_j) = 0, \\ e(a_j \mid x_j)/b(a_j \mid x_j), & b(a_j \mid x_j) \neq 0. \end{cases}$$

Then Definition 16 gives

$$Z_t(k)(o) = y_t \Pi_{t,k}(o), \qquad \Pi_{t,k}(o) := \prod_{j \in \mathcal{W}_{t,k}} \rho_j(o).$$

Since $k \leq t$, the window is the contiguous block

$$\mathcal{W}_{t,k} = \{t-k, t-k+1, \ldots, t\}.$$

2. The window product has observed-law mean one:

$$\mathbb{E}_{\mathrm{obs}}\big[\Pi_{t,k}\big] = 1.$$

Put $r = t - k$. Since $k \le t$, this is an integer in $\{0, \dots, t\}$, and the window from step 1 is exactly the interval $\{r, r+1, \dots, t\}$. First establish the corresponding identity under the full trajectory law. For a full trajectory τ, write

$$\rho_j(\tau) := \begin{cases} 0, & b(A_j(\tau) \mid X_j(\tau)) = 0, \\ e(A_j(\tau) \mid X_j(\tau))/b(A_j(\tau) \mid X_j(\tau)), & b(A_j(\tau) \mid X_j(\tau)) \neq 0, \end{cases}$$

so the full-law block identity to prove is

$$\mathbb{E}\left[\prod_{j=r}^{t} \rho_j(\tau)\right] = 1.$$

The argument uses the theorem hypothesis that the target carrier is a policy vector, so for every observed state x,

$$e(a \mid x) \ge 0 \quad (a \in \mathcal{A}), \qquad \sum_{a \in \mathcal{A}} e(a \mid x) = 1,$$

together with the domination inequality supplied by Assumption 5. Under Assumption 2, conditional on the past and current joint state at epoch m, the action has behavior probabilities $b(\cdot \mid X_m)$. If $b(a \mid x) = 0$, then $0 \le e(a \mid x) \le Lb(a \mid x) = 0$, hence $e(a \mid x) = 0$; therefore, for every observed state x,

$$\sum_{a \in \mathcal{A}} b(a \mid x) \begin{cases} 0, & b(a \mid x) = 0, \\ e(a \mid x)/b(a \mid x), & b(a \mid x) \neq 0 \end{cases} = \sum_{a \in \mathcal{A}} e(a \mid x) = 1.$$

Thus, for every $m \in \{r, \dots, t\}$, conditioning on the history and current joint state up to epoch m gives

$$\mathbb{E}[\rho_m(\tau) \mid X_{0:m}, H_{0:m}, A_{0:m-1}, Y_{0:m-1}] = 1.$$

Peeling the factors one epoch at a time, with Assumption 1 identifying the next history law after each peeled action, gives

$$\mathbb{E}\left[\prod_{j=r}^{m} \rho_j(\tau)\right] = \mathbb{E}\left[\prod_{j=r}^{m-1} \rho_j(\tau)\right], \qquad m = r, \dots, t,$$

where the product over $\{r, \dots, r-1\}$ is the empty product. Iterating from $m = t$ down to $m = r$ leaves the empty product, whose expectation is 1, and hence

$$\mathbb{E}\left[\prod_{j=r}^{t} \rho_j(\tau)\right] = 1.$$

It remains to pass this identity to observed trajectories. Let O be the observation map

$$O(\tau) = ((X_j(\tau), A_j(\tau), Y_j(\tau)))_{j=0}^{T-1},$$

so that the observed law is the pushforward of the full law by O. By definition of $\Pi_{t,k}$ and the interval identity $\mathcal{W}_{t,k} = \{r, \dots, t\}$, for every full trajectory τ,

$$\Pi_{t,k}(O(\tau)) = \prod_{j \in \mathcal{W}_{t,k}} \rho_j(O(\tau)) = \prod_{j=r}^{t} \rho_j(\tau).$$

The usual integration formula for a pushforward measure therefore gives

$$\mathbb{E}_{\mathrm{obs}}\big[\Pi_{t,k}\big] = \int \Pi_{t,k}(o)\, d\mathbb{P}_{\mathrm{obs}}(o) = \int \Pi_{t,k}(O(\tau))\, d\mathbb{P}(\tau) = \int \prod_{j=r}^{t} \rho_j(\tau)\, d\mathbb{P}(\tau) = 1.$$

3. For every observed trajectory o and every j,

$$0 \le \rho_j(o) \le L.$$

The lower bound uses the behavior-policy vector property from Assumption 2 and the theorem hypothesis that the target carrier is a policy vector. The upper bound is immediate in a zero cell and otherwise follows from the domination inequality in Assumption 5 by dividing $e(a_j \mid x_j) \le L b(a_j \mid x_j)$ by the positive number $b(a_j \mid x_j)$. Hence

$$0 \le \Pi_{t,k}(o) \le L^{|\mathcal{W}_{t,k}|}.$$

The map $j \mapsto t - j$ injects $\mathcal{W}_{t,k}$ into $\{0, \dots, k\}$, so

$$|\mathcal{W}_{t,k}| \le k + 1.$$

Since $L \ge 1$,

$$0 \le \Pi_{t,k}(o) \le L^{k+1}.$$

4. By Assumption 4, the full trajectory law has $|Y_t| \le 1$ almost surely; pushing forward to the observed law gives

$$|y_t| \le 1 \qquad \mathbb{P}_{\mathrm{obs}}\text{-almost surely.}$$

Together with the preceding product bound, this also gives

$$|Z_t(k)| \le L^{k+1} \qquad \mathbb{P}_{\mathrm{obs}}\text{-almost surely,}$$

so the square and the comparison function $L^{k+1}\Pi_{t,k}$ are integrable under the observed law.

5. For $\mathbb{P}_{\mathrm{obs}}$-almost every observed trajectory,

$$Z_t(k)^2 = y_t^2 \Pi_{t,k}^2 \le 1 \cdot \Pi_{t,k}^2 \le L^{k+1} \Pi_{t,k},$$

where the last inequality uses $0 \le \Pi_{t,k} \le L^{k+1}$. Integrating and using the normalization from step 2,

$$\mathbb{E}_{\mathrm{obs}}\big[Z_t(k)^2\big] \le \mathbb{E}_{\mathrm{obs}}\big[L^{k+1}\Pi_{t,k}\big] = L^{k+1}\mathbb{E}_{\mathrm{obs}}\big[\Pi_{t,k}\big] = L^{k+1}.$$

□

Lemma 9. [lem:phiw-overlapping-window-covariance] (Overlapping-window covariance) *Let M be a raw POMDP experiment of length T on finite observed and hidden alphabets. Let $k, h \in \mathbb{N}$, let $L \in \mathbb{R}$, and let $t, u \in \{0, \dots, T-1\}$. Assume:*

- ***(Model conditions.)*** *M satisfies Assumption 2, Assumption 5 with overlap constant L, Assumption 1, and Assumption 4.*
- ***(Overlap constant.)*** *$L \geq 1$.*
- ***(Initial window.)*** *$k \leq t$.*
- ***(Lag identity.)*** *$u = t + h$.*
- ***(Overlapping lag.)*** *$1 \leq h \leq k$.*

Then the depth-k scores $Z_t(k)$ of Definition 16, evaluated under the observed trajectory law of M, satisfy

$$\left|\mathrm{Cov}_{\mathrm{obs}}\big(Z_t(k), Z_u(k)\big)\right| \leq 2L^{k+1-h}.$$

Proof of Lemma 9. 1. Let

$$\mu_{\mathrm{obs}} := \mathcal{L}_{\mathrm{obs}}(M)$$

be the observed-trajectory law of M. Since it is the pushforward of the probability law of the full raw experiment under the observation map, μ_{obs} is a probability measure. Also, from $k \leq t$, $u = t + h$, and $h \geq 1$, we have

$$k \leq u.$$

2. We use the following covariance reduction. If μ is a probability law, $\xi, \eta \in L^2(\mu)$, and a number B satisfies

$$\int |\xi|\, d\mu \leq 1, \qquad \int |\eta|\, d\mu \leq 1, \qquad \int |\xi\eta|\, d\mu \leq B, \qquad 1 \leq B,$$

then

$$|\operatorname{Cov}_\mu(\xi, \eta)| \leq 2B.$$

Indeed,

$$|\operatorname{Cov}_\mu(\xi, \eta)| \leq \int |\xi\eta|\, d\mu + \left(\int |\xi|\, d\mu\right)\left(\int |\eta|\, d\mu\right),$$

by expanding the covariance and applying the triangle inequality. The two L^1-bounds give

$$\left(\int |\xi|\, d\mu\right)\left(\int |\eta|\, d\mu\right) \leq 1,$$

because both factors are nonnegative. Hence

$$|\operatorname{Cov}_\mu(\xi, \eta)| \leq B + 1 \leq 2B.$$

3. Apply this reduction with

$$\xi(w) := Z_t(k)(w), \qquad \eta(w) := Z_u(k)(w), \qquad B := L^{k+1-h},$$

for observed trajectories w. The two scores are square-integrable under μ_{obs}. To see this, write the one-step ratio in Definition 16 as

$$\rho_j(w) := \begin{cases} e(A_j(w) \mid X_j(w))/b(A_j(w) \mid X_j(w)), & b(A_j(w) \mid X_j(w)) > 0, \\ 0, & b(A_j(w) \mid X_j(w)) = 0, \end{cases}$$

for every observed trajectory w. The policy-vector parts of Assumptions 2 and 5 give $b(a \mid x), e(a \mid x) \geq 0$. If $b(a \mid x) > 0$, Assumption 5 gives

$$0 \leq \frac{e(a \mid x)}{b(a \mid x)} \leq L.$$

If $b(a \mid x) = 0$, then $e(a \mid x) \leq Lb(a \mid x) = 0$, hence $e(a \mid x) = 0$, and the zero-cell convention sets the ratio equal to zero. Thus, for every observed trajectory w,

$$0 \leq \rho_j(w) \leq L \qquad \text{for every } j.$$

Also $|Y_r| \leq 1$ almost surely by Assumption 4. Hence, for every time r,

$$|Z_r(k)(w)| \leq \prod_{j \in \mathcal{W}_r} \rho_j(w) \leq L^{|\mathcal{W}_r|} \leq L^{k+1}, \qquad \mathcal{W}_r := \{j : j \leq r,\ r - j \leq k\},$$

outside the null set where the reward bound fails. The window $\mathcal{W}_r$ is a subset of the integer interval $\{r - k, \ldots, r\}$, so $|\mathcal{W}_r| \leq k + 1$. The map $w \mapsto Z_r(k)(w)$ is measurable: the observed trajectory space carries real rewards and is not finite, so measurability is not a counting argument, but each coordinate map $w \mapsto (X_j(w), A_j(w))$ and $w \mapsto Y_j(w)$ is measurable, the ratio ρ_j is a measurable function of the first of these by the two-case definition above, and $Z_r(k)$ is a finite product of these measurable maps. Since μ_{obs} is a probability law, the almost-sure bound above implies $Z_r(k) \in L^2(\mu_{\mathrm{obs}})$, and in particular it applies to $r = t$ and $r = u$.

Moreover, Assumption 1 and the initial-window conditions $k \leq t$ and $k \leq u$ give

$$\int |Z_t(k)| \, d\mu_{\mathrm{obs}} \leq 1, \qquad \int |Z_u(k)| \, d\mu_{\mathrm{obs}} \leq 1.$$

For a single time r with $k \leq r$, this last estimate is obtained as follows. For integers $r_{\mathrm{lo}}, r_{\mathrm{hi}}$ with $0 \leq r_{\mathrm{lo}} \leq r_{\mathrm{hi}} < T$, define

$$\Gamma_{r_{\mathrm{lo}}, r_{\mathrm{hi}}}(w) := \prod_{j=r_{\mathrm{lo}}}^{r_{\mathrm{hi}}} \rho_j(w),$$

with the empty product equal to one when the lower endpoint is above the upper endpoint. This chronological block has unit expectation:

$$\int \Gamma_{r_{\mathrm{lo}}, r_{\mathrm{hi}}}(w) \, d\mu_{\mathrm{obs}}(w) = 1.$$

To prove this, pull the integral back from μ_{obs} to the full trajectory law, so that X_j, A_j, and Y_j are full-trajectory coordinates. For $s \in \{r_{\mathrm{lo}}, \dots, r_{\mathrm{hi}}\}$, let

$$\mathscr{G}_s := \sigma(X_0, H_0, A_0, Y_0, \dots, A_{s-1}, Y_{s-1}, X_s, H_s)$$

and

$$\Gamma^-_{r_{\mathrm{lo}},s} := \prod_{j=r_{\mathrm{lo}}}^{s-1} \rho_j,$$

where the product is one for $s = r_{\mathrm{lo}}$. The factor $\Gamma^-_{r_{\mathrm{lo}},s}$ is $\mathscr{G}_s$-measurable. By Assumption 2, conditionally on $\mathscr{G}_s$ the action A_s has law $b(\cdot \mid X_s)$. The zero-cell convention and the implication $b(a \mid x) = 0 \Rightarrow e(a \mid x) = 0$ computed above give

$$\sum_{a \in \mathcal{A}} b(a \mid X_s)\rho(a, X_s) = \sum_{a \in \mathcal{A}} e(a \mid X_s) = 1,$$

and hence

$$\mathbb{E}\big[\Gamma^-_{r_{\mathrm{lo}},s}\rho_s \mid \mathscr{G}_s\big] = \Gamma^-_{r_{\mathrm{lo}},s}.$$

The remaining coordinates after the action at time s are integrated with the probability kernel in Assumption 1; integrating first over the fibre containing (Y_s, X_{s+1}, H_{s+1}) preserves the expectation of any function of $\mathscr{G}_s$. Applying the displayed conditional identity successively for $s = r_{\mathrm{hi}}, r_{\mathrm{hi}} - 1, \dots, r_{\mathrm{lo}}$ removes the factors one at a time and leaves the empty product, whose expectation is one. Taking $r_{\mathrm{lo}} = r - k$ and $r_{\mathrm{hi}} = r$ gives

$$\int \prod_{j=r-k}^{r} \rho_j(w)\, d\mu_{\mathrm{obs}}(w) = 1.$$

Now

$$Z_r(k)(w) = Y_r(w) \prod_{j=r-k}^{r} \rho_j(w),$$

and $|Y_r| \le 1$ almost surely by Assumption 4. Therefore

$$\int |Z_r(k)|\, d\mu_{\mathrm{obs}} \le \int \prod_{j=r-k}^{r} \rho_j(w)\, d\mu_{\mathrm{obs}}(w) = 1.$$

4. It remains to bound the cross moment. For the two overlapping windows define

$$\mathcal{W}_t := \{j \in \{0, \dots, T-1\} : t-k \le j \le t\}, \qquad \mathcal{W}_u := \{j \in \{0, \dots, T-1\} : u-k \le j \le u\}, \qquad \mathcal{W}_\cup := \mathcal{W}_t \cup \mathcal{W}$$

Since t and u are elements of $\{0, \dots, T-1\}$, their integer values are valid finite-time endpoints. The assumptions $k \le t$, $u = t + h$, and $h \le k$ give $t - k \le u$ and identify the union in the formal finite index set as

$$\mathcal{W}_t \cup \mathcal{W}_u = \{j \in \{0, \dots, T-1\} : t - k \le j \le u\}.$$

The same endpoint inequalities give the cardinality bound for the overlap,

$$|\mathcal{W}_t \cap \mathcal{W}_u| \le k + 1 - h.$$

Using the same zero-cell ratio ρ_j as above, the policy-vector and overlap parts of Assumptions 2 and 5, together with $L \geq 1$, give

$$0 \leq \rho_j(w) \leq L \qquad \text{for every } j, w.$$

Therefore

$$\prod_{j \in \mathcal{W}_t \cap \mathcal{W}_u} \rho_j(w) \leq L^{|\mathcal{W}_t \cap \mathcal{W}_u|} \leq L^{k+1-h},$$

the second inequality by the cardinality bound $|\mathcal{W}_t \cap \mathcal{W}_u| \leq k+1-h$ together with the monotonicity of $n \mapsto L^n$ for $L \geq 1$; the intersection need not have exactly $k+1-h$ indices, so this step is an inequality and not an identity.

5. Since $|Y_t| \leq 1$ and $|Y_u| \leq 1$ almost surely,

$$|Z_t(k) Z_u(k)| \leq \left(\prod_{j \in \mathcal{W}_t} \rho_j \right) \left(\prod_{j \in \mathcal{W}_u} \rho_j \right).$$

The product over two windows decomposes into the product over the union times the product over the overlap:

$$\left(\prod_{j \in \mathcal{W}_t} \rho_j \right) \left(\prod_{j \in \mathcal{W}_u} \rho_j \right) = \left(\prod_{j \in \mathcal{W}_\cup} \rho_j \right) \left(\prod_{j \in \mathcal{W}_t \cap \mathcal{W}_u} \rho_j \right).$$

Combining this with the previous display gives the pointwise almost-sure bound

$$|Z_t(k) Z_u(k)| \leq L^{k+1-h} \prod_{j \in \mathcal{W}_\cup} \rho_j.$$

The finite-window identity from the preceding step rewrites the union product in exactly chronological form:

$$\mathcal{W}_\cup = \{j \in \{0, \ldots, T-1\} : t-k \leq j \leq u\}, \qquad \prod_{j \in \mathcal{W}_\cup} \rho_j(w) = \prod_{j=t-k}^{u} \rho_j(w).$$

The interval-normalization argument proved above applies with lower endpoint $t-k$ and upper endpoint u. The lower endpoint is nonnegative because $k \leq t$, the upper endpoint is in $\{0, \ldots, T-1\}$ because u is a time index, and $t-k \leq u$ follows from $u = t+h$. Thus

$$\int \prod_{j \in \mathcal{W}_\cup} \rho_j(w) \, d\mu_{\text{obs}}(w) = \int \prod_{j=t-k}^{u} \rho_j(w) \, d\mu_{\text{obs}}(w) = 1.$$

Hence

$$\int |Z_t(k) Z_u(k)| \, d\mu_{\text{obs}} \leq L^{k+1-h}.$$

6. Finally,

$$1 \leq L^{k+1-h},$$

because $L \geq 1$. The covariance reduction from Step 2, applied with $B = L^{k+1-h}$, yields

$$\left| \text{Cov}_{\text{obs}}(Z_t(k), Z_u(k)) \right| \leq 2L^{k+1-h},$$

which is the claimed bound.

□

The analysis of separated windows uses both the observed law and the full trajectory law. The next definitions fix those probability spaces and the observation map connecting them.

Definition 26. [synth_15] (Observed-trajectory law) *For a raw POMDP experiment* $M = (\mathcal{X}, \mathcal{H}, K, b, e, \mathcal{L}(\mathsf{O}_T))$*, let*

$$\mathbb{P}_M^{\mathrm{obs}} := \mathrm{Law}_M^{\mathrm{obs}}$$

denote the probability law of the observed trajectory

$$\mathsf{O}_T = ((X_t, A_t, Y_t))_{t=0}^{T-1}$$

induced by the behavior-policy experiment M*. Equivalently, under* $\mathbb{P}_M^{\mathrm{obs}}$*,* $S_0 \sim d_b$*,* $A_t \sim b(\cdot \mid X_t)$*, and* K *generates* (Y_t, S_{t+1}) *from* (S_t, A_t)*, with only* $(X_t, A_t, Y_t)_{t=0}^{T-1}$ *retained.*

Definition 27. [synth_17] (Full trajectory law and observation map) *For a raw POMDP experiment* $M = (\mathcal{X}, \mathcal{H}, K, b, e, \mathcal{L}(\mathsf{O}_T))$ *with joint state space* $\mathcal{S} = \mathcal{X} \times \mathcal{H}$*, let* $\mathbb{P}_M$ *denote the full behavior-policy trajectory law on*

$$(S_0, A_0, Y_0, S_1, \ldots, A_{T-1}, Y_{T-1}, S_T).$$

Under $\mathbb{P}_M$*, the initial state satisfies* $S_0 \sim d_b$*, each action is drawn as* $A_t \sim b(\cdot \mid X_t)$*, and the kernel* K *generates* (Y_t, S_{t+1}) *from* (S_t, A_t)*, for* $t = 0, \ldots, T-1$*.*

Let O *be the observation map from a full trajectory to its observed trajectory,*

$$O(S_0, A_0, Y_0, \ldots, A_{T-1}, Y_{T-1}, S_T) := ((X_t, A_t, Y_t))_{t=0}^{T-1},$$

where $S_t = (X_t, H_t)$*. The observed behavior-policy law is the pushforward*

$$\mathbb{P}_M^{\mathrm{obs}} := \mathbb{P}_M \circ O^{-1}.$$

Together, Definitions 26 and 27 let the covariance calculation move between observable scores and latent-state transition operators. The observed law is the sampling law for the estimator, while the full law retains the latent state needed to apply contraction after a gap.

Lemma 10. [lem:phiw-separated-window-peeling] (Separated window peeling) *Let* M *be a raw POMDP experiment with trajectory length* T*, and let* L*, the policy-overlap constant, be real. Let* $Z_s(k)$ *denote the depth-*k *partial-history weighted reward score from Definition 16. Suppose:*

- ***(Model conditions.)*** M *satisfies Assumptions 1, 2, 4 and 5 with overlap constant* L*.*
- ***(Overlap scale.)*** $1 \le L$*.*
- ***(Mature earlier window.)*** *The history depth* k *satisfies* $k \le t$*.*
- ***(Separated future window.)*** *For an epoch* t *and a gap* $\gamma \in \mathbb{N}$*, the epoch*

$$u = t + 1 + \gamma + k$$

lies in the observed horizon, equivalently $t + 1 + \gamma + k < T$*.*

For $p \in \{b, e\}$, let P_p be the policy-induced transition operator obtained from the reward-marginalized kernel under policy p, let

$$g_e(s) = \sum_a e(a \mid x(s)) \int y\, K(dy \times \mathcal{S} \mid s, a),$$

and put

$$\Phi_{\gamma,k}(s) = P_b^\gamma P_e^k g_e(s).$$

Then, with $\mathbb{P}_M^{\mathrm{obs}}$ the observed law induced by the observed trajectory map O,

$$\int Z_t(k)(w) Z_u(k)(w)\, d\mathbb{P}_M^{\mathrm{obs}}(w) = \int Z_t(k)(O(\tau))\, \Phi_{\gamma,k}(S_{t+1}(\tau))\, d\mathbb{P}_M(\tau).$$

Proof of Lemma 10. Set

$$\iota_0 = t + 1, \qquad \iota_1 = t + 1 + \gamma, \qquad \iota_2 = \iota_1 + k = u.$$

The horizon condition gives $\iota_2 < T$, and $k \le t$ makes the earlier score mature. For a full trajectory τ, write the joint state at epoch m as

$$S_m(\tau) := (X_m(\tau), H_m(\tau))$$

and write

$$\mathrm{hist}_m(\tau) = \big((S_j(\tau))_{j<m}, (A_j(\tau), Y_j(\tau))_{j<m}, S_m(\tau)\big)$$

for the state-and-epoch history through epoch m. For an observed trajectory w, write its epoch-j coordinates as

$$w_j = (x_j(w), a_j(w), y_j(w)).$$

Define the totalized one-step ratio, for an observed state x and action a, by

$$\rho(a \mid x) := \begin{cases} 0, & b(a \mid x) = 0, \\ e(a \mid x)/b(a \mid x), & b(a \mid x) > 0, \end{cases} \qquad \rho_j(w) := \rho(a_j(w) \mid x_j(w)).$$

For $0 \le j_{\mathrm{lo}} \le j_{\mathrm{hi}} \le T - 1$, put

$$\mathrm{rat}^{<}_{j_{\mathrm{lo}}:j_{\mathrm{hi}}}(\tau) := \prod_{j=j_{\mathrm{lo}}}^{j_{\mathrm{hi}}-1} \rho_j(O(\tau)), \qquad \mathrm{rat}^{<}_{j_{\mathrm{hi}}:j_{\mathrm{hi}}}(\tau) := 1.$$

Thus the superscript $<$ means that the upper endpoint is excluded: the largest ratio in a nonempty block is $\rho_{j_{\mathrm{hi}}-1}$. In particular, $\mathrm{rat}^{<}_{\iota_1:\iota_2}$ stops at ρ_{u-1}, while the terminal factor ρ_u is written separately below. The target-policy one-step reward regression is

$$g_e(s) = \sum_{a \in \mathcal{A}} e(a \mid x(s)) \int y\, K(dy \times \mathcal{S} \mid s, a),$$

and the separated future function is $\Phi_{\gamma,k} = P_b^\gamma P_e^k g_e$. In this proof the policy-induced transition operator is the reward-marginalized kernel

$$(P_p F)(s) = \sum_{a \in \mathcal{A}} p(a \mid x(s)) \int F(s')\, K(dy \times ds' \mid s, a), \qquad p \in \{b, e\},$$

for every bounded $F : \mathcal{S} \to \mathbb{R}$.

1. Since $\mathbb{P}_M^{\mathrm{obs}}$ is the pushforward of $\mathbb{P}_M$ under O,

$$\int Z_t(k)(w)Z_u(k)(w)\,d\mathbb{P}_M^{\mathrm{obs}}(w) = \int Z_t(k)(O(\tau))Z_u(k)(O(\tau))\,d\mathbb{P}_M(\tau).$$

The measurability needed for this change of variables follows from the observable definition of $Z_t(k)$ in Definition 16.

2. Define lifted carriers on state histories by

$$\Lambda_0(\mathrm{hist}_{\iota_0}(\tau)) := Z_t(k)(O(\tau)),$$

and, for $0 \le \nu < \gamma + k$, let $\Lambda_{\nu+1}$ be the lift of Λ_ν that forgets the last appended action, reward, and next state:

$$\Lambda_{\nu+1}(\mathrm{hist}_{\iota_0+\nu+1}(\tau)) := \Lambda_\nu(\mathrm{hist}_{\iota_0+\nu}(\tau)).$$

The base carrier is well defined because $Z_t(k)$ uses only rewards, actions, and observed states through epoch t, while $t < \iota_0$; hence every lifted carrier satisfies

$$\Lambda_\nu(\mathrm{hist}_{\iota_0+\nu}(\tau)) = Z_t(k)(O(\tau)) \qquad (0 \le \nu \le \gamma + k).$$

The depth-k window in $Z_u(k)$ is inclusive. Since $u = \iota_2$ and $u-k = \iota_1$, the half-open block contains the factors with $\iota_1 \le j < \iota_2$, and the endpoint contributes the separate terminal factor:

$$Z_u(k)(O(\tau)) = Y_{\iota_2}(\tau)\ \mathrm{rat}^{<}_{\iota_1:\iota_2}(\tau)\,\rho_{\iota_2}(O(\tau)).$$

Therefore

$$\int Z_t(k)(O(\tau))Z_u(k)(O(\tau))\,d\mathbb{P}_M(\tau) = \int \Lambda_{\gamma+k}(\mathrm{hist}_{\iota_2}(\tau))\ \mathrm{rat}^{<}_{\iota_1:\iota_2}(\tau)\,\rho_{\iota_2}(O(\tau))Y_{\iota_2}(\tau)\,d\mathbb{P}_M(\tau).$$

The totalized ratios obey $0 \le \rho(a \mid x) \le L$: on a positive behavior cell this is the domination inequality divided by $b(a \mid x) > 0$, and on a zero cell it follows from the definition of ρ. Hence finite state spaces, Assumption 4, and the finite product bound $\prod_j \rho_j \le L^{\#\{j\}}$ make all displayed integrands integrable. Empty behavior or target blocks are interpreted as products equal to one; in particular, when $k = 0$ the block $\mathrm{rat}^{<}_{\iota_1:\iota_2}$ is empty and the endpoint factor $\rho_{\iota_2}(O(\tau))$ is still present.

3. The terminal reward peel is

$$\begin{aligned}\int \Lambda_{\gamma+k}(\mathrm{hist}_{\iota_2})\ \mathrm{rat}^{<}_{\iota_1:\iota_2}\ \rho_{\iota_2}Y_{\iota_2}\,d\mathbb{P}_M \\ = \int \Lambda_{\gamma+k}(\mathrm{hist}_{\iota_2})\ \mathrm{rat}^{<}_{\iota_1:\iota_2}\ g_e(S_{\iota_2})\,d\mathbb{P}_M.\end{aligned}$$

Indeed, condition first on the history and action at ι_2. Assumption 1 supplies the conditional law of $(Y_{\iota_2}, S_{\iota_2+1})$ as $K(\cdot \mid S_{\iota_2}, a)$. Applying this conditional-law identity to the measurable test function $(y, s') \mapsto y$, with integrability supplied by Assumption 4, and then marginalizing over the successor state gives

$$E[Y_{\iota_2} \mid \mathrm{hist}_{\iota_2}, A_{\iota_2} = a] = \int y\,K(dy \times \mathcal{S} \mid S_{\iota_2}, a).$$

By Assumption 2, the action is drawn from $b(\cdot \mid X_{\iota_2})$. For every observed state x and action a, the totalized ratio satisfies

$$b(a \mid x)\rho(a \mid x) = e(a \mid x).$$

Indeed, if $b(a \mid x) > 0$, this is the definition of ρ. If $b(a \mid x) = 0$, then the domination inequality in Assumption 5 gives $e(a \mid x) \le 0$, while the target policy has nonnegative coordinates, so $e(a \mid x) = 0$, matching the zero-cell value. Applying this identity at $x = X_{\iota_2}$ and summing over a gives

$$\sum_{a \in \mathcal{A}} e(a \mid X_{\iota_2}) \int y\, K(dy \times \mathcal{S} \mid S_{\iota_2}, a) = g_e(S_{\iota_2}),$$

by the displayed definition of g_e.

4. For every $\iota_1 \le m < \iota_2$, put $\nu = m - \iota_0$, so $\gamma \le \nu < \gamma + k$. For every bounded $F : \mathcal{S} \to \mathbb{R}$, the target-ratio peeling step is

$$\begin{aligned} &\int \Lambda_{\nu+1}(\mathrm{hist}_{m+1})\ \mathrm{rat}^{<}_{\iota_1:m+1}\ F(S_{m+1})\, d\mathbb{P}_M \\ &\qquad = \int \Lambda_{\nu}(\mathrm{hist}_{m})\ \mathrm{rat}^{<}_{\iota_1:m}\ (P_e F)(S_m)\, d\mathbb{P}_M. \end{aligned}$$

The first target peel has $m = \iota_1$ and lowers the carrier from $\Lambda_{\gamma+1}$ to Λ_γ; the last has $m = \iota_2 - 1$ and lowers $\Lambda_{\gamma+k}$ to $\Lambda_{\gamma+k-1}$. Iterating the displayed identity backward over these values of m therefore removes exactly the k target-window ratios. Indeed, $\Lambda_{\nu+1}$ is the lift of Λ_ν, and

$$\mathrm{rat}^{<}_{\iota_1:m+1} = \mathrm{rat}^{<}_{\iota_1:m}\, \rho_m.$$

Conditioning on the history and action at m, Assumption 1 gives the transition integral over (Y_m, S_{m+1}), and the reward coordinate is integrated out:

$$\sum_{a \in \mathcal{A}} b(a \mid X_m)\rho(a \mid X_m) \int F(s')\, K(dy \times ds' \mid S_m, a) = \sum_{a \in \mathcal{A}} e(a \mid X_m) \int F(s')\, K(dy \times ds' \mid S_m, a).$$

The ratio identity displayed in the terminal peel justifies the replacement of $b\rho$ by e, including zero behavior cells, and the last display is exactly $(P_e F)(S_m)$. Iterating for $m = \iota_2 - 1, \dots, \iota_1$, starting from $F = g_e$, yields

$$\int \Lambda_{\gamma+k}(\mathrm{hist}_{\iota_2})\ \mathrm{rat}^{<}_{\iota_1:\iota_2}\ g_e(S_{\iota_2})\, d\mathbb{P}_M = \int \Lambda_{\gamma}(\mathrm{hist}_{\iota_1})\, (P_e^k g_e)(S_{\iota_1})\, d\mathbb{P}_M.$$

When $k = 0$, this iteration is empty and the displayed identity reduces to the same integrand at $\iota_1 = \iota_2$.

5. For every $\iota_0 \le m < \iota_1$, put $\nu = m - \iota_0$. For every bounded $F : \mathcal{S} \to \mathbb{R}$, the behavior-gap peeling step is

$$\int \Lambda_{\nu+1}(\mathrm{hist}_{m+1}) F(S_{m+1})\, d\mathbb{P}_M = \int \Lambda_{\nu}(\mathrm{hist}_{m})(P_b F)(S_m)\, d\mathbb{P}_M.$$

Here hist_{m+1} is the history obtained from hist_m by appending (A_m, Y_m, S_{m+1}), and $\Lambda_{\nu+1}$ is the lift of Λ_ν, so $\Lambda_{\nu+1}(\mathrm{hist}_{m+1}) = \Lambda_\nu(\mathrm{hist}_m)$. Conditioning on the history and action at

m, Assumption 2 supplies the behavior action law, while Assumption 1 supplies the joint law of (Y_m, S_{m+1}) given (S_m, A_m). Integrating out the reward coordinate gives, for each history through m,

$$\sum_{a\in\mathcal{A}} b(a \mid X_m) \int F(s')\, K(dy \times ds' \mid S_m, a) = (P_b F)(S_m),$$

which is the reward-marginalized behavior transition appearing in the displayed identity. Iterating over $m = \iota_1 - 1, \dots, \iota_0$ with $F = P_e^k g_e$ gives

$$\int \Lambda_\gamma(\mathrm{hist}_{\iota_1})(P_e^k g_e)(S_{\iota_1})\, d\mathbb{P}_M = \int \Lambda_0(\mathrm{hist}_{\iota_0})(P_b^\gamma P_e^k g_e)(S_{\iota_0})\, d\mathbb{P}_M.$$

When $\gamma = 0$, this iteration is empty. Since $\Lambda_0(\mathrm{hist}_{\iota_0}(\tau)) = Z_t(k)(O(\tau))$, $S_{\iota_0} = S_{t+1}$, and $\Phi_{\gamma,k} = P_b^\gamma P_e^k g_e$, combining the preceding displays proves

$$\int Z_t(k)(w) Z_u(k)(w)\, d\mathbb{P}_M^{\mathrm{obs}}(w) = \int Z_t(k)(O(\tau))\, \Phi_{\gamma,k}(S_{t+1}(\tau))\, d\mathbb{P}_M(\tau).$$

□

Lemma 10 rewrites a separated score product as an earlier observable score multiplied by a future continuation value evaluated at S_{t+1}. This identity supplies the point where the gap between windows becomes a behavior-transition power, followed by a target-policy window.

Lemma 11. [lem:behavior-support-operator-congruence] (Behavior-support operator congruence) *Let $M \in \mathcal{M}_T(t_0, \zeta, C)$ satisfy Definition 10, with behavior policy b, target policy e, and the policy-overlap condition in Assumption 5. For $p \in \{b, e\}$, write*

$$P_p(s, s') = \sum_{a\in\{0,1\}} p(a \mid x(s))\, K(\{S' = s'\} \mid s, a),$$

and let $P_p^0 h = h$ and $P_p^{m+1} h = P_p(P_p^m h)$. Let d_b denote the behavior stationary law of P_b.

Assume:

- ***(Policy choice.)*** *The policy p is either b or e.*
- ***(Terminal agreement.)*** *The functions $f, g : \mathcal{S} \to \mathbb{R}$ satisfy*

  $$f(s) = g(s)$$

  for every joint state s with $d_b(s) > 0$.
- ***(Initial support.)*** *The joint state s satisfies $d_b(s) > 0$.*

Then, for every $m \in \mathbb{N}$,

$$(P_p^m f)(s) = (P_p^m g)(s).$$

Proof of Lemma 11. Fix a policy $p \in \{b, e\}$. The argument first records the support-closure fact used at the induction step. For joint states $\sigma, \tau \in \mathcal{S}$ and an action $a \in \{0, 1\}$, write

$$\kappa_{\sigma,\tau}(a) := K(\{S' = \tau\} \mid \sigma, a).$$

Then

$$P_p(\sigma, \tau) = \sum_{a\in\{0,1\}} p(a \mid x(\sigma))\, \kappa_{\sigma,\tau}(a).$$

1. Suppose $d_b(\sigma) > 0$ and $P_p(\sigma, \tau) > 0$. We claim that

$$d_b(\tau) > 0.$$

First show that $P_b(\sigma, \tau) > 0$. If $p = b$, this is immediate. If $p = e$, suppose instead that $P_b(\sigma, \tau) = 0$. By Definition 10, the behavior policy is a probability vector and K is a Markov kernel, so every summand in

$$P_b(\sigma, \tau) = \sum_{a \in \{0,1\}} b(a \mid x(\sigma))\, \kappa_{\sigma,\tau}(a)$$

is nonnegative. Hence, for each action a,

$$b(a \mid x(\sigma))\, \kappa_{\sigma,\tau}(a) = 0.$$

For such an a, if $\kappa_{\sigma,\tau}(a) = 0$ then

$$e(a \mid x(\sigma))\, \kappa_{\sigma,\tau}(a) = 0.$$

If $\kappa_{\sigma,\tau}(a) > 0$, the preceding zero product gives $b(a \mid x(\sigma)) = 0$; by Assumption 5,

$$e(a \mid x(\sigma)) \leq L\, b(a \mid x(\sigma)) = 0,$$

and the target policy is nonnegative by Definition 10, so $e(a \mid x(\sigma)) = 0$. Thus again

$$e(a \mid x(\sigma))\, \kappa_{\sigma,\tau}(a) = 0.$$

Summing over a gives

$$P_e(\sigma, \tau) = \sum_{a \in \{0,1\}} e(a \mid x(\sigma))\, \kappa_{\sigma,\tau}(a) = 0,$$

contradicting $P_p(\sigma, \tau) > 0$ in the case $p = e$. Therefore $P_b(\sigma, \tau) > 0$.

By Definition 10, d_b is stationary for P_b, so

$$d_b(\tau) = \sum_{\omega \in \mathcal{S}} d_b(\omega) P_b(\omega, \tau).$$

All terms in this sum are nonnegative. Therefore

$$0 < d_b(\sigma) P_b(\sigma, \tau) \leq \sum_{\omega \in \mathcal{S}} d_b(\omega) P_b(\omega, \tau) = d_b(\tau),$$

which proves the claim.

2. We now prove the asserted identity by induction on m. For $m = 0$,

$$(P_p^0 f)(s) = f(s) = g(s) = (P_p^0 g)(s),$$

because $d_b(s) > 0$ and $f = g$ on the behavior support.

Assume the statement holds for m: for every joint state σ with $d_b(\sigma) > 0$,

$$(P_p^m f)(\sigma) = (P_p^m g)(\sigma).$$

Let s satisfy $d_b(s) > 0$. Using the recursive definition of the iterated operator,

$$(P_p^{m+1} f)(s) = \sum_{\tau \in \mathcal{S}} P_p(s, \tau)(P_p^m f)(\tau), \qquad (P_p^{m+1} g)(s) = \sum_{\tau \in \mathcal{S}} P_p(s, \tau)(P_p^m g)(\tau).$$

Fix $\tau \in \mathcal{S}$. If $P_p(s, \tau) = 0$, then the two τ-summands are both zero. If $P_p(s, \tau) \neq 0$, the row nonnegativity of P_p gives $P_p(s, \tau) > 0$. The support-closure claim from the previous step then gives $d_b(\tau) > 0$, so the induction hypothesis yields

$$(P_p^m f)(\tau) = (P_p^m g)(\tau).$$

Multiplying by $P_p(s, \tau)$, the two τ-summands agree. Since this holds for every τ, the sums agree:

$$(P_p^{m+1} f)(s) = (P_p^{m+1} g)(s).$$

The induction proves the result for every $m \in \mathbb{N}$.

□

The support congruence in Lemma 11 keeps transition-operator calculations on the behavior support. Because the overlap condition makes target transitions compatible with behavior support, functions that agree on the behavior-stationary support continue to agree after iterating either policy operator from a supported state.

Lemma 12. [`lem:phiw-separated-window-covariance`] (Separated window covariance) *Let $T, n_X, n_H, k, h \in \mathbb{N}$, let $t_0, \zeta, C \in \mathbb{R}$, and let $M \in \mathcal{M}_T(t_0, \zeta, C)$ be a raw POMDP experiment as in Definition 10, with the uniform contraction restriction of Assumption 6. Write the contraction factor as $\alpha = e^{-1/t_0}$. Let $t, u \in \{0, \dots, T-1\}$ satisfy*

$$k \leq t, \qquad u = t + h, \qquad k + 1 \leq h.$$

Then the depth-k scores $Z_t(k)$ of Definition 16, evaluated under the observed trajectory law of M, satisfy

$$\left|\mathrm{Cov}_{\mathrm{obs}}\big(Z_t(k), Z_u(k)\big)\right| \leq 2\alpha^{h-k-1}.$$

Proof of Lemma 12. Put

$$\gamma := h - k - 1.$$

Then $0 \leq \gamma$ by $k + 1 \leq h$, and

$$u = t + h = t + 1 + \gamma + k.$$

Let $\mathbb{P}_M$ denote the full trajectory law and $\mathbb{P}_M^{\mathrm{obs}}$ its pushforward under the observation map O. For $p \in \{b, e\}$, write

$$P_p(s, s') = \sum_{a \in \mathcal{A}} p(a \mid x(s)) K(\{S' = s'\} \mid s, a), \qquad g_e(s) = \sum_{a \in \mathcal{A}} e(a \mid x(s)) \int y \, K(dy \times \mathcal{S} \mid s, a),$$

and define

$$\Phi_{\gamma,k}(s) := P_b^\gamma P_e^k g_e(s).$$

1. Since $L = \exp(\zeta)$ and $\zeta > 0$ in Definition 10, $1 \le L$. The maturity and separation hypotheses give $t + 1 + \gamma + k < T$, so Lemma 10 applies and yields

$$\int Z_t(k)(w) Z_u(k)(w)\, d\mathbb{P}_M^{\mathrm{obs}}(w) = \int Z_t(k)(O(\tau))\, \Phi_{\gamma,k}(S_{t+1}(\tau))\, d\mathbb{P}_M(\tau).$$

For the companion mean identity, let $r_+ := t + 1$ and $q_\gamma := r_+ + \gamma$, so $u = q_\gamma + k$. Since $u = q_\gamma + k$, the depth-k window in $Z_u(k)$ is exactly $q_\gamma, q_\gamma + 1, \ldots, u$. We claim

$$\int Z_u(k)(w)\, d\mathbb{P}_M^{\mathrm{obs}}(w) = \int \Phi_{\gamma,k}(S_{t+1}(\tau))\, d\mathbb{P}_M(\tau).$$

To prove this identity, condition first on S_{q_γ}. Write

$$\rho(a, x) := \begin{cases} e(a \mid x)/b(a \mid x), & b(a \mid x) > 0, \\ 0, & b(a \mid x) = 0. \end{cases}$$

The zero-cell convention and Assumption 5 imply $e(a \mid x) = 0$ whenever $b(a \mid x) = 0$. Hence, for each pre-terminal epoch $m_{\mathrm{win}} \in \{q_\gamma, \ldots, u - 1\}$,

$$\sum_{a \in \mathcal{A}} b(a \mid x(s))\rho(a, x(s))K(\{S' = s'\} \mid s, a) = \sum_{a \in \mathcal{A}} e(a \mid x(s))K(\{S' = s'\} \mid s, a) = P_e(s, s'),$$

while at the terminal epoch

$$\sum_{a \in \mathcal{A}} b(a \mid x(s))\rho(a, x(s)) \int y\, K(dy \times \mathcal{S} \mid s, a) = \sum_{a \in \mathcal{A}} e(a \mid x(s)) \int y\, K(dy \times \mathcal{S} \mid s, a) = g_e(s).$$

Thus the target-weighted terminal window has conditional mean $P_e^k g_e(S_{q_\gamma})$, with the terminal display alone applying when $k = 0$. Conditioning backward over the gap epochs $t+1, \ldots, q_\gamma - 1$, no score ratio appears, and Assumptions 1 and 2 give the behavior propagation P_b^γ; when $\gamma = 0$, this is the identity operator. Bounded rewards and the finite state and action spaces justify the displayed integrals. Therefore the conditional mean given S_{t+1} is $P_b^\gamma P_e^k g_e(S_{t+1}) = \Phi_{\gamma,k}(S_{t+1})$, proving the claim. Also, by definition of $\mathbb{P}_M^{\mathrm{obs}}$,

$$\int Z_t(k)(w)\, d\mathbb{P}_M^{\mathrm{obs}}(w) = \int Z_t(k)(O(\tau))\, d\mathbb{P}_M(\tau).$$

Thus the covariance equals the centered full-law product

$$\mathrm{Cov}_{\mathrm{obs}}(Z_t(k), Z_u(k)) = \int Z_t(k)(O(\tau))\, \Phi_{\gamma,k}(S_{t+1}(\tau))\, d\mathbb{P}_M(\tau) - \Big(\int Z_t(k)(O(\tau))\, d\mathbb{P}_M(\tau)\Big)\Big(\int \Phi_{\gamma,k}(S_{t+1}(\tau))\, d$$

The covariance identity is legitimate because

$$Z_t(k), Z_u(k) \in L^2(\mathbb{P}_M^{\mathrm{obs}}).$$

Indeed, write the window product in $Z_m(k)$ as

$$R_m^{(k)}(w) := \prod_{j \in \{0,\ldots,T-1\}:\, j \le m,\ m-j \le k} \rho_j(w), \qquad m \in \{t, u\}.$$

The product contains at most $k + 1$ factors. The zero-cell convention and Assumption 5 make each factor nonnegative and at most L on observed behavior-support cells, while Assumption 4 gives $|Y_m| \le 1$ almost surely. Hence $|Z_m(k)| \le L^{k+1}$ almost surely under $\mathbb{P}_M^{\mathrm{obs}}$, and the two scores are square-integrable.

2. The first score has unit L^1 norm:

$$\int |Z_t(k)(O(\tau))|\, d\mathbb{P}_M(\tau) \le 1.$$

Indeed, with

$$R_t^{(k)}(w) := \prod_{j=t-k}^{t} \rho_j(w),$$

the zero-cell convention and Assumption 5 give

$$\mathbb{E}_{\mathrm{obs}}\left[R_t^{(k)}\right] = 1$$

by peeling the ratios from $j = t$ down to $j = t - k$: conditionally at each peeled epoch,

$$\sum_{a\in\mathcal{A}} b(a \mid X_j)\rho_j(a, X_j) = \sum_{a:\, b(a|X_j)>0} e(a \mid X_j) = \sum_{a\in\mathcal{A}} e(a \mid X_j) = 1.$$

By Assumption 4, $|Z_t(k)| \le R_t^{(k)}$ almost surely, which proves the displayed L^1 bound after passing between the observed law and the full law by the observation map.

3. Let d_b be the stationary law of P_b, and define the support-truncated regression on all of $\mathcal{S}$ by

$$\widetilde{g}_e(s) := \begin{cases} g_e(s), & d_b(s) > 0, \\ 0, & d_b(s) = 0. \end{cases}$$

For every s with $d_b(s) > 0$,

$$|g_e(s)| \le 1.$$

To see this, fix such an s. For an action a with $b(a \mid x(s)) > 0$, consider epoch 0. The history before epoch 0 is the unique empty history, so the full history-action singleton with current state s and action a has probability

$$d_b(s)b(a \mid x(s)) > 0,$$

by Assumption 3 and the epoch-zero action factorization in Assumption 2. Under the history-next-pair factorization supplied by Assumption 1, Assumption 4 gives, for almost every history-action pair, a kernel row whose reward coordinate is almost surely in $[-1, 1]$. Since this epoch-zero singleton has positive mass, that almost-sure row statement holds at (s, a), and hence

$$\left|\int y\, K(dy \times \mathcal{S} \mid s, a)\right| \le 1.$$

If $e(a \mid x(s)) > 0$, then Assumption 5 forces $b(a \mid x(s)) > 0$. Therefore

$$|g_e(s)| \le \sum_{a\in\mathcal{A}} e(a \mid x(s)) \left|\int y\, K(dy \times \mathcal{S} \mid s, a)\right| \le \sum_{a\in\mathcal{A}} e(a \mid x(s)) = 1.$$

For $d_b(s) = 0$, the definition gives $|\widetilde{g}_e(s)| = 0$. Consequently

$$\sup_{s,s'} |\widetilde{g}_e(s) - \widetilde{g}_e(s')| \le 2.$$

4. Since $t_0 > 0$, $\alpha = \exp(-1/t_0)$ satisfies $0 \le \alpha \le 1$. For a function $f : \mathcal{S} \to \mathbb{R}$, write

$$\operatorname{osc}(f) := \sup_{s,s' \in \mathcal{S}} |f(s) - f(s')|.$$

For each $p_\star \in \{b, e\}$, the rows of $P_{p_\star}$ are probability vectors. Nonnegativity follows from the nonnegativity of the policy probabilities and of the kernel rows. For the row sums,

$$\sum_{s' \in \mathcal{S}} P_{p_\star}(s, s') = \sum_{a \in \mathcal{A}} p_\star(a \mid x(s)) \sum_{s' \in \mathcal{S}} K(\{S' = s'\} \mid s, a) = \sum_{a \in \mathcal{A}} p_\star(a \mid x(s)) = 1.$$

Here Assumption 1 supplies the kernel row mass, Assumption 2 supplies the behavior probability vector when $p_\star = b$, and the target-carrier part of Assumption 5 supplies the target probability vector when $p_\star = e$.

We first convert the total-variation contraction in Assumption 6 into the oscillation contraction used below. Let ν and ν' be probability vectors on the finite state space, and let $f : \mathcal{S} \to \mathbb{R}$. Put

$$f_{\min} := \min_{x \in \mathcal{S}} f(x), \qquad f_0(x) := f(x) - f_{\min}.$$

Then $0 \le f_0(x) \le \operatorname{osc}(f)$ for every x, and, since ν and ν' have equal total mass,

$$\sum_{x \in \mathcal{S}} f(x)(\nu(x) - \nu'(x)) = \sum_{x \in \mathcal{S}} f_0(x)(\nu(x) - \nu'(x)).$$

The layer-cake identity on the finite state space gives

$$\sum_{x \in \mathcal{S}} f_0(x)(\nu(x) - \nu'(x)) = \int_0^{\operatorname{osc}(f)} \left(\nu\{x : f_0(x) > \lambda_{\mathrm{lev}}\} - \nu'\{x : f_0(x) > \lambda_{\mathrm{lev}}\}\right) d\lambda_{\mathrm{lev}},$$

and therefore

$$\left|\sum_{x \in \mathcal{S}} f(x)(\nu(x) - \nu'(x))\right| \le \operatorname{osc}(f)\, \|\nu - \nu'\|_{\mathrm{TV}}.$$

Now fix states s, s'. Applying Assumption 6 to the point masses δ_s and $\delta_{s'}$ gives

$$\|\delta_s P_{p_\star} - \delta_{s'} P_{p_\star}\|_{\mathrm{TV}} \le \alpha\, \|\delta_s - \delta_{s'}\|_{\mathrm{TV}} \le \alpha.$$

Combining this with the preceding dual bound yields the one-step estimate

$$|P_{p_\star} f(s) - P_{p_\star} f(s')| \le \alpha \operatorname{osc}(f).$$

Taking the supremum over s, s' and iterating gives the finite-state oscillation form: whenever $\operatorname{osc}(f) \le B_{\mathrm{osc}}$ for some $B_{\mathrm{osc}} \ge 0$,

$$\operatorname{osc}(P_{p_\star}^{m_{\mathrm{iter}}} f) \le \alpha^{m_{\mathrm{iter}}} B_{\mathrm{osc}}, \qquad m_{\mathrm{iter}} \in \mathbb{N}.$$

Applying this first with $p_\star = e$, $f = \widetilde{g}_e$, and $B_{\mathrm{osc}} = 2$, and then using $\alpha^k \le 1$, gives

$$\sup_{s,s'} |P_e^k \widetilde{g}_e(s) - P_e^k \widetilde{g}_e(s')| \le 2.$$

Applying it next with $p_\star = b$ to $P_e^k \widetilde{g}_e$ gives

$$\sup_{s,s'} \left|P_b^\gamma P_e^k \widetilde{g}_e(s) - P_b^\gamma P_e^k \widetilde{g}_e(s')\right| \le 2\alpha^\gamma.$$

5. Define

$$\widetilde{\Phi}_{\gamma,k}(s) := P_b^{\gamma} P_e^{k} \widetilde{g}_e(s).$$

Because $g_e = \widetilde{g}_e$ on $\{s : d_b(s) > 0\}$, Lemma 11 applied with policy e gives, for every behavior-supported state x,

$$P_e^{k} g_e(x) = P_e^{k} \widetilde{g}_e(x) \qquad \text{whenever } d_b(x) > 0.$$

This pointwise behavior-supported equality is the terminal-agreement hypothesis needed for the second application of Lemma 11, now with policy b. Thus, for every initial state s with $d_b(s) > 0$,

$$\Phi_{\gamma,k}(s) = P_b^{\gamma} P_e^{k} g_e(s) = P_b^{\gamma} P_e^{k} \widetilde{g}_e(s) = \widetilde{\Phi}_{\gamma,k}(s).$$

We next record the behavior-stationarity identity used at time $t+1$. For every bounded $F : \mathcal{S} \to \mathbb{R}$ and every epoch m_{time},

$$\int F(S_{m_{\text{time}}}(\tau))\, d\mathbb{P}_M(\tau) = \sum_{s \in \mathcal{S}} d_b(s) F(s).$$

This follows by induction on m_{time}. At the initial epoch the identity is Assumption 3. For the induction step, Assumption 2 and Assumption 1 give the behavior peeling identity

$$\int F(S_{m_{\text{time}}+1}(\tau))\, d\mathbb{P}_M(\tau) = \int \sum_{x \in \mathcal{S}} P_b(S_{m_{\text{time}}}(\tau), x) F(x)\, d\mathbb{P}_M(\tau).$$

Using the induction hypothesis with the bounded function $s \mapsto \sum_x P_b(s,x)F(x)$, the right-hand side becomes

$$\sum_{s \in \mathcal{S}} d_b(s) \sum_{x \in \mathcal{S}} P_b(s,x) F(x) = \sum_{x \in \mathcal{S}} \left(\sum_{s \in \mathcal{S}} d_b(s) P_b(s,x) \right) F(x) = \sum_{x \in \mathcal{S}} d_b(x) F(x),$$

where the last equality is the stationarity equation for the behavior stationary law d_b, part of the model-class data in Definition 10. Taking the zero-support indicator

$$\chi_{\text{null}}(s) := \mathbf{1}\{d_b(s) = 0\}$$

at $m_{\text{time}} = t+1$ gives

$$\int \mathbf{1}\{d_b(S_{t+1}(\tau)) = 0\}\, d\mathbb{P}_M(\tau) = 0,$$

so $d_b(S_{t+1}(\tau)) > 0$ for $\mathbb{P}_M$-almost every τ. Hence

$$\Phi_{\gamma,k}(S_{t+1}(\tau)) = \widetilde{\Phi}_{\gamma,k}(S_{t+1}(\tau)) \quad \text{for } \mathbb{P}_M\text{-almost every } \tau.$$

6. Put

$$A(\tau) := Z_t(k)(O(\tau)), \qquad \overline{\Phi} := \int \widetilde{\Phi}_{\gamma,k}(S_{t+1}(\tau))\, d\mathbb{P}_M(\tau).$$

Using the almost-sure replacement from the previous step, both

$$\int A(\tau) \Phi_{\gamma,k}(S_{t+1}(\tau))\, d\mathbb{P}_M(\tau) = \int A(\tau) \widetilde{\Phi}_{\gamma,k}(S_{t+1}(\tau))\, d\mathbb{P}_M(\tau)$$

and

$$\int \Phi_{\gamma,k}(S_{t+1}(\tau))\, d\mathbb{P}_M(\tau) = \int \widetilde{\Phi}_{\gamma,k}(S_{t+1}(\tau))\, d\mathbb{P}_M(\tau) = \overline{\Phi}$$

hold; the integrability needed for these replacements follows from the finite state space and the score bound above. The centered representation from the first step therefore gives

$$|\mathrm{Cov}_{\mathrm{obs}}(Z_t(k), Z_u(k))| = \left| \int A(\tau)\big(\widetilde{\Phi}_{\gamma,k}(S_{t+1}(\tau)) - \overline{\Phi}\big)\, d\mathbb{P}_M(\tau) \right|.$$

The oscillation bound in the fourth step implies, for every τ,

$$\left|\widetilde{\Phi}_{\gamma,k}(S_{t+1}(\tau)) - \overline{\Phi}\right| \le 2\alpha^{\gamma},$$

because $\overline{\Phi}$ is an average of values of $\widetilde{\Phi}_{\gamma,k}$. Therefore, using the unit L^1 bound from the second step,

$$|\mathrm{Cov}_{\mathrm{obs}}(Z_t(k), Z_u(k))| \le \int |A(\tau)|\, 2\alpha^{\gamma}\, d\mathbb{P}_M(\tau) \le 2\alpha^{\gamma} = 2\alpha^{h-k-1}.$$

□

Together, Lemmas 8 to 12 decompose the sampling fluctuation into a weighted-window term and a geometric dependence tail. The overlap condition controls the inflation inside the window, while contraction controls pairs whose windows are separated by at least one time step.

The next statement sums those covariance bounds over future times. It is the finite-sample accounting step that turns the local bounds into a variance inequality for the average score.

Lemma 13. [lem:phiw-future-covariance-sum] (Aggregate future-covariance bound) *Let $T \ge 1$, $k \in \mathbb{N}$, $t_0, \zeta, C \in \mathbb{R}$, and let $M \in \mathcal{M}_T(t_0, \zeta, C)$ be as in Definition 10. Write $L = e^{\zeta}$, $\alpha = e^{-1/t_0}$, and $\mathscr{T}_k = \{t : k \le t \le T-1\}$. Then, for every $t \in \mathscr{T}_k$, the depth-k scores of Definition 16 satisfy, under the observed trajectory law of M,*

$$\sum_{\substack{u \in \mathscr{T}_k \\ t<u}} \big|\mathrm{Cov}_{\mathrm{obs}}\big(Z_t(k), Z_u(k)\big)\big| \le \frac{2L^{k+1}}{L-1} + \frac{2}{1-\alpha}.$$

Proof of Lemma 13. Fix $t \in \mathscr{T}_k$. By Definition 10, $0 < \zeta$ and $0 < t_0$, so

$$L = e^{\zeta} > 1, \qquad 0 \le \alpha = e^{-1/t_0} < 1.$$

The same membership supplies the model restrictions used by the covariance estimates below.

1. Define the future index set and its two lag classes by

$$\mathscr{U}_t := \{u \in \mathscr{T}_k : t < u\}, \qquad \lambda_{t,u} := u - t \quad (u \in \mathscr{U}_t),$$

and

$$\mathscr{N}_t := \{u \in \mathscr{U}_t : \lambda_{t,u} \le k\}, \qquad \mathscr{G}_t := \{u \in \mathscr{U}_t : k+1 \le \lambda_{t,u}\}.$$

For each $u \in \mathscr{U}_t$, the lag $\lambda_{t,u}$ is a positive integer, hence exactly one of $\lambda_{t,u} \le k$ and $k+1 \le \lambda_{t,u}$ holds. Thus

$$\mathscr{U}_t = \mathscr{N}_t \dot{\cup} \mathscr{G}_t.$$

Writing

$$\Gamma_t(u) := \left|\mathrm{Cov}_{\mathrm{obs}}\big(Z_t(k), Z_u(k)\big)\right|, \qquad u \in \mathscr{U}_t,$$

we get

$$\sum_{u\in\mathscr{U}_t} \Gamma_t(u) = \sum_{u\in\mathscr{N}_t} \Gamma_t(u) + \sum_{u\in\mathscr{G}_t} \Gamma_t(u).$$

2. For $u \in \mathscr{N}_t$, the lag identity is $u = t + \lambda_{t,u}$ and $1 \le \lambda_{t,u} \le k$. Since $k \le t$ and $L \ge 1$, Lemma 9 gives

$$\Gamma_t(u) \le 2L^{k+1-\lambda_{t,u}}.$$

Therefore

$$\sum_{u\in\mathscr{N}_t} \Gamma_t(u) \le 2L \sum_{u\in\mathscr{N}_t} L^{k-\lambda_{t,u}}.$$

The map

$$j_{\mathrm{near}}(u) := k - \lambda_{t,u}, \qquad u \in \mathscr{N}_t,$$

is injective and takes its values in $\{0, \dots, k-1\}$. Since $L \ge 0$,

$$\sum_{u\in\mathscr{N}_t} L^{k-\lambda_{t,u}} \le \sum_{0\le j<k} L^j.$$

For $L > 1$,

$$\sum_{0\le j<k} L^j = \frac{L^k - 1}{L - 1} \le \frac{L^k}{L-1}.$$

Combining the last three displays yields

$$\sum_{u\in\mathscr{N}_t} \Gamma_t(u) \le \frac{2L^{k+1}}{L-1}.$$

3. For $u \in \mathscr{G}_t$, the lag identity is $u = t + \lambda_{t,u}$ and $k+1 \le \lambda_{t,u}$. Hence Lemma 12 gives

$$\Gamma_t(u) \le 2\alpha^{\lambda_{t,u}-k-1}.$$

Thus

$$\sum_{u\in\mathscr{G}_t} \Gamma_t(u) \le 2 \sum_{u\in\mathscr{G}_t} \alpha^{\lambda_{t,u}-k-1}.$$

The map

$$j_{\mathrm{sep}}(u) := \lambda_{t,u} - k - 1, \qquad u \in \mathscr{G}_t,$$

is injective and takes its values in $\{0, \dots, T-1\}$. Since $\alpha \ge 0$,

$$\sum_{u\in\mathscr{G}_t} \alpha^{\lambda_{t,u}-k-1} \le \sum_{0\le j<T} \alpha^j.$$

Because $0 \le \alpha < 1$,

$$\sum_{0\le j<T} \alpha^j \le \sum_{j=0}^{\infty} \alpha^j = \frac{1}{1-\alpha}.$$

Consequently,

$$\sum_{u\in\mathscr{G}_t} \Gamma_t(u) \le \frac{2}{1-\alpha}.$$

4. Adding the two bounds over the disjoint decomposition of $\mathscr{U}_t$ gives

$$\sum_{\substack{u\in\mathscr{T}_k\\ t<u}} \left|\mathrm{Cov}_{\mathrm{obs}}\big(Z_t(k), Z_u(k)\big)\right| = \sum_{u\in\mathscr{U}_t} \Gamma_t(u) \le \frac{2L^{k+1}}{L-1} + \frac{2}{1-\alpha}.$$

□

Lemma 13 supplies the two terms that reappear in the assembled variance bound. The geometric sum over overlapping lags gives the policy-overlap contribution, and the contraction tail gives the dependence contribution.

Lemma 14. [`lem:phiw-raw-variance`] (Raw variance assembly) *Let T and k be nonnegative integers, let $t_0, \zeta, C \in \mathbb{R}$, and let M be a raw POMDP experiment. Write $L = e^{\zeta}$ for the policy-overlap factor and $\alpha = e^{-1/t_0}$ for the contraction factor. Suppose that*

- ***(Horizon.)*** *$T \ge 1$.*
- ***(Model class.)*** *$M \in \mathcal{M}_T(t_0, \zeta, C)$, with $\mathcal{M}_T(t_0, \zeta, C)$ as in Definition 10.*
- ***(History depth.)*** *$k \le \lfloor T/2 \rfloor$.*

For the depth-k scores $Z_t(k)$ of Definition 16, form the unclipped partial-history average

$$\frac{1}{T-k}\sum_{t=k}^{T-1} Z_t(k).$$

Under the observed trajectory law obtained by projecting the full trajectory law of M onto observed histories, its variance satisfies

$$\mathrm{Var}\left(\frac{1}{T-k}\sum_{t=k}^{T-1} Z_t(k)\right) \le \frac{2}{T}\left\{L^{k+1}\left(1+\frac{4}{L-1}\right) + \frac{4}{1-\alpha}\right\}.$$

Proof of Lemma 14. Let $\mathbb{P}_{\mathrm{obs}}$ denote the observed trajectory law and let $\mathbb{E}_{\mathrm{obs}}$ denote expectation under this law. Let $\mathbb{P}_{\mathrm{full}}$ denote the full trajectory law carried by M, and let $\mathbb{E}_{\mathrm{full}}$ denote expectation under that law. Put

$$\mathscr{T}_k := \{t \in \{0,\dots,T-1\} : k \le t\}, \qquad \mathfrak{B}_k := L^{k+1}\left(1+\frac{4}{L-1}\right) + \frac{4}{1-\alpha}.$$

For $t \in \mathscr{T}_k$, write

$$\Gamma_t^{(k)} := \prod_{\substack{0\le j\le T-1\\ j\le t,\ t-j\le k}} \rho_j, \qquad Z_t(k) = Y_t\Gamma_t^{(k)}.$$

Since $M \in \mathcal{M}_T(t_0, \zeta, C)$, Definition 10 gives $t_0 > 0$, $\zeta > 0$, hence

$$L = e^{\zeta} > 1, \qquad 0 \le \alpha = e^{-1/t_0} < 1.$$

Also $k \le \lfloor T/2 \rfloor$ implies

$$|\mathscr{T}_k| = T - k > 0, \qquad \frac{T}{2} \le T - k.$$

In particular $\mathfrak{B}_k \ge 0$. The raw average is

$$\frac{1}{T-k}\sum_{t=k}^{T-1} Z_t(k) = \frac{1}{T-k}\sum_{t\in\mathscr{T}_k} Z_t(k).$$

1. For each $t \in \mathscr{T}_k$, Lemma 8 applies with the sequential-ignorability, policy-overlap, kernel-law, and bounded-reward clauses supplied by Definition 10, together with $L = e^{\zeta} \geq 1$ and $k \leq t$. It gives
$$\mathbb{E}_{\text{obs}}\left[Z_t(k)^2\right] \leq L^{k+1}.$$
Using the standard identity
$$\text{Var}_{\text{obs}}(Z_t(k)) = \mathbb{E}_{\text{obs}}\left[Z_t(k)^2\right] - \left(\mathbb{E}_{\text{obs}} Z_t(k)\right)^2,$$
we obtain
$$\text{Var}_{\text{obs}}(Z_t(k)) \leq L^{k+1}.$$

2. Fix $t \in \mathscr{T}_k$. The aggregate covariance estimate in Lemma 13 applies to the same model class and the same future index set $\mathscr{T}_k$. Therefore
$$\sum_{\substack{u \in \mathscr{T}_k \\ t<u}} \left|\text{Cov}_{\text{obs}}\left(Z_t(k), Z_u(k)\right)\right| \leq \frac{2L^{k+1}}{L-1} + \frac{2}{1-\alpha}.$$

3. Now expand the variance of the finite sum:
$$\text{Var}_{\text{obs}}\left(\sum_{t \in \mathscr{T}_k} Z_t(k)\right) = \sum_{t \in \mathscr{T}_k} \sum_{u \in \mathscr{T}_k} \text{Cov}_{\text{obs}}\left(Z_t(k), Z_u(k)\right).$$
Bounding each covariance by its absolute value and then using covariance symmetry gives
$$\text{Var}_{\text{obs}}\left(\sum_{t \in \mathscr{T}_k} Z_t(k)\right) \leq \sum_{t \in \mathscr{T}_k} \text{Var}_{\text{obs}}(Z_t(k)) + 2 \sum_{t \in \mathscr{T}_k} \sum_{\substack{u \in \mathscr{T}_k \\ t<u}} \left|\text{Cov}_{\text{obs}}\left(Z_t(k), Z_u(k)\right)\right|.$$
Using the diagonal and future-sum bounds,
$$\text{Var}_{\text{obs}}\left(\sum_{t \in \mathscr{T}_k} Z_t(k)\right) \leq |\mathscr{T}_k| L^{k+1} + 2|\mathscr{T}_k| \left(\frac{2L^{k+1}}{L-1} + \frac{2}{1-\alpha}\right) = (T-k)\mathfrak{B}_k.$$

4. Finally,
$$\text{Var}_{\text{obs}}\left(\frac{1}{T-k} \sum_{t \in \mathscr{T}_k} Z_t(k)\right) = \frac{1}{(T-k)^2} \text{Var}_{\text{obs}}\left(\sum_{t \in \mathscr{T}_k} Z_t(k)\right) \leq \frac{\mathfrak{B}_k}{T-k}.$$
Since $T/2 \leq T-k$, $T > 0$, and $\mathfrak{B}_k \geq 0$,
$$\frac{\mathfrak{B}_k}{T-k} \leq \frac{2\mathfrak{B}_k}{T}.$$
Substituting the definition of $\mathfrak{B}_k$ gives
$$\text{Var}_{\text{obs}}\left(\frac{1}{T-k} \sum_{t=k}^{T-1} Z_t(k)\right) \leq \frac{2}{T}\left\{L^{k+1}\left(1 + \frac{4}{L-1}\right) + \frac{4}{1-\alpha}\right\}.$$

□

The two terms in Lemma 14 have distinct origins: multiplying observable policy ratios over a window produces the L^{k+1} term, while serial dependence contributes the contraction term. The bound is stated for the unclipped average because the displayed variance is the component used before the clipped estimator's squared-error risk is assembled.

The next statement supplies the fixed-radius upper bound for the partial-history importance-weighted estimator. It gives both the rate bound for the balanced depth and the corresponding variance calculation for the unclipped average.

Lemma 15. [lem:phiw-upper] (Balanced PHIW upper bound) *Fix constants satisfying*

- ***(Mixing scale.)*** $t_0 > 0$.
- ***(Policy-overlap scale.)*** $\zeta > 0$.
- ***(Latent-overlap radius.)*** $C \geq 1$.

Let k_T be the balanced history depth in Definition 17, and set

$$\beta = \frac{2}{2 + t_0 \zeta}.$$

Then there exist constants $c^\star > 0$ and $T_\star \in \mathbb{N}$ such that, for every $T \geq T_\star$, the estimator-specific worst-case squared risk from Definition 13 of the observable PHIW estimator $\widehat{\theta}_{k_T}$ from Definition 16 satisfies

$$R_T(t_0, \zeta, C; \widehat{\theta}_{k_T}) \leq c^\star T^{-\beta}.$$

Moreover, for every $T \geq T_\star$, every finite alphabet sizes n_X, n_H, and every raw POMDP experiment $M \in \mathcal{M}_T(t_0, \zeta, C)$, the variance under the observed trajectory law induced by M of the unclipped partial-history average

$$\frac{1}{T - k_T} \sum_{t=k_T}^{T-1} Z_t(k_T)$$

satisfies

$$\mathrm{Var}\left(\frac{1}{T - k_T} \sum_{t=k_T}^{T-1} Z_t(k_T) \right) \leq \frac{2}{T} \left[e^{\zeta(k_T+1)} \left(1 + \frac{4}{e^\zeta - 1} \right) + \frac{4}{1 - e^{-1/t_0}} \right].$$

Proof of Lemma 15. 1. Define

$$\alpha = \exp(-1/t_0), \qquad L = \exp(\zeta), \qquad \beta = \frac{2}{2 + t_0 \zeta}, \qquad q_C = \frac{C - 1}{C}.$$

The assumptions $t_0 > 0$ and $\zeta > 0$ give

$$0 < \alpha < 1, \qquad 1 < L, \qquad 0 < \beta < 1.$$

Also $C \geq 1$ implies

$$0 \leq q_C \leq 1, \qquad q_C^2 \leq 1.$$

Let

$$\mathfrak{d} = 2\log(1/\alpha) + \log L = \frac{2}{t_0} + \zeta.$$

Then $\mathfrak{d} > 0$ and

$$\beta = \frac{2/t_0}{\mathfrak{d}}.$$

Since $\log T = o(T)$, there is an integer $T_{\mathrm{bal}} \geq 1$ such that, for every $T \geq T_{\mathrm{bal}}$,

$$0 \leq \frac{\log T}{\mathfrak{d}} \leq \frac{T}{2}.$$

For such T, Definition 17 gives

$$k_T = \left\lfloor \frac{\log T}{\mathfrak{d}} \right\rfloor,$$

and hence

$$\frac{\log T}{\mathfrak{d}} - 1 < k_T \leq \frac{\log T}{\mathfrak{d}}.$$

Therefore

$$\alpha^{2k_T} = \exp(-2k_T/t_0) \leq \exp(2/t_0)\, T^{-\beta},$$

and

$$\frac{L^{k_T+1}}{T} = \exp\{(k_T+1)\zeta - \log T\} \leq \exp(\zeta)\, T^{-\beta}.$$

With

$$c_{\mathrm{bal}} := \exp(2/t_0) + \exp(\zeta) > 0,$$

we have, for every $T \geq T_{\mathrm{bal}}$,

$$\alpha^{2k_T} + \frac{L^{k_T+1}}{T} \leq c_{\mathrm{bal}} T^{-\beta}.$$

2. Set

$$\mathsf{A} := 1 + \frac{4}{L-1}, \qquad \mathsf{B} := \frac{4}{1-\alpha}, \qquad \mathsf{D} := \max\{4, 2\mathsf{A}\},$$

and define

$$c^\star := \mathsf{D}\, c_{\mathrm{bal}} + 2\mathsf{B}, \qquad T_\star := T_{\mathrm{bal}}.$$

Since $L > 1$ and $\alpha < 1$, the constants $\mathsf{A}, \mathsf{B}, \mathsf{D}$ are positive, and so $c^\star > 0$.

3. Fix $T \geq T_\star$. Then $1 \leq T$, $k_T \leq T/2$, and $k_T < T$. For every $M \in \mathcal{M}_T(t_0, \zeta, C)$, Lemma 34, applied with $k = k_T$, gives

$$\mathbb{E}_M\left[\{\widehat{\theta}_{k_T} - \theta(K, e)\}^2\right] \leq 4q_C^2 \alpha^{2k_T} + \frac{2}{T}\left(L^{k_T+1}\mathsf{A} + \mathsf{B}\right).$$

Using $q_C^2 \leq 1$, $\mathsf{D} \geq 4$, and $\mathsf{D} \geq 2\mathsf{A}$,

$$4q_C^2\alpha^{2k_T} + \frac{2}{T}\left(L^{k_T+1}\mathsf{A} + \mathsf{B}\right) \leq \mathsf{D}\left(\alpha^{2k_T} + \frac{L^{k_T+1}}{T}\right) + 2\mathsf{B}\, T^{-1}.$$

The balance bound from the first step and the inequality $T^{-1} \le T^{-\beta}$, valid because $T \ge 1$ and $0 < \beta < 1$, yield

$$\mathbb{E}_M\left[\{\widehat{\theta}_{k_T} - \theta(K,e)\}^2\right] \le \mathsf{D} c_{\mathrm{bal}} T^{-\beta} + 2\mathsf{B} T^{-\beta} = c^\star T^{-\beta}.$$

There are two cases. If the indexed class $\mathcal{M}_T(t_0,\zeta,C)$ is nonempty, the pointwise bound over every member of the class, together with the estimator-specific risk definition in Definition 13, gives

$$R_T(t_0,\zeta,C;\widehat{\theta}_{k_T}) \le c^\star T^{-\beta}.$$

If the indexed class is empty, the real-valued worst-case-risk functional used for R_T has empty supremum equal to 0. Since $c^\star > 0$, $T \ge 1$, and $T^{-\beta} \ge 0$, the bound also holds in this case.

4. It remains to record the variance assertion. For every $T \ge T_\star$, the definition of $T_\star$ gives $1 \le T$, and the minimum in Definition 17 gives $k_T \le \lfloor T/2 \rfloor$. Apply Lemma 14 with $k = k_T$, $L = e^\zeta$, and $\alpha = e^{-1/t_0}$. Its conclusion is exactly

$$\mathrm{Var}_M\left(\frac{1}{T-k_T}\sum_{t=k_T}^{T-1} Z_t(k_T)\right) \le \frac{2}{T}\left[e^{\zeta(k_T+1)}\left(1+\frac{4}{e^\zeta-1}\right)+\frac{4}{1-e^{-1/t_0}}\right],$$

uniformly over the finite alphabet sizes and every $M \in \mathcal{M}_T(t_0,\zeta,C)$, as required.

□

The bound in Lemma 15 identifies the observable estimator used for the upper side of the fixed-C comparison. The variance display isolates the cost of multiplying observable policy ratios over a window of length k_T and the serial-dependence contribution controlled by contraction. Balancing those terms gives the exponent β under fixed overlap and fixed latent-stationary domination.

For the lower side, the appendix uses the signed-depth alternatives introduced in Definition 18. The next two statements record the contraction and stationary-law calculations that make the construction a member of the same finite-POMDP class.

Lemma 16. [lem:signed-depth-constant-policy-contraction] (Contraction of the constant-policy signed-depth kernel) *Let $t_0 > 0$, $\zeta > 0$, $C \ge 1$, $Q \ge 1$, $v \in \{-1,+1\}$ and $r \in [0,1]$, and write $\alpha = e^{-1/t_0}$. Let $P^{(r)}$ be the transition kernel induced on the joint state space $\mathcal{S}_Q = \{0\} \times (\{0,\dots,Q\} \times \{-1,+1\})$ of the signed-depth alternative of Definition 18 with terminal depth Q and index v by the constant policy a_r that plays action one with probability r. Then, for all probability vectors μ, μ' on $\mathcal{S}_Q$,*

$$\|\mu P^{(r)} - \mu' P^{(r)}\|_{\mathrm{TV}} \le \alpha \|\mu - \mu'\|_{\mathrm{TV}}.$$

Proof of Lemma 16. Let

$$\varepsilon_C = \min\{(C-1)/2, 1/2\}, \qquad \alpha = \exp(-1/t_0).$$

For $\sigma = (0,j,u)$ and $\sigma' = (0,j',u')$ in $\mathcal{S}_Q$, define the refresh law

$$\eta_{C,Q}(\sigma') = \mathbf{1}\{j' = 0\}\,\frac{1+u'\varepsilon_C}{2},$$

and define the residual kernel

$$\Gamma_{C,r,Q}(\sigma,\sigma') = \begin{cases} \eta_{C,Q}(\sigma'), & j = Q, \\ \dfrac{1-r}{2} + r\,\mathbf{1}\{u' = u\}, & j < Q \text{ and } j' = j+1, \\ 0, & \text{otherwise.} \end{cases}$$

Expanding the transition rule in Definition 18 under the constant policy $a_r(1) = r$, $a_r(0) = 1-r$, gives, for every $\sigma, \sigma' \in \mathcal{S}_Q$,

$$P^{(r)}(\sigma,\sigma') = (1-\alpha)\eta_{C,Q}(\sigma') + \alpha\,\Gamma_{C,r,Q}(\sigma,\sigma').$$

Indeed, at terminal depth $j = Q$ both sides equal $\eta_{C,Q}(\sigma')$; at depths $j < Q$, the reset part contributes $(1-\alpha)\eta_{C,Q}(\sigma')$, while the advance-to-$j+1$ part contributes $\alpha[(1-r)/2 + r\mathbf{1}\{u' = u\}]$.

The residual kernel is stochastic. Since $C \geq 1$, one has $\varepsilon_C \in [0, 1/2]$, so the two reset weights are nonnegative and

$$\sum_{u' \in \{-1,+1\}} \frac{1 + u'\varepsilon_C}{2} = 1.$$

If $j = Q$, summing $\Gamma_{C,r,Q}(\sigma,\cdot)$ over $\mathcal{S}_Q$ therefore gives 1. If $j < Q$, the only nonzero entries have $j' = j+1$, and their two sign weights sum to

$$\left(\frac{1-r}{2} + r\right) + \frac{1-r}{2} = 1,$$

with nonnegativity following from $r \in [0,1]$. Thus

$$\Gamma_{C,r,Q}(\sigma,\sigma') \geq 0 \quad \text{and} \quad \sum_{\sigma' \in \mathcal{S}_Q} \Gamma_{C,r,Q}(\sigma,\sigma') = 1$$

for every $\sigma \in \mathcal{S}_Q$.

For any signed vector λ on $\mathcal{S}_Q$, stochasticity of $\Gamma_{C,r,Q}$ gives the total-variation nonexpansiveness bound

$$\begin{aligned} \|\lambda\Gamma_{C,r,Q}\|_{\mathrm{TV}} &= \frac{1}{2}\sum_{\sigma'}\left|\sum_{\sigma}\lambda(\sigma)\Gamma_{C,r,Q}(\sigma,\sigma')\right| \\ &\leq \frac{1}{2}\sum_{\sigma'}\sum_{\sigma}|\lambda(\sigma)|\Gamma_{C,r,Q}(\sigma,\sigma') \\ &= \frac{1}{2}\sum_{\sigma}|\lambda(\sigma)|\sum_{\sigma'}\Gamma_{C,r,Q}(\sigma,\sigma') = \|\lambda\|_{\mathrm{TV}}. \end{aligned}$$

Now take probability vectors μ, μ'. Because $\sum_\sigma \mu(\sigma) = \sum_\sigma \mu'(\sigma) = 1$, the common refresh term cancels:

$$(\mu P^{(r)} - \mu' P^{(r)})(\sigma') = \alpha\sum_{\sigma}(\mu(\sigma) - \mu'(\sigma))\Gamma_{C,r,Q}(\sigma,\sigma').$$

Since $\alpha = \exp(-1/t_0) > 0$, the preceding display and nonexpansiveness yield

$$\|\mu P^{(r)} - \mu' P^{(r)}\|_{\mathrm{TV}} = \alpha\|(\mu - \mu')\Gamma_{C,r,Q}\|_{\mathrm{TV}} \leq \alpha\|\mu - \mu'\|_{\mathrm{TV}}.$$

This is the claimed contraction. □

Lemma 17. [lem:signed-depth-stationary-law] (Closed form of the signed-depth stationary law) *Let $t_0 > 0$, $\zeta > 0$, $C \geq 1$, $Q \geq 1$, $v \in \{-1, +1\}$ and $r \in [0, 1]$. Write $\alpha = e^{-1/t_0}$, $\varepsilon_C = \min\{(C-1)/2, 1/2\}$, and $w_j = (1-\alpha)\alpha^j/(1-\alpha^{Q+1})$. Let $P^{(r)}$ be the kernel induced on $\mathcal{S}_Q$ by the constant policy a_r for the signed-depth alternative of Definition 18 with terminal depth Q and index v, and let d_r be the stationary law selected for $P^{(r)}$. Then, for every hidden state $h = (j(h), u(h))$,*

$$d_r(0, h) = w_{j(h)} \frac{1 + u(h)\varepsilon_C r^{j(h)}}{2}.$$

Proof of Lemma 17. Write

$$\bar{d}_r(0, (j, u)) := w_j \frac{1 + u\varepsilon_C r^j}{2}, \qquad j \in \{0, \dots, Q\}, \quad u \in \{-1, +1\}.$$

We prove that this probability vector is stationary for $P^{(r)}$, and then use the contraction statement to identify it with the selected stationary law.

1. Since $t_0 > 0$, the value $\alpha = e^{-1/t_0}$ lies in $(0, 1)$. Also $C \geq 1$ gives $0 \leq \varepsilon_C \leq 1/2$, and $r \in [0, 1]$ gives $0 \leq r^j \leq 1$. Hence every factor $(1 + u\varepsilon_C r^j)/2$ is nonnegative. For each fixed depth j,

$$\sum_{u \in \{-1,+1\}} w_j \frac{1 + u\varepsilon_C r^j}{2} = w_j,$$

and the geometric normalization gives

$$\sum_{j=0}^{Q} w_j = \frac{1-\alpha}{1-\alpha^{Q+1}} \sum_{j=0}^{Q} \alpha^j = 1.$$

Thus $\bar{d}_r$ is a probability vector on $\mathcal{S}_Q$.

2. The state kernel induced by the constant policy a_r has the following hidden-state transition probabilities, obtained from the signed-depth dynamics in Definition 18 by averaging the action-one and action-zero advance weights. If $j < Q$, then

$$P^{(r)}\big((0, (j, u)), (0, (0, u'))\big) = (1-\alpha)\frac{1 + u'\varepsilon_C}{2},$$

and

$$P^{(r)}\big((0, (j, u)), (0, (j+1, u'))\big) = \alpha \frac{1 + uu'r}{2}.$$

If $j = Q$, then

$$P^{(r)}\big((0, (Q, u)), (0, (0, u'))\big) = \frac{1 + u'\varepsilon_C}{2}.$$

The remaining transitions have weight zero.

For a target state of depth 0, summing the displayed reset probabilities gives

$$(\bar{d}_r P^{(r)})(0, (0, u')) = \frac{1 + u'\varepsilon_C}{2} \left\{ w_Q + (1-\alpha) \sum_{j=0}^{Q-1} w_j \right\}.$$

Since $\sum_{j=0}^{Q} w_j = 1$,

$$w_Q + (1-\alpha)\sum_{j=0}^{Q-1} w_j = (1-\alpha) + \alpha w_Q = (1-\alpha) + \frac{(1-\alpha)\alpha^{Q+1}}{1-\alpha^{Q+1}} = \frac{1-\alpha}{1-\alpha^{Q+1}} = w_0.$$

Therefore

$$(\bar{d}_r P^{(r)})(0,(0,u')) = w_0 \frac{1+u'\varepsilon_C}{2} = \bar{d}_r(0,(0,u')).$$

For a target state of depth $m \in \{1,\dots,Q\}$, only depth $m-1$ can advance to it, so

$$(\bar{d}_r P^{(r)})(0,(m,u')) = \alpha w_{m-1} \sum_{u\in\{-1,+1\}} \frac{1+u\varepsilon_C r^{m-1}}{2}\frac{1+uu'r}{2}.$$

The sign sum is

$$\sum_{u\in\{-1,+1\}} \frac{1+u\varepsilon_C r^{m-1}}{2}\frac{1+uu'r}{2} = \frac{1+u'\varepsilon_C r^m}{2}.$$

Using $w_m = \alpha w_{m-1}$, this yields

$$(\bar{d}_r P^{(r)})(0,(m,u')) = w_m \frac{1+u'\varepsilon_C r^m}{2} = \bar{d}_r(0,(m,u')).$$

Thus $\bar{d}_r P^{(r)} = \bar{d}_r$.

3. The contraction supplied by Lemma 16 implies uniqueness of stationary probability vectors. Indeed, if λ and λ' are stationary for $P^{(r)}$, then

$$\|\lambda - \lambda'\|_{\mathrm{TV}} = \|\lambda P^{(r)} - \lambda' P^{(r)}\|_{\mathrm{TV}} \le \alpha\|\lambda - \lambda'\|_{\mathrm{TV}}.$$

Because $\alpha < 1$ and total variation is nonnegative, the norm is zero, hence $\lambda = \lambda'$ on the finite state space. The selected law d_r is stationary because the stationary vector $\bar{d}_r$ exists, so uniqueness gives $d_r = \bar{d}_r$.

4. Evaluating the identity $d_r = \bar{d}_r$ at the hidden state $h = (j(h), u(h))$ gives

$$d_r(0,h) = w_{j(h)} \frac{1+u(h)\varepsilon_C r^{j(h)}}{2},$$

as claimed.

□

Lemma 16 gives the uniform total-variation contraction of the signed-depth transition under any constant action-one probability. Lemma 17 then gives the stationary law explicitly: the depth mass is a truncated geometric sequence, and the sign imbalance is scaled by the action-one probability to the current depth.

Lemma 18. [lem:signed-depth-membership] (Signed-depth membership) *For every $T \geq 1$, $Q \geq 1$, $t_0 > 0$, $\zeta > 0$, $C > 1$, and $v \in \{-1, +1\}$, let $M^T_{Q,v}$ be the signed-depth experiment of Definition 18. Put*

$$\varepsilon_C = \min\{(C-1)/2, 1/2\}, \qquad L = \exp(\zeta), \qquad w_j = \frac{(1-\alpha(t_0))\alpha(t_0)^j}{1-\alpha(t_0)^{Q+1}},$$

where $\alpha(t_0)$ is the mixing factor and $j = 0, \ldots, Q$. Then:

- ***(Membership.)*** *$M^T_{Q,v} \in \mathcal{M}_T(t_0, \zeta, C)$ in the sense of Definition 10.*
- ***(Behavior stationary law.)*** *For every hidden state h, with decoded depth $j(h) \in \{0, \ldots, Q\}$ and sign $u(h) \in \{-1, +1\}$,*

$$d_b(0, h) = w_{j(h)} \frac{1 + u(h)\varepsilon_C L^{-j(h)}}{2}.$$

- ***(Constant action-one stationary laws.)*** *For every $r \in [0, 1]$, if d_r denotes the stationary law induced by the constant policy that chooses action one with probability r, then for every hidden state h,*

$$d_r(0, h) = w_{j(h)} \frac{1 + u(h)\varepsilon_C r^{j(h)}}{2}.$$

- ***(Target-over-behavior domination.)*** *For every joint state s,*

$$d_e(s) \leq (1 + \varepsilon_C) d_b(s), \qquad 1 + \varepsilon_C \leq C.$$

The stationary target value of Definition 12 is

$$\theta(K, e) = v\, c_0(t_0)\, \varepsilon_C\, w_Q.$$

Proof of Lemma 18. Write $\alpha = \alpha(t_0) = \exp(-1/t_0)$, $L = \exp(\zeta)$, and $\varepsilon = \varepsilon_C$. Since $t_0 > 0$ and $\zeta > 0$, we have $0 < \alpha < 1$ and $1 < L$. Also $C > 1$ gives $C \geq 1$, so

$$0 \leq \varepsilon = \min\{(C-1)/2, 1/2\} \leq \frac{1}{2}.$$

For a hidden state h, abbreviate $j = j(h)$ and $u = u(h) \in \{-1, +1\}$.

1. For $r \in [0, 1]$, let a_r be the constant policy with $\Pr(A_t = 1) = r$, and let $P^{(r)}$ be the induced transition kernel on the joint state space

$$\mathcal{S}_Q = \{(0, h) : h \in \{0, \ldots, Q\} \times \{-1, +1\}\}.$$

   By Lemma 16, for all probability laws μ and μ' on $\mathcal{S}_Q$,

$$\|\mu P^{(r)} - \mu' P^{(r)}\|_{\mathrm{TV}} \leq \alpha \|\mu - \mu'\|_{\mathrm{TV}}.$$

   The behavior policy in Definition 18 is the case $r = L^{-1}$, while the target policy is the case $r = 1$. The displayed inequality therefore gives the contraction required in Assumption 6 for both policies.

2. For $r \in [0,1]$, Lemma 17 identifies the selected stationary law of the constant-policy kernel $P^{(r)}$: for every hidden state h,

$$d_r(0,h) = w_{j(h)} \frac{1 + u(h)\varepsilon r^{j(h)}}{2}.$$

Taking $r = L^{-1}$ gives the behavior stationary law because the behavior policy in Definition 18 chooses action one with probability $1/L$, and $(L^{-1})^{j(h)} = L^{-j(h)}$. Thus

$$d_b(0,h) = w_{j(h)} \frac{1 + u(h)\varepsilon L^{-j(h)}}{2}.$$

Taking $r = 1$ gives the target stationary law

$$d_e(0,h) = w_{j(h)} \frac{1 + u(h)\varepsilon}{2}.$$

3. We now verify the model-class predicates in Definition 10. The embedded experiment is generated by the initial law, the behavior policy, and the time-homogeneous reward-transition kernel in Definition 18. Therefore, for each epoch $t = 0, \dots, T-1$,

$$\mathcal{L}(Y_t, X_{t+1}, H_{t+1} \mid X_{0:t}, H_{0:t}, A_{0:t}, Y_{0:t-1}) = K(\cdot \mid X_t, H_t, A_t),$$

which is Assumption 1. The behavior action law is

$$b(1 \mid 0) = L^{-1}, \qquad b(0 \mid 0) = 1 - L^{-1}, \qquad b(0 \mid 0) + b(1 \mid 0) = 1,$$

and these two masses are nonnegative because $L > 1$. Since the observed state is the singleton $\{0\}$, the chronological law factors at every epoch $t = 0, \dots, T-1$ as

$$\begin{aligned} &\mathbb{P}\big\{\big((X_s, H_s)_{s<t}, (A_s, Y_s)_{s<t}, (X_t, H_t), A_t\big) \in d\mathsf{h}\, da\big\} \\ &\quad = \mathbb{P}\big\{\big((X_s, H_s)_{s<t}, (A_s, Y_s)_{s<t}, (X_t, H_t)\big) \in d\mathsf{h}\big\}\; b(da \mid X_t(\mathsf{h})), \end{aligned}$$

which is Assumption 2. The finite reward symbols are decoded as -1 and $+1$, so

$$\Pr(|Y_t| \le 1) = 1,$$

giving Assumption 4. The policy-overlap inequality is

$$e(1 \mid 0) = 1 = L\, b(1 \mid 0), \qquad e(0 \mid 0) = 0 \le L\, b(0 \mid 0),$$

so Assumption 5 holds with factor L. The initial law is

$$\Pr\{S_0 = (0,h)\} = w_{j(h)} \frac{1 + u(h)\varepsilon L^{-j(h)}}{2} = d_b(0,h),$$

where the last equality is the stationary-law formula from the case $r = L^{-1}$. Thus Assumption 3 holds.

For the latent stationary overlap, first identify the two selected stationary laws and the initial law on the same cells. The target and behavior formulas from Step 2 give

$$d_e(0,h) = w_j \frac{1 + u(h)\varepsilon}{2}, \qquad d_b(0,h) = w_j \frac{1 + u(h)\varepsilon L^{-j}}{2}.$$

The displayed initialization in Definition 18 gives

$$\Pr\{S_0 = (0,h)\} = w_j \frac{1 + u(h)\varepsilon L^{-j}}{2} = d_b(0,h),$$

where the last equality is the behavior stationary-law formula from Step 2. The depth mass is nonnegative: since

$$0 < \alpha < 1, \qquad 1 - \alpha^{Q+1} > 0,$$

we have

$$w_j = \frac{(1-\alpha)\alpha^j}{1-\alpha^{Q+1}} \geq 0.$$

Here $L^{-j} = (L^j)^{-1}$. Since $L > 1$, we have $0 < L^{-j} \leq 1$. For $u(h) = -1$,

$$d_e(0,h) = w_j \frac{1-\varepsilon}{2} \leq w_j \frac{1-\varepsilon L^{-j}}{2} = d_b(0,h) \leq (1+\varepsilon) d_b(0,h),$$

where the first inequality uses $\varepsilon L^{-j} \leq \varepsilon$, and the last uses $\varepsilon \geq 0$ and nonnegativity of the behavior cell. For $u(h) = +1$,

$$d_e(0,h) = w_j \frac{1+\varepsilon}{2} \leq w_j \frac{1+\varepsilon}{2}(1+\varepsilon L^{-j}) = (1+\varepsilon) d_b(0,h),$$

because $w_j \geq 0$ and $1+\varepsilon \geq 0$, so the multiplier $w_j(1+\varepsilon)/2$ is nonnegative, while $1+\varepsilon L^{-j} \geq 1$. Multiplying this last inequality by that nonnegative multiplier gives the displayed inequality, and its right-hand side is $(1+\varepsilon)d_b(0,h)$ by the formula for d_b. Every joint state has the unique observed coordinate 0, so these two sign cases give

$$d_e(s) \leq (1+\varepsilon) d_b(s) \qquad \text{for every } s \in \mathcal{S}_Q.$$

Moreover

$$1 + \varepsilon = 1 + \min\{(C-1)/2, 1/2\} \leq C,$$

because either $\varepsilon = (C-1)/2$, giving $1+\varepsilon = (C+1)/2 \leq C$, or $\varepsilon = 1/2$, which is possible only when $C \geq 2$. Since $d_b(s) \geq 0$,

$$d_e(s) \leq (1+\varepsilon) d_b(s) \leq C d_b(s),$$

so Assumption 7 holds.

It remains to collect the remaining requirements of Definition 10. The horizon requirement $1 \leq T$ is the hypothesis $T \geq 1$ of the present lemma. The observed alphabet of $M^T_{Q,v}$ is the singleton $\{0\}$, the latent alphabet is the $2(Q+1)$-element set $\{0,\dots,Q\} \times \{-1,+1\}$, and the action alphabet is $\{0,1\}$; all three are finite, and the action alphabet has exactly two elements, as the class requires. The parameter requirements $0 < t_0$ and $0 < \zeta$ are hypotheses of the lemma, and they fix $L = \exp(\zeta)$ and $\alpha = \exp(-1/t_0)$ as the overlap and contraction constants used throughout. Together with the seven assumptions verified above – the Markov kernel, sequential ignorability, stationary start, bounded rewards, policy overlap, uniform contraction, and latent stationary overlap – this is every defining property of Definition 10, and therefore $M^T_{Q,v} \in \mathcal{M}_T(t_0, \zeta, C)$.

4. It remains to compute the target value. The two-point alternative label enters the reward law only through its numeric sign: the positive label contributes $+1$, the negative label contributes -1, and in the statement this multiplier is denoted by $v \in \{-1, +1\}$. Under the target policy the action is always one. From the reward weights in Definition 18,

$$g_e(0,h) = \sum_a e(a \mid 0)\, E_K[Y_t \mid S_t = (0,h), A_t = a] = v\, c_0(t_0)\, u(h)\, \mathbf{1}\{j(h) = Q\}.$$

Using Definition 12 and the already computed target stationary law,

$$\theta(K,e) = \sum_h w_{j(h)} \frac{1 + u(h)\varepsilon}{2}\, v\, c_0(t_0)\, u(h)\, \mathbf{1}\{j(h) = Q\}.$$

Only depth Q contributes, and summing over the two signs gives

$$\theta(K,e) = v\, c_0(t_0)\, w_Q \left[\frac{1+\varepsilon}{2} - \frac{1-\varepsilon}{2}\right] = v\, c_0(t_0)\, \varepsilon\, w_Q.$$

This is the claimed value formula.

□

The stationary laws in Lemma 18 show how the lower construction keeps the target-over-behavior ratio bounded while concentrating the signal at terminal depth. The terminal value is proportional to the stationary depth mass w_Q, so increasing Q makes the two alternatives harder to separate from observed data while preserving the fixed latent-overlap radius.

The likelihood calculation rests on finite-prefix posterior bounds, a rare-action second moment, a bootstrap factor that makes the posterior control uniform in the terminal depth, signed-mass identities at terminal depth, exact conditional reward means, and finite-word KL inequalities. The next group develops these ingredients in the order in which the observed-path comparison uses them.

Lemma 19. [`lem:posterior-lower-window-power`] (Posterior window power bound) *Let $N, Q, k, d \in \mathbb{N}$, let $t_0, \zeta, C, b \in \mathbb{R}$, let $v \in \{-1, +1\}$, and let $o = ((a_0, y_0), \ldots, (a_{k-1}, y_{k-1}))$ be an observed prefix for the signed-depth family of Definition 18. Write*

$$c_0(t_0) = \frac{1 - \alpha(t_0)}{4},$$

and, for $0 \le m \le k$, write $o_{\le m}$ for the first m observed pairs. For any observed prefix o', define the raw terminal-depth posterior by

$$\pi_v(o') = \frac{\sum_h \mathbf{1}\{J(h) = Q\} \lambda_v(h; o')}{\sum_h \lambda_v(h; o')}.$$

Assume:

- ***(Nonnegative amplitude.)*** $0 \le c_0(t_0)$.
- ***(Nonnegative base.)*** $0 \le 1 - c_0(t_0) b$.
- ***(Window length.)*** $d \le k$.

- ***(Strict-prefix posterior bound.)*** *For every* $i \in \{0, \dots, k-1\}$,

$$\pi_v(o_{\le i}) \le b.$$

Define the final d*-prefix lower window by*

$$\Lambda_d(o) = \prod_{\substack{0 \le i < k \\ k-d \le i}} \left(1 - c_0(t_0)\pi_v(o_{\le i})\right).$$

Then

$$\left(1 - c_0(t_0)b\right)^d \le \Lambda_d(o).$$

Proof of Lemma 19. We prove the slightly stronger assertion with the prefix length left variable. For a prefix o of length n, write

$$\Lambda_d^{(n)}(o) = \prod_{\substack{0 \le i < n \\ n-d \le i}} \left(1 - c_0(t_0)\pi_v(o_{\le i})\right),$$

with the convention that an empty product is 1.

1. If $d = 0$, then the condition $n \le i < n$ selects no indices, so

$$\Lambda_0^{(n)}(o) = 1 = (1 - c_0(t_0)b)^0.$$

This proves the base case.

2. Assume the claim has been proved for d. Consider a prefix o of length n with $d + 1 \le n$. If $n = 0$, this inequality is impossible, so there is no case to prove. Thus write $n = k + 1$. Let o^- be the length-k prefix obtained by dropping the last observed pair:

$$o^-(i) = o(i), \qquad 0 \le i < k.$$

For every $0 \le i < k$, the first i observations of o^- are exactly the first i observations of o:

$$(o^-)_{\le i} = o_{\le i}.$$

Hence the strict-prefix posterior hypothesis for o gives

$$\pi_v\big((o^-)_{\le i}\big) = \pi_v(o_{\le i}) \le b, \qquad 0 \le i < k.$$

Since $d + 1 \le k + 1$, we have $d \le k$, so the induction hypothesis applied to o^- yields

$$(1 - c_0(t_0)b)^d \le \Lambda_d^{(k)}(o^-).$$

3. The final-window product for length $k + 1$ splits into the length-k final-window product for o^- and the last factor:

$$\Lambda_{d+1}^{(k+1)}(o) = \Lambda_d^{(k)}(o^-)\left(1 - c_0(t_0)\pi_v(o_{\le k})\right).$$

Indeed, the threshold satisfies

$$(k + 1) - (d + 1) = k - d,$$

so the eligible indices before the last one are precisely the eligible indices in the length-k window, and the remaining eligible index is $i = k$.

4. The posterior hypothesis at the last strict prefix gives $\pi_v(o_{\le k}) \le b$. Because $0 \le c_0(t_0)$, multiplying preserves the inequality,

$$c_0(t_0)\pi_v(o_{\le k}) \le c_0(t_0)b,$$

and subtracting from 1 reverses the side on which the smaller product appears:

$$1 - c_0(t_0)b \le 1 - c_0(t_0)\pi_v(o_{\le k}).$$

Together with the assumed nonnegativity $0 \le 1 - c_0(t_0)b$, we have

$$0 \le (1 - c_0(t_0)b)^d.$$

Combining this nonnegativity, the induction bound, and the last-factor bound gives

$$(1 - c_0(t_0)b)^{d+1} = (1 - c_0(t_0)b)^d(1 - c_0(t_0)b) \le \Lambda_d^{(k)}(o^-)\,(1 - c_0(t_0)\pi_v(o_{\le k})) = \Lambda_{d+1}^{(k+1)}(o).$$

This completes the induction, and hence the desired bound for the original prefix of length k. □

Lemma 20. [`lem:terminal-posterior-mature-window`] (Terminal posterior window bound) *Let $N, Q \in \mathbb{N}$, let $t_0, \zeta, C \in \mathbb{R}$, and let $v \in \{-1, +1\}$. Put $\alpha = \exp\{-(1/t_0)\}$. In the finite signed-depth model of Definition 18, for an observed prefix $o = ((a_1, y_1), \dots, (a_k, y_k))$, write*

$$\pi_v(o) = \frac{\sum_{h \in \{0,\dots,Q\}\times\{-1,+1\}} \mathbf{1}\{J(h) = Q\}\lambda_v(h; o)}{\mathsf{M}_v(o)}.$$

For $d \in \mathbb{N}$, define the lower posterior window

$$\Lambda_d(o) = \prod_{\substack{0 \le i < k \\ k - d \le i}} (1 - c_0(t_0)\,\pi_v(o_{\le i}))\,,$$

where $o_{\le i}$ is the prefix of o containing its first i action–reward pairs.

Assume:

- ***(Positive scale.)*** $t_0 > 0$.
- ***(Positive tilt.)*** $\zeta > 0$.
- ***(Overlap level.)*** $C > 1$.
- ***(Mature window.)*** $Q \le k$.

Then the terminal-depth posterior satisfies

$$\pi_v(o) \le \frac{\alpha^Q}{\Lambda_Q(o)}.$$

Proof of Lemma 20. Set $o^- := o_{\le k-Q}$, and introduce

$$\mathsf{W}_{\mathrm{fair}} := \prod_{k-Q\le i<k} b(a_{i+1})\,\frac{1}{2}, \qquad \mathsf{M}_{\mathrm{pre}} := \mathsf{M}_v(o^-),$$

$$\mathsf{U}_{\mathrm{pre}} := \sum_{u\in\{-1,+1\}} \lambda_v((0,u);o^-), \qquad \mathsf{R}_{\mathrm{corr}} := \prod_{k-Q\le i<k} \frac{\mathsf{M}_v(o_{\le i+1})}{b(a_{i+1})\,\mathsf{M}_v(o_{\le i})},$$

and

$$\mathsf{P}_{\mathrm{low}} := \Lambda_Q(o).$$

Here the products are over the final Q action–reward pairs; if $Q=0$, they are empty products and equal one.

1. The terminal-depth numerator after the full prefix factors as

$$\sum_{h\in\{0,\dots,Q\}\times\{-1,+1\}} \mathbf{1}\{J(h)=Q\}\lambda_v(h;o) = \alpha^Q\,\mathsf{W}_{\mathrm{fair}}\,\mathsf{U}_{\mathrm{pre}}.$$

Indeed, summing over the two terminal signs gives the unsigned mass at depth Q. Over a mature window of length Q, the only way to arrive at terminal depth Q is to start the window at depth 0 and advance once at each of the Q updates; this contributes α^Q, while the action and sign-symmetric factors over the same window contribute exactly $\mathsf{W}_{\mathrm{fair}}$. The remaining mass is the unsigned depth-zero mass $\mathsf{U}_{\mathrm{pre}}$ at the beginning of the window.

2. The total mass over the observed prefix factors through the same window as

$$\mathsf{M}_v(o) = \mathsf{W}_{\mathrm{fair}}\,\mathsf{M}_{\mathrm{pre}}\,\mathsf{R}_{\mathrm{corr}}.$$

This follows by iterating the one-step mass recursion across the last Q observations: each step extracts the behavior/fair-sign factor and leaves the corresponding reward-correction factor. For $Q=0$ the window is empty and the identity reduces to $\mathsf{M}_v(o)=\mathsf{M}_v(o)$.

3. The unsigned depth-zero mass at the window start is bounded by the total mass there:

$$\mathsf{U}_{\mathrm{pre}} = \sum_{u\in\{-1,+1\}} \lambda_v((0,u);o^-) \le \sum_{j=0}^{Q}\sum_{u\in\{-1,+1\}} \lambda_v((j,u);o^-) = \mathsf{M}_{\mathrm{pre}}.$$

All summands are nonnegative, since each prefix weight is a sum of nonnegative path weights.

4. The lower posterior window is positive and is dominated by the correction window:

$$0 < \mathsf{P}_{\mathrm{low}} \le \mathsf{R}_{\mathrm{corr}}.$$

For each factor in $\mathsf{P}_{\mathrm{low}}$,

$$0 \le \pi_v(o_{\le i}) \le 1, \qquad 0 \le c_0(t_0) < 1,$$

so $1 - c_0(t_0)\pi_v(o_{\le i}) > 0$. The one-step reward-correction factor at the same prefix is at least

$$1 - c_0(t_0)\pi_v(o_{\le i}),$$

and taking products over the final window gives the displayed inequality.

5. The remaining factors have the needed signs:

$$\mathsf{W}_{\mathrm{fair}} > 0, \qquad \mathsf{M}_{\mathrm{pre}} > 0, \qquad \mathsf{R}_{\mathrm{corr}} > 0, \qquad \alpha^Q \ge 0.$$

The first inequality is the product of positive behavior weights and $1/2$ factors, while the second is positivity of the finite total mass at the window start. The third follows from Step 4 by $0 < \mathsf{P}_{\mathrm{low}} \le \mathsf{R}_{\mathrm{corr}}$. Finally, $\alpha = \exp\{-(1/t_0)\} > 0$, so its Q-th power is nonnegative.

6. Combining the two window factorizations with the monotonicity bounds gives

$$\mathsf{U}_{\mathrm{pre}}\mathsf{P}_{\mathrm{low}} \le \mathsf{M}_{\mathrm{pre}}\mathsf{R}_{\mathrm{corr}},$$

because $0 \le \mathsf{U}_{\mathrm{pre}} \le \mathsf{M}_{\mathrm{pre}}$, $0 < \mathsf{P}_{\mathrm{low}} \le \mathsf{R}_{\mathrm{corr}}$, and $\mathsf{M}_{\mathrm{pre}} \ge 0$. Multiplying by the nonnegative factor $\alpha^Q \mathsf{W}_{\mathrm{fair}}$ yields

$$\alpha^Q \mathsf{W}_{\mathrm{fair}}\mathsf{U}_{\mathrm{pre}}\mathsf{P}_{\mathrm{low}} \le \alpha^Q \mathsf{W}_{\mathrm{fair}}\mathsf{M}_{\mathrm{pre}}\mathsf{R}_{\mathrm{corr}}.$$

Using the identities from Steps 1 and 2, this is

$$\left(\sum_h \mathbf{1}\{J(h) = Q\}\lambda_v(h; o)\right)\mathsf{P}_{\mathrm{low}} \le \alpha^Q\, \mathsf{M}_v(o).$$

Since $\mathsf{M}_v(o) > 0$ and $\mathsf{P}_{\mathrm{low}} > 0$, division gives

$$\pi_v(o) = \frac{\sum_h \mathbf{1}\{J(h) = Q\}\lambda_v(h; o)}{\mathsf{M}_v(o)} \le \frac{\alpha^Q}{\mathsf{P}_{\mathrm{low}}} = \frac{\alpha^Q}{\Lambda_Q(o)},$$

as claimed. □

Lemma 21. [lem:terminal-posterior-initial-window] (Initial-window posterior bound) *Let $N, Q \in \mathbb{N}$, let $t_0, \zeta, C \in \mathbb{R}$, let $v \in \{-1, +1\}$, let $k \in \mathbb{N}$, and let*

$$o : \{0, \dots, k-1\} \to \{0, 1\} \times \{0, 1\}$$

be an observed action/reward prefix for the signed-depth construction of Definition 18. Write $\alpha = e^{-1/t_0}$. For any prefix o', define the raw terminal-depth posterior by

$$\pi_v(o') = \frac{\sum_{h:\, J(h)=Q} \lambda_v(h; o')}{\sum_h \lambda_v(h; o')}.$$

Define the initial lower posterior window over the available prefixes by

$$\Lambda_k(o) = \prod_{i=0}^{k-1} \left(1 - c_0\, \pi_v(o_{\le i})\right),$$

where $o_{\le i}$ is the prefix of o of length i.

Assume:

- ***(Positive scale.)*** $t_0 > 0$.
- ***(Positive overlap exponent.)*** $\zeta > 0$.
- ***(Nontrivial overlap radius.)*** $C > 1$.
- ***(Initial window.)*** $k < Q$.

Then the terminal-depth posterior after observing o satisfies

$$\pi_v(o) \le \frac{\alpha^Q}{\Lambda_k(o)}.$$

Proof of Lemma 21. Write $o_i = (a_i, y_i)$ for $0 \le i < k$, and let $o_{\le 0}$ denote the empty prefix. For any prefix ω, set

$$\mathsf{Mass}_v(\omega) := \sum_h \lambda_v(h;\omega), \qquad \mathsf{Term}_v(\omega) := \sum_{h:\, J(h)=Q} \lambda_v(h;\omega).$$

Let $b_\zeta(a)$ be the behavior-policy probability of action a from Definition 18, and define

$$\mathsf{Fair}_k(o) := \prod_{i=0}^{k-1} b_\zeta(a_i)\,\frac{1}{2}.$$

For a prefix ω and a next observation (a, y), define the reward-correction factor

$$\Gamma_v(\omega,(a,y)) := \frac{2\sum_h \lambda_v(h;\omega)\,\frac{1+(2y-1)vc_0U(h)\mathbf{1}\{J(h)=Q\}}{2}}{\mathsf{Mass}_v(\omega)},$$

whenever $\mathsf{Mass}_v(\omega) > 0$, with any fixed value when the denominator is zero. Along the observed prefixes below the denominator is positive. Finally set

$$\mathsf{Corr}_k(o) := \prod_{i=0}^{k-1} \Gamma_v(o_{\le i}, o_i), \qquad \mathsf{Low}_k(o) := \Lambda_k(o).$$

1. Since $k < Q$, a hidden path contributing to $\mathsf{Term}_v(o)$ starts, after summing over the sign coordinate, from depth $Q-k$ and then makes k successive depth-increment moves. Thus the terminal numerator factors as

   $$\mathsf{Term}_v(o) = \alpha^k w_{Q-k}\,\mathsf{Fair}_k(o),$$

   where

   $$w_j = \frac{(1-\alpha)\alpha^j}{1-\alpha^{Q+1}}, \qquad 0 \le j \le Q,$$

   is the signed-depth initial depth mass from Definition 18. Since $k < Q$,

   $$\alpha^k w_{Q-k} = \alpha^k \frac{(1-\alpha)\alpha^{Q-k}}{1-\alpha^{Q+1}} = \frac{(1-\alpha)\alpha^Q}{1-\alpha^{Q+1}} = w_Q,$$

   and therefore

   $$\mathsf{Term}_v(o) = w_Q\,\mathsf{Fair}_k(o).$$

2. Summing the one-step kernel over the next hidden state leaves total hidden-transition mass one. Hence each extension satisfies

   $$\mathsf{Mass}_v(o_{\le i+1}) = b_\zeta(a_i)\,\frac{1}{2}\,\Gamma_v(o_{\le i}, o_i)\,\mathsf{Mass}_v(o_{\le i}).$$

   Multiplying this identity for $i = 0, \dots, k-1$, with the empty-product convention when $k = 0$, gives

   $$\mathsf{Mass}_v(o) = \mathsf{Fair}_k(o)\,\mathsf{Mass}_v(o_{\le 0})\,\mathsf{Corr}_k(o).$$

3. The empty-prefix mass is one. Indeed, the initial weights sum to

$$\sum_{j=0}^{Q} w_j \sum_{u\in\{-1,+1\}} \frac{1+u\varepsilon_C L^{-j}}{2} = \sum_{j=0}^{Q} w_j = 1.$$

Thus

$$\mathsf{Mass}_v(o_{\le 0}) = 1.$$

4. The lower window is positive and is bounded by the correction window:

$$0 < \mathsf{Low}_k(o), \qquad \mathsf{Low}_k(o) \le \mathsf{Corr}_k(o).$$

To see this, $t_0 > 0$ gives $0 < \alpha < 1$, hence

$$0 \le c_0 = \frac{1-\alpha}{4} < 1.$$

For each observed prefix $o_{\le i}$, the filter weights are nonnegative and their total mass is positive, so

$$0 \le \pi_v(o_{\le i}) \le 1.$$

Consequently every factor $1 - c_0\pi_v(o_{\le i})$ is positive. Moreover, for each i,

$$\Gamma_v(o_{\le i}, o_i) \ge (1-\pi_v(o_{\le i}))\cdot 1 + \pi_v(o_{\le i})(1-c_0) = 1 - c_0\pi_v(o_{\le i}),$$

because the doubled reward factor equals 1 off terminal depth and is at least $1 - c_0$ on terminal depth. Multiplying these pointwise inequalities gives the displayed product bound.

5. The fair-window factor is strictly positive:

$$0 < \mathsf{Fair}_k(o).$$

Each behavior probability $b_\zeta(a_i)$ is positive under $\zeta > 0$, and each reward-symbol base factor $1/2$ is positive; the empty product is 1.

6. The terminal depth mass is bounded by α^Q:

$$w_Q \le \alpha^Q.$$

Indeed,

$$w_Q = \frac{(1-\alpha)\alpha^Q}{1-\alpha^{Q+1}},$$

and $0 < \alpha < 1$ implies $1 - \alpha^{Q+1} > 0$ and $\alpha^{Q+1} \le \alpha$. After multiplying by the positive denominator, the desired inequality is

$$(1-\alpha)\alpha^Q \le (1-\alpha^{Q+1})\alpha^Q,$$

which is exactly $\alpha^{Q+1} \le \alpha$ times the nonnegative factor α^Q.

7. Combining the previous bounds,

$$w_Q \,\mathsf{Low}_k(o) \le \alpha^Q \,\mathsf{Corr}_k(o).$$

Multiplying by the nonnegative factor $\mathsf{Fair}_k(o)$, and using the identities above, yields

$$\mathsf{Term}_v(o)\,\mathsf{Low}_k(o) \le \alpha^Q \,\mathsf{Mass}_v(o).$$

Since $\mathsf{Mass}_v(o) > 0$ and $\mathsf{Low}_k(o) > 0$, division gives

$$\pi_v(o) = \frac{\mathsf{Term}_v(o)}{\mathsf{Mass}_v(o)} \le \frac{\alpha^Q}{\mathsf{Low}_k(o)} = \frac{\alpha^Q}{\Lambda_k(o)}.$$

□

Lemma 22. [lem:terminal-posterior-crude-bound] (Crude terminal posterior bound) *Let $N, Q, k \in \mathbb{N}$, let $t_0, \zeta, C \in \mathbb{R}$, and let $v \in \{-1, +1\}$. Suppose that*

- ***(Positive temperature.)*** *$t_0 > 0$.*
- ***(Positive log-overlap.)*** *$\zeta > 0$.*
- ***(Overlap level.)*** *$C > 1$.*

For the finite signed-depth model of Definition 18, write

$$\alpha = \exp(-1/t_0), \qquad c_0 = \frac{1-\alpha}{4}.$$

For every observed prefix $o \in (\{0,1\} \times \{0,1\})^k$, define the raw terminal-depth posterior by

$$\pi_v(o) = \frac{\sum_h \mathbf{1}\{J(h) = Q\}\,\lambda_v(h;o)}{\mathsf{M}_v(o)}, \qquad \mathsf{M}_v(o) = \sum_h \lambda_v(h;o),$$

where the sums range over the signed-depth hidden states at depth horizon Q. Then

$$\pi_v(o) \le \left(\frac{\alpha}{1-c_0}\right)^Q.$$

Proof of Lemma 22. Write $\alpha = \exp(-1/t_0)$ and $c_0 = (1-\alpha)/4$.

1. Since $t_0 > 0$, one has $0 < \alpha < 1$. Hence

$$0 \le c_0 = \frac{1-\alpha}{4} < 1, \qquad 0 < 1 - c_0, \qquad 0 \le 1 - c_0 \cdot 1.$$

For every observed prefix o', the raw terminal-depth posterior satisfies

$$0 \le \pi_v(o') \le 1.$$

Indeed, the numerator is a sum of nonnegative filter masses over terminal-depth states, the denominator $\mathsf{M}_v(o')$ is positive, and the terminal-depth mass is bounded by the total mass.

2. First suppose $Q \le k$. By Lemma 20,

$$\pi_v(o) \le \frac{\alpha^Q}{\Lambda_Q(o)}.$$

Apply Lemma 19 with $b = 1$ and $d = Q$. The hypotheses are exactly those checked above: $0 \le c_0$, $0 \le 1 - c_0 \cdot 1$, $Q \le k$, and $\pi_v(o_{\le i}) \le 1$ for every strict prefix. Thus

$$(1 - c_0)^Q \le \Lambda_Q(o).$$

Since $\alpha^Q \ge 0$ and $1 - c_0 > 0$, division by the larger positive denominator gives

$$\pi_v(o) \le \frac{\alpha^Q}{(1 - c_0)^Q} = \left(\frac{\alpha}{1 - c_0}\right)^Q.$$

3. It remains to consider $k < Q$. By Lemma 21,

$$\pi_v(o) \le \frac{\alpha^Q}{\Lambda_k(o)}.$$

Apply Lemma 19 with $b = 1$ and $d = k$. Since $k \le k$, and the prefix posterior bound $\pi_v(o_{\le i}) \le 1$ holds for every strict prefix, this gives

$$(1 - c_0)^k \le \Lambda_k(o).$$

Therefore

$$\pi_v(o) \le \frac{\alpha^Q}{(1 - c_0)^k}.$$

Because $0 < 1 - c_0 \le 1$ and $k < Q$, monotonicity of powers on $[0, 1]$ yields

$$(1 - c_0)^Q \le (1 - c_0)^k.$$

Dividing the nonnegative numerator α^Q by these positive denominators gives

$$\frac{\alpha^Q}{(1 - c_0)^k} \le \frac{\alpha^Q}{(1 - c_0)^Q} = \left(\frac{\alpha}{1 - c_0}\right)^Q.$$

Combining the displayed inequalities proves the claimed bound in the initial-window case as well. □

Lemma 23. [`lem:refined-terminal-posterior`] (Refined terminal posterior bound) *Let N, Q, k be nonnegative integers. Suppose that*

- ***(Positive scale.)*** $t_0 > 0$.
- ***(Positive perturbation.)*** $\zeta > 0$.
- ***(Overlap level.)*** $C > 1$.

Write $\alpha = \exp(-1/t_0)$ for the contraction parameter and $c_0 = (1-\alpha)/4$ for the signed-depth reward-amplitude constant. For an alternative $v \in \{-1, +1\}$ in the signed-depth family of Definition 18 and any observed prefix $o = (a_1, y_1, \ldots, a_k, y_k)$, let $\pi_{N,Q,v}(o)$ be the raw terminal-depth posterior: the run-local unnormalized filter mass at latent depth Q, divided by the total run-local unnormalized filter mass of the prefix. Then

$$\pi_{N,Q,v}(o) \le \frac{\alpha^Q}{\left\{1 - c_0 \left(\frac{\alpha}{1-c_0}\right)^Q\right\}^Q}.$$

Proof of Lemma 23. 1. Let

$$\Gamma := \frac{\alpha}{1-c_0}.$$

Since $t_0 > 0$, the contraction parameter satisfies $0 < \alpha < 1$. Hence

$$0 \le c_0 = \frac{1-\alpha}{4} \le \frac{1}{4}, \qquad 1 - c_0 > 0,$$

and therefore $\Gamma \ge 0$. Moreover,

$$\Gamma < 1 \iff \alpha < 1 - \frac{1-\alpha}{4} \iff \alpha < 1.$$

Thus $\Gamma^Q \le 1$. Consequently

$$B_Q := 1 - c_0 \Gamma^Q$$

is strictly positive; indeed $c_0 \Gamma^Q \le c_0 \le 1/4$.

2. For $0 \le d \le k$, define the lower posterior window

$$\Lambda_d(o) := \prod_{s=k-d+1}^{k} \left(1 - c_0\, \pi_{N,Q,v}(o_{\le s-1})\right),$$

where $o_{\le m}$ denotes the first m observed action–reward pairs, with $o_{\le 0}$ the empty prefix; for $d = 0$ this is the empty product and equals 1. If $Q \le k$, Lemma 20 applies under the present hypotheses and gives

$$\pi_{N,Q,v}(o) \le \frac{\alpha^Q}{\Lambda_Q(o)}.$$

If $k < Q$, Lemma 21 applies under the present hypotheses and gives

$$\pi_{N,Q,v}(o) \le \frac{\alpha^Q}{\Lambda_k(o)}.$$

These two estimates are exactly the two window cases needed below.

3. The crude terminal-posterior estimate in Lemma 22 applies to every observed prefix o', of any length, under the same assumptions $t_0 > 0$, $\zeta > 0$, and $C > 1$. With the present notation it states

$$\pi_{N,Q,v}(o') \le \left(\frac{\alpha}{1-c_0}\right)^Q = \Gamma^Q.$$

This supplies the prefix-uniform bootstrap bound used in the lower-window estimate.

4. Apply Lemma 19 with

$$b := \Gamma^Q.$$

Its nonnegative-amplitude hypothesis is $0 \le c_0$, established in the first step, and its nonnegative-base hypothesis is

$$0 \le 1 - c_0\Gamma^Q = B_Q,$$

also established there. Its strict-prefix posterior hypothesis holds because the preceding step gives, for every prefix appearing in the window,

$$\pi_{N,Q,v}(o_{\le s-1}) \le \Gamma^Q.$$

The indexing in Lemma 19 ranges over the final d strict prefixes of o, precisely the factors in $\Lambda_d(o)$. Hence, for every $d \le k$,

$$B_Q^d = \left(1 - c_0\Gamma^Q\right)^d \le \Lambda_d(o).$$

5. If $Q \le k$, the terminal-window bound with $d = Q$ and the preceding display give

$$\pi_{N,Q,v}(o) \le \frac{\alpha^Q}{\Lambda_Q(o)} \le \frac{\alpha^Q}{B_Q^Q}.$$

If $k < Q$, the terminal-window bound with $d = k$ gives

$$\pi_{N,Q,v}(o) \le \frac{\alpha^Q}{\Lambda_k(o)} \le \frac{\alpha^Q}{B_Q^k}.$$

Since $0 < B_Q \le 1$ and $k < Q$, we have $B_Q^Q \le B_Q^k$, hence

$$\frac{\alpha^Q}{B_Q^k} \le \frac{\alpha^Q}{B_Q^Q}.$$

In both cases,

$$\pi_{N,Q,v}(o) \le \frac{\alpha^Q}{\left(1 - c_0\left(\frac{\alpha}{1-c_0}\right)^Q\right)^Q},$$

which is the claimed bound.

□

The posterior bounds in Lemmas 19 to 23 separate the survival of the terminal depth from the normalization accumulated along the observed prefix. The mature-window and initial-window cases cover prefixes on the two sides of the terminal depth Q, the crude bound supplies a uniform starting point, and the refined statement packages these ingredients into the terminal-posterior control used below.

Lemma 24. [`lem:rare-action-second-moment-exact`] (Rare-action second moment) *Let $Q, k \in \mathbb{N}$, let $t_0, \zeta, C \in \mathbb{R}$, and let $v \in \{-1, +1\}$. Assume:*

- ***(Time scale.)*** $t_0 > 0$.
- ***(Overlap exponent.)*** $\zeta > 0$.
- ***(Radius constant.)*** $C > 1$.

Write $L = \exp(\zeta)$, *let* $\ell = \min\{Q, k\}$, *and consider the finite signed-depth model of Definition 18 at horizon* $k+1$. *For an action/reward-symbol prefix*

$$o = ((a_1, y_1), \dots, (a_k, y_k)),$$

let $\pi_k^v(o)$ *be its prefix mass under this model, and define the rare-action-history factor from Definition 25 by*

$$\mathfrak{a}_k(o) = L^{-(Q-\ell)} \prod_{i:\, k-\ell<i\le k} \mathbf{1}\{a_i = 1\},$$

with the empty product equal to one. Then

$$\sum_o \pi_k^v(o)\, \mathfrak{a}_k(o)^2 = L^{-2(Q-\ell)} \left(\frac{1}{L}\right)^{|\{i:\ k-\ell<i\le k\}|},$$

where the sum ranges over all action/reward-symbol prefixes of length k.

Proof of Lemma 24. Let

$$S := \{i \in \{1, \dots, k\} : k - \ell < i \le k\}.$$

Reindex each observed prefix o by its action sequence $a = (a_1, \dots, a_k)$ and reward-symbol sequence $y = (y_1, \dots, y_k)$. This bijection gives

$$\sum_o \pi_k^v(o)\mathfrak{a}_k(o)^2 = \sum_a \sum_y \pi_k^v(a, y) \left(L^{-(Q-\ell)} \prod_{i\in S} \mathbf{1}\{a_i = 1\}\right)^2.$$

For each action sequence a, the product

$$I_S(a) := \prod_{i\in S} \mathbf{1}\{a_i = 1\}$$

is either 0 or 1. Hence $I_S(a)^2 = I_S(a)$, including the empty-product case $S = \varnothing$, where $I_S(a) = 1$. Therefore

$$\sum_o \pi_k^v(o)\mathfrak{a}_k(o)^2 = L^{-2(Q-\ell)} \sum_a I_S(a) \sum_y \pi_k^v(a, y).$$

After summing over the reward symbols with the actions fixed, the observed prefix mass factors as the behavior action probability product:

$$\sum_y \pi_k^v(a, y) = \prod_{i=1}^k b(a_i), \qquad b(1) = \frac{1}{L}, \quad b(0) = 1 - \frac{1}{L}.$$

This is the reward-summed prefix identity for the signed-depth family of Definition 18. Thus

$$\sum_o \pi_k^v(o)\mathfrak{a}_k(o)^2 = L^{-2(Q-\ell)} \sum_a \left(\prod_{i=1}^k b(a_i)\right) \left(\prod_{i\in S} \mathbf{1}\{a_i = 1\}\right).$$

It remains to evaluate the action sum. Since the behavior action law is product-form,

$$\sum_a \left(\prod_{i=1}^k b(a_i)\right)\left(\prod_{i\in S}\mathbf{1}\{a_i=1\}\right) = \prod_{i=1}^k \sum_{a_i\in\{0,1\}} b(a_i)\begin{cases}\mathbf{1}\{a_i=1\}, & i\in S,\\ 1, & i\notin S.\end{cases}$$

For $i \in S$, the inner sum is $b(1) = 1/L$; for $i \notin S$, it is $b(0) + b(1) = 1$. Hence

$$\sum_a \left(\prod_{i=1}^k b(a_i)\right)\left(\prod_{i\in S}\mathbf{1}\{a_i=1\}\right) = \left(\frac{1}{L}\right)^{|S|}.$$

Combining the preceding displays and recalling $S = \{i : k-\ell < i \le k\}$ gives

$$\sum_o \pi_k^v(o)\mathfrak{a}_k(o)^2 = L^{-2(Q-\ell)}\left(\frac{1}{L}\right)^{|\{i:\ k-\ell<i\le k\}|},$$

as claimed. □

Lemma 24 gives the exact second moment of the visible rare-action factor. The displayed expression isolates the contribution of the forced terminal-depth action pattern, and this is the L^{-Q} term that enters the observed KL calculation after summing conditional mean differences.

Lemma 25. [lem:finite-signed-depth-kl] (Finite observed KL bound) *Let $T, Q \in \mathbb{N}$ and $t_0, \zeta, C, K_0 \in \mathbb{R}$. Write $\alpha = e^{-1/t_0}$, $L = e^{\zeta}$, $c_0 = (1-\alpha)/4$, and $\varepsilon_C = \min\{(C-1)/2, 1/2\}$. Suppose that*

- ***(Positive scale.)*** *$t_0 > 0$, $\zeta > 0$, $C > 1$, and $K_0 > 0$.*
- ***(Uniform terminal posterior.)*** *For every horizon N, every alternative index $v \in \{-1, +1\}$, every time $t < N$, and every observed prefix o of actions and two-symbol rewards before t, the finite conditional terminal-depth probability $q_t^v(o)$ from Definition 25 satisfies*

$$q_t^v(o) \le K_0\alpha^Q.$$

Then the finite observed-word laws of the signed-depth alternatives of Definition 18, taken before the reward symbols are decoded as real rewards, satisfy

$$\mathrm{KL}\left(\mathbb{P}^{\mathrm{fin},T}_{Q,+1}\middle\|\mathbb{P}^{\mathrm{fin},T}_{Q,-1}\right) \le \frac{4c_0^2K_0^2}{1-c_0^2}\,T\,\varepsilon_C^2\,\alpha^{2Q}\,L^{-Q}.$$

Proof of Lemma 25. Let

$$D := \frac{4c_0^2K_0^2}{1-c_0^2}.$$

Since $t_0 > 0$, $\alpha = \exp(-1/t_0)$ lies in $(0,1)$, and hence

$$0 < c_0 = \frac{1-\alpha}{4} < 1, \qquad 0 \le 1-c_0^2, \qquad D \ge 0.$$

Also $\varepsilon_C^2 \ge 0$, $\alpha^{2Q} \ge 0$, and $L^{-Q} \ge 0$. We prove the claimed bound by induction on the horizon T.

Step 1. The zero-horizon case. For $T = 0$, the observed word has a single possible value. The finite-word chain rule gives

$$\mathrm{KL}\Big(\mathbb{P}^{\mathrm{fin},0}_{Q,+1}\Big\|\mathbb{P}^{\mathrm{fin},0}_{Q,-1}\Big) = 0,$$

so the desired inequality is immediate.

Step 2. Full support and projectivity. Fix $k \geq 0$ and assume the bound at horizon k. Every observed word

$$w = ((x_1, a_1, y_1), \ldots, (x_{k+1}, a_{k+1}, y_{k+1}))$$

has positive mass under each signed-depth alternative. Indeed, the observed coordinate has only one value, and one may choose a hidden trajectory with the displayed observed actions and reward symbols. The initial weights are positive because $0 < \alpha < 1$ and $0 < \varepsilon_C \leq 1/2$; the behavior weights are positive because $L = \exp(\zeta) > 1$; and the selected hidden transitions have positive reset or continuation weight, with the case $Q = 0$ handled directly by the reset transition. Thus

$$\mathbb{P}^{\mathrm{fin},k+1}_{Q,v}(w) > 0, \qquad v \in \{-1, +1\}.$$

Moreover, summing the chronological law over the last observed action-reward symbol gives the prefix law:

$$\mathrm{pref}_k\, \mathbb{P}^{\mathrm{fin},k+1}_{Q,v} = \mathbb{P}^{\mathrm{fin},k}_{Q,v}.$$

Consequently the last-coordinate chain rule applies and yields

$$\begin{aligned}
&\mathrm{KL}\Big(\mathbb{P}^{\mathrm{fin},k+1}_{Q,+1}\Big\|\mathbb{P}^{\mathrm{fin},k+1}_{Q,-1}\Big) \\
&\quad = \mathrm{KL}\Big(\mathbb{P}^{\mathrm{fin},k}_{Q,+1}\Big\|\mathbb{P}^{\mathrm{fin},k}_{Q,-1}\Big) + \sum_{w} \mathbb{P}^{\mathrm{fin},k}_{Q,+1}(w)\ \mathrm{KL}\Big(\widetilde{P}^{\mathrm{next}}_{+,w}\Big\|\widetilde{P}^{\mathrm{next}}_{-,w}\Big),
\end{aligned}$$

where $\widetilde{P}^{\mathrm{next}}_{v,w}$ is the conditional law of the next full observed symbol $(x, a, r) \in \{0\} \times \{0, 1\} \times \{0, 1\}$ after prefix w.

Step 3. The one-step increment. For a prefix w of length k, write $o(w) = ((a_1, y_1), \ldots, (a_k, y_k))$ for its action/reward-symbol projection and define

$$\mathfrak{a}_k(o) := L^{-(Q-\min\{Q,k\})} \prod_{i:\, k-\min\{Q,k\}<i\leq k} \mathbf{1}\{a_i = 1\}.$$

For $v \in \{-1, +1\}$, let $P^{\mathrm{ar}}_{v,w}$ be the image of $\widetilde{P}^{\mathrm{next}}_{v,w}$ under the bijection

$$\{0\} \times \{0, 1\} \times \{0, 1\} \longrightarrow \{0, 1\} \times \{0, 1\}, \qquad (0, a, r) \longmapsto (a, r).$$

The KL divergence is unchanged by this relabelling, so

$$\mathrm{KL}\Big(\widetilde{P}^{\mathrm{next}}_{+,w}\Big\|\widetilde{P}^{\mathrm{next}}_{-,w}\Big) = \mathrm{KL}\big(P^{\mathrm{ar}}_{+,w}\big\|P^{\mathrm{ar}}_{-,w}\big)\,.$$

For $a \in \{0, 1\}$ and reward symbol $r \in \{0, 1\}$, write $\bar{y}(r) = -1$ if $r = 0$ and $\bar{y}(r) = 1$ if $r = 1$. Let

$$m_v(w) := \mathbb{E}_v[Y_{k+1} \mid \mathcal{F}^{\mathrm{obs}}_k = o(w)]$$

with the conditional ratio evaluated on the finite prefix mass; full support from Step 2 makes this ordinary conditional expectation on the present prefix. Summing the chronological law over all hidden states at time $k+1$ gives, for every (a, r),

$$\begin{aligned} P^{\mathrm{ar}}_{v,w}(a,r) &= \frac{\Pr_v\{\mathcal{F}^{\mathrm{obs}}_k = o(w),\ A_{k+1} = a,\ Y_{k+1} = \bar{y}(r)\}}{\Pr_v\{\mathcal{F}^{\mathrm{obs}}_k = o(w)\}} \\ &= b(a)\ \Pr_v\{Y_{k+1} = \bar{y}(r) \mid \mathcal{F}^{\mathrm{obs}}_k = o(w)\} \\ &= b(a)\,\frac{1 + \bar{y}(r) m_v(w)}{2}, \end{aligned}$$

where $b(1) = L^{-1}$ and $b(0) = 1 - L^{-1}$. The second equality uses the behavior policy in Definition 18, which chooses the next action with mass $b(a)$ independently of the alternative after the observed prefix. The last equality is the two-symbol reward law determined by its mean: for a variable taking values $\{-1, +1\}$, mass $(1+m)/2$ is assigned to $+1$ and mass $(1-m)/2$ to -1. The positivity assumptions give $|m_v(w)| < 1$, so all displayed masses are positive.

By Lemma 30, these means are

$$m_+(w) = c_0 \varepsilon_C \mathfrak{a}_k(o(w)) q^{+1}_{k+1}(o(w)), \qquad m_-(w) = -c_0 \varepsilon_C \mathfrak{a}_k(o(w)) q^{-1}_{k+1}(o(w)).$$

Using the preceding mass formula and writing out the finite KL sum,

$$\begin{aligned} \mathrm{KL}\big(P^{\mathrm{ar}}_{+,w} \big\| P^{\mathrm{ar}}_{-,w}\big) &= \sum_{a\in\{0,1\}} \sum_{r\in\{0,1\}} b(a) \frac{1 + \bar{y}(r) m_+(w)}{2} \log \frac{b(a)(1 + \bar{y}(r) m_+(w))/2}{b(a)(1 + \bar{y}(r) m_-(w))/2} \\ &= \sum_{r\in\{0,1\}} \frac{1 + \bar{y}(r) m_+(w)}{2} \log \frac{1 + \bar{y}(r) m_+(w)}{1 + \bar{y}(r) m_-(w)}. \end{aligned}$$

Here $\sum_a b(a) = 1$, and the positive common factor $b(a)$ cancels inside the logarithm. The standard two-point comparison, valid for $|m_+(w)| < 1$ and $|m_-(w)| < 1$, yields

$$\mathrm{KL}\Big(\widetilde{P}^{\mathrm{next}}_{+,w} \Big\| \widetilde{P}^{\mathrm{next}}_{-,w}\Big) \le \frac{(m_+(w) - m_-(w))^2}{1 - m_-(w)^2}.$$

Step 4. The conditional chi-square calibration. Set

$$d(w) := c_0 \varepsilon_C \mathfrak{a}_k(o(w)).$$

The signed-depth factors are nonnegative: $c_0 > 0$, $\varepsilon_C \ge 0$, and $\mathfrak{a}_k(o(w)) \ge 0$. The action factor also obeys

$$0 \le \mathfrak{a}_k(o(w)) \le 1, \qquad \varepsilon_C \mathfrak{a}_k(o(w)) \le 1.$$

Hence

$$0 \le d(w) \quad \text{and} \quad d(w)^2 = c_0^2 \big(\varepsilon_C \mathfrak{a}_k(o(w))\big)^2 \le c_0^2.$$

The terminal posteriors are probabilities, so

$$0 \le q^v_{k+1}(o(w)) \le 1, \qquad v \in \{-1, +1\}.$$

Together with the assumed posterior bound,

$$q^v_{k+1}(o(w)) \le K_0 \alpha^Q, \qquad v \in \{-1, +1\},$$

this gives

$$q_{k+1}^{+1}(o(w)) + q_{k+1}^{-1}(o(w)) \le 2K_0\alpha^Q.$$

Substituting the conditional-mean identities from Step 3,

$$\begin{aligned}(m_+(w) - m_-(w))^2 &= d(w)^2\big(q_{k+1}^{+1}(o(w)) + q_{k+1}^{-1}(o(w))\big)^2 \\ &= c_0^2\varepsilon_C^2\mathfrak{a}_k(o(w))^2\big(q_{k+1}^{+1}(o(w)) + q_{k+1}^{-1}(o(w))\big)^2 \\ &\le 4c_0^2K_0^2\varepsilon_C^2\mathfrak{a}_k(o(w))^2\alpha^{2Q}.\end{aligned}$$

The denominator is compared using both $d(w)^2 \le c_0^2$ and $q_{k+1}^{-1}(o(w))^2 \le 1$:

$$\begin{aligned}1 - m_-(w)^2 &= 1 - d(w)^2 q_{k+1}^{-1}(o(w))^2 \\ &\ge 1 - c_0^2.\end{aligned}$$

Since $0 < c_0 < 1$, the lower bound $1 - c_0^2$ is positive. Applying the preceding numerator and denominator comparisons to the one-step KL bound from Step 3 gives

$$\begin{aligned}\mathrm{KL}\Big(\widetilde{P}_{+,w}^{\mathrm{next}}\Big\|\widetilde{P}_{-,w}^{\mathrm{next}}\Big) &\le \frac{4c_0^2K_0^2\varepsilon_C^2\mathfrak{a}_k(o(w))^2\alpha^{2Q}}{1 - c_0^2} \\ &= D\,\varepsilon_C^2\alpha^{2Q}\mathfrak{a}_k(o(w))^2.\end{aligned}$$

Step 5. Averaging over prefixes. For an action/reward-symbol prefix o of length k, define its finite prefix mass using the horizon $k+1$ model by

$$\pi_{k,k+1}^{+1}(o) := \Pr_{Q,+1}^{\mathrm{fin},k+1}\big((A_1,Y_1),\dots,(A_k,Y_k) = o\big).$$

For every observed word w of length k, the observed state coordinate is fixed, and the mass of w under the horizon-k observed law is

$$\mathbb{P}_{Q,+1}^{\mathrm{fin},k}(w) = \pi_{k,k+1}^{+1}(o(w)).$$

Thus the average increment is bounded by

$$\begin{aligned}&\sum_w \mathbb{P}_{Q,+1}^{\mathrm{fin},k}(w)\ \mathrm{KL}\Big(\widetilde{P}_{+,w}^{\mathrm{next}}\Big\|\widetilde{P}_{-,w}^{\mathrm{next}}\Big) \\ &\quad\le D\,\varepsilon_C^2\alpha^{2Q}\sum_w \pi_{k,k+1}^{+1}(o(w))\mathfrak{a}_k(o(w))^2.\end{aligned}$$

The map $w \mapsto o(w)$ is a bijection from length-k observed words to length-k action/reward-symbol prefixes, because the observed state coordinate has the unique value. Reindexing the finite sum gives

$$\sum_w \pi_{k,k+1}^{+1}(o(w))\mathfrak{a}_k(o(w))^2 = \sum_o \pi_{k,k+1}^{+1}(o)\mathfrak{a}_k(o)^2.$$

By Lemma 24, with $\ell = \min\{Q,k\}$,

$$\sum_o \pi_{k,k+1}^{+1}(o)\mathfrak{a}_k(o)^2 = L^{-2(Q-\ell)}\left(\frac{1}{L}\right)^{|\{i:\ k-\ell<i\le k\}|}.$$

The index set $\{i:\ k-\ell<i\le k\}$ contains exactly the final ℓ positions of the length-k prefix; if $\ell=0$ it is empty, and otherwise the shift $j\mapsto k-\ell+j$, $1\le j\le \ell$, is a bijection. Therefore

$$\sum_{o}\pi^{+1}_{k,k+1}(o)\mathfrak{a}_k(o)^2 = L^{-2(Q-\ell)}L^{-\ell} = L^{-2Q+\ell}\le L^{-Q},$$

because $\ell\le Q$ and $L\ge 1$. Hence the increment is bounded by

$$D\,\varepsilon_C^2\alpha^{2Q}L^{-Q}.$$

Step 6. Closing the induction. Using the induction hypothesis and the increment bound,

$$\begin{aligned}\mathrm{KL}\Big(\mathbb{P}^{\mathrm{fin},k+1}_{Q,+1}\Big\|\mathbb{P}^{\mathrm{fin},k+1}_{Q,-1}\Big) &\le D\,k\,\varepsilon_C^2\alpha^{2Q}L^{-Q}+D\,\varepsilon_C^2\alpha^{2Q}L^{-Q}\\ &= D\,(k+1)\,\varepsilon_C^2\alpha^{2Q}L^{-Q}.\end{aligned}$$

Substituting the definition of D gives the asserted bound for $T=k+1$, completing the induction. □

Lemma 26. [lem:signed-depth-bootstrap-factor] (Uniform bootstrap factor) *Let $t_0>0$ and write $\alpha=e^{-1/t_0}$, $c_0=(1-\alpha)/4$, $\eta_0=\alpha/(1-c_0)$. Then there exists $K_0>0$ such that*

$$\frac{1}{\big(1-c_0\big(\alpha/(1-c_0)\big)^Q\big)^Q}\le K_0\qquad(Q\in\mathbb{N}).$$

Proof of Lemma 26. Let

$$\alpha=e^{-1/t_0},\qquad c_0=\frac{1-\alpha}{4},\qquad \eta_0=\frac{\alpha}{1-c_0}.$$

We prove the stated uniform bound by constructing the constant explicitly.

1. Since $t_0>0$, one has $0<\alpha<1$. Hence

   $$0\le c_0=\frac{1-\alpha}{4}\le\frac14,\qquad 1-c_0=\frac{3+\alpha}{4}>0,$$

   and therefore

   $$0\le\eta_0=\frac{\alpha}{1-c_0}<1,$$

   because $\alpha<1-c_0$ is equivalent to $\alpha<(3+\alpha)/4$, or $\alpha<1$.

2. The sequence $Q\eta_0^Q$ is bounded above. To see this in the same normalization, set

   $$\bar\eta=\frac{1+\eta_0}{2}.$$

   Then $0\le\bar\eta<1$ and $|\eta_0|<\bar\eta$. Writing

   $$Q\eta_0^Q=\big(Q(\eta_0/\bar\eta)^Q\big)\bar\eta^Q,$$

   the first factor tends to 0 by the ratio test, since

   $$\frac{(Q+1)(\eta_0/\bar\eta)^{Q+1}}{Q(\eta_0/\bar\eta)^Q}=\Big(1+\frac1Q\Big)\frac{\eta_0}{\bar\eta}\longrightarrow\frac{\eta_0}{\bar\eta}<1,$$

   while $\bar\eta^Q\to0$. Thus $Q\eta_0^Q\to0$, and in particular there is a finite $M_\star\ge0$ such that

   $$Q\eta_0^Q\le M_\star\qquad(Q\in\mathbb{N}).$$

3. Define
$$K_0 = \exp(2c_0 M_\star).$$
Then $K_0 > 0$. Fix $Q \in \mathbb{N}$ and put
$$x_Q = c_0 \eta_0^Q.$$
From $0 \le c_0 \le 1/4$ and $0 \le \eta_0 < 1$, we have
$$0 \le x_Q \le \frac{1}{4}, \qquad 1 - x_Q > 0.$$
For every $x \in [0, 1/4]$,
$$(1-x)^{-1} \le 1 + 2x \le e^{2x}.$$
Indeed, the first inequality follows after multiplication by $1 - x > 0$, since
$$(1+2x)(1-x) = 1 + x - 2x^2 \ge 1,$$
and the second is the standard bound $1 + y \le e^y$ with $y = 2x$. Applying this with $x = x_Q$ and raising to the Q-th power gives
$$\big((1-x_Q)^{-1}\big)^Q \le \big(e^{2x_Q}\big)^Q = e^{2Qx_Q}.$$

4. Finally,
$$2Qx_Q = 2c_0\, Q\eta_0^Q \le 2c_0 M_\star,$$
so monotonicity of the exponential yields
$$e^{2Qx_Q} \le e^{2c_0 M_\star} = K_0.$$
Combining the preceding displays,
$$\frac{1}{\big(1 - c_0\eta_0^Q\big)^Q} = \big((1-x_Q)^{-1}\big)^Q \le K_0.$$
Since $Q \in \mathbb{N}$ was arbitrary, this K_0 proves the claim.

□

Lemma 27. [lem:signed-depth-terminal-sign-invariant] (Terminal signed mass invariant) *Let $N, Q, k \in \mathbb{N}$, let $t_0, \zeta, C \in \mathbb{R}$, let $v \in \{-1, +1\}$, and consider the finite signed-depth model of Definition 18. For an observed prefix $o = ((a_i, y_i))_{i=1}^k$, write $\lambda_v(h; o)$ for the unnormalized filter mass of terminal hidden state h, write $J(h) \in \{0, \dots, Q\}$ and $U(h) \in \{-1, +1\}$ for its depth and sign coordinates, and set*
$$\varepsilon_C = \min\left\{\frac{C-1}{2}, \frac{1}{2}\right\}.$$

Assume:

- ***(Positive initial scale.)*** $t_0 > 0$.
- ***(Positive overlap exponent.)*** $\zeta > 0$.

- ***(Nontrivial overlap radius.)*** $C > 1$.

With $L = e^{\zeta}$, define the depth-Q action-history factor

$$a^{(Q)}(o) = L^{-\max\{Q-k,0\}} \prod_{\substack{1\le i\le k \\ i>k-\min\{Q,k\}}} \mathbf{1}\{a_i = 1\}.$$

Then

$$\sum_{h:\, J(h)=Q} U(h)\lambda_v(h;o) = \varepsilon_C\, a^{(Q)}(o) \sum_{h:\, J(h)=Q} \lambda_v(h;o).$$

Proof of Lemma 27. Put $\alpha = e^{-1/t_0}$ and $L = e^{\zeta}$. For a prefix $\omega = ((\tilde{a}_i, \tilde{y}_i))_{i=1}^{m}$ and a depth $j \in \{0, \dots, Q\}$, define

$$\mathrm{Un}_j(\omega) := \sum_{h:\, J(h)=j} \lambda_v(h;\omega), \qquad \mathrm{Sg}_j(\omega) := \sum_{h:\, J(h)=j} U(h)\lambda_v(h;\omega),$$

and

$$A_j(\omega) := L^{-(j-\min\{j,m\})} \prod_{i=m-\min\{j,m\}+1}^{m} \mathbf{1}\{\tilde{a}_i = 1\},$$

with the empty product equal to one. Thus $A_Q(o) = a^{(Q)}(o)$.

1. We first prove the depthwise invariant

   $$\mathrm{Sg}_j(\omega) = \varepsilon_C A_j(\omega)\,\mathrm{Un}_j(\omega) \qquad \text{for every } m,\ \omega,\ \text{and } j \in \{0, \dots, Q\}.$$

   For $m = 0$, the initial mass at depth j and sign $u \in \{-1, +1\}$ is

   $$w_j \frac{1 + u\varepsilon_C L^{-j}}{2}.$$

   Summing over $u = +1$ and $u = -1$ gives

   $$\mathrm{Un}_j(\omega) = w_j, \qquad \mathrm{Sg}_j(\omega) = w_j \varepsilon_C L^{-j} = \varepsilon_C A_j(\omega)\,\mathrm{Un}_j(\omega),$$

   since $A_j(\omega) = L^{-j}$ when the prefix is empty.

2. Suppose the invariant holds for every prefix of length m, and let $\omega^+ = ((\tilde{a}_i, \tilde{y}_i))_{i=1}^{m+1}$. Write ω for its first m observations and $a_+ = \tilde{a}_{m+1}$. At reset depth 0, direct summation of the two reset signs in Definition 18 gives, for each previous hidden state h,

   $$\sum_{u' \in \{-1,+1\}} u'\, p\{J' = 0, U' = u' \mid h, a_+\} = \varepsilon_C \sum_{u' \in \{-1,+1\}} p\{J' = 0, U' = u' \mid h, a_+\}.$$

   After multiplying by the previous unnormalized filter mass, the behavior-policy weight, and the reward-emission weight, and then summing over h, this yields

   $$\mathrm{Sg}_0(\omega^+) = \varepsilon_C\, \mathrm{Un}_0(\omega^+).$$

   Because $A_0(\omega^+) = 1$, the invariant holds at depth 0.

3. For a positive target depth, write it as $j+1$ with $j < Q$. The last observation of ω^+ is $(a_+, \tilde{y}_{m+1})$, and the preceding prefix ω is the restriction of ω^+ to its first m coordinates. Let $\Gamma_{a_+}((\delta,u),(\delta',u'))$ be the hidden-transition weight from (δ,u) to (δ',u') under action a_+, for $\delta,\delta' \in \{0,\dots,Q\}$ and $u,u' \in \{-1,+1\}$. The raw filter recursion gives

$$\mathrm{Un}_{j+1}(\omega^+) = b(a_+)\sum_{\delta=0}^{Q}\sum_{u\in\{-1,+1\}}\lambda_v((\delta,u);\omega)\frac{1+\tilde{y}_{m+1}vc_0u\mathbf{1}\{\delta=Q\}}{2}\sum_{u'\in\{-1,+1\}}\Gamma_{a_+}((\delta,u),(j+1,u')),$$

where $b(a_+)$ is the behavior-policy probability of the observed action and $c_0 = (1-\alpha)/4$. By the transition weights in Definition 18, reaching the positive depth $j+1$ is possible exactly by advancing from depth j, and

$$\sum_{u'\in\{-1,+1\}}\Gamma_{a_+}((\delta,u),(j+1,u')) = \alpha\mathbf{1}\{\delta=j\}.$$

Because $j < Q$, the reward factor on the surviving terms is $1/2$. Hence

$$\mathrm{Un}_{j+1}(\omega^+) = b(a_+)\,\alpha\,\frac{1}{2}\,\mathrm{Un}_j(\omega).$$

4. With the same notation, the signed advancing aggregate is obtained by inserting the new sign in the inner sum:

$$\mathrm{Sg}_{j+1}(\omega^+) = b(a_+)\sum_{\delta=0}^{Q}\sum_{u\in\{-1,+1\}}\lambda_v((\delta,u);\omega)\frac{1+\tilde{y}_{m+1}vc_0u\mathbf{1}\{\delta=Q\}}{2}\sum_{u'\in\{-1,+1\}}u'\,\Gamma_{a_+}((\delta,u),(j+1,u')).$$

Again only $\delta = j$ survives. For that depth, the transition weights in Definition 18 give

$$\sum_{u'\in\{-1,+1\}}u'\,\Gamma_{a_+}((\delta,u),(j+1,u')) = \alpha\mathbf{1}\{\delta=j\}\mathbf{1}\{a_+=1\}u.$$

Indeed, action 1 preserves the old sign on an advancing transition, while action 0 assigns equal total advancing weight to the two possible new signs. Using once more that the reward factor is $1/2$ at depth $j < Q$, we obtain

$$\mathrm{Sg}_{j+1}(\omega^+) = b(a_+)\,\alpha\,\frac{1}{2}\,\mathbf{1}\{a_+=1\}\,\mathrm{Sg}_j(\omega).$$

5. The action-history factor has the corresponding one-step recursion. From the definition of A_j,

$$A_{j+1}(\omega^+) = L^{-((j+1)-\min\{j+1,m+1\})}\prod_{i=m+2-\min\{j+1,m+1\}}^{m+1}\mathbf{1}\{\tilde{a}_i=1\}.$$

The identities

$$\min\{j+1,m+1\} = \min\{j,m\}+1, \qquad (j+1)-\min\{j+1,m+1\} = j-\min\{j,m\},$$

and

$$m+2-\min\{j+1,m+1\} = m+1-\min\{j,m\}$$

show that the displayed product is the product defining $A_j(\omega)$, followed by the last factor $\mathbf{1}\{a_+=1\}$. Thus

$$A_{j+1}(\omega^+) = A_j(\omega)\mathbf{1}\{a_+=1\},$$

with the same empty-product convention as before.

6. Applying the induction hypothesis at depth j to the preceding prefix ω gives

$$\mathrm{Sg}_{j+1}(\omega^+) = b(a_+)\,\alpha\,\frac{1}{2}\,\mathbf{1}\{a_+ = 1\}\,\varepsilon_C A_j(\omega)\,\mathrm{Un}_j(\omega).$$

The unsigned recursion and the action-factor recursion give

$$\varepsilon_C A_{j+1}(\omega^+)\,\mathrm{Un}_{j+1}(\omega^+) = \varepsilon_C A_j(\omega)\mathbf{1}\{a_+ = 1\}\,b(a_+)\,\alpha\,\frac{1}{2}\,\mathrm{Un}_j(\omega),$$

which is the same quantity. Equivalently, splitting on the last action, when $a_+ = 1$ the common factor is $b(a_+)\alpha/2$ times the induction hypothesis, and when $a_+ = 0$ both sides are zero. This completes the induction.

7. Finally take $j = Q$ and $\omega = o$. Reindexing the hidden-state sum by its depth and sign coordinates gives

$$\sum_{h:\,J(h)=Q} \lambda_v(h;o) = \mathrm{Un}_Q(o), \qquad \sum_{h:\,J(h)=Q} U(h)\lambda_v(h;o) = \mathrm{Sg}_Q(o).$$

The depthwise invariant at $j = Q$, together with $A_Q(o) = a^{(Q)}(o)$, therefore gives

$$\sum_{h:\,J(h)=Q} U(h)\lambda_v(h;o) = \varepsilon_C\,a^{(Q)}(o) \sum_{h:\,J(h)=Q} \lambda_v(h;o),$$

as claimed.

□

Lemma 28. [lem:signed-depth-next-reward-numerator] (Next reward numerator) *Let $U, k, N, Q \in \mathbb{N}$, let $v \in \{-1, +1\}$, and consider the finite signed-depth model of Definition 18 with horizon N, terminal depth Q, initial weights p_v, behavior policy b, and reward-symbol/hidden-state kernel K_v. Suppose that*

- ***(Positive time scale.)*** *$t_0 > 0$.*
- ***(Positive behavior perturbation.)*** *$\zeta > 0$.*
- ***(Overlap radius.)*** *$C > 1$.*

Write

$$c_0(t_0) = \frac{1-\alpha}{4},$$

where α is the signed-depth mixing parameter, and write $\phi(0) = -1$, $\phi(1) = 1$. For any observed action/reward-symbol prefix $o \in (\{0,1\} \times \{0,1\})^k$, let

$$\lambda_v(h;o)$$

denote the unnormalized signed-depth filter mass, after the prefix o, assigned to observed state 0 and packed hidden state $h \in \{0, \ldots, 2(Q+1)-1\}$. Decode the packed hidden state by

$$J(h) = \lfloor h/2 \rfloor, \qquad U(h) = \begin{cases} +1, & h \equiv 1 \pmod 2, \\ -1, & h \equiv 0 \pmod 2. \end{cases}$$

Then

$$\sum_{\tau} \mathbf{1}\Big\{((a_i, r_i))_{i=1}^{k}(\tau) = o\Big\}\, \omega_v(\tau)\, \phi(r_{k+1}(\tau)) = v\, c_0(t_0) \sum_{h=0}^{2(Q+1)-1} \mathbf{1}\{J(h) = Q\}\, U(h)\, \lambda_v(h; o),$$

where the outer sum is over all finite signed-depth trajectories τ of horizon $U+k+1$, and $\omega_v(\tau)$ is their chronological path weight under p_v, b, and K_v.

Proof of Lemma 28. 1. Let Σ_k be the set of joint-state prefixes

$$\sigma = (\sigma_0, \dots, \sigma_k), \qquad \sigma_i = (0, h_i), \quad h_i \in \{0, \dots, 2(Q+1)-1\}.$$

For such a prefix, write $W_o(\sigma)$ for the chronological weight of realizing the observed prefix o and ending with hidden state h_k. Summing first over all continuations beyond the observed prefix gives

$$\sum_{\tau} \mathbf{1}\{((a_i, r_i))_{i=1}^{k}(\tau) = o\}\omega_v(\tau)\phi(r_{k+1}(\tau)) = \sum_{\sigma \in \Sigma_k} W_o(\sigma) \sum_{a \in \{0,1\}} b(a) \sum_{y \in \{0,1\}} \phi(y) \sum_{s'} K_v((0, h_k), a)(y, s').$$

Indeed, partition trajectories by their first k observed action–reward symbols and by the joint-state prefix through time k. The chronological product factors into $W_o(\sigma)$, the next behavior probability, and the next kernel probability; summing over the future hidden state s' and then over a, y gives the displayed identity.

2. For every packed hidden state h, the signed-depth reward kernel in Definition 18 satisfies

$$\sum_{a \in \{0,1\}} b(a) \sum_{y \in \{0,1\}} \phi(y) \sum_{s'} K_v((0, h), a)(y, s') = v\, c_0(t_0)\, U(h)\, \mathbf{1}\{J(h) = Q\}.$$

To see this, use the product form of the kernel: the reward-symbol weight depends on h and v, while the hidden-transition weights sum to one for each current state and action. Hence the left-hand side reduces to

$$\Big(\sum_{a \in \{0,1\}} b(a)\Big) \sum_{y \in \{0,1\}} \phi(y)\, \frac{1 + \phi(y)\, v\, c_0(t_0)\, U(h) \mathbf{1}\{J(h) = Q\}}{2}.$$

The behavior probabilities sum to one, and since $\phi(0) = -1$, $\phi(1) = 1$,

$$\sum_{y \in \{0,1\}} \phi(y) \frac{1}{2} = 0, \qquad \sum_{y \in \{0,1\}} \phi(y)^2 \frac{v\, c_0(t_0)\, U(h) \mathbf{1}\{J(h) = Q\}}{2} = v\, c_0(t_0)\, U(h) \mathbf{1}\{J(h) = Q\}.$$

This proves the displayed one-step payoff identity.

3. Substituting the one-step identity into the prefix decomposition gives

$$\sum_{\tau} \mathbf{1}\{((a_i, r_i))_{i=1}^{k}(\tau) = o\}\omega_v(\tau)\phi(r_{k+1}(\tau)) = v\, c_0(t_0) \sum_{\sigma \in \Sigma_k} W_o(\sigma)\, U(h_k) \mathbf{1}\{J(h_k) = Q\}.$$

The observed state alphabet has one element, so each terminal joint state is uniquely of the form $(0, h_k)$; this is the only bookkeeping hidden in the notation above.

4. By definition of the unnormalized filter mass,

$$\lambda_v(h; o) = \sum_{\sigma \in \Sigma_k :\, h_k = h} W_o(\sigma).$$

Therefore

$$\sum_{h=0}^{2(Q+1)-1} \mathbf{1}\{J(h)=Q\}\,U(h)\,\lambda_v(h;o) = \sum_{\sigma\in\Sigma_k} W_o(\sigma)\,U(h_k)\mathbf{1}\{J(h_k)=Q\}.$$

This is just regrouping the preceding prefix sum by its terminal packed hidden state.

5. Combining the last two displays yields

$$\sum_{\tau} \mathbf{1}\{((a_i,r_i))_{i=1}^{k}(\tau)=o\}\,\omega_v(\tau)\,\phi(r_{k+1}(\tau)) = v\,c_0(t_0) \sum_{h=0}^{2(Q+1)-1} \mathbf{1}\{J(h)=Q\}\,U(h)\,\lambda_v(h;o),$$

as claimed. □

Lemma 29. [lem:signed-depth-terminal-prefix-mass] (Terminal prefix mass identity) *Let $U,k,Q\in\mathbb{N}$ with $U\ge 1$, let $t_0,\zeta,C\in\mathbb{R}$, let $v\in\{-1,+1\}$, and consider the finite signed-depth model of Definition 18 at horizon $U+k$. For an action/reward-symbol prefix o of length k, define the joint prefix-and-terminal-depth mass at the cut after the first k epochs by*

$$m_v^Q(o) = \Pr_v\{(A_1,R_1),\dots,(A_k,R_k)=o,\ J_{k+1}=Q\}.$$

For each hidden coordinate $h\in\{0,\dots,2(Q+1)-1\}$, let $J(h)=\lfloor h/2\rfloor$, and let $\lambda_v(h;o)$ be the unnormalized filter mass, as in Definition 25, assigned after observing o to the cut state with hidden coordinate h. Then

$$m_v^Q(o) = \sum_{h=0}^{2(Q+1)-1} \mathbf{1}\{J(h)=Q\}\,\lambda_v(h;o).$$

Proof of Lemma 29. 1. Let Γ_{k+1} denote the set of joint-state prefixes

$$\gamma=(S_1,\dots,S_{k+1}),\qquad S_i\in\{0\}\times\{0,\dots,2(Q+1)-1\}.$$

For such a prefix, write $W_v(\gamma;o)$ for the real chronological weight of the prefix whose observed action/reward word is o. Summing the full trajectory law over all continuations after the cut gives

$$m_v^Q(o) = \sum_{\gamma\in\Gamma_{k+1}} \mathbf{1}\{J(S_{k+1}^{\mathrm{hid}})=Q\}\,W_v(\gamma;o).$$

This is exactly the fixed-prefix expansion of the terminal-depth event at the cut after k observed epochs; the hypothesis $U\ge 1$ ensures that this cut is a valid time in the horizon $U+k$.

2. For a joint state $s\in\{0\}\times\{0,\dots,2(Q+1)-1\}$, define

$$\Lambda_v(s;o) := \sum_{\gamma\in\Gamma_{k+1}} \mathbf{1}\{S_{k+1}=s\}\,W_v(\gamma;o).$$

Then

$$\sum_{s} \mathbf{1}\{J(s^{\mathrm{hid}})=Q\}\Lambda_v(s;o) = \sum_{\gamma\in\Gamma_{k+1}} \mathbf{1}\{J(S_{k+1}^{\mathrm{hid}})=Q\}W_v(\gamma;o).$$

Indeed, interchange the two finite sums. For each fixed prefix γ, the inner sum over s has exactly one nonzero contribution, namely $s=S_{k+1}$, and therefore returns the displayed indicator and weight.

3. Combining the preceding two displays yields

$$m_v^Q(o) = \sum_{s \in \{0\} \times \{0,\dots,2(Q+1)-1\}} \mathbf{1}\{J(s^{\mathrm{hid}}) = Q\} \Lambda_v(s; o).$$

The observed-state coordinate has the singleton value 0, so this sum is

$$m_v^Q(o) = \sum_{h=0}^{2(Q+1)-1} \mathbf{1}\{J(h) = Q\} \Lambda_v((0,h); o).$$

For the signed-depth filter at the cut after the word o, the unnormalized mass assigned to hidden coordinate h is the terminal-state prefix mass at the joint state $(0,h)$:

$$\lambda_v(h; o) := \sum_{\gamma \in \Gamma_{k+1}} \mathbf{1}\{S_{k+1} = (0,h)\}\, W_v(\gamma; o).$$

Thus $\lambda_v(h; o) = \Lambda_v((0,h); o)$ for every h, and the claimed identity follows. □

Lemma 30. [`lem:signed-depth-conditional-reward-mean`] (Conditional reward mean identity) *Let $T, Q \in \mathbb{N}$, let $t_0, \zeta, C \in \mathbb{R}$, and let v be a Boolean alternative, written through*

$$\sigma(v) = \begin{cases} 1, & v = \text{true}, \\ -1, & v = \text{false}. \end{cases}$$

Assume

- ***(Positive baseline.)*** $t_0 > 0$.
- ***(Positive policy scale.)*** $\zeta > 0$.
- ***(Overlap radius.)*** $C > 1$.

For every epoch $t < T$ and every observed prefix $o \in \{\,\text{false}, \text{true}\,\}^t \times \{0,1\}^t$, the conditional next-reward mean satisfies

$$\mathbb{E}_v[Y_t \mid \mathcal{F}_{t-1}^{\mathrm{obs}} = o] = \sigma(v)\, c_0(t_0)\, \varepsilon_C(C)\, a_t^{(Q)}(o)\, q_t^v(o),$$

where the left-hand side is the totalized ratio defining the finite-prefix conditional reward mean, $c_0(t_0) = (1-\alpha)/4$, $\varepsilon_C(C) = \min\{(C-1)/2, 1/2\}$, and $a_t^{(Q)}(o)$ and $q_t^v(o)$ are the rare-action-history factor and terminal posterior of Definition 25.

Proof of Lemma 30. 1. Write the epoch as a finite ordinal $n_{\mathrm{pre}} < T$, and set

$$n_{\mathrm{tail}} := T - (n_{\mathrm{pre}} + 1).$$

Then $T = n_{\mathrm{tail}} + (n_{\mathrm{pre}} + 1)$. Thus it is enough to prove the identity at the canonical cut after an observed prefix of length n_{pre}, with at least the next epoch present.

2. Fix such a canonical horizon and write the observed prefix as $o = ((a_i, r_i))_{i=1}^{k}$, where $a_i \in \{0,1\}$, $r_i \in \{0,1\}$, and the real reward map is

$$\phi(0) = -1, \qquad \phi(1) = 1.$$

Let $\mathcal{T}^{+}_{n_{\mathrm{tail}},k}$ be the finite trajectory space with horizon $n_{\mathrm{tail}}+(k+1)$. Its elements record the initial hidden state, the behavior actions, reward symbols, and subsequent hidden states through the remaining epochs of the finite signed-depth weights displayed below. For every $Q \in \mathbb{N}$, the hidden coordinates are $(J,U) \in \{0,\dots,Q\} \times \{-1,+1\}$. When $Q \geq 1$, these weights are the signed-depth construction of Definition 18; when $Q = 0$, the same formulas are read on the one-point depth set $\{0\}$, so only the terminal-depth reset line can occur. For $\tau \in \mathcal{T}^{+}_{n_{\mathrm{tail}},k}$, define

$$\mathrm{Obs}_{\mathrm{pre}}(\tau) := ((a_1(\tau), r_1(\tau)), \dots, (a_k(\tau), r_k(\tau))), \qquad r_{\mathrm{next}}(\tau) := r_{k+1}(\tau),$$

and let $\omega_v(\tau)$ be the product of the signed-depth initial hidden-state weight, the behavior action probabilities, and the reward-symbol/hidden-transition weights along τ. Thus, if $h_0, \dots, h_{n_{\mathrm{tail}}+k+1}$ are the hidden states in τ,

$$\omega_v(\tau) = \mu_v(h_0) \prod_{i=1}^{n_{\mathrm{tail}}+k+1} \pi_\zeta(a_i(\tau))\, \mathfrak{m}_v(r_i(\tau), h_i \mid h_{i-1}, a_i(\tau)),$$

where μ_v is the initial hidden-state law with weights

$$\mu_v(j,u) = \frac{(1-\alpha)\alpha^j}{1-\alpha^{Q+1}} \frac{1 + u\varepsilon_C(C)L^{-j}}{2}, \qquad 0 \leq j \leq Q, \quad u \in \{-1,+1\},$$

$\pi_\zeta(1) = L^{-1}$, $\pi_\zeta(0) = 1 - L^{-1}$, and $\mathfrak{m}_v(r, h' \mid h, a)$ is the following one-step reward-symbol and next-hidden-state weight. If $h = (j,u)$, $h' = (j',u')$, and $\phi(0) = -1$, $\phi(1) = 1$, then

$$\mathfrak{m}_v(r, h' \mid h, a) = \frac{1 + \phi(r)\sigma(v)c_0(t_0)u\mathbf{1}\{j = Q\}}{2}\, \kappa(h' \mid h, a),$$

where $\kappa(0, u' \mid Q, u, a) = (1 + u'\varepsilon_C(C))/2$; for $j < Q$,

$$\kappa(0, u' \mid j, u, a) = (1-\alpha)\frac{1 + u'\varepsilon_C(C)}{2},$$

and

$$\kappa(j+1, u' \mid j, u, 1) = \alpha\mathbf{1}\{u' = u\}, \qquad \kappa(j+1, u' \mid j, u, 0) = \frac{\alpha}{2},$$

with all other hidden transitions assigned weight zero. These formulas are valid for all $Q \in \mathbb{N}$; if $Q = 0$, only the reset line $j = Q$ occurs. Summing the two reward weights gives one, and the displayed hidden-transition weights sum to one for each hidden state and action. Hence

$$\sum_{r \in \{0,1\}} \sum_{h'} \mathfrak{m}_v(r, h' \mid h, a) = 1 \qquad \text{for every hidden state } h \text{ and action } a.$$

The numerator of the totalized finite-prefix conditional reward mean is

$$N_v(o) := \sum_{\tau \in \mathcal{T}^{+}_{n_{\mathrm{tail}},k}} \mathbf{1}\{\mathrm{Obs}_{\mathrm{pre}}(\tau) = o\}\, \phi(r_{\mathrm{next}}(\tau))\, \omega_v(\tau).$$

For a hidden prefix $\eta = (h_0, \dots, h_k)$, define its observed-prefix weight by

$$\Omega_v(\eta; o) := \mu_v(h_0) \prod_{i=1}^{k} \pi_\zeta(a_i)\, \mathfrak{m}_v(r_i, h_i \mid h_{i-1}, a_i),$$

with value $\mu_v(h_0)$ when $k = 0$. The unnormalized filter mass at the cut is

$$\lambda_v(h; o) := \sum_{\eta:\, h_k = h} \Omega_v(\eta; o), \qquad \Lambda_v(o) := \sum_h \lambda_v(h; o), \qquad \Lambda_{v,Q}(o) := \sum_{h:\, J(h) = Q} \lambda_v(h; o),$$

where $J(h) \in \{0, \dots, Q\}$ and $U(h) \in \{-1, +1\}$ are the depth and sign coordinates of the hidden state h.

We next decompose the finite trajectories at the cut. For a hidden state h and integer $m \geq 0$, set

$$\Xi_m(h) := \sum_{(a_1, r_1, h_1), \dots, (a_m, r_m, h_m)} \prod_{j=1}^{m} \pi_\zeta(a_j)\, \mathfrak{m}_v(r_j, h_j \mid h_{j-1}, a_j), \qquad h_0 = h,$$

with $\Xi_0(h) = 1$. Since $\sum_a \pi_\zeta(a) = 1$ and the displayed normalization of $\mathfrak{m}_v$ holds at each state-action pair, induction gives

$$\Xi_m(h) = 1 \qquad \text{for every } m \geq 0 \text{ and every } h.$$

Splitting $N_v(o)$ according to the hidden prefix η, the next action a, the next reward symbol r, and the next hidden state h', then summing the remaining suffix by $\Xi_{n_{\mathrm{tail}}}(h') = 1$, yields

$$N_v(o) = \sum_\eta \Omega_v(\eta; o) \sum_{a \in \{0,1\}} \pi_\zeta(a) \sum_{r \in \{0,1\}} \phi(r) \sum_{h'} \mathfrak{m}_v(r, h' \mid h_k, a).$$

This is a partition of the same finite set of trajectories: the data (η, a, r, h') give the first $k+1$ epochs, and the suffix sum runs over all chronological continuations from h'.

For fixed h and a, the displayed one-step formula factors $\mathfrak{m}_v(r, h' \mid h, a)$ into the reward weight and a hidden-transition probability. Hence

$$\sum_{h'} \mathfrak{m}_v(r, h' \mid h, a) = \frac{1 + \phi(r)\sigma(v)c_0(t_0)U(h)\mathbf{1}\{J(h) = Q\}}{2}.$$

Using $\sum_a \pi_\zeta(a) = 1$, $\sum_r \phi(r)/2 = 0$, and $\sum_r \phi(r)^2/2 = 1$, the inner one-step payoff is

$$\begin{aligned}\sum_a \pi_\zeta(a) \sum_r \phi(r) \sum_{h'} \mathfrak{m}_v(r, h' \mid h, a) &= \sum_a \pi_\zeta(a) \sum_r \phi(r) \frac{1 + \phi(r)\sigma(v)c_0(t_0)U(h)\mathbf{1}\{J(h) = Q\}}{2} \\ &= \sigma(v)c_0(t_0)U(h)\mathbf{1}\{J(h) = Q\}.\end{aligned}$$

Therefore

$$\begin{aligned}N_v(o) &= \sigma(v)c_0(t_0) \sum_\eta \Omega_v(\eta; o) U(h_k)\mathbf{1}\{J(h_k) = Q\} \\ &= \sigma(v)c_0(t_0) \sum_{h:\, J(h) = Q} U(h)\lambda_v(h; o).\end{aligned}$$

This is the numerator identity supplied by Lemma 28; the displayed derivation records the finite-trajectory regrouping used at the canonical cut. The denominator and the terminal-depth quotient are identified below.

3. Lemma 27 applies under $t_0 > 0$, $\zeta > 0$, and $C > 1$, and gives the prefix-level identity

$$\sum_{h:\, J(h)=Q} U(h)\lambda_v(h;o) = \varepsilon_C(C)\, a_{\mathrm{pre}}^{(Q)}(o) \sum_{h:\, J(h)=Q} \lambda_v(h;o),$$

where, for the length-k prefix $o = ((a_i, r_i))_{i=1}^{k}$,

$$a_{\mathrm{pre}}^{(Q)}(o) = L^{-\max\{Q-k,0\}} \prod_{\substack{1\le i\le k \\ i>k-\min\{Q,k\}}} \mathbf{1}\{a_i = 1\}.$$

At the canonical cut, the time-indexed factor of Definition 25 is the same quantity:

$$a_t^{(Q)}(o) = L^{-(Q-\min\{Q,k\})} \prod_{\substack{1\le i\le k \\ i>k-\min\{Q,k\}}} \mathbf{1}\{a_i = 1\} = a_{\mathrm{pre}}^{(Q)}(o),$$

since $Q - \min\{Q,k\} = \max\{Q-k, 0\}$. Thus

$$\sum_{h:\, J(h)=Q} U(h)\lambda_v(h;o) = \varepsilon_C(C)\, a_t^{(Q)}(o)\, \Lambda_{v,Q}(o).$$

The empty-prefix case is included by the empty-product convention in Definition 25.

4. It remains to identify the denominator and terminal-depth numerator. For a trajectory $\tau \in \mathcal{T}^{+}_{n_{\mathrm{tail}},k}$, let $h_{\mathrm{cut}}(\tau) := h_k(\tau)$ be the hidden state after the first k observed epochs. In the one-based indexing used for the signed-depth path statement, this same cut state is H_{k+1}; hence the event $J(h_{\mathrm{cut}}(\tau)) = Q$ is exactly the event written there as $J_{k+1} = Q$. Define the finite prefix mass and the terminal-depth prefix mass by

$$\Pi_v(o) := \sum_{\tau \in \mathcal{T}^{+}_{n_{\mathrm{tail}},k}} \mathbf{1}\{\mathrm{Obs}_{\mathrm{pre}}(\tau) = o\}\, \omega_v(\tau),$$

and

$$\Pi_{v,Q}(o) := \sum_{\tau \in \mathcal{T}^{+}_{n_{\mathrm{tail}},k}} \mathbf{1}\{\mathrm{Obs}_{\mathrm{pre}}(\tau) = o\}\, \mathbf{1}\{J(h_{\mathrm{cut}}(\tau)) = Q\}\, \omega_v(\tau).$$

The same cut decomposition as above, now without weighting by the next reward, gives

$$\begin{aligned}\Pi_v(o) &= \sum_{\eta} \Omega_v(\eta;o) \sum_{(a,r,h')} \pi_\zeta(a)\, \mathfrak{m}_v(r, h' \mid h_k, a)\, \Xi_{n_{\mathrm{tail}}}(h') \\ &= \sum_{\eta} \Omega_v(\eta;o) = \sum_{h} \lambda_v(h;o) = \Lambda_v(o).\end{aligned}$$

Indeed, the first sum partitions trajectories by their length-k hidden prefix, next action, next reward symbol, and next hidden state; the continuation has total mass $\Xi_{n_{\mathrm{tail}}}(h') = 1$, and the next one-step mass sums to one. Carrying the terminal-depth indicator at the cut through the same partition gives

$$\begin{aligned}\Pi_{v,Q}(o) &= \sum_{\eta} \Omega_v(\eta;o)\mathbf{1}\{J(h_k) = Q\} \sum_{(a,r,h')} \pi_\zeta(a)\, \mathfrak{m}_v(r, h' \mid h_k, a)\, \Xi_{n_{\mathrm{tail}}}(h') \\ &= \sum_{\eta} \Omega_v(\eta;o)\mathbf{1}\{J(h_k) = Q\} \\ &= \sum_{h:\, J(h)=Q} \lambda_v(h;o) = \Lambda_{v,Q}(o).\end{aligned}$$

Thus $\Pi_v(o)$ is the mass of the observed-prefix event and $\Pi_{v,Q}(o)$ is the mass of the same event together with terminal depth at the cut. The latter identity is exactly the terminal-prefix mass identification in Lemma 29 for the canonical horizon.

The finite-prefix conditional reward mean is the totalized quotient

$$\mathbb{E}_v[Y_t \mid \mathcal{F}_{t-1}^{\mathrm{obs}} = o] = \frac{N_v(o)}{\Pi_v(o)},$$

where real division is totalized by $x/0 = 0$. The terminal posterior in Definition 25 is read at this finite prefix through the same totalized convention:

$$q_t^v(o) = \frac{\Pi_{v,Q}(o)}{\Pi_v(o)}.$$

If $\Pi_v(o) = 0$, then $\Pi_{v,Q}(o) = 0$ because the terminal-depth prefix event is contained in the prefix event; for $\Pi_v(o) > 0$, these quotients are the usual conditional averages. Combining the numerator identity, Lemma 27, and $\Pi_{v,Q}(o) = \Lambda_{v,Q}(o)$, we obtain

$$\begin{aligned}\mathbb{E}_v[Y_t \mid \mathcal{F}_{t-1}^{\mathrm{obs}} = o] &= \frac{\sigma(v)c_0(t_0)\varepsilon_C(C)\, a_t^{(Q)}(o)\, \Pi_{v,Q}(o)}{\Pi_v(o)} \\ &= \sigma(v)c_0(t_0)\varepsilon_C(C)\, a_t^{(Q)}(o)\, q_t^v(o).\end{aligned}$$

This is the claimed identity for the original T and $t < T$.

□

Lemma 26 turns the finite-prefix normalization in Lemma 23 into a constant depending on the mixing scale. Lemmas 27 to 29 connect the signed terminal mass, the next reward numerator, and the prefix mass at terminal depth. These identities convert the latent terminal sign into the observable conditional reward mean in Lemma 30, while Lemma 25 converts the resulting terminal-posterior control into a finite observed-word divergence bound. The product $\alpha^{2Q}L^{-Q}$ reflects the joint requirement that terminal depth survives and that the visible action history carries the rare sequence needed to reveal it.

The likelihood calculation next fixes the observed signed-depth law and combines the exact observed reward mean with a bound on the divergence between the two signs.

Definition 28. [synth_18] (Observed signed-depth law) *For $T, Q \in \mathbb{N}$, $t_0, \zeta, C \in \mathbb{R}$, and two-point alternative index v, define*

$$\mathbb{P}_{Q,v}^{\mathrm{obs},T}$$

to be the observed-path law obtained by applying the observed-coordinate projection to the trajectory law of the embedded finite signed-depth model with horizon T, parameters (t_0, ζ, C, Q), and alternative index v.

The law in Definition 28 is the distribution used in the two-point information comparison. It records the observable image of the embedded signed-depth experiment, so the divergence calculation is stated on the same data observed by an estimator.

Lemma 31. [lem:observed-path-kl] (Observed-path KL bound) *Fix $t_0 > 0$, and write $\alpha = \exp(-1/t_0)$. There exists a constant $K_0 > 0$, depending on t_0, such that whenever the following conditions hold:*

- ***(Overlap exponent.)*** *$\zeta > 0$, and $L = \exp(\zeta)$.*
- ***(Reset amplitude.)*** *$C > 1$, and $\varepsilon_C = \min\{(C-1)/2, 1/2\}$.*
- ***(Sequential ignorability.)*** *For every horizon T, terminal depth $Q \geq 1$, and alternative $v \in \{-1, +1\}$, the signed-depth experiment in Definition 18 satisfies Assumption 2.*
- ***(Stationary start.)*** *For every horizon T, terminal depth $Q \geq 1$, and alternative $v \in \{-1, +1\}$, the signed-depth experiment in Definition 18 satisfies Assumption 3.*

there exists a constant $B_0 > 0$, depending on t_0, ζ, C, such that for every T and every $Q \geq 1$, with $\mathbb{P}^{\mathrm{obs},T}_{Q,v}$ denoting the observed-path law of the signed-depth alternative v,

$$\mathrm{KL}\Big(\mathbb{P}^{\mathrm{obs},T}_{Q,+1} \,\Big\|\, \mathbb{P}^{\mathrm{obs},T}_{Q,-1}\Big) \leq B_0\, T\, \alpha^{2Q} L^{-Q}.$$

Moreover, for each $v \in \{-1, +1\}$, each $t < T$, and each observed prefix $o \in \mathcal{F}^{\mathrm{obs}}_{t-1}$, the conditional reward mean satisfies the exact finite-prefix identity

$$\mathbb{E}_v[Y_t \mid \mathcal{F}^{\mathrm{obs}}_{t-1} = o] = v\, c_0\, \varepsilon_C\, a^{(Q)}_t(o)\, q^v_t(o),$$

where $c_0 = (1-\alpha)/4$, and $a^{(Q)}_t$ and q^v_t are the rare-action-history factor and terminal-depth posterior from Definition 25. The posterior satisfies

$$q^v_t(o) \leq K_0\, \alpha^Q.$$

The same observed-path laws also satisfy

$$\mathrm{KL}\Big(\mathbb{P}^{\mathrm{obs},T}_{Q,+1} \,\Big\|\, \mathbb{P}^{\mathrm{obs},T}_{Q,-1}\Big) \leq \frac{4c_0^2 K_0^2}{1-c_0^2}\, T\, \varepsilon_C^2\, \alpha^{2Q} L^{-Q}.$$

Proof of Lemma 31. Fix $t_0 > 0$, put $\alpha = \exp(-1/t_0)$, and set $c_0 = (1-\alpha)/4$.

1. By Lemma 26, there is a constant $K_0 > 0$, depending only on t_0, such that for every $Q \in \mathbb{N}$,

$$\frac{1}{\{1 - c_0(\alpha/(1-c_0))^Q\}^Q} \leq K_0.$$

Fix $\zeta > 0$, $C > 1$, and the two stated model-class hypotheses. Let

$$B_0 := \frac{4c_0^2 K_0^2}{1-c_0^2}.$$

Since $0 < \alpha < 1$, we have $0 < c_0 < 1$, hence $1 - c_0^2 > 0$, and therefore $B_0 > 0$.

2. Fix $T \in \mathbb{N}$ and $Q \geq 1$. The refined posterior estimate in Lemma 23, combined with the choice of K_0 in step 1, gives the uniform terminal-depth posterior bound

$$q^v_t(o) \leq K_0 \alpha^Q$$

for every auxiliary horizon N, every $v \in \{-1, +1\}$, every $t < N$, and every observed prefix $o \in \mathcal{F}^{\mathrm{obs}}_{t-1}$. In particular, for the present horizon T,

$$q^v_t(o) \leq K_0 \alpha^Q, \qquad v \in \{-1, +1\},\ t < T,\ o \in \mathcal{F}^{\mathrm{obs}}_{t-1}.$$

3. Applying Lemma 25 with the posterior bound from step 2 yields the finite observed-word inequality

$$\mathrm{KL}\Big(\mathbb{P}^{\mathrm{fin},T}_{Q,+1}\,\Big\|\,\mathbb{P}^{\mathrm{fin},T}_{Q,-1}\Big) \le \frac{4c_0^2K_0^2}{1-c_0^2}\,T\,\varepsilon_C^2\,\alpha^{2Q}L^{-Q}.$$

The observed-path law is obtained from the finite observed-word law by the deterministic reward decoding map in Definition 18. The data-processing inequality for this embedding gives

$$\mathrm{KL}\Big(\mathbb{P}^{\mathrm{obs},T}_{Q,+1}\,\Big\|\,\mathbb{P}^{\mathrm{obs},T}_{Q,-1}\Big) \le \frac{4c_0^2K_0^2}{1-c_0^2}\,T\,\varepsilon_C^2\,\alpha^{2Q}L^{-Q}.$$

Equivalently, by the definition of B_0,

$$\mathrm{KL}\Big(\mathbb{P}^{\mathrm{obs},T}_{Q,+1}\,\Big\|\,\mathbb{P}^{\mathrm{obs},T}_{Q,-1}\Big) \le B_0\,T\,\varepsilon_C^2\,\alpha^{2Q}L^{-Q}.$$

4. Since $C > 1$, the reset amplitude satisfies

$$0 \le \varepsilon_C = \min\{(C-1)/2, 1/2\} \le 1.$$

Also $L = \exp(\zeta) > 0$, so $L^{-Q} \ge 0$, and $\alpha^{2Q} \ge 0$. Hence

$$B_0\,T\,\varepsilon_C^2\,\alpha^{2Q}L^{-Q} \le B_0\,T\,\alpha^{2Q}L^{-Q}.$$

Monotonicity of $x \mapsto \mathrm{ofReal}(x)$ on this real inequality gives

$$\mathrm{KL}\Big(\mathbb{P}^{\mathrm{obs},T}_{Q,+1}\,\Big\|\,\mathbb{P}^{\mathrm{obs},T}_{Q,-1}\Big) \le B_0\,T\,\alpha^{2Q}L^{-Q}.$$

5. Finally, Lemma 30 supplies, for each $v \in \{-1,+1\}$, each $t < T$, and each observed prefix $o \in \mathcal{F}^{\mathrm{obs}}_{t-1}$, the exact identity

$$\mathbb{E}_v[Y_t \mid \mathcal{F}^{\mathrm{obs}}_{t-1} = o] = v\,c_0\,\varepsilon_C\,a_t^{(Q)}(o)\,q_t^v(o).$$

Together with the posterior bound in step 2 and the sharper KL estimate from step 3, this gives all three asserted conclusions:

$$q_t^v(o) \le K_0\alpha^Q$$

and

$$\mathrm{KL}\Big(\mathbb{P}^{\mathrm{obs},T}_{Q,+1}\,\Big\|\,\mathbb{P}^{\mathrm{obs},T}_{Q,-1}\Big) \le \frac{4c_0^2K_0^2}{1-c_0^2}\,T\,\varepsilon_C^2\,\alpha^{2Q}L^{-Q}.$$

□

The identity in Lemma 31 gives the exact observed conditional mean in terms of a rare visible action run and a terminal-depth posterior. The KL bound then combines the survival factor α^{2Q} with the policy-overlap factor L^{-Q}, matching the fixed-overlap exponent used in the minimax comparison. This is the point at which latent stationarity and observed histories enter the lower-bound calculation together.

The upper-bound argument also uses the stationary mean of each PHIW score. With the observed law already fixed in Definition 26, the next statement bridges the observable score to the target transition operator that controls the truncation term.

Lemma 32. [lem:phiw-score-stationary-mean] (Stationary PHIW score mean) *Let $T, n_X, n_H, k \in \mathbb{N}$, let $t_0, \zeta, C \in \mathbb{R}$, and let M be a raw POMDP experiment. Suppose:*

- ***(Model class.)*** *$M \in \mathcal{M}_T(t_0, \zeta, C)$, as in Definition 10.*
- ***(Mature time.)*** *$t \in \{0, \dots, T-1\}$ and $k \le t$.*

Then the depth-k score $Z_t(k)$ of Definition 16, evaluated under the observed trajectory law of M, satisfies

$$\int Z_t(k)\, d\mathbb{P}_M^{\mathrm{obs}} = \sum_s d_b(s)\, (P_e^k g_e)(s).$$

Here d_b is the stationary law selected for the behavior-policy transition matrix

$$P_b(s, s') = \sum_{a \in \{0,1\}} b(a \mid x_s)\, K(s, a)\{(r, s'') : \ s'' = s'\},$$

and $P_e^k g_e$ is the k-fold iterate of the target-policy transition operator applied to the target-policy reward regression

$$g_e(s) = \sum_{a \in \{0,1\}} e(a \mid x_s) \int r\, K(s, a)(dr, ds').$$

Proof of Lemma 32. Write S_j for the joint state at zero-based epoch j, on the same time scale as $Z_t(k)$. For a policy $p \in \{b, e\}$ and a function φ on joint states, set

$$(P_p\varphi)(s) := \sum_{s'} P_p(s, s')\varphi(s'), \qquad P_p(s, s') = \sum_{a \in \{0,1\}} p(a \mid x_s)\, K(s, a)\{(y, s'') : \ s'' = s'\}.$$

Also write the observable ratio, with its zero-cell convention, as

$$\rho_j = \rho(X_j, A_j) := \begin{cases} e(A_j \mid X_j)/b(A_j \mid X_j), & b(A_j \mid X_j) > 0, \\ 0, & b(A_j \mid X_j) = 0. \end{cases}$$

By Definition 10, the model supplies Assumptions 1 to 5, with $L = \exp(\zeta)$. Since $0 < \zeta$, $1 \le L$.

Let

$$\iota := t - k.$$

The hypothesis $k \le t$ makes ι a valid epoch, and $\iota + k = t < T$. We first compute the mean of the score from the state at the beginning of its ratio window:

$$\int Z_t(k)\, d\mathbb{P}_M^{\mathrm{obs}} = \int (P_e^k g_e)(S_\iota)\, d\mathbb{P}_M.$$

Indeed, the score is a function of the observed trajectory, so its observed-law integral is the corresponding full-trajectory integral,

$$\int Z_t(k)\, d\mathbb{P}_M^{\mathrm{obs}} = \int Y_t \prod_{j=\iota}^{t} \rho_j\, d\mathbb{P}_M.$$

For any bounded φ and any epoch $j < T - 1$, Assumptions 1, 2 and 5 give

$$\mathbb{E}_M[\rho_j\, \varphi(S_{j+1}) \mid S_j = s] = \sum_{a \in \{0,1\}} e(a \mid x_s) \int \varphi(s')\, K(s, a)(dy, ds') = (P_e\varphi)(s).$$

The same conditional calculation at the terminal reward gives

$$\mathbb{E}_M[\rho_t\, Y_t \mid S_t = s] = \sum_{a\in\{0,1\}} e(a \mid x_s) \int y\, K(s,a)(dy, ds') = g_e(s).$$

The zero-cell convention is compatible with the displayed equalities because $b(a \mid x_s) = 0$ and $e(a \mid x_s) \le L b(a \mid x_s)$ imply $e(a \mid x_s) = 0$. Iterating the preceding identities backward from t to ι, with boundedness supplied by Assumption 4 and finite state spaces from Definition 10, yields the displayed score identity.

It remains to identify the law of S_ι. For every function φ on joint states,

$$\int \varphi(S_\iota)\, d\mathbb{P}_M = \sum_s d_b(s)\varphi(s).$$

At the initial epoch this is Assumption 3. If the identity holds at epoch j, then Assumptions 1 and 2 give

$$\int \varphi(S_{j+1})\, d\mathbb{P}_M = \int (P_b\varphi)(S_j)\, d\mathbb{P}_M = \sum_s d_b(s) \sum_{s'} P_b(s,s')\varphi(s').$$

The selected behavior stationary law is stationary for P_b, so for every s',

$$\sum_s d_b(s) P_b(s,s') = d_b(s'),$$

and therefore

$$\int \varphi(S_{j+1})\, d\mathbb{P}_M = \sum_{s'} d_b(s')\varphi(s').$$

Induction gives the claim at ι.

Applying the last identity with $\varphi = P_e^k g_e$ and substituting into the score identity gives

$$\int Z_t(k)\, d\mathbb{P}_M^{\mathrm{obs}} = \sum_s d_b(s)\, (P_e^k g_e)(s),$$

which is the asserted equality. □

Lemma 32 states the exact stationary target of each mature score. Under stationarity, a length-k observable reweighting transports the behavior stationary distribution through k target-policy transitions, so the remaining discrepancy is the contraction-controlled distance between this finite-lag quantity and the stationary target value.

The bounded-reward normalization also pins down the range of the target value. This range fact is used when risks are compared over estimators whose outputs are clipped to the same unit interval.

Lemma 33. [lem:target-value-unit-interval] (Target value is bounded) *Let $T \ge 1$, let $t_0, \zeta, C \in \mathbb{R}$, and let $M = (\mathcal{X}, \mathcal{H}, K, b, e, \mathcal{L}(\mathsf{O}_T)) \in \mathcal{M}_T(t_0, \zeta, C)$ be a model in Definition 10. Assume the model satisfies the stationary-start, sequential-ignorability, POMDP-kernel, bounded-reward, latent-stationary-overlap, and policy-overlap conditions of Assumptions 1 to 5 and 7. Then the stationary target-policy mean reward $\theta(K, e)$ of Definition 12 lies in the unit interval:*

$$\theta(K, e) \in [-1, 1].$$

Proof of Lemma 33. Let P_b and P_e be the joint-state transition kernels induced by the behavior and target policies, and let d_b and d_e be their selected stationary laws. For a joint state $s = (x, h)$ and action $a \in \mathcal{A}$, define the one-step reward mean

$$\bar{r}(s, a) := \int r \, K(dr, dx', dh' \mid s, a).$$

We also use the local notation

$$g_e(s) := \sum_{a \in \mathcal{A}} e(a \mid x) \bar{r}(s, a),$$

which is the target-policy reward regression appearing inside $\theta(K, e)$. With this notation, Definition 12 gives

$$\theta(K, e) = \sum_s d_e(s) g_e(s).$$

1. Since $T \geq 1$, consider epoch 0. Let η_0 denote the unique empty pre-history value at epoch 0, and define the finite state-history and state-action-history spaces

$$\mathsf{H}_0^{\mathrm{st}} := \{(\eta_0, s) : s \in \mathcal{S}\}, \qquad \mathsf{H}_0^{\mathrm{act}} := \mathsf{H}_0^{\mathrm{st}} \times \mathcal{A}.$$

For $\upsilon = (\eta_0, s) \in \mathsf{H}_0^{\mathrm{st}}$, write $\mathrm{cur}(\upsilon) = s$. Let

$$\Gamma_0^{\mathrm{st}} := (\eta_0, S_0), \qquad \Gamma_0^{\mathrm{act}} := (\Gamma_0^{\mathrm{st}}, A_0), \qquad \mu_0 := \mathcal{L}(\Gamma_0^{\mathrm{act}}).$$

For every $\gamma = (\upsilon, a) \in \mathsf{H}_0^{\mathrm{act}}$ and every measurable $B \subseteq \mathbb{R} \times \mathcal{S}$, define the row selected by the whole state-action history view as

$$\mathsf{row}_0(\gamma; B) := K(B \mid \mathrm{cur}(\upsilon), a).$$

By the probability-kernel part of Assumption 1, $\mathsf{row}_0(\gamma; \cdot)$ is a probability measure for every γ. Define the total function

$$J_0(\upsilon, a) := \int r \, \mathsf{row}_0((\upsilon, a); dr, dx', dh') = \int r \, K(dr, dx', dh' \mid \mathrm{cur}(\upsilon), a),$$

for all $(\upsilon, a) \in \mathsf{H}_0^{\mathrm{act}}$; bounds for J_0 will be asserted only on the full-measure set obtained below. Thus, when $\upsilon = (\eta_0, s)$, one has $J_0(\upsilon, a) = \bar{r}(s, a)$.

By Assumption 4, $|Y_0| \leq 1$ holds almost surely under the trajectory law, so the event

$$\mathcal{R}_0 := \{(\gamma, (r, s')) \in \mathsf{H}_0^{\mathrm{act}} \times (\mathbb{R} \times \mathcal{S}) : |r| \leq 1\}$$

has full mass under the law of $(\Gamma_0^{\mathrm{act}}, (Y_0, S_1))$. After both sides are pushed forward to the epoch-0 conditioning view, Assumption 1 gives the composition-product identity

$$\mathcal{L}\{(\Gamma_0^{\mathrm{act}}, (Y_0, S_1)) \in d\gamma \, d(r, s')\} = \mu_0(d\gamma) \, \mathsf{row}_0(\gamma; dr, ds').$$

The section argument is now the standard disintegration fact for a composition product: if a set has full mass under the displayed law, then for μ_0-almost every γ its section has full mass under the row $\mathsf{row}_0(\gamma; \cdot)$. Applying this to $\mathcal{R}_0$ gives

$$\mathsf{row}_0\big(\gamma; \{(r, x', h') : \ |r| \leq 1\}\big) = 1 \quad \text{for } \mu_0\text{-almost every } \gamma.$$

For every such $\gamma = (\upsilon, a)$, the row is a probability measure and the coordinate function $(r, x', h') \mapsto r$ is bounded by 1 in absolute value almost surely under that row. The usual norm bound for the integral of an almost-surely bounded real function under a probability measure therefore gives

$$\begin{aligned} |J_0(\upsilon, a)| &= \left| \int r \, \mathsf{row}_0((\upsilon, a); dr, dx', dh') \right| \\ &\le \int |r| \, \mathsf{row}_0((\upsilon, a); dr, dx', dh') \le 1. \end{aligned}$$

Equivalently, using the definition of row_0,

$$|J_0(\upsilon, a)| = \left| \int r \, K(dr, dx', dh' \mid \mathrm{cur}(\upsilon), a) \right| \le 1 \quad \text{for } \mu_0\text{-almost every } (\upsilon, a).$$

2. Fix $s = (x, h)$ and $a \in \mathcal{A}$ with $d_b(s) > 0$ and $b(a \mid x) > 0$. Put

$$\upsilon_s := (\eta_0, s) \in \mathsf{H}_0^{\mathrm{st}}, \qquad \gamma_{s,a} := (\upsilon_s, a) \in \mathsf{H}_0^{\mathrm{act}}.$$

The singleton mass of $\gamma_{s,a}$ is obtained from three separate ingredients, none of which gives it alone. The first is Assumption 2, which says only that the law μ_0 of Γ_0^{act} factors as the law of Γ_0^{st} followed by the behavior kernel $b(\cdot \mid X_0)$, so that

$$\mu_0\{\gamma_{s,a}\} = \Pr\{\Gamma_0^{\mathrm{st}} = \upsilon_s\} \, b(a \mid x).$$

The second ingredient is the epoch-0 history projection: the pre-history coordinate at epoch 0 is forced to be η_0, so the event $\{\Gamma_0^{\mathrm{st}} = \upsilon_s\}$ is exactly $\{S_0 = s\}$. The third is Assumption 3, which identifies the law of S_0 and gives

$$\Pr\{\Gamma_0^{\mathrm{st}} = \upsilon_s\} = \Pr(S_0 = s) = d_b(s).$$

Combining the last two displays with the two strict-positivity hypotheses yields

$$\mu_0\{\gamma_{s,a}\} = d_b(s) b(a \mid x) > 0.$$

If a full-μ_0-measure set omitted this singleton, its complement would have positive μ_0-mass. Hence the almost-sure bound from the previous step holds at $\gamma_{s,a}$, and therefore

$$|\bar{r}(s, a)| \le 1 \qquad \text{whenever } d_b(s) > 0 \text{ and } b(a \mid x) > 0.$$

3. The selected behavior and target stationary laws are probability vectors: by Definition 6 each d_p lies in $\Delta(\mathcal{S})$, the simplex of nonnegative probability vectors on the joint-state alphabet, fixed by the corresponding policy-induced kernel. Unpacking this simplex membership gives

$$d_b(s) \ge 0, \qquad d_e(s) \ge 0 \quad \text{for all } s, \qquad \sum_s d_e(s) = 1.$$

Now suppose $d_e(s) > 0$ and $e(a \mid x) > 0$. By Assumption 7,

$$d_e(s) \le C \, d_b(s).$$

If $d_b(s)$ were not positive, its nonnegativity would force $d_b(s) = 0$, and substituting that into the displayed overlap inequality would give $0 < d_e(s) \le C \cdot 0 = 0$, a contradiction. This uses the displayed overlap inequality, the nonnegativity $d_b(s) \ge 0$, and the local hypothesis $d_e(s) > 0$; the radius condition $C \ge 1$ plays no further role in this step. Thus $d_b(s) > 0$.

The action-coordinate argument uses the local hypothesis $e(a \mid x) > 0$, the nonnegativity of the behavior carrier, and the policy-overlap inequality. The inequality comes from Assumption 5, which for the policy-overlap constant L gives

$$e(a \mid x) \le L\, b(a \mid x);$$

the nonnegativity $b(a \mid x) \ge 0$ comes from Assumption 2, which is where b is declared a probability vector. If $b(a \mid x)$ were not positive, that nonnegativity would force $b(a \mid x) = 0$, whence $0 < e(a \mid x) \le L \cdot 0 = 0$, a contradiction. Thus $b(a \mid x) > 0$. The row bound from the previous step applies, and hence

$$|\bar{r}(s,a)| \le 1 \qquad \text{whenever } d_e(s) > 0 \text{ and } e(a \mid x) > 0.$$

4. For any $s = (x,h)$ with $d_e(s) > 0$, the target reward regression satisfies

$$\begin{aligned} |g_e(s)| &= \left| \sum_{a \in \mathcal{A}} e(a \mid x) \bar{r}(s,a) \right| \\ &\le \sum_{a \in \mathcal{A}} |e(a \mid x) \bar{r}(s,a)| \\ &= \sum_{a \in \mathcal{A}} e(a \mid x) |\bar{r}(s,a)| \\ &\le \sum_{a \in \mathcal{A}} e(a \mid x) = 1. \end{aligned}$$

The equality in the third line uses $e(a \mid x) \ge 0$; the last inequality uses the preceding step when $e(a \mid x) > 0$, while the corresponding summand is zero when $e(a \mid x) = 0$; and the final equality is the target-policy probability-vector property in Assumption 5. Two distinct facts meet in that last equality, and they come from different places. Which set the sum runs over is a carrier fact: Definition 3 fixes the action coordinate as the Boolean two-point alphabet identified with $\mathcal{A} = \{0,1\}$, and it imposes nothing else — in particular no nonnegativity and no normalization. That the two values at x are nonnegative and add to one is the separate probability-vector restriction on e, which is imposed in Assumption 5. So the normalization is the sum of exactly two action probabilities at x, and no larger action space is involved.

5. Finally,

$$\begin{aligned}|\theta(K,e)| &= \left|\sum_s d_e(s) g_e(s)\right| \\ &\le \sum_s |d_e(s) g_e(s)| \\ &= \sum_s d_e(s) |g_e(s)| \\ &\le \sum_s d_e(s) = 1.\end{aligned}$$

Here $d_e(s) \ge 0$ gives the third line; in the fourth line the summand is zero when $d_e(s) = 0$, and otherwise the regression bound from the previous step applies. Thus

$$-1 \le \theta(K,e) \le 1,$$

which is exactly membership of $\theta(K,e)$ in the closed real interval $[-1,1]$.

□

Lemma 33 records the bounded target range implied by the model-class reward normalization and the target stationary distribution. This range is the common scale for the clipped PHIW estimator and for the bounded observable estimators used in the minimax comparison.

The radius-sensitive upper calculation refines the estimator bound by keeping the latent-overlap distance q_C explicit.

Lemma 34. [`lem:radius-sensitive-phiw`] (Radius-sensitive PHIW risk) *For every trajectory length T, observable alphabet size n_X, latent alphabet size n_H, and history depth k, let*

$$\alpha = \exp\{-(1/t_0)\}, \qquad L = \exp\{\zeta\}, \qquad q_C = \frac{C-1}{C}.$$

Suppose that

- ***(Mixing scale.)*** $0 < t_0$.
- ***(Overlap scale.)*** $0 < \zeta$.
- ***(Radius parameter.)*** $1 \le C$, *with q_C as in Definition 19.*
- ***(Trajectory length.)*** $1 \le T$.
- ***(Model class.)*** *The experiment M belongs to $\mathcal{M}_T(t_0, \zeta, C)$ in the sense of Definition 10.*
- ***(History depth.)*** $k \le T/2$.

Then the clipped partial-history importance-weighted estimator $\widehat{\theta}_k$ from Definition 16, evaluated under the observed trajectory law induced by M, satisfies

$$\mathbb{E}_M\left[\{\widehat{\theta}_k - \theta(K,e)\}^2\right] \le 4q_C^2\alpha^{2k} + \frac{2}{T}\left\{L^{k+1}\left(1 + \frac{4}{L-1}\right) + \frac{4}{1-\alpha}\right\},$$

where $\theta(K,e)$ is the stationary target-policy mean reward from Definition 12.

Proof of Lemma 34. 1. Write $\mathbb{P}_M^{\mathrm{obs}}$ for the observed trajectory law induced by M, and let

$$q_C := \frac{C-1}{C}$$

be the overlap-distance radius from Definition 19. Define the unclipped partial-history average

$$\widetilde{\theta}_k := \frac{1}{T-k}\sum_{t=k}^{T-1} Z_t(k), \qquad \widehat{\theta}_k = \mathrm{clip}_{[-1,1]}(\widetilde{\theta}_k),$$

with $Z_t(k)$ as in Definition 16. Since $1 \le T$ and $k \le \lfloor T/2 \rfloor$, one has $k < T$, so $T - k \ge 1$. The conditions defining $M \in \mathcal{M}_T(t_0, \zeta, C)$ in Definition 10 give $L = e^{\zeta} > 1$ and $0 < \alpha = e^{-1/t_0} < 1$. The finite observation space makes each score $Z_t(k)$ measurable. Moreover $L = e^{\zeta} \ge 1$, and the sequential-ignorability, policy-overlap, and bounded-reward restrictions in Definition 10 imply the window envelope

$$|Z_t(k)| \le L^{k+1} \qquad \mathbb{P}_M^{\mathrm{obs}}\text{-a.s.},$$

for every t. Thus each score is square-integrable under $\mathbb{P}_M^{\mathrm{obs}}$, and so is $\widetilde{\theta}_k$. Moreover, Lemma 33 applies to the present horizon and model class membership, giving

$$\theta(K, e) \in [-1, 1].$$

2. For every $u \in \mathbb{R}$ and every $\vartheta \in [-1, 1]$,

$$\big(\mathrm{clip}_{[-1,1]}(u) - \vartheta\big)^2 \le (u - \vartheta)^2.$$

Applying this pointwise with $u = \widetilde{\theta}_k$ and $\vartheta = \theta(K, e)$, and then using the usual bias–variance identity for the square-integrable random variable $\widetilde{\theta}_k$, yields

$$\mathbb{E}_M\Big[\{\widehat{\theta}_k - \theta(K, e)\}^2\Big] \le \mathrm{Var}_M(\widetilde{\theta}_k) + \Big(\mathbb{E}_M[\widetilde{\theta}_k] - \theta(K, e)\Big)^2.$$

3. We next bound the bias term. Let d_b and d_e be the stationary laws of the behavior- and target-policy transition kernels, and let g_e be the target-policy reward regression appearing in Definition 12. Define

$$g_e^{\circ}(s) := \begin{cases} g_e(s), & d_b(s) > 0, \\ 0, & d_b(s) = 0, \end{cases} \qquad \Psi_k(s) := (P_e^k g_e^{\circ})(s).$$

On states with $d_b(s) > 0$, the horizon condition and the model-class restrictions give $|g_e(s)| \le 1$; on states with $d_b(s) = 0$, the definition gives $g_e^{\circ}(s) = 0$. Hence $|g_e^{\circ}(s)| \le 1$ for every s, and therefore

$$\mathrm{osc}(g_e^{\circ}) := \sup_{s,s'} |g_e^{\circ}(s) - g_e^{\circ}(s')| \le 2.$$

The kernel and policy restrictions in Definition 10 make P_e a Markov transition kernel, and Assumption 6 supplies

$$\|\lambda P_e - \lambda' P_e\|_{\mathrm{TV}} \le \alpha \|\lambda - \lambda'\|_{\mathrm{TV}}$$

for all probability laws λ, λ' on $\mathcal{S}$. Consequently, for any bounded function $r : \mathcal{S} \to \mathbb{R}$ and any states s, s', the dual total-variation inequality gives

$$\left|(P_e r)(s) - (P_e r)(s')\right| \le \operatorname{osc}(r)\, \|\delta_s P_e - \delta_{s'} P_e\|_{\mathrm{TV}} \le \alpha \operatorname{osc}(r).$$

Iterating this one-step oscillation contraction yields

$$\operatorname{osc}(P_e^k r) \le \alpha^k \operatorname{osc}(r).$$

Applying the last display to $r = g_e^\circ$ gives

$$\operatorname{osc}(\Psi_k) = \operatorname{osc}(P_e^k g_e^\circ) \le 2\alpha^k.$$

By Lemma 32, every score in the average defining $\widetilde{\theta}_k$ has mean

$$\mathbb{E}_M[Z_t(k)] = \sum_{s \in \mathcal{S}} d_b(s)(P_e^k g_e)(s), \qquad t = k, \dots, T-1.$$

Therefore

$$\mathbb{E}_M[\widetilde{\theta}_k] = \sum_{s \in \mathcal{S}} d_b(s)(P_e^k g_e)(s).$$

For each state s with $d_b(s) > 0$, the functions g_e and g_e° agree on the behavior support, so Lemma 11, applied with the target policy and initial state s, gives

$$(P_e^k g_e)(s) = \Psi_k(s).$$

For states with $d_b(s) = 0$, the common prefactor $d_b(s)$ makes both weighted summands vanish, regardless of the two function values. Hence

$$\sum_s d_b(s)(P_e^k g_e)(s) = \sum_s d_b(s)\Psi_k(s).$$

By Assumption 7, included among the restrictions in Definition 10, $d_e(s) \le C d_b(s)$. Thus $d_e(s) > 0$ implies $d_b(s) > 0$, so

$$\theta(K, e) = \sum_s d_e(s) g_e(s) = \sum_s d_e(s) g_e^\circ(s).$$

By the meaning of the target-policy stationary law used in Definition 12, d_e is invariant for P_e. Thus, for every function $r_{\mathrm{test}} : \mathcal{S} \to \mathbb{R}$ and every integer $m \ge 0$,

$$\sum_s d_e(s) r_{\mathrm{test}}(s) = \sum_s d_e(s)(P_e^m r_{\mathrm{test}})(s).$$

Applying this identity with $r_{\mathrm{test}} = g_e^\circ$ and $m = k$ yields

$$\sum_s d_e(s) g_e^\circ(s) = \sum_s d_e(s)\Psi_k(s).$$

Combining the preceding displays,

$$\mathbb{E}_M[\widetilde{\theta}_k] - \theta(K, e) = \sum_s \{d_b(s) - d_e(s)\}\Psi_k(s).$$

For any function with oscillation at most $2\alpha^k$,

$$\left|\sum_s \{d_b(s) - d_e(s)\}\Psi_k(s)\right| \le 2\alpha^k \, \|d_b - d_e\|_{\mathrm{TV}}.$$

It remains to prove the stationary-law distance subclaim

$$\|d_b - d_e\|_{\mathrm{TV}} \le q_C.$$

This subclaim uses the pointwise comparison $d_e(s) \le C d_b(s)$ from Assumption 7. Throughout this step,

$$\|\mu - \eta\|_{\mathrm{TV}} := \frac{1}{2}\sum_{s\in\mathcal{S}} |\mu(s) - \eta(s)| = \sup_{B\subseteq\mathcal{S}} |\mu(B) - \eta(B)|$$

for probability laws on the finite joint-state space. With this convention, any two probability laws have total-variation distance at most 1, since

$$\frac{1}{2}\sum_s |\mu(s) - \eta(s)| \le \frac{1}{2}\sum_s \{\mu(s) + \eta(s)\} = 1.$$

If $C = 1$, then $d_e(s) \le d_b(s)$ for every s, and the two probability laws have equal total mass, so $d_b = d_e$. Hence

$$\|d_b - d_e\|_{\mathrm{TV}} = 0 = q_C.$$

If $C > 1$, define

$$\eta_C(s) := \frac{C d_b(s) - d_e(s)}{C - 1}, \qquad s \in \mathcal{S}.$$

The latent stationary overlap condition gives $C d_b(s) - d_e(s) \ge 0$ for every s, and

$$\sum_s \eta_C(s) = \frac{C\sum_s d_b(s) - \sum_s d_e(s)}{C - 1} = 1,$$

so η_C is a probability law. The identity

$$d_b = \frac{1}{C} d_e + q_C \eta_C, \qquad q_C = \frac{C - 1}{C},$$

then gives

$$d_b - d_e = q_C(\eta_C - d_e).$$

Therefore

$$\|d_b - d_e\|_{\mathrm{TV}} = q_C \|\eta_C - d_e\|_{\mathrm{TV}} \le q_C.$$

Thus in all cases,

$$\left|\mathbb{E}_M[\widetilde{\theta}_k] - \theta(K, e)\right| \le 2 q_C \alpha^k.$$

4. The variance estimate supplied by Lemma 14, with $L = e^{\zeta}$ and $\alpha = e^{-1/t_0}$, gives

$$\mathrm{Var}_M(\widetilde{\theta}_k) \le \frac{2}{T}\left\{L^{k+1}\left(1 + \frac{4}{L-1}\right) + \frac{4}{1-\alpha}\right\}.$$

5. Since $q_C \geq 0$ and $\alpha \geq 0$,

$$\left(\mathbb{E}_M[\widetilde{\theta}_k] - \theta(K,e)\right)^2 \leq (2q_C\alpha^k)^2 = 4q_C^2\alpha^{2k}.$$

Substituting this and the variance bound into the clipped bias–variance inequality proves

$$\mathbb{E}_M\left[\{\widehat{\theta}_k - \theta(K,e)\}^2\right] \leq 4q_C^2\alpha^{2k} + \frac{2}{T}\left\{L^{k+1}\left(1 + \frac{4}{L-1}\right) + \frac{4}{1-\alpha}\right\}.$$

□

In Lemma 34, the first term is the squared truncation contribution after k observable lags, scaled by the radius q_C. The second term is the sampling contribution from estimating with length-k products of observable policy ratios. Choosing the radius-calibrated depth in Definition 20 balances these terms in the shrinking-overlap results.

The matching radius-explicit KL statement records the lower-bound calculation with the same radius normalization.

Lemma 35. [`lem:radius-explicit-observed-kl`] (Observed KL radius bound) *Let $t_0 > 0$. Define*

$$\alpha(t_0) = \exp(-1/t_0), \qquad c_0(t_0) = \frac{1-\alpha(t_0)}{4}.$$

Then there exists a constant $K_0 = K_0(t_0) > 0$ such that, for every $\zeta > 0$, $C > 1$, $T \in \mathbb{N}$, and $Q \in \mathbb{N}$ with $Q \geq 1$, if

$$\varepsilon_C = \min\left\{\frac{C-1}{2}, \frac{1}{2}\right\}, \qquad L(\zeta) = \exp(\zeta),$$

and if q_C is the radius in Definition 19, then

$$\frac{q_C}{2} \leq \varepsilon_C \leq q_C$$

and the observed-path laws $\mathbb{P}^{\mathrm{obs},T}_{Q,+1}$ and $\mathbb{P}^{\mathrm{obs},T}_{Q,-1}$ of the signed-depth alternatives satisfy

$$\mathrm{KL}\left(\mathbb{P}^{\mathrm{obs},T}_{Q,+1} \,\middle\|\, \mathbb{P}^{\mathrm{obs},T}_{Q,-1}\right) \leq \frac{4c_0(t_0)^2K_0^2}{1-c_0(t_0)^2}\, T\, \varepsilon_C^2\, \alpha(t_0)^{2Q}\, L(\zeta)^{-Q}.$$

Proof of Lemma 35. Fix $t_0 > 0$.

1. Apply Lemma 31 with this value of t_0. It gives a constant $K_0 = K_0(t_0) > 0$, and it remains only to supply the sequential-ignorability and stationary-start hypotheses for the signed-depth experiments over the range requested there. For every auxiliary horizon $T' \in \mathbb{N}$, terminal depth $Q' \geq 1$, and sign alternative $v \in \{-1, +1\}$, the finite construction underlying Definition 18 has the observed-state behavior kernel

$$\Pr_v(A_t = a \mid X_{1:t}, H_{1:t}, A_{1:t-1}, Y_{1:t-1}) = b(a \mid X_t),$$

and the embedding into the observable-data experiment preserves this factorization, so Assumption 2 holds. The same construction starts the joint state from its behavior stationary law,

$$S_1 \sim d_b,$$

so Assumption 3 holds for the same (T', Q', v), including the zero-horizon case. Hence Lemma 31 supplies, for every $\zeta > 0$, $C > 1$, $T \in \mathbb{N}$, and $Q \geq 1$, the sharpened observed-path bound

$$\mathrm{KL}\Big(\mathbb{P}^{\mathrm{obs},T}_{Q,+1} \,\Big\|\, \mathbb{P}^{\mathrm{obs},T}_{Q,-1}\Big) \leq \frac{4c_0(t_0)^2 K_0^2}{1 - c_0(t_0)^2}\, T\, \varepsilon_C^2\, \alpha(t_0)^{2Q}\, L(\zeta)^{-Q}.$$

This KL estimate is one conjunct of the conclusion; the radius comparison is verified separately.

2. It remains to prove

$$\frac{q_C}{2} \leq \varepsilon_C \leq q_C, \qquad q_C = \frac{C-1}{C}, \qquad \varepsilon_C = \min\left\{\frac{C-1}{2}, \frac{1}{2}\right\}.$$

Since $C > 1$, we have $C > 0$ and $C - 1 \geq 0$. First suppose $C \leq 2$. Then

$$\varepsilon_C = \frac{C-1}{2}.$$

For the lower bound, $q_C/2 \leq \varepsilon_C$ is

$$\frac{C-1}{2C} \leq \frac{C-1}{2},$$

which follows from $C > 0$, $C \geq 1$, and $C - 1 \geq 0$. For the upper bound, $\varepsilon_C \leq q_C$ is

$$\frac{C-1}{2} \leq \frac{C-1}{C}.$$

Multiplying by $2C > 0$, this becomes

$$C(C-1) \leq 2(C-1),$$

and this follows from $C - 1 \geq 0$ and $C \leq 2$.

3. Now suppose $C > 2$. Then

$$\varepsilon_C = \frac{1}{2}.$$

For the lower bound, $q_C/2 \leq \varepsilon_C$ is

$$\frac{C-1}{2C} \leq \frac{1}{2},$$

equivalently $C - 1 \leq C$, which is immediate. For the upper bound, $\varepsilon_C \leq q_C$ is

$$\frac{1}{2} \leq \frac{C-1}{C}.$$

Multiplying by $2C > 0$, this becomes $C \leq 2(C-1)$, equivalently $2 \leq C$, which follows from $C > 2$.

Combining the two cases gives the asserted radius comparison, and the KL bound is the one obtained in Step 1 with $L(\zeta) = \exp(\zeta)$ and $\alpha(t_0) = \exp(-1/t_0)$. □

The inequalities in Lemma 35 connect the construction amplitude ε_C to the normalized radius from Definition 19. Consequently, the lower calculation can be stated on the same radius scale as the estimator bound in Lemma 34.

The final group records the parametric two-experiment comparison used to anchor the risk at the usual T^{-1} scale across all radii in the model class. The first two statements compute the separation in target value and the observed KL budget for the singleton-state pair; the following bounded-risk and information inequalities convert those calculations into risk floors.

Lemma 36. [lem:parametric-pair-target-value] (Parametric pair target value) *Let $T \in \mathbb{N}$, let $v \in \{0,1\}$, and define the sign*

$$\sigma(v) = \begin{cases} +1, & v = 1, \\ -1, & v = 0. \end{cases}$$

Put

$$\eta_T = \frac{1}{4\sqrt{T}},$$

with value 0 when $T = 0$. Suppose the scalar parameters satisfy:

- ***(Initial scale.)*** $t_0 > 0$.
- ***(Overlap scale.)*** $\zeta > 0$.
- ***(Radius scale.)*** $C \geq 1$.

Let M_T^v be the singleton-state finite experiment with one observed state and one hidden state, fair behavior and target policies $b(a \mid x_\star) = e(a \mid x_\star) = 1/2$, reward support $\{-1, +1\}$, and reward-transition kernel that assigns, from every state-action pair, probability

$$\frac{1 + \sigma(v) y \eta_T}{2}$$

to reward $y \in \{-1, +1\}$ and next state $(x_\star, h_\star)$. Then the stationary target-policy mean reward of Definition 12, evaluated in this experiment, satisfies

$$\theta(K_{M_T^v}, e_{M_T^v}) = \sigma(v)\eta_T.$$

Proof of Lemma 36. Let $F = M_T^v$, viewed first as the finite reward model before applying the embedding into the raw experiment. Write

$$d_T^v(s) := d_e(s)$$

for the target-policy stationary law on the singleton joint-state space $\mathcal{S} = \{(x_\star, h_\star)\}$ associated with this finite model. The construction of M_T^v, under $t_0 > 0$, $\zeta > 0$, and $C \geq 1$, belongs to $\mathcal{M}_T(t_0, \zeta, C)$; in particular the target stationary law is a probability vector. Thus

$$d_T^v(s) \geq 0 \quad \text{for all } s \in \mathcal{S}, \qquad \sum_{s \in \mathcal{S}} d_T^v(s) = 1.$$

The equality between the finite encoding and the raw experiment preserves this stationary law, so the same d_T^v is the law used in $\theta(K_{M_T^v}, e_{M_T^v})$ as defined in Definition 12.

For every joint state $s \in \mathcal{S}$ and action $a \in \{0,1\}$, the conditional reward mean in the finite kernel is

$$\sum_{y \in \{-1,+1\}} y \Pr(Y = y \mid S = s, A = a) = (-1)\frac{1 - \sigma(v)\eta_T}{2} + (+1)\frac{1 + \sigma(v)\eta_T}{2} = \sigma(v)\eta_T.$$

The displayed calculation is exactly the reward-kernel mean, with the encoding $0 \mapsto -1$ and $1 \mapsto +1$.

By Definition 12 and the finite-to-raw value identity for this encoding,

$$\theta(K_{M_T^v}, e_{M_T^v}) = \sum_{s\in\mathcal{S}} d_T^v(s) \sum_{a\in\{0,1\}} e_{M_T^v}(a \mid x(s)) \sum_{y\in\{-1,+1\}} y \Pr(Y = y \mid S = s, A = a).$$

Substituting the preceding reward-mean identity gives

$$\theta(K_{M_T^v}, e_{M_T^v}) = \sum_{s\in\mathcal{S}} d_T^v(s) \sum_{a\in\{0,1\}} e_{M_T^v}(a \mid x(s))\, \sigma(v)\eta_T.$$

The target policy in M_T^v is fair, hence for every s,

$$\sum_{a\in\{0,1\}} e_{M_T^v}(a \mid x(s)) = 1.$$

Therefore

$$\theta(K_{M_T^v}, e_{M_T^v}) = \sum_{s\in\mathcal{S}} d_T^v(s)\, \sigma(v)\eta_T = \sigma(v)\eta_T \sum_{s\in\mathcal{S}} d_T^v(s) = \sigma(v)\eta_T,$$

using the total mass identity for d_T^v. □

Lemma 37. [lem:parametric-pair-observed-kl] (Observed KL budget) *Let $T \in \mathbb{N}$ satisfy $T \geq 1$, and set $\eta_T = 1/(4\sqrt{T})$. Let M_T^- and M_T^+ be the two finite singleton-state experiments with one observed state $x_\star$, one hidden state, initial mass at the single joint state, fair behavior and target policies $b = e$ on $\{0,1\}$, and reward code $\mathsf{r} \in \{0,1\}$ decoded as $2\mathsf{r} - 1$. Under M_T^σ, for $\sigma \in \{-1,+1\}$, the reward-transition kernel returns to the single joint state and assigns reward $y \in \{-1,+1\}$ probability*

$$\frac{1 + \sigma y \eta_T}{2}.$$

Writing $\mathbb{P}^{\mathrm{obs}}_{M_T^\sigma}$ for the image of the decoded trajectory law under the observed-path projection, the required KL budget holds in the direction

$$\mathrm{KL}\left(\mathbb{P}^{\mathrm{obs}}_{M_T^-} \,\middle\|\, \mathbb{P}^{\mathrm{obs}}_{M_T^+}\right) \leq \frac{1}{4}.$$

Proof of Lemma 37. Write F_- and F_+ for the two finite reward-code models with Boolean indices false and true, so that the associated signs are -1 and $+1$. Let

$$\eta_T = \frac{1}{4\sqrt{T}}.$$

Since $T \geq 1$, we have $0 < \sqrt{T}$ and $0 \leq \eta_T \leq 1/4$.

1. The two finite models use the same reward decoder $r \mapsto 2r - 1$. Therefore applying the decoder and then projecting to the observed path is a common measurable map, and KL contracts under this map:

$$\mathrm{KL}\left(\mathbb{P}^{\mathrm{obs}}_{M_T^-} \,\middle\|\, \mathbb{P}^{\mathrm{obs}}_{M_T^+}\right) \leq \mathrm{KL}\left((F_-)^{\mathrm{obs}} \,\middle\|\, (F_+)^{\mathrm{obs}}\right),$$

where $(F_\sigma)^{\mathrm{obs}}$ denotes the finite observed-path law before decoding the reward code into $\{-1,+1\}$.

2. For one epoch define probability laws μ_- and μ_+ on $\{x_\star\} \times \{0,1\} \times \{0,1\}$ by

$$\mu_\sigma(x_\star, a, r) = \frac{1}{2} \frac{1 + \sigma(2r-1)\eta_T}{2}, \qquad \sigma \in \{-1,+1\}, \quad a, r \in \{0,1\}.$$

The first factor is the fair action law, and the second factor is the reward-code law. Because $0 \le \eta_T \le 1/4$, every displayed cell mass is positive.

3. Multiplying the singleton initial mass, the fair action probabilities, and the displayed time-homogeneous reward-transition probabilities gives the product form

$$(F_\sigma)^{\mathrm{obs}} = \bigotimes_{t=1}^{T} \mu_\sigma, \qquad \sigma \in \{-1,+1\}.$$

The singleton observed state contributes unit mass at each time, so the only nontrivial coordinates in each factor are the fair action and the reward code.

4. The one-epoch divergence is bounded by

$$\mathrm{KL}(\mu_- \| \mu_+) \le 4\eta_T^2.$$

Let A be the fair law on $\{0,1\}$, let $\mathrm{Ber}_{\{0,1\}}(u)$ be the law on $\{0,1\}$ assigning mass u to reward code 1, and set

$$u_- := \frac{1-\eta_T}{2}, \qquad u_+ := \frac{1+\eta_T}{2}.$$

Since $0 \le \eta_T \le 1/4$, both u_- and u_+ lie in $[3/8, 5/8]$. The law μ_σ is obtained from $\mathrm{Ber}_{\{0,1\}}(u_\sigma) \otimes A$ by the bijective coordinate map $(r,a) \mapsto (x_\star, a, r)$. Hence invariance under this reordering and the common action coordinate give

$$\mathrm{KL}(\mu_- \| \mu_+) = \mathrm{KL}\big(\mathrm{Ber}_{\{0,1\}}(u_-) \otimes A \,\big\|\, \mathrm{Ber}_{\{0,1\}}(u_+) \otimes A\big) = \mathrm{KL}\big(\mathrm{Ber}_{\{0,1\}}(u_-) \,\big\|\, \mathrm{Ber}_{\{0,1\}}(u_+)\big)\,.$$

The Boolean Bernoulli laws are images of the two-point real laws $(1-u)\delta_0 + u\delta_1$ under the measurable map that records whether the point is 1, so data processing gives the upper bound by the corresponding real two-point divergence. Evaluating that finite divergence yields

$$\begin{aligned}\mathrm{KL}\big(\mathrm{Ber}_{\{0,1\}}(u_-) \,\big\|\, \mathrm{Ber}_{\{0,1\}}(u_+)\big) &\le u_- \log \frac{u_-}{u_+} + (1-u_-) \log \frac{1-u_-}{1-u_+} \\ &= \eta_T \log \frac{1+\eta_T}{1-\eta_T}.\end{aligned}$$

For $0 \le s \le 1/4$, the function $4s - \log((1+s)/(1-s))$ has value 0 at $s = 0$ and derivative

$$4 - \frac{2}{1-s^2} \ge 4 - \frac{2}{1-(1/4)^2} > 0.$$

Thus $\log((1+s)/(1-s)) \le 4s$ on this interval, and substituting $s = \eta_T$ gives the displayed one-epoch bound.

5. Since every one-epoch cell under μ_+ has positive mass, $\mu_- \ll \mu_+$; on this finite space the log-likelihood ratio is integrable. Tensorization of KL for the T-fold product laws gives

two finite-valued conclusions: the product divergence is finite, and after applying the usual real value map to the extended nonnegative divergence,

$$\mathrm{KL}\left(\bigotimes_{t=1}^{T}\mu_{-}\,\middle\|\,\bigotimes_{t=1}^{T}\mu_{+}\right)_{\mathbb{R}}\le T\ \mathrm{KL}(\mu_{-}\|\mu_{+})_{\mathbb{R}}.$$

The one-epoch bound from the previous step is finite, so the same real-value comparison gives

$$\mathrm{KL}(\mu_{-}\|\mu_{+})_{\mathbb{R}}\le 4\eta_T^2.$$

The product representation of the finite observed-path law therefore yields

$$\mathrm{KL}\Big((F_{-})^{\mathrm{obs}}\,\Big\|\,(F_{+})^{\mathrm{obs}}\Big)_{\mathbb{R}}\le T\cdot 4\eta_T^2.$$

Because the displayed product divergence is finite, comparison with any finite numerical bound is equivalent to comparison after taking this real value.

6. Finally,

$$T\cdot 4\eta_T^2=T\cdot 4\left(\frac{1}{4\sqrt{T}}\right)^2=\frac{1}{4},$$

using $0<\sqrt{T}$ and $(\sqrt{T})^2=T$. The finite-valued comparison from the preceding step gives

$$\mathrm{KL}\Big((F_{-})^{\mathrm{obs}}\,\Big\|\,(F_{+})^{\mathrm{obs}}\Big)\le\frac{1}{4}.$$

The contraction inequality from the first step and the finite observed-path bound from the last step give

$$\mathrm{KL}\Big(\mathbb{P}^{\mathrm{obs}}_{M_T^{-}}\,\Big\|\,\mathbb{P}^{\mathrm{obs}}_{M_T^{+}}\Big)\le\frac{1}{4},$$

as claimed. □

Before applying the two-point inequality, the next statement records a uniform boundedness fact for clipped observable-data estimators. It uses the target-value range from Lemma 33 and the estimator range built into Definition 14.

Lemma 38. [`lem:observed-risk-bounded-by-four`] (Observed risk bound) *Let $T\ge 1$ and let $t_0,\zeta,C\in\mathbb{R}$. Let $\widehat{\theta}$ be an observable-data estimator of the admissible class used in Definition 14: for each observed-alphabet size and policy pair, $\widehat{\theta}$ is a measurable real-valued function of the observed path taking values in $[-1,1]$. For every*

$$M=(\mathcal{X},\mathcal{H},K,b,e,\mathcal{L}(\mathsf{O}_T))\in\mathcal{M}_T(t_0,\zeta,C)$$

as in Definition 10, with $\widehat{\theta}$ evaluated at the observed-alphabet size and policies of M, the squared risk for the target value $\theta(K,e)$ of Definition 12 satisfies

$$E_M\left[(\widehat{\theta}-\theta(K,e))^2\right]\le 4.$$

Proof of Lemma 38. 1. Fix

$$M = (\mathcal{X}, \mathcal{H}, K, b, e, \mathcal{L}(\mathsf{O}_T)) \in \mathcal{M}_T(t_0, \zeta, C),$$

and write an observed path as

$$\omega_{\mathrm{obs}} = \big((X_t, A_t, Y_t)\big)_{t=0}^{T-1} \in \mathsf{O}_T.$$

For this path set

$$r_{\mathrm{err}}(\omega_{\mathrm{obs}}) := \widehat{\theta}(n_X, b, e, \omega_{\mathrm{obs}}) - \theta(K, e).$$

The admissibility condition in Definition 14 gives

$$-1 \le \widehat{\theta}(n_X, b, e, \omega_{\mathrm{obs}}) \le 1,$$

and Lemma 33 gives

$$-1 \le \theta(K, e) \le 1.$$

Hence

$$-2 \le r_{\mathrm{err}}(\omega_{\mathrm{obs}}) \le 2.$$

Equivalently,

$$0 \le r_{\mathrm{err}}(\omega_{\mathrm{obs}}) + 2, \qquad 0 \le 2 - r_{\mathrm{err}}(\omega_{\mathrm{obs}}).$$

Multiplying these two nonnegative quantities yields

$$0 \le \big(r_{\mathrm{err}}(\omega_{\mathrm{obs}}) + 2\big)\big(2 - r_{\mathrm{err}}(\omega_{\mathrm{obs}})\big) = 4 - r_{\mathrm{err}}(\omega_{\mathrm{obs}})^2,$$

so

$$\big(\widehat{\theta}(n_X, b, e, \omega_{\mathrm{obs}}) - \theta(K, e)\big)^2 \le 4 \qquad \text{for every } \omega_{\mathrm{obs}} \in \mathsf{O}_T.$$

2. The function

$$\omega_{\mathrm{obs}} \longmapsto \big(\widehat{\theta}(n_X, b, e, \omega_{\mathrm{obs}}) - \theta(K, e)\big)^2$$

is measurable because $\widehat{\theta}(n_X, b, e, \cdot)$ is measurable by admissibility and the target value is constant in the observed path. The observed-path law $\mathcal{L}_M^{\mathrm{obs}}$ is a probability law, and the pointwise bound from the previous step gives integrability together with

$$\int \big(\widehat{\theta}(n_X, b, e, \omega_{\mathrm{obs}}) - \theta(K, e)\big)^2 \, d\mathcal{L}_M^{\mathrm{obs}}(\omega_{\mathrm{obs}}) \le \int 4 \, d\mathcal{L}_M^{\mathrm{obs}}(\omega_{\mathrm{obs}}) = 4.$$

By the squared-risk definition in Definition 14, the left-hand side is exactly

$$E_M\big[(\widehat{\theta} - \theta(K, e))^2\big].$$

Therefore

$$E_M\big[(\widehat{\theta} - \theta(K, e))^2\big] \le 4.$$

□

Lemma 38 places all admissible estimator losses on a common bounded scale. This supplies the integrability and range control used when the two-point comparison is embedded into the minimax risk.

Lemma 39. [lem:two-point-risk-floor] (Two-point risk floor) *Let $T \geq 1$, $t_0, \zeta, C \in \mathbb{R}$, and let $M_0, M_1 \in \mathcal{M}_T(t_0, \zeta, C)$ as in Definition 10. After identifying their common observed alphabet, write $\mathbb{P}_j$ for the observed-data law of M_j and θ_j for the target value of M_j from Definition 12, $j \in \{0, 1\}$. Let s be a nonnegative separation radius and let χ be a nonnegative KL upper bound. Assume:*

- ***(Common observable structure.)*** *The two models have the same observed-alphabet cardinality, the same behavior policy, and the same target policy.*
- ***(Information bound.)*** $\mathbb{P}_0 \ll \mathbb{P}_1$,
$$\mathrm{KL}(\mathbb{P}_0 \| \mathbb{P}_1) \leq \chi, \qquad \mathrm{KL}(\mathbb{P}_0 \| \mathbb{P}_1) < \infty.$$
- ***(Target separation.)***
$$2s \leq |\theta_0 - \theta_1|.$$

Then every raw observable-data estimator $\widehat{\theta}$, viewed as a real-valued measurable function of the observed-alphabet size, the policy pair, and the observed path, with $(\widehat{\theta} - \theta_j)^2$ integrable under $\mathbb{P}_j$ at M_j for $j = 0, 1$, satisfies

$$\frac{s^2 \exp(-\chi)}{4} \leq \max\left\{ \mathbb{E}_{M_0}\left[(\widehat{\theta} - \theta_0)^2\right], \mathbb{E}_{M_1}\left[(\widehat{\theta} - \theta_1)^2\right] \right\}.$$

Consequently, the minimax squared-error risk of Definition 14 satisfies

$$\frac{s^2 \exp(-\chi)}{4} \leq R_T(t_0, \zeta, C).$$

Proof of Lemma 39. Write the common observed alphabet size as n_X, and use the common observable structure to identify the behavior and target policy carriers of the two models. Put

$$P_j := \mathbb{P}_j, \qquad \theta_j := \theta(M_j), \qquad j \in \{0, 1\},$$

where $\theta(M_j)$ is the target value from Definition 12. The information bound and $\chi \geq 0$ give

$$\mathrm{KL}(P_0 \| P_1)_{\mathbb{R}} \leq \chi,$$

where $\mathrm{KL}(P_0 \| P_1)_{\mathbb{R}}$ denotes the real value of the finite extended-real KL divergence. Hence

$$\exp(-\chi) \leq \exp\{-\mathrm{KL}(P_0 \| P_1)_{\mathbb{R}}\}.$$

By the Bretagnolle–Huber affinity inequality, using $P_0 \ll P_1$ and finite KL divergence,

$$\frac{1}{2} \exp\{-\mathrm{KL}(P_0 \| P_1)_{\mathbb{R}}\} \leq 1 - \mathrm{TV}(P_0, P_1).$$

Combining the two displays yields

$$\frac{\exp(-\chi)}{4} \leq \frac{1 - \mathrm{TV}(P_0, P_1)}{2}. \tag{1}$$

Fix a raw observable-data estimator $\widehat{\theta}$ satisfying the measurability and integrability hypotheses. With the common observable inputs fixed, define

$$\widehat{f}(o) := \widehat{\theta}(n_X, b, e, o), \qquad o \in \mathsf{O}_T,$$

and define the two error events

$$E_j := \{o \in \mathsf{O}_T : \ s \le |\widehat{f}(o) - \theta_j|\}, \qquad j \in \{0,1\}.$$

The separation condition $2s \le |\theta_0 - \theta_1|$ and the two-point testing bound for real-valued parameters give

$$\frac{1 - \mathrm{TV}(P_0, P_1)}{2} \le \max\{P_0(E_0), P_1(E_1)\}. \tag{2}$$

Since $s \ge 0$, for each $j \in \{0,1\}$,

$$E_j = \{o \in \mathsf{O}_T : \ s^2 \le (\widehat{f}(o) - \theta_j)^2\}.$$

The squared-error integrability assumptions therefore allow the elementary tail-to-mean comparison

$$s^2 P_j(E_j) \le \int (\widehat{f}(o) - \theta_j)^2 \, dP_j(o), \qquad j \in \{0,1\}. \tag{3}$$

Combining Equations (1) and (2) gives

$$\frac{\exp(-\chi)}{4} \le \max\{P_0(E_0), P_1(E_1)\}.$$

If $P_0(E_0) \le P_1(E_1)$, then the last maximum is $P_1(E_1)$, and Equation (3) gives

$$\frac{s^2 \exp(-\chi)}{4} \le s^2 P_1(E_1) \le \int (\widehat{f}(o) - \theta_1)^2 \, dP_1(o) \le \max_{j \in \{0,1\}} \int (\widehat{f}(o) - \theta_j)^2 \, dP_j(o).$$

If $P_1(E_1) \le P_0(E_0)$, the same argument gives

$$\frac{s^2 \exp(-\chi)}{4} \le s^2 P_0(E_0) \le \int (\widehat{f}(o) - \theta_0)^2 \, dP_0(o) \le \max_{j \in \{0,1\}} \int (\widehat{f}(o) - \theta_j)^2 \, dP_j(o).$$

These two cases prove

$$\frac{s^2 \exp(-\chi)}{4} \le \max\left\{\mathbb{E}_{M_0}\left[(\widehat{\theta} - \theta_0)^2\right], \mathbb{E}_{M_1}\left[(\widehat{\theta} - \theta_1)^2\right]\right\}. \tag{4}$$

It remains to pass from the fixed two models to the minimax risk in Definition 14. The observable-estimator class is nonempty, for example it contains the constant-zero estimator. We first record the bounded-risk input needed for the minimax reduction. Fix an admissible observable estimator $\widetilde{\theta}$. For every model $M \in \mathcal{M}_T(t_0, \zeta, C)$, with observed alphabet size n_X^M, behavior carrier b_M, target carrier e_M, observed law P_M, and target value $\theta(M)$, Lemma 38 gives

$$\mathbb{E}_M\left[\left(\widetilde{\theta}(n_X^M, b_M, e_M, \mathsf{O}_T) - \theta(M)\right)^2\right] = \int \left(\widetilde{\theta}(n_X^M, b_M, e_M, o) - \theta(M)\right)^2 dP_M(o) \le 4.$$

Consequently,

$$\left\{\mathbb{E}_M\left[\left(\widetilde{\theta}(n_X^M, b_M, e_M, \mathsf{O}_T) - \theta(M)\right)^2\right] : M \in \mathcal{M}_T(t_0, \zeta, C)\right\} \subset (-\infty, 4],$$

which is the boundedness hypothesis needed in the two-point minimax reduction. For such an estimator, the two integrability requirements used above hold at M_0 and M_1. These are two

separate statements, one for each model, and each is taken at that model's own observable inputs before any identification of carriers: at M_j the estimator is measurable by admissibility and its range condition gives

$$|\widehat{\theta}(n_X, b_{M_j}, e_{M_j}, o)| \le 1, \qquad o \in \mathsf{O}_T,$$

so the fixed-model squared error is dominated by the constant $(1 + |\theta(M_j)|)^2$, since

$$|\widehat{\theta}(n_X, b_{M_j}, e_{M_j}, o) - \theta(M_j)| \le 1 + |\theta(M_j)|, \qquad j \in \{0, 1\}.$$

Only after both integrability statements are in hand do the common-policy hypotheses $b_{M_0} = b_{M_1}$ and $e_{M_0} = e_{M_1}$ rewrite the two carriers into the single pair (b, e) used for $\widehat{f}$ above, so that the two integrals are integrals of one function against the two observed laws. Applying Equation (4) to each admissible observable estimator and then the standard two-point reduction to the minimax value yields

$$\frac{s^2 \exp(-\chi)}{4} \le R_T(t_0, \zeta, C),$$

which is the claimed minimax lower bound. □

Lemma 36 sets the value separation for the singleton-state Rademacher pair, and Lemma 37 gives the corresponding observed KL budget. The bounded-risk statement in Lemma 38 fixes the loss scale for admissible clipped estimators, and the generic reduction in Lemma 39 then translates a common-policy, low-divergence pair with target separation into a lower bound for every observable-data estimator.

Lemma 40. [`lem:uniform-parametric-floor`] (Uniform parametric floor) *Fix a positive mixing scale $t_0 > 0$ and a positive log-overlap scale $\zeta > 0$. For every horizon $T \in \mathbb{N}$ with $T \ge 1$ and every radius $C \ge 1$, let M_T^- and M_T^+ be the two raw POMDP experiments generated by the singleton-state, singleton-observation Rademacher construction with rewards in $\{-1, +1\}$, fair behavior and target policies $b = e$, and the two opposite signs of the construction's parametric reward tilt. Then:*

- ***(Class membership.)*** *Both M_T^- and M_T^+ belong to the latent-overlap model class $\mathcal{M}_T(t_0, \zeta, C)$ of Definition 10.*
- ***(Common policies.)*** *Each experiment has equal behavior and target policies, and the two experiments share these policies:*

$$b_{M_T^-} = e_{M_T^-}, \qquad b_{M_T^+} = e_{M_T^+}, \qquad b_{M_T^-} = b_{M_T^+}, \qquad e_{M_T^-} = e_{M_T^+}.$$

- ***(Uniform minimax floor.)*** *The minimax squared-error risk $R_T(t_0, \zeta, C)$ of Definition 14 satisfies*

$$\frac{\exp(-1/4)}{64T} \le R_T(t_0, \zeta, C).$$

- ***(Pointwise two-experiment floor.)*** *For every observable-data estimator $\widehat{\theta}$, viewed as a measurable function of the revealed alphabet and the policies, under the usual squared-error integrability conditions for the observed trajectory laws of M_T^- and M_T^+,*

$$\frac{\exp(-1/4)}{64T} \le \max\left\{ \mathbb{E}^{\mathrm{obs}}_{M_T^-}\left[\left(\widehat{\theta}(1, b_{M_T^-}, e_{M_T^-}; \mathsf{O}_T) - \theta(K_{M_T^-}, e_{M_T^-})\right)^2\right], \mathbb{E}^{\mathrm{obs}}_{M_T^+}\left[\left(\widehat{\theta}(1, b_{M_T^+}, e_{M_T^+}; \mathsf{O}_T) - \theta(K_{M_T^+},\right.\right.$$

where $\theta(K, e)$ is the stationary target-policy mean reward of Definition 12.

Proof of Lemma 40. 1. Fix $T \geq 1$ and put

$$\eta_T := \frac{1}{4\sqrt{T}}.$$

Since $T \geq 1$, $0 \leq \eta_T \leq 1/4$. For each sign $\sigma \in \{-1, +1\}$, define a singleton-state experiment with $\mathcal{X} = \{x_\star\}$, $\mathcal{H} = \{h_\star\}$, fair policies

$$b_\sigma(a \mid x_\star) = e_\sigma(a \mid x_\star) = \frac{1}{2}, \qquad a \in \{0, 1\},$$

and reward-transition kernel

$$\Pr_\sigma\{Y_t = y,\ X_{t+1} = x_\star,\ H_{t+1} = h_\star \mid X_t = x_\star, H_t = h_\star, A_t = a\} = \frac{1 + \sigma y \eta_T}{2}, \qquad y \in \{-1, +1\}.$$

Let M_T^- be the experiment with $\sigma = -1$, and M_T^+ the experiment with $\sigma = +1$. The weights above are nonnegative and sum to one. The hypotheses give the positive parameter-scale requirements $0 < t_0$ and $0 < \zeta$ appearing in Definition 10. The Markov-kernel, sequential-ignorability, stationary-start, and bounded-reward requirements in Assumptions 1 to 4 hold by this generated singleton construction. The policy-overlap condition in Assumption 5 holds because $b_\sigma = e_\sigma$ and $\exp(\zeta) \geq 1$. For either policy $p \in \{b_\sigma, e_\sigma\}$,

$$\nu P_p = \nu' P_p = \delta_{(x_\star, h_\star)}$$

for all probability laws ν, ν' on the joint state space. Put

$$\alpha_{t_0} := \exp(-1/t_0).$$

Since $t_0 > 0$, $\alpha_{t_0} > 0$, and therefore $\alpha_{t_0}\|\nu - \nu'\|_{\mathrm{TV}} \geq 0$. Thus

$$\|\nu P_p - \nu' P_p\|_{\mathrm{TV}} = 0 \leq \alpha_{t_0}\|\nu - \nu'\|_{\mathrm{TV}},$$

which is the uniform-contraction requirement in Assumption 6. Finally, the latent stationary overlap condition in Assumption 7 is immediate: $d_{e_\sigma} = d_{b_\sigma} = \delta_{(x_\star, h_\star)}$, and hence $d_{e_\sigma}(s) \leq C\, d_{b_\sigma}(s)$ for every state whenever $C \geq 1$. Thus both alternatives belong to $\mathcal{M}_T(t_0, \zeta, C)$ as defined in Definition 10. The same displayed definition of the policies also gives

$$b_{M_T^-} = e_{M_T^-}, \qquad b_{M_T^+} = e_{M_T^+}, \qquad b_{M_T^-} = b_{M_T^+}, \qquad e_{M_T^-} = e_{M_T^+}.$$

2. The singleton alternatives just constructed are the two alternatives covered by Lemma 36. With the sign convention of Step 1, that result gives

$$\theta_- := \theta(K_{M_T^-}, e_{M_T^-}) = -\eta_T, \qquad \theta_+ := \theta(K_{M_T^+}, e_{M_T^+}) = \eta_T.$$

Consequently

$$2\eta_T = |\theta_- - \theta_+|.$$

For the divergence calculation, let $\mathsf{r} \in \{0, 1\}$ be the reward code and set

$$\mathsf{y}(\mathsf{r}) := 2\mathsf{r} - 1, \qquad \mathsf{y}(0) = -1, \qquad \mathsf{y}(1) = +1.$$

Define the one-epoch coded observed law $\Lambda_T^\sigma \in \Delta(\{x_\star\} \times \{0,1\} \times \{0,1\})$ by

$$\Lambda_T^\sigma\{(x_\star, a, \mathsf{r})\} := \frac{1}{2}\,\frac{1+\sigma \mathsf{y}(\mathsf{r})\eta_T}{2}, \qquad a, \mathsf{r} \in \{0,1\}.$$

Since $0 \le \eta_T \le 1/4$, every displayed cell has positive mass. Thus $\Lambda_T^- \ll \Lambda_T^+$, and the log likelihood ratio is integrable. If Ξ_T^- and Ξ_T^+ denote the reward-code marginals of Λ_T^- and Λ_T^+, then

$$\Xi_T^-\{1\} = \frac{1-\eta_T}{2}, \qquad \Xi_T^+\{1\} = \frac{1+\eta_T}{2}.$$

The fair action factor is common to the two signs, so the product rule for KL with a common second factor gives

$$\mathrm{KL}(\Lambda_T^- \| \Lambda_T^+) = \mathrm{KL}(\Xi_T^- \| \Xi_T^+).$$

Writing out the two Bernoulli masses gives

$$\mathrm{KL}(\Xi_T^- \| \Xi_T^+) = \frac{1-\eta_T}{2}\log\frac{1-\eta_T}{1+\eta_T} + \frac{1+\eta_T}{2}\log\frac{1+\eta_T}{1-\eta_T} = \eta_T \log\frac{1+\eta_T}{1-\eta_T}.$$

For $0 \le u \le 1/4$,

$$\log\frac{1+u}{1-u} = \log(1+u) - \log(1-u) \le u + \frac{u}{1-u} \le 4u,$$

using $\log(1+u) \le u$ and $-\log(1-u) \le u/(1-u)$. Applying this with $u = \eta_T$ yields the one-epoch bound

$$\mathrm{KL}(\Lambda_T^- \| \Lambda_T^+) \le 4\eta_T^2.$$

Let

$$\Pi_{\sigma,T}^{\mathrm{code}} := \bigotimes_{i=1}^{T} \Lambda_T^\sigma \quad \text{on} \quad (\{x_\star\} \times \{0,1\} \times \{0,1\})^T.$$

The generated finite trajectory construction has this product as its coded observed-word law. By finite-product KL tensorization, with the absolute continuity and integrability supplied by the positive one-epoch masses,

$$\mathrm{KL}(\Pi_{-,T}^{\mathrm{code}} \| \Pi_{+,T}^{\mathrm{code}}) \le T\,\mathrm{KL}(\Lambda_T^- \| \Lambda_T^+) \le 4T\eta_T^2.$$

Because $T \ge 1$, $\sqrt{T} > 0$, and $\eta_T = 1/(4\sqrt{T})$,

$$4T\eta_T^2 = \frac{1}{4}.$$

Finally, let $\mathfrak{d}_T$ be the decoding map that replaces each reward code r by $\mathsf{y}(\mathsf{r})$ and keeps the observed state and action coordinates. The observed laws in the statement are the pushforwards

$$\mathbb{P}_{M_T^-}^{\mathrm{obs}} = (\mathfrak{d}_T)_\# \Pi_{-,T}^{\mathrm{code}}, \qquad \mathbb{P}_{M_T^+}^{\mathrm{obs}} = (\mathfrak{d}_T)_\# \Pi_{+,T}^{\mathrm{code}}.$$

KL divergence is monotone under measurable pushforward, so the preceding coded product bound gives

$$\mathrm{KL}\left(\mathbb{P}_{M_T^-}^{\mathrm{obs}} \,\middle\|\, \mathbb{P}_{M_T^+}^{\mathrm{obs}}\right) \le \mathrm{KL}(\Pi_{-,T}^{\mathrm{code}} \| \Pi_{+,T}^{\mathrm{code}}) \le \frac{1}{4}.$$

This is the observed KL budget in Lemma 37. The positive one-epoch masses give $\Pi_{-,T}^{\mathrm{code}} \ll \Pi_{+,T}^{\mathrm{code}}$, hence also $\mathbb{P}_{M_T^-}^{\mathrm{obs}} \ll \mathbb{P}_{M_T^+}^{\mathrm{obs}}$, and the displayed upper bound gives finite KL divergence.

3. Let $\widehat{\theta}$ be any observable-data estimator satisfying the measurability and two integrability assumptions in the statement, and write

$$\varphi(o) := \widehat{\theta}(1, b_{M_T^-}, e_{M_T^-}; o) = \widehat{\theta}(1, b_{M_T^+}, e_{M_T^+}; o),$$

where the equality uses the common policies from Step 1. The two-point risk inequality applied here is Lemma 39, used in its raw-estimator form for the estimator $\widehat{\theta}$. Its structural hypotheses are available here: the observed alphabets both have one state and the behavior and target policies are common by Step 1. Step 2 supplies

$$2\eta_T \le |\theta_- - \theta_+|, \qquad \mathbb{P}^{\mathrm{obs}}_{M_T^-} \ll \mathbb{P}^{\mathrm{obs}}_{M_T^+}, \qquad \mathrm{KL}\Big(\mathbb{P}^{\mathrm{obs}}_{M_T^-} \Big\| \mathbb{P}^{\mathrm{obs}}_{M_T^+}\Big) \le \frac{1}{4},$$

together with finite KL divergence. The present step supplies the required measurability of $\widehat{\theta}(n_X, b, e; \cdot)$ for every input (n_X, b, e), and the two squared-error integrability assumptions for the two displayed risks.

Apply Lemma 39 directly to the raw estimator $\widehat{\theta}$, the two observed laws

$$\mathbb{P}^{\mathrm{obs}}_{M_T^-}, \qquad \mathbb{P}^{\mathrm{obs}}_{M_T^+},$$

the target values

$$\theta_- = \theta(K_{M_T^-}, e_{M_T^-}), \qquad \theta_+ = \theta(K_{M_T^+}, e_{M_T^+}),$$

and the numerical choices

$$s_\star = \eta_T, \qquad \chi_{\mathrm{KL}} = \frac{1}{4}.$$

The separation condition is supplied by Step 2, and the absolute-continuity and finite-KL hypotheses are supplied by the final display of Step 2. Hence

$$\frac{\eta_T^2 \exp(-1/4)}{4} \le \max\left\{\mathbb{E}^{\mathrm{obs}}_{M_T^-}\left[\{\varphi(\mathsf{O}_T) - \theta_-\}^2\right], \mathbb{E}^{\mathrm{obs}}_{M_T^+}\left[\{\varphi(\mathsf{O}_T) - \theta_+\}^2\right]\right\}.$$

Finally,

$$\frac{\eta_T^2 \exp(-1/4)}{4} = \frac{\exp(-1/4)}{64T},$$

which is exactly the pointwise two-experiment floor in the statement.

4. It remains to record the minimax lower bound. For the supplied radius C, the two singleton alternatives with the same kernels and policies satisfy the hypotheses of the minimax conclusion in Lemma 39: Step 1 gives their common observable cardinality, common behavior policy, common target policy, and membership in $\mathcal{M}_T(t_0, \zeta, C)$; Step 2 gives

$$2\eta_T \le |\theta_- - \theta_+|, \qquad \mathbb{P}^{\mathrm{obs}}_{M_T^-} \ll \mathbb{P}^{\mathrm{obs}}_{M_T^+}, \qquad \mathrm{KL}\Big(\mathbb{P}^{\mathrm{obs}}_{M_T^-} \Big\| \mathbb{P}^{\mathrm{obs}}_{M_T^+}\Big) \le \frac{1}{4},$$

with finite KL divergence. Applying Lemma 39 with

$$s_\star = \eta_T, \qquad \chi_{\mathrm{KL}} = \frac{1}{4}$$

yields

$$\frac{\eta_T^2 \exp(-1/4)}{4} \le R_T(t_0, \zeta, C),$$

where $R_T(t_0, \zeta, C)$ is the observable minimax risk of Definition 14. Since

$$\frac{\eta_T^2 \exp(-1/4)}{4} = \frac{\exp(-1/4)}{64T},$$

the desired minimax floor follows. Together with the membership and policy identities from Step 1 and the pointwise inequality from Step 3, this proves all four asserted conclusions. □

Lemma 40 complements the signed-depth lower construction by giving a radius-uniform benchmark at the parametric scale. The experiments share behavior and target policies, so the displayed floor reflects value uncertainty in the observed rewards themselves and applies throughout the latent-overlap model class.

Verification note This note records the machine-verification scope of the paper.

What is checked. Every displayed assumption, definition, lemma, proposition, and theorem of the main text and of this appendix is either mapped to a declaration of an accompanying Lean 4 development, listed under *Statement map* below, or flagged under *Presentation-level definitions* as notation introduced for the prose alone. Each displayed proof renders the corresponding Lean proof. The Lean development is the verified artifact; the displayed statements and proofs are human-readable transcriptions of it, and the constants, exponents, and inequalities they display are the ones that appear in the machine-checked declarations.

Toolchain and artifact version. The development compiles under the Lean 4 toolchain `leanprover/lean4:v4.33.0` against `mathlib4` at revision `db584cd6d46c92f209a44c0f1c829460d327499d`, in the module tree `CausalSmith/Stat/STAT_PomdpLatentOverlapMinimax_Research`, from the artifact source tree at commit `f2e7e979167a2f3f9242e93c680c7ba1667b6dcc`, which is the commit recorded in the verification contract accompanying this paper.

What the receipts certify. The receipts behind this note are declaration-level: for each declaration listed below, the recorded check is that it compiles under the stated toolchain and that its source carries no `sorry` and no `admit`. That is the exact scope of the trust claim made here. Readers who additionally want the kernel-level axiom footprint of a particular declaration can obtain it directly by running `#print axioms` on that declaration in the accompanying development.

External inputs. No result in this paper is stated conditionally on a published theorem: the cited literature supplies framing, motivation, and comparison only, and no bibliographic item enters any displayed statement as a hypothesis.

Declarations from outside the module tree. Lemma 5 is carried by `Causalean.Mathlib.Probability.CertifiedFiniteMarkovExpectation.stationaryRewardInterval_sound_of_chunked`, reached through `Helpers/InsulinGrid.lean`. It belongs to the same verified library and compiles under the same toolchain.

Statement map. Each anchored object and the declaration it is mapped to, with the file of the module tree that contains it:

- Definition 1: `policyFactor` in `Basic.lean`.
- Definition 2: `FullTrajectory` in `Basic.lean`.
- Assumption 1: `PomdpKernelLaw` in `Basic.lean`.
- Definition 3: `Policy` in `Basic.lean`.

- Assumption 2: `SequentialIgnorability` in `Basic.lean`.
- Definition 4: `policyKernel` in `Basic.lean`.
- Definition 5: `JointState` in `Basic.lean`.
- Assumption 3: `StationaryStart` in `Basic.lean`.
- Assumption 4: `BoundedReward` in `Basic.lean`.
- Assumption 5: `PolicyOverlap` in `Basic.lean`.
- Definition 7: `mixingAlpha` in `Basic.lean`.
- Assumption 6: `UniformContraction` in `Basic.lean`.
- Assumption 7: `LatentStationaryOverlap` in `Basic.lean`.
- Definition 8: `ObsView` in `Basic.lean`.
- Definition 10: `LatentOverlapClass` in `Basic.lean`.
- Definition 11: `rewardRegression` in `Basic.lean`.
- Definition 12: `targetValue` in `Basic.lean`.
- Definition 14: `minimaxRisk` in `Basic.lean`.
- Definition 15: `ratio` in `Basic.lean`.
- Definition 16: `phiwEstimator` in `Basic.lean`.
- Definition 17: `historyDepth` in `Basic.lean`.
- Definition 18: `signedDepthFamily` in `Helpers/SignedDepth.lean`.
- Definition 19: `overlapRadius` in `Basic.lean`.
- Definition 20: `radiusAdaptiveDepth` in `Basic.lean`.
- Theorem 1: `uniform_overlap_frontier` in `TUniformOverlapFrontier.lean`.
- Theorem 2: `fixed_c_minimax` in `TFixedCMinimax.lean`.
- Theorem 3: `unit_overlap_boundary` in `TUnitOverlapBoundary.lean`.
- Theorem 4: `shrinking_overlap_frontier` in `TShrinkingOverlapFrontier.lean`.
- Definition 24: `insulinGrid` in `Helpers/InsulinGridCore/Model.lean`.
- Lemma 1: `insulinPolicy_contraction` in `Helpers/InsulinGridCore/Model.lean`.
- Lemma 2: `insulinStationaryLaw_isStationary` in `Helpers/InsulinGridCore/Model.lean`.
- Lemma 3: `insulinTargetRewardRegression` in `Helpers/InsulinGridCore/Semantic.lean`.
- Lemma 4: `insulinImmediateRewardRegression` in `Helpers/InsulinGridCore/Semantic.lean`.
- Lemma 6: `insulinBiasInterval_bounds` in `Helpers/InsulinChunkAssembly.lean`.
- Lemma 7: `insulinGrid_bias_certificate` in `Helpers/InsulinGrid.lean`.
- Theorem 5: `finite_insulin_demonstration` in `TFiniteInsulinDemonstration.lean`.
- Definition 25: `filterQuantities` in `Helpers/ObservedFilter.lean`.
- Lemma 8: `integral_sq_phiwScore_le` in `Helpers/PhiwWindowMoments.lean`.
- Lemma 9: `abs_covariance_phiwScore_le_overlapDecay` in `Helpers/PhiwWindowMoments.lean`.

- Lemma 10: `integral_phiwScore_mul_eq_futureFunction` in `Helpers/PhiwFutureEndpoint.lean`.
- Lemma 11: `markovOperatorIter_congr_on_behaviorSupport` in `Helpers/StationaryRewardSupport.le`
- Lemma 12: `abs_covariance_phiwScore_le_disjoint_sharp` in `Helpers/PhiwFutureEndpoint.lean`.
- Lemma 13: `sum_abs_covariance_future_le` in `Helpers/PhiwVariance.lean`.
- Lemma 14: `variance_phiwRaw_le` in `Helpers/PhiwVariance.lean`.
- Lemma 15: `phiw_upper` in `Helpers/PhiwUpper.lean`.
- Lemma 16: `signedDepth_uniformContraction_actionOne` in `Helpers/SignedDepthStationarity.lean`.
- Lemma 17: `signedDepth_stationaryLaw_apply` in `Helpers/SignedDepthStationarity.lean`.
- Lemma 18: `signedDepth_membership` in `Helpers/SignedDepthMembership.lean`.
- Lemma 19: `pow_one_sub_le_signedDepthPosteriorLowerWindow` in `Helpers/TerminalPosteriorBound.`
- Lemma 20: `signedDepthRawPosterior_le_product_of_le` in `Helpers/TerminalPosteriorBound.lean`.
- Lemma 21: `signedDepthRawPosterior_le_product_of_lt` in `Helpers/TerminalPosteriorBound.lean`.
- Lemma 22: `signedDepthRawPosterior_le_ratio_pow` in `Helpers/TerminalPosteriorBound.lean`.
- Lemma 23: `signedDepthRawPosterior_le_refined` in `Helpers/TerminalPosteriorBound.lean`.
- Lemma 24: `rareAction_secondMoment_eq` in `Helpers/SignedDepthObservedChain.lean`.
- Lemma 25: `signedDepth_finiteObs_klDiv_le` in `Helpers/ObservedKL.lean`.
- Lemma 26: `exists_uniform_signedDepth_bootstrap_factor` in `Helpers/TerminalPosteriorBound.lea`
- Lemma 27: `signedDepth_terminalMass_invariant` in `Helpers/SignedDepthFilterInvariant.lean`.
- Lemma 28: `signedDepth_prefix_nextRewardNumerator` in `Helpers/SignedDepthFilterRecursion.lean`
- Lemma 29: `terminalPrefixMass_eq_sum_signedDepthFilterMass` in `Helpers/ObservedFilter.lean`.
- Lemma 30: `conditionalRewardMean_signedDepth_all` in `Helpers/ObservedFilter.lean`.
- Definition 28: `signedDepthObsLaw` in `Helpers/ObservedKL.lean`.
- Lemma 31: `observed_path_kl` in `Helpers/ObservedKL.lean`.
- Lemma 32: `integral_phiwScore_eq_init_targetIter` in `Helpers/BiasVariance.lean`.
- Lemma 33: `targetValue_mem_unit` in `Helpers/Kernels.lean`.
- Lemma 34: `radius_sensitive_phiw` in `Helpers/BiasVariance.lean`.
- Lemma 35: `radius_explicit_observed_kl` in `Helpers/ObservedKL.lean`.
- Lemma 36: `parametricPair_targetValue` in `Helpers/TwoPoint.lean`.
- Lemma 37: `parametricPair_observed_klDiv_le` in `Helpers/TwoPoint.lean`.
- Lemma 38: `observedRisk_le_four` in `Helpers/Kernels.lean`.
- Lemma 39: `two_point_floor_explicit` in `Helpers/TwoPoint.lean`.
- Lemma 40: `uniform_parametric_floor` in `Helpers/TwoPoint.lean`.

Presentation-level definitions. The following definitions correspond to no single Lean declaration. They name objects the prose needs in order to display the argument, and which the Lean development carries inline rather than through a named definition; no claim is attached to them beyond the notation they fix: Definition 6, Definition 9, Definition 13, Definition 21, Definition 22, Definition 23, Definition 26, Definition 27.

B Proofs of the main results

Proof of Theorem 1. Write

$$\alpha = \exp(-1/t_0), \qquad L = \exp(\zeta), \qquad \beta = \frac{2}{2 + t_0\zeta}.$$

Then $0 < \alpha < 1$, $1 < L$, and $0 < \beta < 1$.

Step 1. Adaptive-depth calibration. Set

$$\mathcal{D} = 2\log(1/\alpha) + \log L = \frac{2}{t_0} + \zeta, \qquad q = q_C, \qquad \xi = Tq^2.$$

There are $c_{\mathrm{rad}} > 0$ and $T_{\mathrm{rad}} \geq 1$, depending only on t_0, ζ, such that, for all $T \geq T_{\mathrm{rad}}$ and all $C \geq 1$,

$$q_C^2 \alpha^{2\kappa_T^{\mathrm{rad}}(C)} + \frac{L^{\kappa_T^{\mathrm{rad}}(C)}}{T} \leq c_{\mathrm{rad}} \left\{ T^{-1} + T^{-\beta} q_C^{2(1-\beta)} \right\}.$$

Indeed, since $\log T = o(T)$, enlarge T_{rad} so that

$$|\log T| \leq \frac{\mathcal{D}}{2} T \qquad (T \geq T_{\mathrm{rad}}).$$

If $\xi \leq 1$, then $\kappa_T^{\mathrm{rad}}(C) = 0$ by Definition 20, and $q_C^2 \leq T^{-1}$, hence

$$q_C^2 + \frac{1}{T} \leq 2T^{-1} \leq 2 \left\{ T^{-1} + T^{-\beta} q_C^{2(1-\beta)} \right\}.$$

If $\xi > 1$, then $q_C > 0$, $0 < \xi \leq T$, and the same logarithmic bound gives

$$\left\lfloor \frac{\log \xi}{\mathcal{D}} \right\rfloor \leq \frac{T}{2}.$$

Thus $\kappa_T^{\mathrm{rad}}(C) = \lfloor \log \xi / \mathcal{D} \rfloor$. The floor bounds

$$\frac{\log \xi}{\mathcal{D}} - 1 < \kappa_T^{\mathrm{rad}}(C) \leq \frac{\log \xi}{\mathcal{D}}$$

give

$$\alpha^{2\kappa_T^{\mathrm{rad}}(C)} \leq e^{2/t_0} \xi^{-\beta}, \qquad L^{\kappa_T^{\mathrm{rad}}(C)} \leq \xi^{1-\beta}.$$

Consequently,

$$q_C^2 \alpha^{2\kappa_T^{\mathrm{rad}}(C)} + \frac{L^{\kappa_T^{\mathrm{rad}}(C)}}{T} \leq \{e^{2/t_0} + 1\} T^{-\beta} q_C^{2(1-\beta)}.$$

Taking $c_{\mathrm{rad}} = \max\{2, e^{2/t_0} + 1\}$ proves the calibration.

Step 2. Upper bound. Let

$$A = \max\left\{4,\, 2L\left(1 + \frac{4}{L-1}\right),\, \frac{8}{1-\alpha}\right\}, \qquad c_{\mathrm{att}} = A(c_{\mathrm{rad}} + 1).$$

For $T \geq T_{\mathrm{rad}}$, put $k = \kappa_T^{\mathrm{rad}}(C)$. By Definition 20, $k \leq T/2$, and Lemma 34 gives, uniformly over $M \in \mathcal{M}_T(t_0, \zeta, C)$,

$$\mathbb{E}_M[(\widehat{\theta}_k - \theta(K, e))^2] \leq 4q_C^2 \alpha^{2k} + \frac{2}{T}\left\{ L^{k+1}\left(1 + \frac{4}{L-1}\right) + \frac{4}{1-\alpha} \right\}.$$

By the definition of A,

$$4q_C^2\alpha^{2k} + \frac{2}{T}\left\{L^{k+1}\left(1+\frac{4}{L-1}\right)+\frac{4}{1-\alpha}\right\} \le A\left\{q_C^2\alpha^{2k}+\frac{L^k}{T}+\frac{1}{T}\right\}.$$

The calibration in Step 1 and the inequality

$$T^{-1} \le T^{-1} + T^{-\beta}q_C^{2(1-\beta)}$$

therefore imply

$$\sup_{M\in\mathcal{M}_T(t_0,\zeta,C)} \mathbb{E}_M[(\widehat{\theta}_{\kappa_T^{\mathrm{rad}}(C)} - \theta(K,e))^2] \le c_{\mathrm{att}}\mathfrak{r}_T(t_0,\zeta,C).$$

Since $R_T(t_0,\zeta,C)$ is the infimum over observable-data estimators in Definition 14, this also yields

$$R_T(t_0,\zeta,C) \le c_{\mathrm{att}}\mathfrak{r}_T(t_0,\zeta,C).$$

Step 3. Constants for the lower comparison. From Lemma 35, choose $K_0 = K_0(t_0) > 0$. With $c_0 = (1-\alpha)/4$, define

$$\mathsf{B} = \frac{4c_0^2K_0^2}{1-c_0^2} > 0, \qquad x_\star = \max\left\{1, \frac{1}{4\mathsf{B}}\right\}.$$

Then $1 < 8\mathsf{B}x_\star$. Put

$$c_{\mathrm{par}} = \frac{e^{-1/4}}{64}, \qquad c_{\mathrm{hid}} = \left\{\frac{c_0(1-\alpha)\alpha}{2}\right\}^2 (8\mathsf{B})^{-\beta}\frac{e^{-1/8}}{4},$$

$$s_\star = 1 + x_\star^{1-\beta}, \qquad c_{\mathrm{lo}} = \min\left\{\frac{c_{\mathrm{par}}}{2s_\star}, \frac{c_{\mathrm{hid}}}{2}\right\}.$$

All these constants are positive and depend only on t_0, ζ.

Step 4. Lower bound in the small-radius branch. By Lemma 40,

$$\frac{c_{\mathrm{par}}}{T} \le R_T(t_0,\zeta,C) \qquad (T\ge 1,\ C\ge 1).$$

If $\xi = Tq_C^2 < x_\star$, then

$$T^{-\beta}q_C^{2(1-\beta)} = T^{-1}\xi^{1-\beta} \le T^{-1}x_\star^{1-\beta},$$

with both sides equal to 0 when $q_C = 0$. Hence

$$\mathfrak{r}_T(t_0,\zeta,C) \le \frac{s_\star}{T},$$

and therefore

$$c_{\mathrm{lo}}\mathfrak{r}_T(t_0,\zeta,C) \le \frac{c_{\mathrm{par}}}{T} \le R_T(t_0,\zeta,C).$$

Step 5. Lower bound in the large-radius branch. Assume $\xi \ge x_\star$. Then $C > 1$. Define

$$Q = \left\lceil \frac{\log(8\mathsf{B}\xi)}{\mathcal{D}} \right\rceil.$$

Because $8\mathsf{B}\xi \geq 8\mathsf{B}x_\star > 1$, we have $Q \geq 1$, and

$$\mathcal{D}Q \geq \log(8\mathsf{B}\xi), \qquad Q < \frac{\log(8\mathsf{B}\xi)}{\mathcal{D}} + 1.$$

Let M_+ and M_- be the two signed-depth alternatives of depth Q from Definition 18, and write θ_+ and θ_- for their target values. By Lemma 18, both alternatives belong to $\mathcal{M}_T(t_0, \zeta, C)$. With w_Q denoting the terminal depth mass in that construction, the same result gives

$$\theta_+ = c_0 \varepsilon_C w_Q, \qquad \theta_- = -c_0 \varepsilon_C w_Q.$$

Define the separation radius

$$\sigma_Q := c_0 \varepsilon_C w_Q.$$

The target-value identities give the separation condition needed for the two-point bound,

$$2\sigma_Q \leq |\theta_+ - \theta_-|.$$

Also, by Lemma 35,

$$\frac{q_C}{2} \leq \varepsilon_C \leq q_C$$

and, orienting the two observed laws as $\mathbb{P}^{\mathrm{obs},T}_{Q,+}$ against $\mathbb{P}^{\mathrm{obs},T}_{Q,-}$,

$$\mathrm{KL}(\mathbb{P}^{\mathrm{obs},T}_{Q,+} \,\|\, \mathbb{P}^{\mathrm{obs},T}_{Q,-}) \leq \mathsf{B}\, T \varepsilon_C^2 \alpha^{2Q} L^{-Q}.$$

Since $\alpha^{2Q} L^{-Q} = e^{-\mathcal{D}Q}$, the choice of Q gives

$$\mathsf{B}\, T \varepsilon_C^2 \alpha^{2Q} L^{-Q} \leq \mathsf{B}\, T q_C^2 (8\mathsf{B}\, T q_C^2)^{-1} = \frac{1}{8}.$$

The two alternatives have the same observed alphabet and the same revealed policies by their construction; the displayed KL bound, in the same orientation, supplies the information and finiteness hypotheses with $\chi = 1/8$. Applying Lemma 39 therefore gives

$$R_T(t_0, \zeta, C) \geq \sigma_Q^2 \frac{e^{-1/8}}{4} = (c_0 \varepsilon_C w_Q)^2 \frac{e^{-1/8}}{4}.$$

The depth mass in Lemma 18 satisfies

$$w_Q = \frac{(1-\alpha)\alpha^Q}{1-\alpha^{Q+1}} \geq (1-\alpha)\alpha^Q,$$

and the ceiling upper bound implies

$$\alpha^Q \geq \alpha (8\mathsf{B}\xi)^{-1/(t_0 \mathcal{D})}.$$

Thus

$$(c_0 \varepsilon_C w_Q)^2 \frac{e^{-1/8}}{4} \geq c_{\mathrm{hid}}\, q_C^2 \xi^{-\beta} = c_{\mathrm{hid}}\, T^{-\beta} q_C^{2(1-\beta)}.$$

Combining this hidden lower bound with the parametric floor and using $c_{\mathrm{lo}} \leq c_{\mathrm{par}}/2$ and $c_{\mathrm{lo}} \leq c_{\mathrm{hid}}/2$, we get

$$c_{\mathrm{lo}} \mathfrak{r}_T(t_0, \zeta, C) \leq \frac{1}{2} R_T(t_0, \zeta, C) + \frac{1}{2} R_T(t_0, \zeta, C) = R_T(t_0, \zeta, C).$$

Step 6. Assemble the frontier. Set

$$c_{\mathrm{hi}} = \max\{c_{\mathrm{att}}, c_{\mathrm{lo}}\}, \qquad T_\star = T_{\mathrm{rad}}.$$

Here $c_{\mathrm{att}} = A(c_{\mathrm{rad}} + 1)$ uses the same calibration constant fixed in Step 1. Thus $0 < c_{\mathrm{lo}} \le c_{\mathrm{hi}}$ and $c_{\mathrm{att}} \le c_{\mathrm{hi}}$. Since $\mathfrak{r}_T(t_0, \zeta, C) \ge 0$, the two upper bounds from Step 2 with constant c_{att} also hold with constant c_{hi}. Together, Steps 2, 4, and 5 prove, for all $C \ge 1$ and $T \ge T_\star$,

$$c_{\mathrm{lo}}\mathfrak{r}_T(t_0, \zeta, C) \le R_T(t_0, \zeta, C) \le c_{\mathrm{hi}}\mathfrak{r}_T(t_0, \zeta, C),$$

together with

$$\sup_{M \in \mathcal{M}_T(t_0,\zeta,C)} \mathbb{E}_M[(\widehat{\theta}_{\kappa_T^{\mathrm{rad}}(C)} - \theta(K, e))^2] \le c_{\mathrm{hi}}\mathfrak{r}_T(t_0, \zeta, C).$$

Step 7. Immediate estimator in the $Tq_C^2 = O(1)$ regime. Let $C_T \ge 1$ and suppose there is $D_{\mathrm{imm}} \ge 0$ such that, eventually,

$$Tq_{C_T}^2 \le D_{\mathrm{imm}}.$$

Define

$$V = L\left(1 + \frac{4}{L-1}\right) + \frac{4}{1-\alpha}, \qquad c_{\mathrm{imm}} = 4D_{\mathrm{imm}} + 2V > 0.$$

For all sufficiently large T, apply Lemma 34 with $k = 0$. Since $q_{C_T}^2 \le D_{\mathrm{imm}}/T$,

$$\sup_{M \in \mathcal{M}_T(t_0,\zeta,C_T)} \mathbb{E}_M[(\widehat{\theta}_0 - \theta(K, e))^2] \le 4q_{C_T}^2 + \frac{2}{T}V \le \frac{4D_{\mathrm{imm}} + 2V}{T} = \frac{c_{\mathrm{imm}}}{T}.$$

This is the asserted eventual parametric bound for the immediate estimator. □

Proof of Theorem 2. 1. Set

$$\alpha := \exp(-1/t_0), \qquad L := \exp(\zeta), \qquad D := \frac{2}{t_0} + \zeta.$$

By $t_0 > 0$ and $\zeta > 0$, the constants satisfy

$$0 < \alpha < 1, \qquad L > 1, \qquad D > 0,$$

and, using the definitions in Definitions 1 and 7,

$$\log \alpha = -\frac{1}{t_0}, \qquad \log L = \zeta.$$

The exponent in the theorem can be written as

$$\beta = \frac{2}{2 + t_0\zeta} = \frac{2/t_0}{D}.$$

Apply Lemma 15 with $C \ge 1$. It gives constants $c_{\mathrm{up},0} > 0$ and T_{up} such that, for all $T \ge T_{\mathrm{up}}$,

$$\sup_{M \in \mathcal{M}_T(t_0,\zeta,C)} \mathbb{E}_M\left[\{\widehat{\theta}_{k_T} - \theta(M)\}^2\right] \le c_{\mathrm{up},0}T^{-\beta}.$$

For the signed-depth family, the behavior policy in Definition 18 is a memoryless function of the single observed state, independent of the latent coordinate and the past. Consequently,

for every horizon T, terminal depth $Q \geq 1$, and sign $v \in \{-1, +1\}$, the action kernel satisfies Assumption 2. The same alternatives start from the behavior stationary law for every horizon, including the zero-horizon case. Indeed, for

$$r := L^{-1},$$

the signed-depth constant-policy stationary law at action-one probability r has masses

$$w_j \frac{1 + u\varepsilon_C r^j}{2}, \qquad w_j = \frac{(1-\alpha)\alpha^j}{1 - \alpha^{Q+1}},$$

and the behavior policy in Definition 18 chooses action one with probability $L^{-1} = r$. Hence its stationary law is precisely the initial hidden-state law

$$w_j \frac{1 + u\varepsilon_C L^{-j}}{2}.$$

The chronological path law is generated from this initial distribution, so the time-zero joint-state marginal is this stationary law; when $T = 0$, the same identity is the whole marginal check, and for $T \geq 1$ it agrees with the behavior-stationary masses displayed in Lemma 18. Thus Assumption 3 holds for every horizon. These are the two structural hypotheses needed for Lemma 31. Therefore that lemma gives a constant $B_0 > 0$, depending on t_0, ζ, C, such that for all T and $Q \geq 1$,

$$\mathrm{KL}\left(\mathbb{P}^{\mathrm{obs},T}_{Q,+1} \,\middle\|\, \mathbb{P}^{\mathrm{obs},T}_{Q,-1}\right) \leq B_0 T \alpha^{2Q} L^{-Q}.$$

Let

$$\varepsilon_C := \min\left\{\frac{C-1}{2}, \frac{1}{2}\right\}, \qquad c_0 := \frac{1-\alpha}{4},$$

and define

$$c_\star := \{c_0 \varepsilon_C (1-\alpha)\alpha\}^2 (8B_0)^{-\beta} \frac{\exp(-1/8)}{4}.$$

Since $C > 1$, every factor in this product is positive, so $c_\star > 0$. Finally set

$$c^\star := \max\{c_{\mathrm{up},0}, c_\star\}.$$

Since $c_{\mathrm{up},0} > 0$ and $c_\star > 0$, the maximum is positive; moreover the second argument is bounded by the maximum. Thus

$$0 < c^\star, \qquad 0 < c_\star \leq c^\star.$$

2. Define

$$T_{\mathrm{pos}} := \left\lceil \frac{1}{8B_0} \right\rceil + 1, \qquad T_\star := \max\{T_{\mathrm{up}}, T_{\mathrm{pos}}, 1\}.$$

Fix an integer $T \geq T_\star$. Then $T \geq 1$, $T \geq T_{\mathrm{up}}$, and

$$1 < 8B_0 T.$$

Indeed,

$$\frac{1}{8B_0} \leq \left\lceil \frac{1}{8B_0} \right\rceil < T,$$

and multiplication by $8B_0 > 0$ gives the displayed inequality. Put

$$\tau_T := \frac{\log(8B_0T)}{D}, \qquad Q := \lceil \tau_T \rceil.$$

Since $\log(8B_0T) > 0$ and $D > 0$, one has $Q \geq 1$. Moreover,

$$\tau_T \leq Q < \tau_T + 1$$

and therefore

$$DQ \geq \log(8B_0T).$$

3. The preceding depth choice makes the two signed-depth alternatives statistically close. First,

$$\alpha^{2Q} L^{-Q} = \exp\{Q(2\log\alpha - \log L)\} = \exp(-DQ),$$

because $D = 2/t_0 + \zeta$, $\log\alpha = -1/t_0$, and $\log L = \zeta$. Hence

$$\exp(-DQ) \leq \exp\{-\log(8B_0T)\} = (8B_0T)^{-1}.$$

Combining this inequality with the KL bound from Lemma 31 gives

$$\mathrm{KL}\Big(\mathbb{P}^{\mathrm{obs},T}_{Q,+1} \,\Big\|\, \mathbb{P}^{\mathrm{obs},T}_{Q,-1}\Big) \leq \frac{1}{8}.$$

4. Let $M^T_{Q,+1}$ and $M^T_{Q,-1}$ be the two signed-depth experiments of Definition 18. They have one observed state and $2(Q+1)$ latent states. Their revealed behavior and target policies are common, because the policy part of the construction is independent of the sign v. By Lemma 18, both experiments belong to $\mathcal{M}_T(t_0, \zeta, C)$, and their target values are

$$\theta(M^T_{Q,+1}) = c_0\varepsilon_C w_Q, \qquad \theta(M^T_{Q,-1}) = -c_0\varepsilon_C w_Q,$$

where

$$w_Q := \frac{(1-\alpha)\alpha^Q}{1-\alpha^{Q+1}}.$$

Set

$$s_T := c_0\varepsilon_C w_Q.$$

Then $s_T \geq 0$, and the two target values satisfy the separation bound

$$2s_T \leq \big|\theta(M^T_{Q,+1}) - \theta(M^T_{Q,-1})\big|.$$

The two embedded observed laws have the same finite real-reward support,

$$\Omega^\circ_T := \big\{((0, a_t, y_t))_{t=1}^T : a_t \in \{0,1\},\ y_t \in \{-1,+1\}\big\}.$$

They assign total mass one to Ω°_T, because the construction has one observed state, Boolean actions, and rewards in $\{-1,+1\}$. We now check that every atom in this common support has positive mass under both signs. Fix $o = ((0, a_t, y_t))_{t=1}^T \in \Omega^\circ_T$, and choose a fixed latent sign $u_0 \in \{-1,+1\}$. Consider the full hidden lift that keeps the latent state at depth zero and sign u_0:

$$S_t = (0, (0, u_0)), \qquad A_t = a_t, \qquad Y_t = y_t, \qquad t = 1, ldots, T.$$

Since $0 < \alpha < 1$ and $Q \geq 1$,

$$1 - \alpha^{Q+1} > 0, \qquad w_0 = \frac{1-\alpha}{1-\alpha^{Q+1}} > 0.$$

Also $0 < \varepsilon_C \leq 1/2$, so the initial mass of the chosen hidden state is positive:

$$w_0 \frac{1 + u_0 \varepsilon_C}{2} > 0.$$

For each action in the observed word,

$$b(1 \mid 0) = \frac{1}{L} > 0, \qquad b(0 \mid 0) = 1 - \frac{1}{L} > 0,$$

because $L > 1$. Along the displayed hidden lift, the current depth is $0 < Q$. Hence the reward factor is

$$\frac{1 + y_t v c_0 u_0 \mathbf{1}\{0 = Q\}}{2} = \frac{1}{2} > 0,$$

and the hidden transition that resets to the same depth-zero sign has weight

$$(1-\alpha)\frac{1 + u_0 \varepsilon_C}{2} > 0.$$

Thus this single compatible hidden lift has strictly positive finite-path probability for each sign $v \in \{-1, +1\}$, and summing over all lifts gives

$$\mathbb{P}^{\mathrm{obs},T}_{Q,v}(\{o\}) > 0, \qquad o \in \Omega^{\circ}_T, \qquad v \in \{-1, +1\}.$$

The reward decoding used in the embedding is the same for the two signs, so the embedded laws preserve this common atomic support. Therefore any measurable set with zero mass under one embedded observed law contains no atom of Ω°_T, and consequently has zero mass under the other embedded observed law. Hence

$$\mathbb{P}^{\mathrm{obs},T}_{Q,+1} \ll \mathbb{P}^{\mathrm{obs},T}_{Q,-1}, \qquad \mathbb{P}^{\mathrm{obs},T}_{Q,-1} \ll \mathbb{P}^{\mathrm{obs},T}_{Q,+1},$$

and the finite KL displayed above is an ordinary KL bound between mutually absolutely continuous observed laws.

5. Apply Lemma 39 to the present pair with $s = s_T$, taking the $+1$ alternative as the lemma's first model and the -1 alternative as its second: $M_0 = M^T_{Q,+1}$ and $M_1 = M^T_{Q,-1}$. With this assignment the lemma's divergence hypothesis $\mathrm{KL}(\mathbb{P}_0 \| \mathbb{P}_1) \leq \chi$ reads

$$\mathrm{KL}\Big(\mathbb{P}^{\mathrm{obs},T}_{Q,+1} \,\Big\|\, \mathbb{P}^{\mathrm{obs},T}_{Q,-1}\Big) \leq \frac{1}{8},$$

which is exactly the direction bounded above, and its absolute-continuity hypothesis reads $\mathbb{P}^{\mathrm{obs},T}_{Q,+1} \ll \mathbb{P}^{\mathrm{obs},T}_{Q,-1}$, which the preceding step supplies. The common observable structure was checked in the preceding step, the KL bound gives the information hypothesis with $\chi = 1/8$, and the displayed target-value calculation gives

$$2s_T \leq \big|\theta(M^T_{Q,+1}) - \theta(M^T_{Q,-1})\big|.$$

Therefore every measurable observable estimator $\widehat{\theta}$, under the stated squared-error integrability conditions for both alternatives, satisfies

$$s_T^2 \frac{\exp(-1/8)}{4} \le \max\left\{ \mathbb{E}_{M_{Q,+1}^T}\left[\{\widehat{\theta} - \theta(M_{Q,+1}^T)\}^2\right], \mathbb{E}_{M_{Q,-1}^T}\left[\{\widehat{\theta} - \theta(M_{Q,-1}^T)\}^2\right] \right\}.$$

Commuting the two entries of the maximum, this is exactly

$$s_T^2 \frac{\exp(-1/8)}{4} \le \max_{v\in\{-1,+1\}} \mathbb{E}_{M_{Q,v}^T}\left[\{\widehat{\theta} - \theta(M_{Q,v}^T)\}^2\right].$$

The two signed-depth alternatives belong to $\mathcal{M}_T(t_0, \zeta, C)$ and share the revealed inputs (n_X, b, e). For the witness clause of the theorem the two squared-error integrability conditions are hypotheses, imposed on the estimator and discharged by whoever supplies it; nothing above derives them. They are derived only when $\widehat{\theta}$ is further restricted to the admissible carrier of Definition 14, whose members are measurable and uniformly $[-1, 1]$-valued by definition: for such an estimator the squared error at either alternative is a measurable function bounded by $(1 + |\theta_\pm|)^2 \le 4$ on a probability space, hence integrable, so the preceding pointwise lower bound applies to every admissible estimator. Taking the supremum over the model class, then the infimum over these admissible estimators, yields

$$s_T^2 \frac{\exp(-1/8)}{4} \le R_T(t_0, \zeta, C).$$

6. It remains to compare s_T with $T^{-\beta/2}$. Since $0 < \alpha < 1$ and $Q \ge 1$, also $Q + 1 \ge 1$. Thus

$$0 < \alpha^{Q+1} < 1,$$

so the denominator in the depth mass is strictly positive and satisfies

$$0 < 1 - \alpha^{Q+1} \le 1.$$

The numerator $(1 - \alpha)\alpha^Q$ is nonnegative, and division by a positive number at most one gives

$$w_Q = \frac{(1-\alpha)\alpha^Q}{1-\alpha^{Q+1}} \ge (1-\alpha)\alpha^Q.$$

The upper ceiling inequality $Q < \tau_T + 1$ and the identity $\log \alpha = -1/t_0$ give

$$\alpha^Q = \exp\left(-\frac{Q}{t_0}\right) \ge \exp\left(-\frac{\tau_T + 1}{t_0}\right) = \alpha(8B_0T)^{-(1/t_0)/D}.$$

Therefore

$$s_T \ge c_0\varepsilon_C(1-\alpha)\alpha(8B_0T)^{-(1/t_0)/D}.$$

Squaring and using $\beta = (2/t_0)/D$ yields

$$s_T^2 \ge \{c_0\varepsilon_C(1-\alpha)\alpha\}^2(8B_0T)^{-\beta} = \{c_0\varepsilon_C(1-\alpha)\alpha\}^2(8B_0)^{-\beta}T^{-\beta}.$$

Multiplying by $\exp(-1/8)/4$ gives

$$c_\star T^{-\beta} \le s_T^2 \frac{\exp(-1/8)}{4}.$$

Combining this with the two-point floor proves

$$c_\star T^{-\beta} \le R_T(t_0, \zeta, C),$$

and, for every measurable observable-data estimator satisfying the two integrability conditions,

$$c_\star T^{-\beta} \le \max_{v \in \{-1,+1\}} \mathbb{E}_{M^T_{Q,v}} \left[\{\widehat{\theta} - \theta(M^T_{Q,v})\}^2 \right].$$

7. The upper bound follows from the estimator bound in the first step. Since $T \ge T_{\mathrm{up}}$,

$$\sup_{M \in \mathcal{M}_T(t_0,\zeta,C)} \mathbb{E}_M \left[\{\widehat{\theta}_{k_T} - \theta(M)\}^2 \right] \le c_{\mathrm{up},0} T^{-\beta} \le c^\star T^{-\beta}.$$

By Definition 14, the minimax risk is bounded above by the worst-case risk of any admissible observable estimator, in particular by that of $\widehat{\theta}_{k_T}$. Thus

$$R_T(t_0, \zeta, C) \le c^\star T^{-\beta}.$$

Together with the lower bound from the preceding step, this proves all asserted inequalities for every $T \ge T_\star$, and the chosen $Q = \lceil \log(8B_0T)/D \rceil$ supplies the stated signed-depth witnesses with one observed state and $2(Q+1)$ latent states.

□

Proof of Theorem 3. Write

$$\alpha := \exp(-1/t_0), \qquad L := \exp(\zeta).$$

Since $t_0 > 0$ and $\zeta > 0$, one has $0 < \alpha < 1$ and $1 < L$.

1. Define

$$c_{\mathrm{imm}} := 2 \left\{ L \left(1 + \frac{4}{L-1} \right) + \frac{4}{1-\alpha} \right\}.$$

The preceding inequalities imply $c_{\mathrm{imm}} > 0$. For any $T \ge 1$, Lemma 34 applied at depth $k = 0$ and overlap constant $C = 1$ gives, uniformly over $M \in \mathcal{M}_T(t_0, \zeta, 1)$,

$$\mathbb{E}_M \left[\left(\widehat{\theta}_0 - \theta(K, e) \right)^2 \right] \le \frac{2}{T} \left\{ L \left(1 + \frac{4}{L-1} \right) + \frac{4}{1-\alpha} \right\} = \frac{c_{\mathrm{imm}}}{T}.$$

Taking the supremum over the model class yields

$$\sup_{M \in \mathcal{M}_T(t_0,\zeta,1)} \mathbb{E}_M \left[\left(\widehat{\theta}_0 - \theta(K, e) \right)^2 \right] \le \frac{c_{\mathrm{imm}}}{T}.$$

2. Let $M \in \mathcal{M}_T(t_0, \zeta, 1)$. Write d_b and d_e for the behavior and target stationary laws on the finite joint state space $\mathcal{S}$, and define

$$\|d_b - d_e\|_{\mathrm{TV}} := \frac{1}{2} \sum_{s \in \mathcal{S}} |d_b(s) - d_e(s)|.$$

The stationary-law total-variation bound at overlap radius q_C gives

$$\|d_b - d_e\|_{\mathrm{TV}} \le q_1,$$

and at unit overlap Definition 19 gives

$$q_1 = \frac{1-1}{1} = 0.$$

Thus

$$\|d_b - d_e\|_{\mathrm{TV}} \le 0.$$

Since the displayed total-variation quantity is a half-sum of absolute values, it is nonnegative, so

$$\|d_b - d_e\|_{\mathrm{TV}} = 0.$$

Unfolding the definition of total variation yields

$$\sum_{s\in\mathcal{S}} |d_b(s) - d_e(s)| = 0.$$

For each joint state s, the nonnegative summand is bounded by the whole finite sum,

$$0 \le |d_b(s) - d_e(s)| \le \sum_{u\in\mathcal{S}} |d_b(u) - d_e(u)| = 0,$$

and hence

$$|d_b(s) - d_e(s)| = 0.$$

Therefore $d_b(s) = d_e(s)$ for every $s \in \mathcal{S}$, and consequently $d_e = d_b$.

3. Set

$$c_{\mathrm{lo}} := \frac{\exp(-1/4)}{64}, \qquad c_{\mathrm{hi}} := \max\{c_{\mathrm{imm}}, c_{\mathrm{lo}}\}, \qquad T_\star := 1.$$

Then $c_{\mathrm{lo}} > 0$, $c_{\mathrm{lo}} \le c_{\mathrm{hi}}$, and therefore $c_{\mathrm{hi}} > 0$; also $T_\star \ge 1$. For every $T \ge T_\star$, Lemma 40 at $C = 1$ gives

$$\frac{\exp(-1/4)}{64T} \le R_T(t_0, \zeta, 1),$$

which is

$$\frac{c_{\mathrm{lo}}}{T} \le R_T(t_0, \zeta, 1).$$

For the upper bound, the minimax risk is bounded above by the worst-case risk of any admissible estimator with nonnegative squared risk; applying this to $\widehat{\theta}_0$ and using Step 1 gives

$$R_T(t_0, \zeta, 1) \le \sup_{M\in\mathcal{M}_T(t_0,\zeta,1)} \mathbb{E}_M\left[\left(\widehat{\theta}_0 - \theta(K, e)\right)^2\right] \le \frac{c_{\mathrm{imm}}}{T} \le \frac{c_{\mathrm{hi}}}{T}.$$

4. Finally,

$$t_0\zeta > 0, \qquad \text{hence} \qquad 2 + t_0\zeta > 2.$$

Therefore

$$\frac{2}{2 + t_0\zeta} < 1,$$

which is the asserted bound on the hidden-state exponent. □

Proof of Theorem 4. Put

$$\alpha = \exp(-1/t_0), \qquad L = \exp(\zeta), \qquad \beta = \frac{2}{2 + t_0\zeta}, \qquad \gamma = 2(1-\beta).$$

Since $t_0 > 0$ and $\zeta > 0$, one has $0 < \alpha < 1$, $1 < L$, $0 < \beta < 1$, and hence $\gamma > 0$.

1. A radius-depth calibration. Let

$$\Lambda := 2\log(1/\alpha) + \log L = \frac{2}{t_0} + \zeta.$$

Then $\Lambda > 0$,

$$\beta = \frac{2/t_0}{\Lambda}, \qquad 1 - \beta = \frac{\zeta}{\Lambda}.$$

There are $c_R > 0$ and $T_R \geq 1$, depending only on t_0, ζ, such that for every $T \geq T_R$ and every $C \geq 1$,

$$q_C^2 \alpha^{2\kappa_T^{\mathrm{rad}}(C)} + \frac{L^{\kappa_T^{\mathrm{rad}}(C)}}{T} \leq c_R \left\{ T^{-1} + T^{-\beta} q_C^{2(1-\beta)} \right\}. \tag{5}$$

Indeed, choose T_R so that $|\log T| \leq (\Lambda/2)T$ for all $T \geq T_R$. Fix such a T, put $q = q_C$, and set $x = Tq^2$. The relation $C \geq 1$ and Definition 19 give $0 \leq q \leq 1$, hence $0 \leq x \leq T$. If $x \leq 1$, then $\kappa_T^{\mathrm{rad}}(C) = 0$ by Definition 20, and $q^2 \leq T^{-1}$, so

$$q^2 + \frac{1}{T} \leq 2T^{-1} \leq 2\left\{ T^{-1} + T^{-\beta} q^{2(1-\beta)} \right\}.$$

If $x > 1$, then $q > 0$. With

$$k = \left\lfloor \frac{\log x}{\Lambda} \right\rfloor,$$

the choice of T_R gives $k \leq T/2$, so the cap in Definition 20 is inactive and $\kappa_T^{\mathrm{rad}}(C) = k$. The floor bounds

$$\frac{\log x}{\Lambda} - 1 < k \leq \frac{\log x}{\Lambda}$$

imply

$$\alpha^{2k} = \exp\left(-\frac{2k}{t_0}\right) \leq \exp(2/t_0) x^{-\beta}, \qquad \frac{L^k}{T} \leq \frac{x^{1-\beta}}{T}.$$

Since

$$q^2 x^{-\beta} = T^{-\beta} q^{2(1-\beta)}, \qquad \frac{x^{1-\beta}}{T} = T^{-\beta} q^{2(1-\beta)},$$

the $x > 1$ case is bounded by

$$\big(\exp(2/t_0) + 1\big) T^{-\beta} q^{2(1-\beta)}.$$

Taking

$$c_R = \max\{2, \exp(2/t_0) + 1\}$$

proves Equation (5).

2. Constants. Apply Theorem 1 to obtain constants $c_{\mathrm{lo}}, c_{\mathrm{hi}} > 0$, with $c_{\mathrm{lo}} \leq c_{\mathrm{hi}}$, and a threshold $T_U \geq 1$, all depending only on t_0, ζ, such that its two-sided risk comparison and radius-calibrated estimator bound hold for all $C \geq 1$ and $T \geq T_U$. Define

$$\eta_{\mathrm{rad}} := 2^{-\gamma},$$

and

$$\mathsf{A}_{\mathrm{bd}} := \max\left\{4,\, 2L\left(1+\frac{4}{L-1}\right),\, \frac{8}{1-\alpha}\right\}.$$

Finally set

$$c_{\mathrm{loc}} := c_{\mathrm{lo}}\eta_{\mathrm{rad}}, \qquad c_{\mathrm{att}} := \mathsf{A}_{\mathrm{bd}}(c_R+1), \qquad C_{\mathrm{loc}} := \max\{c_{\mathrm{hi}}, c_{\mathrm{att}}\}.$$

Then $0 < c_{\mathrm{loc}} \le C_{\mathrm{loc}}$, because $0 < \eta_{\mathrm{rad}} \le 1$, $c_{\mathrm{lo}} \le c_{\mathrm{hi}}$, and $\mathsf{A}_{\mathrm{bd}} > 0$.

3. Comparing the two rate surfaces. Fix a deterministic sequence $(\delta_T)_{T\ge 1}$ with $\delta_T \ge 0$ and $\delta_T \to 0$. Choose T_Δ such that

$$0 \le \delta_T < 1 \qquad \text{for all } T \ge T_\Delta.$$

Let $T_\star = \max\{T_U, T_R, T_\Delta\}$. For $T \ge T_\star$, put $C_T = 1+\delta_T$. Then $C_T \ge 1$, and

$$q_{C_T} = \frac{C_T - 1}{C_T} = \frac{\delta_T}{1+\delta_T}.$$

Since $0 \le \delta_T < 1$,

$$0 \le q_{C_T} \le \delta_T, \qquad \frac{\delta_T}{2} \le q_{C_T}.$$

Raising these inequalities to the positive power γ gives

$$\eta_{\mathrm{rad}}\delta_T^\gamma \le q_{C_T}^\gamma \le \delta_T^\gamma.$$

Thus the theorem's local rate, evaluated along the fixed sequence (δ_T), and the fixed-radius rate at C_T satisfy

$$\eta_{\mathrm{rad}}\left\{T^{-1} + T^{-\beta}\delta_T^\gamma\right\} \le \mathfrak{r}_T(t_0,\zeta,C_T) \le T^{-1} + T^{-\beta}\delta_T^\gamma. \tag{6}$$

Equivalently, since $r_T(\delta)$ in the statement denotes this fixed-sequence quantity,

$$\eta_{\mathrm{rad}}\, r_T(\delta) \le \mathfrak{r}_T(t_0,\zeta,C_T) \le r_T(\delta).$$

Combining Equation (6) with the two-sided comparison from Theorem 1 gives

$$c_{\mathrm{loc}}\, r_T(\delta) \le R_T(t_0,\zeta,1+\delta_T) \le C_{\mathrm{loc}}\, r_T(\delta). \tag{7}$$

4. Attainment by the displayed shrinking depth. For $T \ge T_\star$, define

$$C_T^\dagger := \frac{1}{1-\delta_T}.$$

Then $C_T^\dagger \ge 1$, and

$$q_{C_T^\dagger} = \frac{C_T^\dagger - 1}{C_T^\dagger} = \delta_T.$$

Therefore Definition 20 gives

$$\kappa_T^{\mathrm{rad}}(C_T^\dagger) = \kappa_T. \tag{8}$$

where κ_T is the depth displayed in the theorem statement. In particular $\kappa_T \le T/2$. Applying Equation (5) with $C = C_T^\dagger$ yields

$$\delta_T^2\alpha^{2\kappa_T} + \frac{L^{\kappa_T}}{T} \le c_R r_T(\delta). \tag{9}$$

Now fix $M \in \mathcal{M}_T(t_0, \zeta, 1+\delta_T)$. By Lemma 34, the estimator from Definition 16 satisfies

$$\mathbb{E}_M\Big[\big(\widehat{\theta}_{\kappa_T} - \theta(M)\big)^2\Big] \le 4q_{C_T}^2 \alpha^{2\kappa_T} + \frac{2}{T}\left\{L^{\kappa_T+1}\left(1 + \frac{4}{L-1}\right) + \frac{4}{1-\alpha}\right\}.$$

Using $q_{C_T} \le \delta_T$ and the definition of A_{bd},

$$\mathbb{E}_M\Big[\big(\widehat{\theta}_{\kappa_T} - \theta(M)\big)^2\Big] \le \mathsf{A}_{\mathrm{bd}}\left\{\delta_T^2 \alpha^{2\kappa_T} + \frac{L^{\kappa_T}}{T} + T^{-1}\right\}.$$

Together with Equation (9) and $T^{-1} \le r_T(\delta)$, this gives

$$\mathbb{E}_M\Big[\big(\widehat{\theta}_{\kappa_T} - \theta(M)\big)^2\Big] \le c_{\mathrm{att}}\, r_T(\delta) \le C_{\mathrm{loc}}\, r_T(\delta).$$

Taking the supremum over $M \in \mathcal{M}_T(t_0, \zeta, 1+\delta_T)$ proves the asserted estimator bound.

5. The bounded local regime. Assume that there is a constant $D_{\mathrm{bd}} \ge 0$ such that

$$T\delta_T^2 \le D_{\mathrm{bd}}$$

for all sufficiently large T. Since $q_{1+\delta_T} \le \delta_T$, the same tail of T's satisfies

$$Tq_{1+\delta_T}^2 \le D_{\mathrm{bd}}.$$

Applying the final assertion of Theorem 1 to the sequence $C_T = 1+\delta_T$ gives a constant $c_{\mathrm{imm}} > 0$ such that, for all sufficiently large T,

$$\sup_{M \in \mathcal{M}_T(t_0,\zeta,1+\delta_T)} \mathbb{E}_M\Big[\big(\widehat{\theta}_0 - \theta(M)\big)^2\Big] \le \frac{c_{\mathrm{imm}}}{T}.$$

This is the claimed immediate-depth conclusion. □

Proof of Lemma 1. Fix T, a policy p, and probability vectors ν, ν' on $\mathsf{S}_{\mathrm{grid}}$. We prove the displayed total-variation bound; the ℓ^1 formulation follows from $\|\eta\|_1 = 2\|\eta\|_{\mathrm{TV}}$.

1. Let $\mathsf{U}(s') = 1/360$ be the uniform law on $\mathsf{S}_{\mathrm{grid}}$. For $s = (g, d_1, d_2, x_1, x_2, a_1)$, $a \in \{0,1\}$, and $s' = (g', d_1', d_2', x_1', x_2', a_1')$, define the non-refresh transition weight

$$\mathsf{D}(s, a; s') = \gamma_{s,a}(g')\, \pi_d(d_1')\, \pi_x(x_1')\, \mathbf{1}\{d_2' = d_1\}\, \mathbf{1}\{x_2' = x_1\}\, \mathbf{1}\{a_1' = a\},$$

where $\pi_d(0) = 4/5$, $\pi_d(78) = 1/5$, $\pi_x(0) = 2/5$, $\pi_x(31) = 2/5$, $\pi_x(850) = 1/5$, and

$$\gamma_{s,a}(g^\star) = \Pr\Big\{\mathrm{round}_{\{50,100,150,200,250\}}\, \mathrm{clip}_{[50,250]}\big(10 + 0.9g + 0.1d_1 + 0.1d_2 - 0.01x_1 - 0.01x_2 - 2a - 4a_1 + \xi\big) = g^\star$$

for $\xi \sim N(0, 25)$. The refresh construction in Definition 24 gives, after marginalizing over the reward,

$$K_T\{(q, s'') :\ s'' = s' \mid s, a\} = \frac{1}{720} + \frac{1}{2}\,\mathsf{D}(s, a; s').$$

Therefore the policy-induced matrix can be written as

$$P_p^{(T)}(s, s') = \frac{1}{720} + \frac{1}{2} R_p(s, s'), \qquad R_p(s, s') = \sum_{a \in \{0,1\}} p(a \mid x(s))\, \mathsf{D}(s, a; s').$$

Equivalently,

$$P_p^{(T)}(s,s') = \Big(1-\frac{1}{2}\Big)\mathsf{U}(s') + \frac{1}{2}R_p(s,s').$$

2. For each fixed s, $R_p(s,\cdot)$ is a probability vector. Indeed, all factors in $\mathsf{D}(s,a;s')$ are nonnegative. Also,

$$\sum_{s'\in\mathsf{S}_{\text{grid}}} \mathsf{D}(s,a;s') = \sum_{g^\star}\gamma_{s,a}(g^\star)\sum_{d^\star}\pi_d(d^\star)\sum_{x^\star}\pi_x(x^\star) = 1,$$

because the rounded Gaussian cells partition the clipped glucose outcome, the diet probabilities sum to 1, the activity probabilities sum to 1, and the three indicator factors merely select the shifted memory coordinates. Thus

$$\sum_{s'\in\mathsf{S}_{\text{grid}}} R_p(s,s') = \sum_{a\in\{0,1\}} p(a\mid x(s)) \sum_{s'\in\mathsf{S}_{\text{grid}}} \mathsf{D}(s,a;s') = \sum_{a\in\{0,1\}} p(a\mid x(s)) = 1.$$

3. It remains to apply the common-refresh contraction calculation. For any finite state space, any probability vectors η,η', and any stochastic matrix R, a kernel of the form

$$M_\alpha(s,t) = (1-\alpha)u(t) + \alpha R(s,t), \qquad \alpha\ge 0,$$

satisfies

$$(\eta M_\alpha - \eta' M_\alpha)(t) = \alpha\sum_s(\eta(s)-\eta'(s))R(s,t),$$

since the refresh term contributes $(1-\alpha)u(t)(\sum_s\eta(s)-\sum_s\eta'(s)) = 0$. Hence

$$\begin{aligned}\|\eta M_\alpha - \eta' M_\alpha\|_{\text{TV}} &= \frac{1}{2}\sum_t\Big|\alpha\sum_s(\eta(s)-\eta'(s))R(s,t)\Big| \\ &\le \alpha\,\frac{1}{2}\sum_s|\eta(s)-\eta'(s)|\sum_t R(s,t) \\ &= \alpha\|\eta-\eta'\|_{\text{TV}}.\end{aligned}$$

With $\alpha = 1/2$, $u = \mathsf{U}$, $R = R_p$, and $\eta = \nu, \eta' = \nu'$, this gives

$$\|\nu P_p^{(T)} - \nu' P_p^{(T)}\|_{\text{TV}} \le \frac{1}{2}\|\nu-\nu'\|_{\text{TV}}.$$

Multiplying by 2 yields the stated ℓ^1 version. □

Proof of Lemma 2. Write

$$P = P_p^{(T)}$$

for the transition matrix induced by the policy p on the finite state space S_{grid}. We verify the hypotheses of the finite-state contraction fixed-point criterion and then identify the selected fixed point.

1. Since $p(\cdot\mid x)$ is a probability law on $\{0,1\}$ for each observed grid state x, and since the kernel of $\mathcal{I}_{\text{grid}}$ is a Markov kernel, the policy mixture

$$P(s,s') = \sum_{a\in\{0,1\}} p(a\mid x(s))\,K_T(s,a;s')$$

is stochastic. Thus, for every probability vector ν on $\mathsf{S}_{\mathrm{grid}}$,

$$(\nu P)(s') = \sum_{s \in \mathsf{S}_{\mathrm{grid}}} \nu(s) P(s, s')$$

is again a probability vector on $\mathsf{S}_{\mathrm{grid}}$. The initial law specified for $\mathcal{I}_{\mathrm{grid}}$ is also a probability vector.

2. By the stochasticity verified in the preceding step, the map $\nu \mapsto \nu P$ sends the probability simplex into itself. The displayed hypothesis of Lemma 2 gives

$$\|\nu P - \nu' P\|_{\mathrm{TV}} \leq \frac{1}{2} \|\nu - \nu'\|_{\mathrm{TV}}$$

for all probability vectors ν, ν'. The constant $1/2$ is nonnegative and strictly smaller than 1. Since the probability simplex over the finite set $\mathsf{S}_{\mathrm{grid}}$ is complete in total variation distance, the contraction mapping theorem gives a probability vector d such that

$$dP = d.$$

Equivalently,

$$d(s') = \sum_{s \in \mathsf{S}_{\mathrm{grid}}} d(s) P(s, s'), \qquad s' \in \mathsf{S}_{\mathrm{grid}}.$$

3. The selected law d_p is defined as the stationary law chosen for P whenever such a stationary probability vector exists. The preceding step supplies that existence, so the selected law is itself a probability vector and satisfies

$$d_p P = d_p.$$

Reading this coordinatewise yields

$$d_p(s') = \sum_{s \in \mathsf{S}_{\mathrm{grid}}} d_p(s) P_p^{(T)}(s, s'), \qquad s' \in \mathsf{S}_{\mathrm{grid}},$$

which is the asserted stationarity equation.

□

Proof of Lemma 3. Fix $s = (g, d_1, d_2, x_1, x_2, a_1) \in \mathsf{S}_{\mathrm{grid}}$ and abbreviate

$$g_e(s) := \sum_{a \in \{0,1\}} e(a \mid x(s)) \int p_1 \, K_T(dp \mid s, a).$$

By the definition of the target-policy reward regression, the left-hand side of the claimed identity is $g_e(s)$.

For each $a \in \{0, 1\}$, the embedded insulin kernel assigns the reward coordinate according to the current glucose component of s. Hence its conditional reward mean is

$$\int p_1 \, K_T(dp \mid s, a) = r_{\mathrm{ins}}(s).$$

Indeed, Definition 24 makes the reward coordinate a deterministic function of the present state, with the displayed values defining r_{ins}; the subsequent randomization affects the next state coordinates but not this reward coordinate.

Substituting this identity into the regression gives

$$g_e(s) = \sum_{a\in\{0,1\}} e(a \mid x(s))\, r_{\mathrm{ins}}(s) = \left(\sum_{a\in\{0,1\}} e(a \mid x(s)) \right) r_{\mathrm{ins}}(s).$$

The target policy in Definition 24 is a probability vector on $\{0,1\}$: if $g \geq 125$, then $e(1 \mid x(s)) = 1$ and $e(0 \mid x(s)) = 0$, while if $g < 125$, then $e(1 \mid x(s)) = 0$ and $e(0 \mid x(s)) = 1$. Therefore

$$\sum_{a\in\{0,1\}} e(a \mid x(s)) = 1,$$

and consequently $g_e(s) = r_{\mathrm{ins}}(s)$.

This is the asserted identity. □

Proof of Lemma 4. Fix $s = (g, d_1, d_2, x_1, x_2, a_1)$ and write $x = x(s)$.

1. Let $\mathrm{cat}_{\mathrm{rew}}(s) \in \{0,1,2,3\}$ be the reward category selected by the current glucose coordinate, and let

$$\mathrm{val}_{\mathrm{rew}}(0) = -1, \qquad \mathrm{val}_{\mathrm{rew}}(1) = -\frac{2}{3}, \qquad \mathrm{val}_{\mathrm{rew}}(2) = -\frac{1}{3}, \qquad \mathrm{val}_{\mathrm{rew}}(3) = 0.$$

   Thus $\mathrm{val}_{\mathrm{rew}}(\mathrm{cat}_{\mathrm{rew}}(s)) = r_{\mathrm{ins}}(s)$. In the finite insulin kernel of Definition 24, for each action a the next-state mass may be written as $\omega_{s,a}(\sigma)$, with

$$\omega_{s,a}(\sigma) \geq 0, \qquad \sum_{\sigma\in\mathsf{S}_{\mathrm{grid}}} \omega_{s,a}(\sigma) = 1,$$

   and the reward coordinate is deterministic at the current-state reward category:

$$K_T\big(\{(\mathrm{val}_{\mathrm{rew}}(\mathrm{cat}_{\mathrm{rew}}(s)), \sigma) : \sigma \in \mathsf{S}_{\mathrm{grid}}\} \mid s, a\big) = 1.$$

   Therefore finite-space integration gives

$$\begin{aligned}\int p_1\, K_T(dp \mid s, a) &= \sum_{\sigma\in\mathsf{S}_{\mathrm{grid}}} \omega_{s,a}(\sigma)\ \mathrm{val}_{\mathrm{rew}}(\mathrm{cat}_{\mathrm{rew}}(s)) \\ &= \mathrm{val}_{\mathrm{rew}}(\mathrm{cat}_{\mathrm{rew}}(s)) \sum_{\sigma\in\mathsf{S}_{\mathrm{grid}}} \omega_{s,a}(\sigma) = r_{\mathrm{ins}}(s).\end{aligned}$$

2. The behavior policy in Definition 24 satisfies

$$b(1 \mid x) = \frac{3}{10}, \qquad b(0 \mid x) = \frac{7}{10},$$

   so $b(a \mid x) > 0$ for both $a \in \{0,1\}$. By the zero-cell convention in the definition of ρ, the positive-denominator branch applies and hence

$$b(a \mid x)\rho(x,a) = b(a \mid x)\frac{e(a \mid x)}{b(a \mid x)} = e(a \mid x).$$

3. Substituting the two preceding identities into the weighted regression gives

$$\sum_{a\in\{0,1\}} b(a \mid x)\rho(x,a) \int p_1 \, K_T(dp \mid s,a) = \sum_{a\in\{0,1\}} e(a \mid x)\, r_{\text{ins}}(s)$$
$$= \left(\sum_{a\in\{0,1\}} e(a \mid x) \right) r_{\text{ins}}(s).$$

The target policy in Definition 24 injects exactly when $g \geq 125$, so

$$e(1 \mid x) = \mathbf{1}\{g \geq 125\}, \qquad e(0 \mid x) = 1 - \mathbf{1}\{g \geq 125\}, \qquad \sum_{a\in\{0,1\}} e(a \mid x) = 1.$$

Consequently

$$\sum_{a\in\{0,1\}} b(a \mid x)\rho(x,a) \int p_1 \, K_T(dp \mid s,a) = r_{\text{ins}}(s),$$

as claimed.

□

Proof of Lemma 5. 1. Fix $k = 0, \ldots, m$ and $s' \in \mathsf{S}$. For this pair (k, s'), write the chunked-recurrence data as

$$\mathcal{G}_1^{k,s'}, \ldots, \mathcal{G}_{J_{k,s'}}^{k,s'}, \qquad \Xi_1^{k,s'}(s'), \ldots, \Xi_{J_{k,s'}}^{k,s'}(s'), \qquad \Sigma_1^{k,s'}, \ldots, \Sigma_{J_{k,s'}+1}^{k,s'}.$$

These are the block partition, block intervals, and partial-sum intervals attached to this fixed (k, s'). The chunked recurrence first gives the ordinary interval recurrence

$$\sum_{s\in\mathsf{S}} \mathfrak{T}_k(s) \cdot \mathfrak{A}(s,s') \subseteq \mathfrak{T}_{k+1}(s').$$

Indeed, take arbitrary choices from the interval summands on the left and group them by the disjoint covering blocks $\mathcal{G}_1^{k,s'}, \ldots, \mathcal{G}_{J_{k,s'}}^{k,s'}$. The contribution of block $\mathcal{G}_j^{k,s'}$ lies in $\Xi_j^{k,s'}(s')$. Starting from the tail value $0 \in \Sigma_{J_{k,s'}+1}^{k,s'}$, the inclusions

$$\Xi_j^{k,s'}(s') + \Sigma_{j+1}^{k,s'} \subseteq \Sigma_j^{k,s'}, \qquad j = J_{k,s'}, \ldots, 1,$$

show by backward induction that the sum of the block contributions from j through $J_{k,s'}$ lies in $\Sigma_j^{k,s'}$. At $j = 1$ this is the full sum, and $\Sigma_1^{k,s'} = \mathfrak{T}_{k+1}(s')$. If the block list is empty, the same argument is just the terminal condition $0 \in \Sigma_1^{k,s'} = \mathfrak{T}_{k+1}(s')$. Thus the initial row together with these coordinate recurrences is the ordinary finite-iterate certificate with rows $\mathfrak{T}_0, \ldots, \mathfrak{T}_{m+1}$ used in the remaining argument.

2. Define the true iterates $\eta_n : \mathsf{S} \to \mathbb{R}$, for all $n \in \mathbb{N}$, by

$$\eta_0(s) = \mu_0(s), \qquad \eta_{n+1}(s') = \sum_{s\in\mathsf{S}} \eta_n(s) P(s,s').$$

Here μ_0 is the given initial probability vector, while η_n denotes its n-step image under P. The initial condition gives $\eta_0(s) \in \mathfrak{T}_0(s)$. If $\eta_n(s) \in \mathfrak{T}_n(s)$ for every s, then $P(s,s') \in \mathfrak{A}(s,s')$ and exact interval multiplication give

$$\eta_n(s)P(s,s') \in \mathfrak{T}_n(s) \cdot \mathfrak{A}(s,s'),$$

and exact interval summation together with the recurrence from the previous step gives

$$\eta_{n+1}(s') \in \mathfrak{T}_{n+1}(s').$$

Thus, for $n = 0, \ldots, m+1$,

$$\eta_n(s) \in \mathfrak{T}_n(s) \qquad \text{for every } s \in \mathsf{S}.$$

In particular,

$$\eta_m(s) \in \mathfrak{T}_m(s), \qquad \eta_{m+1}(s) \in \mathfrak{T}_{m+1}(s).$$

Since μ_0 is a probability vector and P is stochastic, every η_n is a probability vector.

3. The terminal interval expectation contains the true reward expectation at time m:

$$\sum_{s \in \mathsf{S}} \eta_m(s) r(s) \in \sum_{s \in \mathsf{S}} \mathfrak{T}_m(s) \cdot \mathfrak{r}(s) = \mathfrak{E}.$$

This follows coordinatewise from $\eta_m(s) \in \mathfrak{T}_m(s)$ and $r(s) \in \mathfrak{r}(s)$, followed by exact interval products and exact finite interval addition.

4. The one-step residual of η_m is bounded by δ. Since $\eta_{m+1} = \eta_m P$, exact interval subtraction gives

$$\eta_m(s) - \eta_{m+1}(s) \in \mathfrak{T}_m(s) \ominus \mathfrak{T}_{m+1}(s),$$

and hence

$$|\eta_m(s) - \eta_{m+1}(s)| \le \left|\mathfrak{T}_m(s) \ominus \mathfrak{T}_{m+1}(s)\right|_{\max}.$$

Summing over s yields

$$\|\eta_m - \eta_m P\|_1 = \|\eta_m - \eta_{m+1}\|_1 \le \delta.$$

Because π is stationary and both η_m and π are probability vectors,

$$\begin{aligned} \|\eta_m - \pi\|_1 &\le \|\eta_m - \eta_m P\|_1 + \|\eta_m P - \pi P\|_1 \\ &\le \delta + \rho \|\eta_m - \pi\|_1. \end{aligned}$$

Since $1 - \rho > 0$, this gives

$$\|\eta_m - \pi\|_1 \le \frac{\delta}{1-\rho}.$$

5. For every state s,

$$|r(s)| \le \max\{|\ell(\mathfrak{r}(s))|, |u(\mathfrak{r}(s))|\} = |\mathfrak{r}(s)|_{\max} \le B.$$

Therefore

$$\left|\sum_{s\in\mathsf{S}}\pi(s)r(s)-\sum_{s\in\mathsf{S}}\eta_m(s)r(s)\right| = \left|\sum_{s\in\mathsf{S}}(\pi(s)-\eta_m(s))r(s)\right| \le \sum_{s\in\mathsf{S}}|\pi(s)-\eta_m(s)|\,|r(s)| \le B\|\pi-\eta_m\|_1 \le \frac{B\delta}{1-\rho}.$$

6. Combining the preceding two enclosures, let

$$x_m=\sum_{s\in\mathsf{S}}\eta_m(s)r(s), \qquad x_\pi=\sum_{s\in\mathsf{S}}\pi(s)r(s).$$

From $x_m\in\mathfrak{E}$ and

$$|x_\pi-x_m|\le\frac{B\delta}{1-\rho},$$

we obtain

$$\ell(\mathfrak{E})-\frac{B\delta}{1-\rho}\le x_\pi\le u(\mathfrak{E})+\frac{B\delta}{1-\rho}.$$

By the definition of $\mathfrak{J}$, this is exactly

$$\sum_{s\in\mathsf{S}}\pi(s)r(s)\in\mathfrak{J}.$$

□

Proof of Lemma 6. All arithmetic below is exact rational interval arithmetic with the operations described in Lemma 5.

1. The residual sums over the 360 states of Definition 24 evaluate to

$$\delta_{\mathrm{b}}=\sum_{s\in\mathsf{S}_{\mathrm{grid}}}\left|\mathfrak{T}_{\mathrm{b},0}(s)\ominus\mathfrak{T}_{\mathrm{b},1}(s)\right|_{\max}=\frac{2596399843}{200000000000000},$$

and

$$\delta_{\mathrm{e}}=\sum_{s\in\mathsf{S}_{\mathrm{grid}}}\left|\mathfrak{T}_{\mathrm{e},0}(s)\ominus\mathfrak{T}_{\mathrm{e},1}(s)\right|_{\max}=\frac{16950307061}{1000000000000000}.$$

2. Since $\mathfrak{r}_{\mathrm{ins}}(s)=[r_{\mathrm{ins}}(s),r_{\mathrm{ins}}(s)]$, the interval expectations computed from the terminal rows are

$$\mathfrak{E}_{\mathrm{b}}=\sum_{s\in\mathsf{S}_{\mathrm{grid}}}\mathfrak{T}_{\mathrm{b},0}(s)\cdot\mathfrak{r}_{\mathrm{ins}}(s)=\left[-\frac{157888864566941}{300000000000000},-\frac{63155545824091}{120000000000000}\right],$$

and

$$\mathfrak{E}_{\mathrm{e}}=\sum_{s\in\mathsf{S}_{\mathrm{grid}}}\mathfrak{T}_{\mathrm{e},0}(s)\cdot\mathfrak{r}_{\mathrm{ins}}(s)=\left[-\frac{312236715379949}{600000000000000},-\frac{312236715366659}{600000000000000}\right].$$

3. Because $1 - 1/2 = 1/2$, the expansion radius in each case is $2\delta_\star$. Hence

$$\mathfrak{J}_{\mathrm{b}} = \left[-\frac{157888864566941}{300000000000000} - 2\frac{2596399843}{2000000000000000}, -\frac{63155545824091}{120000000000000} + 2\frac{2596399843}{2000000000000000} \right] = \left[-\frac{15789665}{30000000} \right.$$

and

$$\mathfrak{J}_{\mathrm{e}} = \left[-\frac{312236715379949}{600000000000000} - 2\frac{16950307061}{1000000000000000}, -\frac{312236715366659}{600000000000000} + 2\frac{16950307061}{1000000000000000} \right] = \left[-\frac{156128}{300000} \right.$$

4. Taking the interval difference,

$$\mathfrak{B} = \mathfrak{J}_{\mathrm{b}} \ominus \mathfrak{J}_{\mathrm{e}} = \left[\ell(\mathfrak{J}_{\mathrm{b}}) - u(\mathfrak{J}_{\mathrm{e}}), u(\mathfrak{J}_{\mathrm{b}}) - \ell(\mathfrak{J}_{\mathrm{e}}) \right],$$

so

$$\mathfrak{B} = \left[-\frac{17884662673771}{3000000000000000}, -\frac{2920912477479}{500000000000000} \right].$$

The endpoint comparisons are then the rational inequalities

$$-\frac{6}{1000} = -\frac{18000000000000}{3000000000000000} < -\frac{17884662673771}{3000000000000000} = \ell(\mathfrak{B}),$$

and

$$u(\mathfrak{B}) = -\frac{2920912477479}{500000000000000} < -\frac{2900000000000}{500000000000000} = -\frac{58}{10000}.$$

This proves both asserted endpoint bounds.

□

Proof of Lemma 7. Fix $T \in \mathbb{N}$. Write $P_b^{(T)}$ and $P_e^{(T)}$ for the policy-induced transition matrices on $\mathsf{S}_{\mathrm{grid}}$ obtained from Lemma 1 by taking $p = b$ and $p = e$, respectively, and write $d_b^{(T)}$ and $d_e^{(T)}$ for their selected stationary laws.

1. The behavior and target kernels both satisfy the one-half contraction estimate

$$\|\nu P_b^{(T)} - \nu' P_b^{(T)}\|_1 \le \frac{1}{2}\|\nu - \nu'\|_1, \qquad \|\nu P_e^{(T)} - \nu' P_e^{(T)}\|_1 \le \frac{1}{2}\|\nu - \nu'\|_1$$

for all probability vectors ν, ν' on $\mathsf{S}_{\mathrm{grid}}$. This is exactly the ℓ^1-form supplied by Lemma 1, instantiated first at $p = b$ and then at $p = e$. The corresponding selected laws are probability vectors and satisfy

$$d_b^{(T)} P_b^{(T)} = d_b^{(T)}, \qquad d_e^{(T)} P_e^{(T)} = d_e^{(T)},$$

by Lemma 2.

2. Let

$$r_{\mathrm{ins}}(s) = \begin{cases} -1, & g \le 70, \\ -2/3, & g > 150, \\ -1/3, & 70 < g \le 80 \text{ or } 120 < g \le 150, \\ 0, & \text{otherwise}, \end{cases} \qquad s = (g, d_1, d_2, x_1, x_2, a_1).$$

For $\star \in \{\mathrm{b}, \mathrm{e}\}$, with $P_\star^{(T)}$ denoting $P_b^{(T)}$ or $P_e^{(T)}$, let $\mathfrak{A}_\star$ be the exact rational interval matrix recorded by the finite-grid certificate. Its entrywise enclosure is

$$P_\star^{(T)}(s, s') \in \mathfrak{A}_\star(s, s'), \qquad s, s' \in \mathsf{S}_{\mathrm{grid}}.$$

The same certificate supplies an exact rational initial probability row $\mu_{\star,0}$, interval rows $\mathfrak{T}_{\star,0}$ and $\mathfrak{T}_{\star,1}$, and an exact chunked recurrence from the first interval row to the second. In the notation of Lemma 5, its rational inclusions verify

$$[\mu_{\star,0}(s), \mu_{\star,0}(s)] \subseteq \mathfrak{T}_{\star,0}(s), \qquad s \in \mathsf{S}_{\mathrm{grid}}.$$

For every output state, the recorded blocks partition $\mathsf{S}_{\mathrm{grid}}$, every block has at most 40 states, and the block and partial-sum intervals satisfy exactly the inclusion chain in Lemma 5. Because the certificate has zero approximation steps, this verified recurrence links its initial interval row directly to its successor row. Passing from the chunked certificate to the finite recurrence used by that lemma preserves both rows and all the rational inclusions. For the reward certificate, put

$$\mathfrak{r}_{\mathrm{ins}}(s) = [r_{\mathrm{ins}}(s), r_{\mathrm{ins}}(s)].$$

Since $|r_{\mathrm{ins}}(s)| \le 1$ for every grid state, the reward bound in Lemma 5 is $B = 1$. Define

$$\delta_\star = \sum_{s \in \mathsf{S}_{\mathrm{grid}}} \left|\mathfrak{T}_{\star,0}(s) \ominus \mathfrak{T}_{\star,1}(s)\right|_{\max}, \qquad \mathfrak{E}_\star = \sum_{s \in \mathsf{S}_{\mathrm{grid}}} \mathfrak{T}_{\star,0}(s) \cdot \mathfrak{r}_{\mathrm{ins}}(s).$$

After that conversion, the terminal row is $\mathfrak{T}_{\star,0}$ and the successor row is $\mathfrak{T}_{\star,1}$, so the residual and stationary-reward interval are

$$\delta_\star = \sum_{s \in \mathsf{S}_{\mathrm{grid}}} \left|\mathfrak{T}_{\star,0}(s) \ominus \mathfrak{T}_{\star,1}(s)\right|_{\max}, \qquad \mathfrak{J}_\star = \left[\ell(\mathfrak{E}_\star) - \frac{\delta_\star}{1 - 1/2}, u(\mathfrak{E}_\star) + \frac{\delta_\star}{1 - 1/2}\right].$$

The initial-row, recurrence, reward-bound, contraction, and stationarity hypotheses of Lemma 5 are therefore satisfied for both policies after passing each chunked recurrence certificate to its associated finite recurrence certificate. Under this conversion, the terminal interval row is $\mathfrak{T}_{\star,0}$, the successor interval row is $\mathfrak{T}_{\star,1}$, and the stationary-reward interval produced by Lemma 5 is exactly

$$\mathfrak{J}_\star = \left[\ell(\mathfrak{E}_\star) - \frac{\delta_\star}{1 - 1/2}, u(\mathfrak{E}_\star) + \frac{\delta_\star}{1 - 1/2}\right],$$

which is the chunked interval $\mathfrak{J}_\star$ used here. Hence

$$\sum_{s \in \mathsf{S}_{\mathrm{grid}}} d_b^{(T)}(s) r_{\mathrm{ins}}(s) \in \mathfrak{J}_{\mathrm{b}}, \qquad \sum_{s \in \mathsf{S}_{\mathrm{grid}}} d_e^{(T)}(s) r_{\mathrm{ins}}(s) \in \mathfrak{J}_{\mathrm{e}}.$$

3. Define the subtraction interval

$$\mathfrak{B} := \mathfrak{J}_{\mathrm{b}} \ominus \mathfrak{J}_{\mathrm{e}} = \big[\ell(\mathfrak{J}_{\mathrm{b}}) - u(\mathfrak{J}_{\mathrm{e}}), u(\mathfrak{J}_{\mathrm{b}}) - \ell(\mathfrak{J}_{\mathrm{e}})\big].$$

Interval subtraction is sound, so the two inclusions from the previous step give

$$\sum_s d_b^{(T)}(s) r_{\mathrm{ins}}(s) - \sum_s d_e^{(T)}(s) r_{\mathrm{ins}}(s) \in \mathfrak{B}.$$

4. The reward-regression identities in Lemmas 3 and 4 give, for every $s \in \mathsf{S}_{\mathrm{grid}}$,

$$\sum_{a\in\{0,1\}} b(a \mid x(s))\,\rho(x(s),a) \int p_1\, K_T(dp \mid s,a) = r_{\mathrm{ins}}(s),$$

and

$$\sum_{a\in\{0,1\}} e(a \mid x(s)) \int p_1\, K_T(dp \mid s,a) = r_{\mathrm{ins}}(s).$$

The finite-grid embedding preserves the policy-induced kernels entry by entry:

$$P_b^{(T)}(s,s') = \sum_a b(a \mid x(s)) K_T(\mathbb{R}\times\{s'\} \mid s,a), \qquad P_e^{(T)}(s,s') = \sum_a e(a \mid x(s)) K_T(\mathbb{R}\times\{s'\} \mid s,a).$$

Consequently the stationary laws in the interval certificate are exactly the behavior- and target-policy stationary laws entering the two regression identities. Combining these kernel identifications, the two regression identities, and Definition 12 yields

$$\Delta_{\mathrm{imm}}(T) = \sum_s d_b^{(T)}(s) r_{\mathrm{ins}}(s) - \sum_s d_e^{(T)}(s) r_{\mathrm{ins}}(s).$$

Therefore

$$\Delta_{\mathrm{imm}}(T) \in \mathfrak{B}.$$

5. By the exact rational interval evaluation in Lemma 6, the endpoints of the subtraction interval $\mathfrak{B}$ satisfy

$$-\frac{6}{1000} < \ell(\mathfrak{B}), \qquad u(\mathfrak{B}) < -\frac{58}{10000}.$$

Since $\Delta_{\mathrm{imm}}(T) \in \mathfrak{B}$, we have

$$\ell(\mathfrak{B}) \le \Delta_{\mathrm{imm}}(T) \le u(\mathfrak{B}).$$

The last two displays imply

$$-\frac{6}{1000} < \Delta_{\mathrm{imm}}(T) < -\frac{58}{10000},$$

as claimed.

□

Proof of Theorem 5. Fix T. We verify the five asserted clauses.

1. By Definition 24, the observed coordinate has

$$|\mathcal{X}| = |\{50,100,150,200,250\}|\,|\{0,31,850\}|^2\,|\{0,1\}| = 5\cdot 3^2\cdot 2 = 90,$$

and the latent diet-memory coordinate has

$$|\mathcal{H}| = |\{0,78\}|^2 = 2^2 = 4.$$

Thus the joint-state alphabet $\mathcal{S} = \mathcal{X}\times\mathcal{H}$ has

$$|\mathcal{S}| = 90\cdot 4 = 360.$$

2. For a memoryless policy $p \in \{b, e\}$, let $P_p^{(T)} : \mathcal{S} \times \mathcal{S} \to [0, 1]$ be the reward-marginalized joint-state transition matrix

$$P_p^{(T)}(s, s') := \sum_{a \in \{0,1\}} p(a \mid s_X)\, K_T\{(r_{\text{out}}, \bar{s}) : \bar{s} = s' \mid s, a\}, \qquad s, s' \in \mathcal{S}.$$

By the construction in Definition 24, the embedded policy kernel of the refreshed insulin experiment is this matrix. Hence, for every pair of probability laws ν, ν' on $\mathcal{S}$, Lemma 1 gives

$$\|\nu P_p^{(T)} - \nu' P_p^{(T)}\|_{\text{TV}} \le \frac{1}{2}\|\nu - \nu'\|_{\text{TV}}.$$

Applying this with $p = b$ and $p = e$ is exactly Assumption 6 with $\alpha = 1/2$.

3. The behavior probabilities are

$$b(1 \mid x) = \frac{3}{10}, \qquad b(0 \mid x) = \frac{7}{10},$$

and the target policy is a probability law with $0 \le e(a \mid x) \le 1$. Therefore

$$e(1 \mid x) \le 1 = \frac{10}{3}\,\frac{3}{10} = \frac{10}{3} b(1 \mid x),$$

and

$$e(0 \mid x) \le 1 \le \frac{7}{3} = \frac{10}{3}\,\frac{7}{10} = \frac{10}{3} b(0 \mid x).$$

Thus Assumption 5 holds with $L = 10/3$.

4. Let d_b and d_e be the selected stationary laws of $P_b^{(T)}$ and $P_e^{(T)}$. The behavior and target policies in Definition 24 are probability vectors; together with the contraction estimate supplied by Lemma 1, Lemma 2 gives

$$d_p(s') = \sum_{s \in \mathcal{S}} d_p(s) P_p^{(T)}(s, s'), \qquad \sum_{s \in \mathcal{S}} d_p(s) = 1, \qquad d_p(s) \ge 0,$$

for $p \in \{b, e\}$ and all $s' \in \mathcal{S}$. Write S_t for the joint state at time t. For $a \in \{0, 1\}$ and $s, s' \in \mathcal{S}$, define the nonrefresh action kernel as follows. If

$$s = (g, d_1, d_2, x_1, x_2, a_1), \qquad s' = (g', d_1', d_2', x_1', x_2', a_1'),$$

let $\varphi_{s,a}(g')$ be the probability that the Gaussian clipping-and-rounding rule in Definition 24 produces next glucose value g', and set this probability to zero when g' is outside the glucose grid. Let

$$\pi_D(0) = \frac{4}{5}, \qquad \pi_D(78) = \frac{1}{5}, \qquad \pi_E(0) = \pi_E(31) = \frac{2}{5}, \qquad \pi_E(850) = \frac{1}{5},$$

with π_D and π_E equal to zero off their displayed supports. Conditional on the nonrefresh branch and action a, Definition 24 draws the new diet and activity coordinates with laws π_D and π_E, shifts the old memory coordinates, and stores the current action. Hence the nonrefresh probability of moving from s to s' under action a is

$$\omega_a(s, s') := \varphi_{s,a}(g')\, \pi_D(d_1')\, \pi_E(x_1')\, \mathbf{1}\{d_2' = d_1\}\, \mathbf{1}\{x_2' = x_1\}\, \mathbf{1}\{a_1' = a\}.$$

All factors in this display are nonnegative, so $\omega_a(s, s') \geq 0$. The refresh branch in Definition 24 has probability $1/2$ and draws the next full state uniformly from the 360-point set $\mathcal{S}$, so its contribution to any fixed next state s' is

$$\frac{1}{2} \cdot \frac{1}{360} = \frac{1}{720}.$$

The remaining branch has probability $1/2$, and the policy-induced kernel averages the action-specific nonrefresh probabilities with weights $p(a \mid s_X)$. Thus, for every memoryless policy p on $\{0, 1\}$,

$$P_p^{(T)}(s, s') = \frac{1}{720} + \frac{1}{2} \sum_{a \in \{0,1\}} p(a \mid s_X)\omega_a(s, s'), \qquad s, s' \in \mathcal{S}.$$

Since the behavior policy has nonnegative action probabilities, this identity gives the pointwise minorization

$$P_b^{(T)}(s, s') = \frac{1}{720} + \frac{1}{2} \sum_{a \in \{0,1\}} b(a \mid s_X)\omega_a(s, s') \geq \frac{1}{720}, \qquad s, s' \in \mathcal{S}.$$

The stationary-minorization calculation is then

$$d_b(s') = \sum_{s \in \mathcal{S}} d_b(s) P_b^{(T)}(s, s') \geq \sum_{s \in \mathcal{S}} d_b(s) \frac{1}{720} = \frac{1}{720}, \qquad s' \in \mathcal{S}.$$

Since d_e is a probability vector,

$$d_e(s') \leq \sum_{\bar{s} \in \mathcal{S}} d_e(\bar{s}) = 1.$$

Combining the last two displays yields

$$d_e(s') \leq 1 \leq 720\, d_b(s'), \qquad s' \in \mathcal{S},$$

and $720 \geq 1$. This proves Assumption 7 with $C = 720$.

5. For the same horizon T, write K for the transition kernel of $\mathcal{I}_{\text{grid}}$, b and e for its behavior and target policies, and d_b for the stationary law induced by b. The stationary immediate-weighting bias diagnostic is

$$\Delta_{\text{imm}}(T) = \sum_s d_b(s) \sum_{a \in \{0,1\}} b(a \mid s_X) \frac{e(a \mid s_X)}{b(a \mid s_X)} \int r\, K(ds', dr \mid s, a) - \theta(K, e).$$

For this diagnostic on the refreshed finite insulin adaptation $\mathcal{I}_{\text{grid}}$ of Definition 24, Lemma 7 supplies the interval certificate

$$-\frac{6}{1000} < \Delta_{\text{imm}}(T) < -\frac{58}{10000},$$

which is the displayed bound

$$-0.006 < \Delta_{\text{imm}}(T) < -0.0058.$$

The five displayed conclusions are precisely the conjunction asserted for the refreshed insulin experiment. □

References

Andrew Bennett and Nathan Kallus. Proximal reinforcement learning: Efficient off-policy evaluation in partially observed markov decision processes, 2021. URL https://arxiv.org/abs/2110.15332.

Yuchen Hu and Stefan Wager. Off-policy evaluation in partially observed markov decision processes under sequential ignorability. *The Annals of Statistics*, 51(4):1561–1585, 2023. URL https://arxiv.org/abs/2110.12343.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 652–661. PMLR, 2016. URL https://arxiv.org/abs/1511.03722.

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. doi: 10.1016/S0004-3702(98)00023-X.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167): 1–63, 2020. URL https://arxiv.org/abs/1908.08526.

Qi Kuang, Jiayi Wang, Fan Zhou, and Zhengling Qi. Breaking the order barrier: Off-policy evaluation for confounded POMDPs. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://openreview.net/forum?id=8WCftxn2Uu.

Peng Liao, Predrag Klasnja, and Susan A. Murphy. Off-policy estimation of long-term average outcomes with applications to mobile health. *Journal of the American Statistical Association*, 116(533):382–391, 2021. doi: 10.1080/01621459.2020.1807993. URL https://arxiv.org/abs/1912.13088.

Peng Liao, Zhengling Qi, Runzhe Wan, Predrag Klasnja, and Susan A. Murphy. Batch policy learning in average reward markov decision processes. *The Annals of Statistics*, 50(6):3364–3387, 2022. doi: 10.1214/22-AOS2231. URL https://arxiv.org/abs/2007.11771.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://arxiv.org/abs/1810.12429.

Daniel J. Luckett, Eric B. Laber, Anna R. Kahkoska, David M. Maahs, Elizabeth J. Mayer-Davis, and Michael R. Kosorok. Estimating dynamic treatment regimes in mobile health using V-learning. *Journal of the American Statistical Association*, 115(530):692–706, 2020. doi: 10.1080/01621459.2018.1537919. URL https://arxiv.org/abs/1611.03531.

David M. Maahs, Elizabeth J. Mayer-Davis, Freya K. Bishop, Lin Wang, Margo Mangan, and Robert G. McMurray. Outpatient assessment of determinants of glucose excursions in adolescents with type 1 diabetes: Proof of concept. *Diabetes Technology & Therapeutics*, 14(8):658–664, 2012. doi: 10.1089/dia.2012.0053. URL https://doi.org/10.1089/dia.2012.0053.

Mohammad Mehrabi and Stefan Wager. Off-policy evaluation in markov decision processes under weak distributional overlap, 2025. URL https://arxiv.org/abs/2402.08201.

Rui Miao, Zhengling Qi, and Xiaoke Zhang. Off-policy evaluation for episodic partially observable markov decision processes under non-parametric models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2209.10064.

Susan A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003. doi: 10.1111/1467-9868.00389.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1906.04733.

Inbal Nahum-Shani, Shawna N. Smith, Bonnie J. Spring, Linda M. Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A. Murphy. Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, 52(6):446–462, 2018. doi: 10.1007/s12160-016-9830-8. URL https://doi.org/10.1007/s12160-016-9830-8.

Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 759–766, 2000. URL https://dl.acm.org/doi/10.5555/645529.657980.

James M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9–12):1393–1512, 1986. doi: 10.1016/0270-0255(86)90088-6.

James M. Robins, Miguel A. Hernán, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000. doi: 10.1097/00001648-200009000-00011.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. doi: 10.1093/biomet/70.1.41.

Chengchun Shi, Masatoshi Uehara, Jiawei Huang, and Nan Jiang. A minimax learning approach to off-policy evaluation in confounded partially observable markov decision processes. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 20057–20094. PMLR, 2022. URL https://arxiv.org/abs/2111.06784.

Richard D. Smallwood and Edward J. Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973. doi: 10.1287/opre.21.5.1071.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, second edition, 2018. ISBN 9780262039246. URL https://mitpress.mit.edu/9780262039246/reinforcement-learning/.

Guy Tennenholtz, Uri Shalit, and Shie Mannor. Off-policy evaluation in partially observable environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34,

pages 10276–10283, 2020. doi: 10.1609/aaai.v34i06.6590. URL https://arxiv.org/abs/1909.03739.

Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2139–2148. PMLR, 2016. URL https://arxiv.org/abs/1604.00923.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and Q-function learning for off-policy evaluation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9659–9668. PMLR, 2020. URL https://arxiv.org/abs/1910.12809.

Masatoshi Uehara, Haruka Kiyohara, Andrew Bennett, Victor Chernozhukov, Nan Jiang, Nathan Kallus, Chengchun Shi, and Wen Sun. Future-dependent value-based off-policy evaluation in POMDPs. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2207.13081.

Yuheng Zhang and Nan Jiang. On the curses of future and history in future-dependent value functions for off-policy evaluation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2402.14703.

Yuheng Zhang and Nan Jiang. Statistical tractability of off-policy evaluation of history-dependent policies in POMDPs. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2503.01134.