CHOOSING WHAT TO REMEMBER: ROBUST OFFLINE REINFORCEMENT LEARNING WITH HISTORY REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust offline reinforcement learning uses recorded data alone to learn policies that perform reliably under uncertainty. With delays or partial observations, policies often use history representations that map observed histories to represented states. The joint distribution of reward and next represented state can then depend on the policy, even at a fixed represented state and action in an unchanged environment. We develop a framework for conditional operator robustness under history representations and propose Conditional Operator Robust Fitted Q-Iteration (COR-FQI). Its robust Bellman recursion chooses actions that maximize the worst-case expected sum of immediate reward and continuation value over uncertainty sets. These sets are constructed to account for differences in conditional Bellman expectations among histories sharing a represented state, in addition to finite-sample estimation error and specified environmental changes. Under stated assumptions, we establish lower confidence bounds on the learned policies' worst-case expected returns without requiring the represented process to be Markov. The guarantees cover tabular settings and function approximation, with additional coverage conditions and uniform error bounds for the latter. These bounds hold simultaneously across a finite family of representations and provide a performance guarantee for choosing the representation–policy pair with the largest lower confidence bound. Tabular experiments show informative bounds and selection of longer memory with more data, while implementations based on function approximation show empirical robustness to perturbations, delays, and distribution shift.

1 INTRODUCTION

Learning a control policy by interacting with a physical system can be costly. Offline reinforcement learning (RL) instead uses recorded trajectories (Levine et al., 2020), while robust RL seeks reliable performance under specified changes in the environment (Iyengar, 2005; Wiesemann et al., 2013). Delays and partial observations introduce another choice: how much of the observed history should the policy retain? Longer memory can distinguish situations that require different actions, but it also divides the available data among more inputs. A useful choice must account for both the information retained and the reliability of the policy learned from it.

A *history representation* maps observation–action–reward histories to *represented states*, such as memory windows or recurrent encodings fixed before calibration. Histories mapped to the same state can have different joint distributions of reward and next represented state under the same action. We call this *conditional variation*. For example, a delayed controller may observe the same physical state while different commands are waiting to execute. Its logged outcome distribution is a mixture of the distributions at those histories, with weights determined by the data-collection policy. A learned policy can change these weights without changing the environment. Accurate estimation of the logged mixture therefore does not, by itself, justify the learned policy's value.

Robust Fitted Q-Iteration (RFQI) addresses environmental uncertainty for Markov inputs (Panaganti et al., 2022). Robust non-Markov offline learning can instead estimate dynamics conditioned on full histories (Huang et al., 2025). Approximate information states and agent-state analyses quantify the loss from compressing histories (Subramanian et al., 2022; Dong et al., 2022; Chandak et al., 2024).

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

Representation selection also trades approximation against estimation error, for Markov candidates online (Maillard et al., 2011) and state abstractions from batch data (Jiang et al., 2015). We study a specific additional requirement: construct uncertainty sets around estimates made from compressed histories that contain every relevant history's distribution, account for deployment changes, and support selection using the same offline dataset. Related approaches learn transformer beliefs for delays (Zhan et al., 2026), use sensitivity sets for omitted confounders (Bruns-Smith & Zhou, 2023), select value-function classes (Lee et al., 2022), or use kernels for robust learning (Lim & Autef, 2019); see Appendix E.

Our approach. Conditional Operator Robust Fitted Q-Iteration (COR-FQI) enlarges a total variation (TV) ball around the estimated mixture by estimation error, a justified history-variation bound, and deployment uncertainty. Robust dynamic programming then learns a policy and bounds its worst-case return. Simultaneous calibration preserves this guarantee after representation selection. The backup is established; the contribution is the calibration that makes it valid after history compression.

1.1 Statement of Contributions

1. We characterize the smallest TV ball containing all reweightings of a cell's histories and show that its radius is not identified by mixture statistics alone (Proposition 1). Finer logged information supplies upper calibration bounds and lower bounds that can establish insufficiency (Proposition 4).
2. We construct simultaneous confidence sets from dependent trajectory outcomes at random visit counts (Theorem 3). History-wise comparison then yields policy certificates and upper bounds on the full-history optimum, without requiring a Markov representation or a rectangular deployment family (Theorem 5).
3. These bounds give a computed optimality-gap bound after representation selection (Theorem 6). We characterize when refinement orders certificates and extend the comparison to approximate backups under explicit uniform-error premises (Proposition 7, Theorem 8).

FrozenLake experiments test the complete finite procedure: informative certificates, selection with limited data, calibration from finer logs, and selection without a sufficient representation. They also separate improved certification from improved policy return. Approximate control and transfer experiments provide empirical use cases; they do not establish the conditions required for certification with function approximation.

2 Preliminaries

All spaces are standard Borel and all maps and kernels are measurable (Appendix A.1). Time subscripts and representation/environment superscripts are used unless stated otherwise; nominal environment superscripts are omitted. Lowercase letters denote realizations; $[x]_+ = \max\{x, 0\}$ denotes the positive part.

2.1 Histories, environments, and returns

Fix a horizon $T \in \mathbb{N}$ and a nonempty finite action set $\mathcal{A}$. Write $[n] = \{1, \dots, n\}$ for integers $n \geq 1$. Observations are $O_t \in \mathcal{O}_t$ for $t \in [T+1]$. At stage $t \in [T]$, the action is $A_t \in \mathcal{A}$ and the reward is $R_t \in [0, 1]$. The pre-action history is

$$H_t = (O_1, A_1, R_1, \dots, O_{t-1}, A_{t-1}, R_{t-1}, O_t) \in \mathsf{H}_t, \tag{1}$$

with H_t the product history space. An environment M specifies history-conditional kernels for (R_t, O_{t+1}). All environments share a known initial distribution μ for $H_1 = O_1$. Let $\mathfrak{M}$ be a nonempty deployment set and let M_0 be the nominal environment that generated the offline data under a behavior policy. The logging reset distribution may differ from μ; the nominal transition kernels are unchanged. A history-dependent policy assigns an action distribution $\pi_t(\cdot \mid h)$ to each history h. For such a policy π, let $\mathbb{E}_{M,\pi}$ denote expectation under its trajectory distribution in M, and define

$$J_M(\pi) = \mathbb{E}_{M,\pi}\left[\sum_{t=1}^{T} R_t\right], \qquad J(\pi) = \inf_{M \in \mathfrak{M}} J_M(\pi).$$

The deployment family is fixed before training. The learner receives supports and bounds valid throughout this family (Section 4), rather than its kernels; these bounds need not determine the family uniquely.

2.2 History representations and policy classes

A *history representation* is a sequence of measurable maps $e = \{e_t : t \in [T+1]\}$ with $e_t : \mathsf{H}_t^\circ \to \mathcal{Z}_t^e$, where H_t° is the admissible history domain and $Z_t^e = e_t(H_t)$ is the *represented state*. The admissible domain contains the initial supports and is closed under every action and specified environment. Let $\mathcal{E}$ be a nonempty finite family of representations fixed before calibration; each $\mathcal{Z}_t^e$ contains only attained values. An *e-policy* is a sequence of probability kernels $\pi_t(da \mid z)$ on $\mathcal{A}$; it can use H_t only through Z_t^e. For deterministic policies, $\pi_t(z)$ denotes the selected action. Let $\Pi(e)$ be this class and Π the class of all admissible history-dependent policies; an e-policy is identified with the history policy it induces. Its class value and the full-history value are

$$J^\star(e) = \sup_{\pi \in \Pi(e)} J(\pi), \qquad J^\star = \sup_{\pi \in \Pi} J(\pi).$$

Write $\lambda(e) = \mu \circ e_1^{-1}$ for the initial represented-state distribution, where $\mu \circ F^{-1}$ denotes the distribution of $F(H)$ for $H \sim \mu$.

3 Problem description

3.1 Conditional variation

Write $\mathcal{L}_M$ for conditional distribution under $M \in \mathfrak{M} \cup \{M_0\}$. Define the one-step represented outcome distribution

$$\nu_t^{e,M}(\cdot \mid h, a) := \mathcal{L}_M\big((R_t, e_{t+1}(H_{t+1})) \mid H_t = h, A_t = a\big), \tag{2}$$

Write ν_t^e for the nominal kernel ν_t^{e,M_0}. The cell $C_t^e(z) = \{h \in \mathsf{H}_t^\circ : e_t(h) = z\}$ contains the histories that e aggregates into z. At a behavior-supported (t, z, a), let $\psi_t^e(\cdot \mid z, a)$ be the conditional distribution of H_t under the logging law, and define the logged mixture distribution

$$\bar{\nu}_t^e(\cdot \mid z, a) = \int_{C_t^e(z)} \nu_t^e(\cdot \mid h, a)\, \psi_t^e(dh \mid z, a).$$

This population mixture describes the logging distribution; the concentration argument below instead uses the average of the conditionals at the actual visits. A representation is *predictively sufficient* if $\nu_t^{e,M}(\cdot \mid h, a)$ is constant over each cell for every stage, action, and specified environment; equivalently, in every specified environment the represented state and the action determine the joint law of the reward and the next represented state. Otherwise it exhibits conditional variation, and changing the policy can change the mixture within a cell.

Write $\mathscr{P}(S)$ for the probability measures on a measurable space S, $\nu[g] = \int g\, d\nu$ for bounded measurable g, and $\mathrm{TV}(\nu_1, \nu_2) = \sup_{S' \subseteq S} |\nu_1(S') - \nu_2(S')|$ over measurable S' for $\nu_1, \nu_2 \in \mathscr{P}(S)$.

Proposition 1 (Mixture statistics do not identify conditional variation)**.** *Fix a behavior-supported (t, e, z, a), abbreviate $\nu_h = \nu_t^e(\cdot \mid h, a)$ and $\bar{\nu} = \bar{\nu}_t^e(\cdot \mid z, a)$, and let $\alpha_t^e(z, a) = \sup_{h \in C_t^e(z)} \mathrm{TV}(\nu_h, \bar{\nu})$.*

(i) For every probability measure ψ on $C_t^e(z)$, let $\bar{\nu}_\psi = \int \nu_h\, \psi(dh)$. Then

$$\sup_{\psi \in \mathscr{P}(C_t^e(z))} \mathrm{TV}(\bar{\nu}_\psi, \bar{\nu}) = \alpha_t^e(z, a). \tag{3}$$

Hence $\alpha_t^e(z, a)$ is the smallest radius of a TV ball at the pooled distribution that contains every reweighting of the cell's histories, including every policy-induced one.

(ii) There exist two nominal models with the same joint distribution of represented state, action, and pooled one-step outcome, but different values of $\alpha_t^e(z, a)$.

The supremum ranges over all cell reweightings (Appendix A.2); a fixed process can realize fewer through its policies. Thus $\alpha_t^e(z, a)$ suffices for policy-induced mixtures, but need not be their smallest radius. Calibrating all histories requires additional structure or finer logged information (Section 4.3).

3.2 CERTIFIED SELECTION

Let $\mathcal{D}$ be the random offline dataset, $\mathbb{P}_{\mathcal{D}}$ its sampling distribution, and $\delta \in (0,1)$ a failure probability. From $\mathcal{D}$, $\mathcal{E}$, and $\mathfrak{M}$, a learner returns for each $e \in \mathcal{E}$ an e-policy $\widehat{\pi}^e$ and its *certificate* $L(e)$; the pair is *certified* when $L(e) \le J(\widehat{\pi}^e)$, and the candidates are *jointly certified* at level $1-\delta$ when

$$\mathbb{P}_{\mathcal{D}}\left(\bigcap_{e\in\mathcal{E}}\{L(e) \le J(\widehat{\pi}^e)\}\right) \ge 1-\delta. \tag{4}$$

The *certificate gap* $J^\star - L(e)$ decomposes as

$$J^\star - L(e) = \Delta(e) + \big(J^\star(e) - L(e)\big), \qquad \Delta(e) = J^\star - J^\star(e) \ge 0, \tag{5}$$

the *representation loss* $\Delta(e)$ plus the *certificate shortfall* within its policy class. A loose certificate need not imply that the representation prevents a good policy. Section 5 bounds the total gap by computed quantities.

The problem. Given the dataset and a fixed representation family, construct jointly certified representation–policy pairs without requiring a sufficient representation. Select the largest certificate and provide a computed bound on the selected policy's loss relative to the full-history optimum. This choice maximizes a guarantee; it need not maximize the unknown policy return.

4 APPROACH: CONDITIONAL CALIBRATION AND COR-FQI

4.1 CALIBRATED SETS AND THE ROBUST BACKUP

Definition 2 (Conditional calibration)**.** The local sets $\{\widehat{\mathcal{U}}_t^e(z,a)\}$ are calibrated at level $1-\delta$ if there is a dataset event $\mathcal{C}$ with $\mathbb{P}_{\mathcal{D}}(\mathcal{C}) \ge 1-\delta$ on which

$$\nu_t^{e,M}(\cdot \mid h,a) \in \widehat{\mathcal{U}}_t^e(e_t(h),a) \tag{6}$$

for every $e \in \mathcal{E}$, $t \in [T]$, $h \in \mathsf{H}_t^\circ$, $a \in \mathcal{A}$, and $M \in \mathfrak{M}$.

For a bounded measurable $v : \mathcal{Z}_{t+1}^e \to \mathbb{R}$, assume each local set is nonempty and consists of probability measures on $[0,1] \times \mathcal{Z}_{t+1}^e$. Define the robust backup

$$(\mathcal{T}_t^e v)(z,a) = \inf_{\nu \in \widehat{\mathcal{U}}_t^e(z,a)} \mathbb{E}_{(R,Z')\sim\nu}[R + v(Z')]. \tag{7}$$

This is the robust Bellman backup (Iyengar, 2005; Wiesemann et al., 2013; Panaganti et al., 2022) on joint reward–next-state distributions. It permits an independent adverse distribution at each local input, a *rectangular* outer relaxation of $\mathfrak{M}$. The resulting value can be conservative even with exact estimates, but history-wise containment still makes it a lower bound.

4.2 CALIBRATION FROM RECORDED OUTCOMES

In the *tabular setting*, represented states and outcome supports are finite. Each tuple (e,t,z,a) collects the transitions assigned to its representation, stage, state, and action. Its empirical distribution, count, support, and radius are denoted by $\widehat{\nu}_t^e$, n_t^e, $\mathcal{Y}_t^e$, and ε_t^e. Let G count all distinct calibrated tuples, including auxiliary finer labels used below. Stage sharing is allowed only under the conditions in Assumption 1.

Assumption 1 (Finite supports and variation bounds)**.** The sets $\mathcal{Z}_t^e$ and the known outcome supports $\mathcal{Y}_t^e(z,a)$ are finite. For all $h,h' \in C_t^e(z)$ and $M \in \mathfrak{M}$, the learner is given bounds $b_t^e(z,a), \rho_t^e(z,a) \in [0,1]$ such that

$$\mathrm{TV}(\nu_t^e(\cdot \mid h,a), \nu_t^e(\cdot \mid h',a)) \le b_t^e(z,a),$$
$$\mathrm{TV}(\nu_t^{e,M}(\cdot \mid h,a), \nu_t^e(\cdot \mid h,a)) \le \rho_t^e(z,a).$$

Every nominal and deployment conditional is supported on $\mathcal{Y}_t^e(z,a)$; logged outcomes use the nominal kernels, and the tuple to which each outcome belongs is determined before the outcome is observed. If stages share counts, the first inequality must also hold across the stages sharing them (Appendix A.1).

Here b is the variation bound and ρ the deployment radius.

Let $\widehat{\nu}_t^e(\cdot \mid z, a)$ be the empirical distribution of the $n = n_t^e(z, a)$ outcomes of a tuple and let $k = |\mathcal{Y}_t^e(z, a)|$. For $n \geq 1$ and $k > 1$, set

$$u(n, k) = \min\left\{1, \sqrt{\frac{\log((2^k - 2)Gn(n+1)/\delta)}{2n}}\right\}. \tag{8}$$

Set $u(0, k) = 1$ for $k > 1$ and $u(n, 1) = 0$, with any center on the known support at zero visits, and count each shared tuple once in G. Define

$$\varepsilon_t^e(z, a) = \min\{1, u(n, k) + b_t^e(z, a) + \rho_t^e(z, a)\}, \tag{9}$$

$$\widehat{\mathcal{U}}_t^e(z, a) = \{\nu \in \mathscr{P}(\mathcal{Y}_t^e(z, a)) : \mathrm{TV}(\nu, \widehat{\nu}_t^e(\cdot \mid z, a)) \leq \varepsilon_t^e(z, a)\}. \tag{10}$$

The radius adds the estimation radius, the variation bound, and the deployment radius. The first and third terms are the estimation and deployment radii of robust offline RL in the Markov case (Zhou et al., 2021; Blanchet et al., 2023; Shi & Chi, 2024); the second is the term that pooling by a representation requires. The number of candidates enters only through G in $u(n, k)$.

Theorem 3 (Simultaneous calibration of separately pooled estimates)**.** *Under Assumption 1, the sets in* (10) *are conditionally calibrated. One event of probability at least* $1 - \delta$ *serves all candidates, admissible histories, deployment environments, and observed visit counts, including counts chosen by a stopping rule.*

The proof (Appendix A.4) centers a martingale concentration bound at the average of the visited conditional distributions and uses b to reach any history's distribution from that average; the pooled outcomes need not be independent or identically distributed. Combining more visits can reduce estimation uncertainty, while discarding history can increase the variation bound. Neither effect alone determines the best certificate.

4.3 Sources of the variation bound

Mixture statistics alone do not identify the required history-to-center radius. A valid bound can come from system structure: in a delayed controller, the probability that an omitted command executes bounds its effect (Lemma B.8). Alternatively, a finer label $x = \vartheta_t(h)$ recorded in the log partitions each represented state into smaller cells. The following result uses their estimates on the *coarse outcome coordinates*.

Proposition 4 (Two-sided variation bounds from finer logged information)**.** *Fix a visited cell* $C = C_t^e(z)$ *at action* a *and a finite, exhaustive finer partition* $\{C_x : x \in \mathcal{X}_C\}$ *into nonempty parts, fixed before calibration. Include its tuples in* G*. Let* $\widehat{\nu}, \widehat{\nu}_x$ *estimate the same coarse outcome with radii* u_0, u_x *from* (8) *at their respective counts and support sizes. Let* α_{vis} *be the supremum TV distance of nominal history-conditionals from the coarse visit average, and let* $\varsigma(x) \geq 0$ *bound their TV distance from the fine visit average within* C_x*. Define*

$$\varepsilon_{\mathrm{fine}} = \min\{1, \max_{x \in \mathcal{X}_C}[\varsigma(x) + u_x + \mathrm{TV}(\widehat{\nu}_x, \widehat{\nu})]\}.$$

On the common concentration event,

$$\max_{x \in \mathcal{X}_C}[\mathrm{TV}(\widehat{\nu}_x, \widehat{\nu}) - u_x - u_0]_+ \leq \alpha_{\mathrm{vis}} \leq \min\{1, \varepsilon_{\mathrm{fine}} + u_0\}.$$

The ball at $\widehat{\nu}$ *with radius* $\min\{1, \rho_t^e(z, a) + \varepsilon_{\mathrm{fine}}\}$ *contains every deployment conditional in* C*, simultaneously over the calibrated tuples. Predictive sufficiency of the finer label permits* $\varsigma(x) = 0$*. Unvisited non-singleton finer tuples use* $u_x = 1$*, making the upper bound uninformative.*

A positive lower bound establishes insufficiency (Appendix A.5). The fine-log radius replaces the basic radius; it still requires an exhaustive partition and valid residual bounds, including at unvisited histories. The visit-average quantity α_{vis} is distinct from the population-mixture radius in Proposition 1.

4.4 THE COR-FQI PROCEDURE

Algorithm 1 uses Assumption 1 and known μ. Write $\widehat{Q}_t^e, \widehat{V}_t^e$ for its pessimistic action and state values. Replacing the infimum in (7) by a supremum, retaining the action maximum, gives optimistic values $\widehat{V}_t^{+,e}$ and $U(e) = \lambda(e)[\widehat{V}_1^{+,e}]$ in (A.4). Updates cover every represented state; ties use a fixed ordering and backups use exact mass transfer (Appendix B.6).

Algorithm 1 COR-FQI, exact tabular specialization

Input: $\mathcal{D}, \mathcal{E}, T, \delta, \mu$; supports and calibration inputs from Assumption 1 or Proposition 4

for each $e \in \mathcal{E}$ **do**
  Calibrate: estimate local distributions; build sets by (9)–(10) or Proposition 4.
  Initialize $\widehat{V}_{T+1}^e \leftarrow 0$.
  for $t = T, \dots, 1$ **do**
    Robust backup: $\widehat{Q}_t^e \leftarrow \mathcal{T}_t^e \widehat{V}_{t+1}^e$ using (7).
    $\widehat{\pi}_t^e(z) \leftarrow \arg\max_a \widehat{Q}_t^e(z,a)$; $\widehat{V}_t^e(z) \leftarrow \max_a \widehat{Q}_t^e(z,a)$.
  end for
  Certify: $L(e) \leftarrow \mathbb{E}_{Z\sim\lambda(e)}[\widehat{V}_1^e(Z)]$.
  Upper audit: compute the optimistic control value $U(e)$ by (A.4).
end for
Select: $\widehat{e} \leftarrow \arg\max_{e\in\mathcal{E}} L(e)$.
return $(\widehat{e}, \widehat{\pi}^{\widehat{e}})$, $L(\widehat{e})$, and the interval $[\max_e L(e), \min_e U(e)]$.

5 THEORETICAL GUARANTEES

5.1 POLICY CERTIFICATES

Let $\tau_t = T - t + 1$ for $t \in [T+1]$. Define the set diameters and their horizon-weighted *accumulated diameter*

$$d_t^e(z,a) = \sup_{\nu_1,\nu_2 \in \widehat{\mathcal{U}}_t^e(z,a)} \mathrm{TV}(\nu_1,\nu_2), \qquad \bar{d}_t^e = \sup_{z,a} d_t^e(z,a),$$

$$B(e) = \min\left\{T, \sum_{t=1}^{T} \tau_t \bar{d}_t^e\right\}.$$

Theorem 5 (Simultaneous policy certificate and full-history comparison)**.** *Suppose $\mathcal{Z}_t^e$ and all local outcome supports are finite, and the nonempty compact local sets are jointly calibrated on $\mathcal{C}$. The exact recursion in Algorithm 1 then satisfies, simultaneously for every $e \in \mathcal{E}$,*

$$0 \le L(e) \le J(\widehat{\pi}^e) \le J^\star(e) \le J^\star \le U(e), \qquad U(e) - L(e) \le B(e), \tag{11}$$

Hence (4) *holds with probability at least $1-\delta$ and $J^\star \in [L(e), U(e)]$ for every e; no rectangularity of $\mathfrak{M}$ is required.*

The proof (Appendix A.6) works backward along full histories. The true conditional is in the local set, so the infimum bounds the returned policy from below, and the supremum bounds every full-history policy from above. The represented process need not be Markov. The computed width is at most $B(e)$ and bounds the total gap in (5); the same sets can certify a policy learned by another method (Corollary B.1).

5.2 REPRESENTATION SELECTION

Theorem 6 (Certificate-based representation selection)**.** *Let $\widehat{e} = \arg\max_{e\in\mathcal{E}} L(e)$, ties broken as in Algorithm 1. On the same event as Theorem 5,*

$$L(\widehat{e}) \le J(\widehat{\pi}^{\widehat{e}}), \quad J^\star - J(\widehat{\pi}^{\widehat{e}}) \le \min_{e\in\mathcal{E}} U(e) - \max_{e\in\mathcal{E}} L(e) \le \min_{e\in\mathcal{E}}\{U(e) - L(e)\} \le \min_{e\in\mathcal{E}} B(e). \tag{12}$$

On the common event, $J^\star \le \min_e U(e)$ and $J(\widehat{\pi}^{\widehat{e}}) \ge \max_e L(e)$; subtraction proves the result (Appendix A.7). The interval remains valid after selection, even without a sufficient member of the family. Its width measures what is certified, not the actual regret against the best learned policy. Environmental uncertainty and the rectangular relaxation can prevent the width from vanishing with more data.

Refinement does not by itself order certificates. A representation e_f *refines* e_c if measurable maps κ_t satisfy $e_{c,t}(h) = \kappa_t(e_{f,t}(h))$ for every admissible history; let $\bar{\kappa}_t(r, z'_f) = (r, \kappa_{t+1}(z'_f))$ map fine outcomes to coarse ones.

Proposition 7 (What refinement orders)**.** *If e_f refines e_c, then (i) $\Pi(e_c) \subseteq \Pi(e_f)$ and $J^\star(e_c) \le J^\star(e_f)$; (ii) the pairwise TV diameter of a fine cell's conditionals, after mapping outcomes by $\bar{\kappa}_t$, is at most that of the coarse cell containing it; (iii) if $\{\nu \circ (\bar{\kappa}_t)^{-1} : \nu \in \widehat{\mathcal{U}}_t^{e_f}(z_f, a)\} \subseteq \widehat{\mathcal{U}}_t^{e_c}(\kappa_t(z_f), a)$ for every (t, z_f, a), then $L(e_c) \le L(e_f)$; (iv) without this nesting, refinement alone does not order the certificates.*

With separately estimated sets (Appendix A.8), fewer samples per fine tuple can offset its smaller variation bound and permit a coarse-to-fine switch. Calibration from shared reference balls instead forces the reference certificate to dominate every coarsening (Proposition B.2).

Relation to RFQI. With a predictively sufficient representation, known rewards, and matched estimates, sets, and optimization, COR-FQI and RFQI have identical exact updates (Proposition B.3). A three-stage separation example shows why correction can matter: an uncorrected mixture update reports .75 for a policy returning .05; calibrated learning instead returns a policy with both value and certificate .59. This is a constructed example, not an empirical result, and another policy in the same class returns .9; calibration ensures validity rather than exact optimality.

5.3 Approximate backups

Theorem 8 (Certification under approximate backups)**.** *Fix e and a realized dataset. Let $\widetilde{\pi}_t(z)$ be a measurable deterministic e-policy, and let $\widetilde{Q}_t(z, a)$ be bounded measurable functions. Set $\widetilde{V}_t(z) = \widetilde{Q}_t(z, \widetilde{\pi}_t(z))$ and $\widetilde{V}_{T+1} = 0$. Suppose nonnegative constants c_t satisfy, for every admissible h, a, M with $z = e_t(h)$, $\widetilde{Q}_t(z, a) \le \mathbb{E}_{(R,Z')\sim\nu_t^{e,M}(\cdot|h,a)}[R + \widetilde{V}_{t+1}(Z')] + c_t$. Then*

$$\mathbb{E}_{Z\sim\lambda(e)}[\widetilde{V}_1(Z)] - \sum_{t=1}^{T} c_t \le J(\widetilde{\pi}). \tag{13}$$

Joint premises over $\mathcal{E}$ on a dataset event of probability $1 - \delta$ give joint bounds that survive data-dependent selection (Appendix A.9).

The fitted-iteration bound (Theorem B.4) separates uniform inner-solver, regression, and action-selection errors; a corresponding result (Proposition B.5) gives a discounted extension. These are conditional error-propagation results. Training loss and held-out prediction error do not establish their premises, and no data-derived uniform error bound is supplied for the approximate implementations below.

6 Experimental results

The finite theory proves validity; FrozenLake tests whether the bounds are informative enough to guide control. Exact return evaluation reveals how memory selection changes with data and why certificates remain conservative. We then test empirical robustness with function approximation.

6.1 FrozenLake: useful bounds and the choice of memory

The agent has ten steps to reach a goal on a slippery 4×4 grid while avoiding holes. Commands execute with ages zero, one, or two with probabilities $(.6, .3, .1)$. Representation e_m, $m \in \{0, 1, 2\}$, retains the tile and last m commands. Only e_2 is predictively sufficient; its larger state space makes estimation harder.

We use 200 independent uniform-command datasets, each with eight nested trajectory counts $N = 10^3$–3×10^6. Calibration combines estimation error u, omitted-history variation $(b(0), b(1), b(2)) = (.4, .1, 0)$, and deployment radius $\rho = .023$ for altered slip and command execution. Supports and variation bounds are supplied; confidence is 95% (Appendix C.1).

Reading the bounds. Figure 1(a) compares e_2's lower certificate $L = L(e_2)$ with its exact worst-case goal-reaching probability $J = J(\widehat{\pi}^{e_2})$. The optimistic control value $U = U(e_2)$ bounds the full-history optimum from above (Theorem 5). Increasing data should raise L and narrow $[L, U]$.

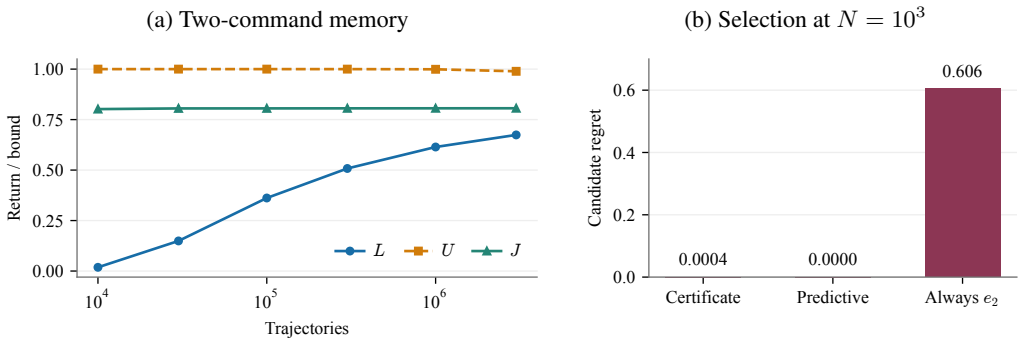


Figure 1: FrozenLake, means over 200 datasets. (a) L lower-bounds policy return J; U upper-bounds the full-history optimum. (b) Return lost relative to the best learned policy at $N = 10^3$; lower is better. "Certificate" selects the largest $L(e)$.

All tested calibrated sets contain the relevant history-conditional distributions, and all certificates satisfy $L \leq J$. These are implementation checks; Theorem 3 supplies the probability guarantee. At $N = 3 \times 10^6$, e_2 gives $L = .674$ and $J = .806$, while $W = \min_e U(e) - \max_e L(e) = .315$ bounds the selected policy's full-history optimality gap (Theorem 6). Thus *more data improves certification even when return barely changes*, but the bound does not establish near-optimality ($J^\star \approx .80648$).

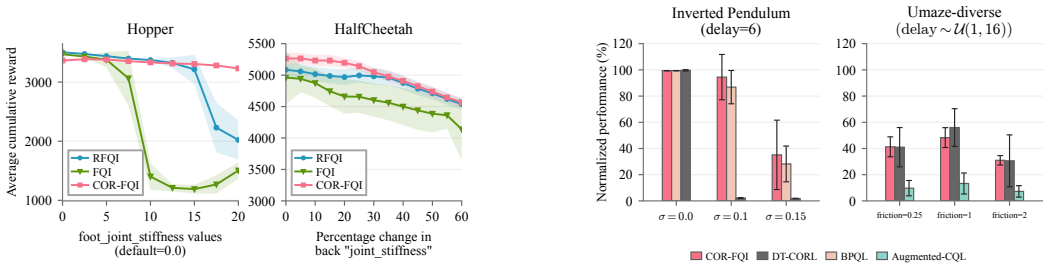
Choosing memory with limited data. Figure 1(b) reports regret against the best learned candidate, whereas W compares with the full-history optimum. At $N = 10^3$, always choosing e_2 loses .606; maximizing the certificate loses .0004. Selection switches from e_1 to e_2 at $N = 3 \times 10^4$: *more history need not improve a finite-data certificate* (Proposition 7). Predictive selection chooses the shortest memory within 5% of the best held-out next-tile log loss.

Table 1: Calibration at $N = 3 \times 10^6$. Values average 200 uniform datasets. Containment of the relevant history-conditional laws holds in all (Yes) or none (No).

Construction	Memory m	Reported value	Return J	Containment
No history term $(u + \rho)$	1	.652	.806	No
Calibrated $(u + b + \rho)$	1	.333	.806	Yes
Mixture-model DP	2	.847	.806	No
Calibrated $(u + b + \rho)$	2	.674	.806	Yes

What calibration contributes. Table 1 separates validity from performance. For e_1, e_2, calibration leaves policies unchanged: *its benefit here is a justified bound*, also obtainable afterward (Corollary B.1). Without delay, matched calibrated RFQI and COR-FQI agree (Proposition B.3). Finer e_2 logs certify e_1 at .343 (Proposition 4); doubling b lowers the supplied-bound certificate from .333 to .133 but leaves the fine-log certificate unchanged. Excluding e_2 only from selection, with calibration data and confidence allocation fixed, retains e_1's .333 certificate and .806 return: *no selectable representation need be sufficient* (Theorem 6). Exact nominal laws and zero estimation error yield $L(e_2) = .764 < .806$: conservatism persists without sampling error, and *loose certificates need not mean poor control*.

6.2 Function approximation: robustness on control and memory benchmarks



(a) MDP environments

(b) Delayed environments

Figure 2: Fixed-policy robustness evaluation. (a) *Hopper* and *HalfCheetah* under joint-stiffness perturbations. (b) *Inverted Pendulum*: action-noise standard deviation σ, observation delay 6; *Umaze-diverse*: friction perturbations under stochastic delays. *Inverted Pendulum* returns are normalized by the maximum episode return and expressed as percentages.

We test COR-FQI policies under physical perturbations, delays, and distribution shift, obtaining competitive performance against the tested baselines (Figures 2–3).

Robustness to parameter uncertainty. On *Hopper* and *HalfCheetah*, we compare with RFQI and FQI at zero action delay. We vary *Hopper*'s foot-joint stiffness from 0 to 20 and *HalfCheetah*'s back-joint stiffness by 0–60% from nominal values. COR-FQI maintains high returns over a wider stiffness range on *Hopper*, while its advantage over RFQI narrows as perturbations increase on *HalfCheetah*.

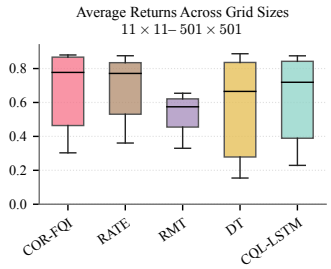


Figure 3: *MiniGrid* returns across various grid sizes.

Robustness to delays. We evaluate delayed control under action noise and friction changes. On *Inverted Pendulum*, we compare with BPQL (Kim et al., 2023) and DT-CORL (Zhan et al., 2026) under a six-step delay and Gaussian action noise with standard deviation $\sigma \in \{0, 0.1, 0.15\}$. On delayed *Umaze-diverse*, we compare with DT-CORL and Augmented-CQL, which applies CQL (Kumar et al., 2020) to an augmented state representation, at sliding-friction coefficients $\{0.25, 1, 2\}$. COR-FQI remains competitive with the tested baselines.

Robustness to distribution shift. *MiniGrid Memory* requires retaining an initial cue through a corridor; larger grids increase navigation length and memory duration. We train on 41×41 grids and test sizes 11×11–501×501 without retraining, with a 500-step limit, to assess transfer under distribution shift in a partially observed task. We compare with RATE (Cherepanov et al., 2026), RMT (Bulatov et al., 2022), DT (Chen et al., 2021), and CQL-LSTM. Our implementation uses a RATE encoder and a heuristic prediction-disagreement penalty. Figure 3 shows that COR-FQI's median return across grid sizes is comparable to RATE's and higher than those of the other baselines, indicating competitive transfer under partial observability.

7 Conclusion

In this work, we presented COR-FQI for choosing a history representation and a policy from offline data. We established simultaneous return certificates and a computable bound on the selected policy's loss relative to the full-history optimum, without requiring a predictively sufficient representation. Experiments demonstrated informative certificates and memory selection on FrozenLake, together with competitive empirical robustness on control and memory benchmarks. Our guarantees require justified bounds on history variation, and the certification conditions for function approximation remain unverified in our experiments. Future work will focus on validating these conditions for certified representation–policy selection and developing less conservative calibration methods.

AI USE STATEMENT

In this work, the authors used generative AI tools for grammar correction, sentence rephrasing, and language polishing of the manuscript's abstract, introduction, and conclusion sections. We used generative AI tools to help fix errors in Python code snippets. We did not use generative AI tools to conceptualize core research ideas, experiments, or to prove theoretical claims. All AI-assisted text and code outputs were thoroughly reviewed, manually tested, and edited by the human authors, who take full responsibility for the accuracy and integrity of the final work.

REFERENCES

Jose Blanchet, Miao Lu, Tong Zhang, and Han Zhong. Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *Advances in Neural Information Processing Systems*, 36:66845–66859, 2023.

David Bruns-Smith and Angela Zhou. Robust fitted-Q-evaluation and iteration under sequentially exogenous unobserved confounders. *arXiv preprint arXiv:2302.00662*, 2023.

Aydar Bulatov, Yury Kuratov, and Mikhail Burtsev. Recurrent memory transformer. *Advances in neural information processing systems*, 35:11079–11091, 2022.

Siddharth Chandak, Pratik Shah, Vivek S Borkar, and Parth Dodhia. Reinforcement learning in non-Markovian environments. *Systems & Control Letters*, 185:105751, 2024.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

Egor Cherepanov, Aleksei Staroverov, Alexey Kovalev, and Aleksandr Panov. Recurrent action transformer with memory. In *International Conference on Learning Representations*, volume 2026, pp. 129379–129407, 2026.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 2978–2988, 2019.

Aditya Dave, Andreas A Malikopoulos, et al. Approximate information states for worst case control and learning in uncertain systems. *IEEE Transactions on Automatic Control*, 70(1):127–142, 2024.

Shi Dong, Benjamin Van Roy, and Zhengyuan Zhou. Simple agent, complex environment: Efficient reinforcement learning with agent states. *Journal of Machine Learning Research*, 23(255):1–54, 2022.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Ruiquan Huang, Yingbin Liang, and Jing Yang. Robust offline reinforcement learning for non-Markovian decision processes. *IEEE Transactions on Information Theory*, 71(9):7208–7228, 2025.

Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.

Nan Jiang, Alex Kulesza, and Satinder Singh. Abstraction selection in model-based reinforcement learning. In *International Conference on Machine Learning*, pp. 179–188. PMLR, 2015.

Jangwon Kim, Hangyeol Kim, Jiwook Kang, Jongchan Baek, and Soohee Han. Belief projection-based reinforcement learning for environments with delayed feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33:1179–1191, 2020.

Jonathan N Lee, George Tucker, Ofir Nachum, Bo Dai, and Emma Brunskill. Oracle inequalities for model selection in offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35:28194–28207, 2022.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Shiau Hong Lim and Arnaud Autef. Kernel-based reinforcement learning in robust Markov decision processes. In *International Conference on Machine Learning*, pp. 3973–3981. PMLR, 2019.

Odalric-Ambrym Maillard, Daniil Ryabko, and Rémi Munos. Selecting the state-representation in reinforcement learning. *Advances in Neural Information Processing Systems*, 24, 2011.

Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Robust reinforcement learning using offline data. *Advances in neural information processing systems*, 35:32211–32224, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Laixi Shi and Yuejie Chi. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *Journal of Machine Learning Research*, 25(200):1–91, 2024.

Jayakumar Subramanian, Amit Sinha, Raihan Seraj, and Aditya Mahajan. Approximate information state for approximate planning and reinforcement learning in partially observed systems. *Journal of Machine Learning Research*, 23(12):1–83, 2022.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

Simon Zhan, Qingyuan Wu, Zhaofeng Wang, Frank Yang, Xiangyu Shi, Chao Huang, and Qi Zhu. Belief-based offline reinforcement learning for delay-robust policy optimization. In *International Conference on Learning Representations*, volume 2026, pp. 134796–134820, 2026.

Zhengqing Zhou, Zhengyuan Zhou, Qinxun Bai, Linhai Qiu, Jose Blanchet, and Peter Glynn. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3331–3339. PMLR, 2021.

A APPENDIX

Appendix A proves the main results, Appendix B gives the supporting structural and approximation results, Appendix C specifies the experimental protocols and implementation scope, Appendix D gives individual environment and implementation details, and Appendix E compares related works.

A.1 CONVENTIONS USED IN THE PROOFS

All observation, history, and representation spaces are standard Borel, with their Borel sigma-algebras, and all maps and kernels are measurable. Finite spaces have the discrete sigma-algebra. An environment specifies kernels $P_t^M(dr, do' \mid h, a)$; together with a policy and the initial distribution, these define the trajectory distribution. Write $h \oplus (a, r, o')$ for the appended history. The admissible domains H_t° have full initial probability under both μ and the logging reset distribution and are forward closed under every action and every nominal or deployment kernel. Kernel versions, including those at zero-probability histories, are fixed on these domains. Thus every history-wise comparison below applies along every deployment trajectory. Conditional one-step expectations $\mathbb{E}_M[\cdot \mid h, a]$ integrate this kernel with $H_{t+1} = h \oplus (a, R_t, O_{t+1})$. Unless specified otherwise, action sums range over $\mathcal{A}$, environment infima over $\mathfrak{M}$, representation extrema over $\mathcal{E}$, and finite-horizon sums over the indicated stages.

The true conditional return is $V_t^{\pi,M}(h) = \mathbb{E}_{M,\pi}[\sum_{s=t}^{T} R_s \mid H_t = h]$, with $V_{T+1}^{\pi,M} = 0$.

Write δ_x for unit mass at x, $\mathbf{1}\{E\}$ for the indicator of an event E, and $\|v\|_\infty = \sup_x |v(x)|$ on the stated domain. Vector norms $\|\cdot\|_1, \|\cdot\|_2$ are the usual ℓ_1 and Euclidean norms; $[\cdot]_+$ acts coordinatewise on vectors. The notation $\mathrm{Ber}(p)$ denotes the Bernoulli distribution with success probability p. Function inequalities are pointwise. For outcome measures, R, Z' denote reward and successor coordinates: $\nu[R + v(Z')] = \int (r + v(z'))\nu(dr, dz')$.

The deployment family $\mathfrak{M}$ is fixed independently of the data. Supports and variation bounds are inputs valid throughout that family; they need not identify it or its cross-time restrictions. Local independent choices from the calibrated sets form an outer relaxation, not an exact representation of a general nonrectangular family.

Predictive sufficiency is imposed separately within each environment. For a nonrectangular deployment set, it need not imply $J^\star(e) = J^\star$: additional history may reveal which environment is operating. Conditional calibration and representation loss are therefore distinct.

For the finite construction, outcome supports are specified before observing the data and contain every nominal and deployment outcome in the corresponding cell. Stages may share counts only when their outcome coordinates agree and the variation bound applies between every contributing nominal conditional distribution and every target history's distribution. Otherwise, counts are stage specific. The terminal value is zero on the entire terminal representation space. In the experiments, e_m indexes memory length; the stage index in the general map e_t remains t.

Table 2: Objects that must remain distinct in the analysis.

Notation	Meaning
$\bar{\nu}$, α	Population logged mixture and its history-to-center radius
α_{vis}	Radius about the random visit-average conditional distribution
b, ρ, ε	Nominal pairwise variation bound, deployment bound, and total set radius
$\widehat{V}^e$, $\widehat{V}^{+,e}$	Pessimistic and optimistic control values over the calibrated sets
$\widetilde{V}^e$	Fitted value; a certificate requires explicit approximation control
$J^\star$, $U(e)$	Full-history optimal worst-case return and its optimistic upper bound
W, $V_t^{\pi,M}$	Selected optimality-gap bound and true history-conditional return

For probability measures p, q on a common space, a bounded measurable function g, and a measurable map F, write $p[g] = \int g\, dp$ and $p \circ F^{-1}$ for its image under F. Let $\mathrm{span}(g) = \sup g - \inf g$. We use

$$|p[g] - q[g]| \le \mathrm{span}(g)\, \mathrm{TV}(p, q), \tag{A.1}$$

$$\mathrm{TV}(p \circ F^{-1}, q \circ F^{-1}) \le \mathrm{TV}(p, q). \tag{A.2}$$

The first inequality follows by subtracting $(\sup g + \inf g)/2$ and applying the dual definition of TV; the second follows by taking inverse images in its setwise definition. Probabilities of calibration events refer to the data distribution $\mathbb{P}_{\mathcal{D}}$; deployment comparisons condition on the realized dataset.

A.2 Proof of Proposition 1

Proof. (i) Exact radius. Suppress the fixed indices, and set $C = C_t^e(z)$ and $\alpha = \sup_{h\in C} \mathrm{TV}(\nu_h, \bar{\nu})$. Here B ranges over measurable outcome sets. Convexity and admissibility of point masses give

$$\mathrm{TV}(\bar{\nu}_\psi, \bar{\nu}) = \sup_B \left| \int_C [\nu_h(B) - \bar{\nu}(B)]\, \psi(dh) \right| \le \int_C \mathrm{TV}(\nu_h, \bar{\nu})\, \psi(dh) \le \alpha,$$
$$\sup_{\psi \in \mathscr{P}(C)} \mathrm{TV}(\bar{\nu}_\psi, \bar{\nu}) \ge \sup_{h\in C} \mathrm{TV}(\bar{\nu}_{\delta_h}, \bar{\nu}) = \alpha.$$

Thus (3) holds. This is minimality over all reweightings; the policy-induced subclass may be smaller.

(ii) Nonidentification. Fix one represented state and action, two histories with weights $(1/2, 1/2)$, and a fixed successor state, identifying each joint outcome with its Bernoulli reward. The two models satisfy

$$(\nu_{h_1}, \nu_{h_2}) = (\mathrm{Ber}(.2), \mathrm{Ber}(.8)), \qquad \bar{\nu} = \mathrm{Ber}(.5), \qquad \alpha = .3,$$
$$(\nu'_{h_1}, \nu'_{h_2}) = (\mathrm{Ber}(.5), \mathrm{Ber}(.5)), \qquad \bar{\nu}' = \mathrm{Ber}(.5), \qquad \alpha' = 0.$$

The represented state–action–outcome laws coincide, but the required radii differ. □

A.3 Concentration for trajectory outcomes

Index the fixed collection of calibrated tuples by $\ell \in [G]$ and all logged transitions, including transitions from successive episodes, chronologically by $j = 1, 2, \ldots$. Let $\mathscr{F}_{j-1}$ contain completed outcomes before transition j, let $\mathscr{F}_j^-$ additionally contain the current reset observation, history, issued action, and logging randomization already used, and let $\mathscr{F}_j$ include the resulting reward and next observation. Thus $\mathscr{F}_{j-1} \subseteq \mathscr{F}_j^- \subseteq \mathscr{F}_j \subseteq \mathscr{F}_{j+1}^-$. The indicator $I_{j,\ell} = \mathbf{1}\{\text{transition } j \text{ belongs to tuple } \ell\}$ is $\mathscr{F}_j^-$-measurable. Conditional on $\mathscr{F}_j^-$, a visited outcome $Y_{j,\ell}$ has distribution $p_{j,\ell}$ given by the fixed nominal kernel and lies in the known support of size k_ℓ. Adaptive actions and resets are allowed provided this conditional-kernel identity holds; anticipatory labels depending on future outcomes are not allowed. Write $N_{j,\ell} = \sum_{i=1}^{j} I_{i,\ell}$. For $n \ge 1$, let $\sigma_{n,\ell} = \inf\{j : N_{j,\ell} = n\}$, with $\inf \varnothing = \infty$. When $\sigma_{n,\ell} < \infty$, define

$$\widehat{p}_{n,\ell} = \frac{1}{n} \sum_{j \le \sigma_{n,\ell}} I_{j,\ell} \delta_{Y_{j,\ell}}, \qquad \bar{p}_{n,\ell} = \frac{1}{n} \sum_{j \le \sigma_{n,\ell}} I_{j,\ell} p_{j,\ell}.$$

These are the empirical distribution and the average of the visited conditional distributions; the latter need not be a fixed behavior mixture. Values on nonvisits may be assigned arbitrarily.

Lemma A.1 (Simultaneous concentration at random visit counts)**.** *With probability at least* $1 - \delta$*, simultaneously for every tuple and every finite positive visit count,*

$$\mathrm{TV}(\widehat{p}_{n,\ell}, \bar{p}_{n,\ell}) \le u_{n,\ell} := u(n, k_\ell), \tag{A.3}$$

Here u *is defined in* (8)*, including its singleton convention. A singleton support has radius zero; an unvisited nonsingleton tuple uses radius one and any fixed supported center.*

Proof. Fix ℓ, a nonempty proper subset B of its support, and $\zeta > 0$. Set

$$X_j = I_{j,\ell}\big(\mathbf{1}_{\{Y_{j,\ell} \in B\}} - p_{j,\ell}(B)\big), \qquad Z_j^{\exp} = \exp\left(\zeta \sum_{i \le j} X_i - \frac{\zeta^2}{8} N_{j,\ell} \right).$$

Conditional mean zero and range length $I_{j,\ell}$ imply, by Hoeffding's lemma,

$$\mathbb{E}[X_j \mid \mathscr{F}_j^-] = 0, \quad \mathbb{E}[e^{\zeta X_j} \mid \mathscr{F}_j^-] \le e^{\zeta^2 I_{j,\ell}/8}, \quad \mathbb{E}[Z_j^{\exp} \mid \mathscr{F}_{j-1}] \le Z_{j-1}^{\exp}, \qquad Z_0^{\exp} = 1.$$

For $n \geq 1$, $s > 0$, and $\zeta = 4s$, Ville's inequality yields

$$\mathbb{P}(\sigma_{n,\ell} < \infty,\ \widehat{p}_{n,\ell}(B) - \bar{p}_{n,\ell}(B) > s)$$
$$\leq \mathbb{P}\left(\sup_j Z_j^{\exp} > e^{2ns^2}\right) \leq e^{-2ns^2},$$
$$\mathbb{P}(\sigma_{n,\ell} < \infty,\ \mathrm{TV}(\widehat{p}_{n,\ell}, \bar{p}_{n,\ell}) > s) \leq (2^{k_\ell} - 2)e^{-2ns^2}.$$

The last step uses $\mathrm{TV}(p, q) = \max_B [p(B) - q(B)]$ on a finite support. Therefore

$$\mathbb{P}(\exists \ell, n :\ \sigma_{n,\ell} < \infty,\ \mathrm{TV}(\widehat{p}_{n,\ell}, \bar{p}_{n,\ell}) > u_{n,\ell}) \leq \sum_{\ell=1}^{G} \sum_{n=1}^{\infty} \frac{\delta}{Gn(n+1)} = \delta.$$

Terms with $u_{n,\ell} = 1$ cannot fail; singleton distributions coincide. No conditioning on the realized visit count is used. □

A.4 Proof of Theorem 3

Proof. On the event of Lemma A.1, fix any e, t, h, a, M and the index ℓ of tuple $(e, t, e_t(h), a)$, with shared tuples identified when applicable. Let $z = e_t(h)$, $n = n_t^e(z, a)$, $b = b_t^e(z, a)$, $\rho = \rho_t^e(z, a)$, $\nu_0 = \nu_t^e(\cdot \mid h, a)$, and $\nu_M = \nu_t^{e,M}(\cdot \mid h, a)$. For $n \geq 1$, Assumption 1 gives

$$\mathrm{TV}(\nu_0, \bar{p}_{n,\ell}) \leq \frac{1}{n} \sum_{j \leq \sigma_{n,\ell}} I_{j,\ell}\, \mathrm{TV}(\nu_0, p_{j,\ell}) \leq b,$$
$$\mathrm{TV}(\nu_M, \widehat{p}_{n,\ell}) \leq \mathrm{TV}(\nu_M, \nu_0) + \mathrm{TV}(\nu_0, \bar{p}_{n,\ell}) + \mathrm{TV}(\bar{p}_{n,\ell}, \widehat{p}_{n,\ell})$$
$$\leq \rho + b + u_{n,\ell}.$$

At this realized count, $\widehat{p}_{n,\ell} = \widehat{\nu}_t^e(\cdot \mid z, a)$. Together with $\mathrm{TV} \leq 1$ and the support premise, this implies

$$\nu_M \in \mathscr{P}(\mathcal{Y}_t^e(e_t(h), a)), \qquad \mathrm{TV}(\nu_M, \widehat{\nu}_t^e(\cdot \mid z, a)) \leq \varepsilon_t^e(e_t(h), a) \implies \nu_M \in \widehat{\mathcal{U}}_t^e(e_t(h), a).$$

At zero visits a nonsingleton set is the full supported simplex; a singleton support determines the distribution exactly. Since the event is common and the indices were arbitrary, (6) holds simultaneously with probability at least $1 - \delta$. □

A.5 Proof of Proposition 4

The finer labels are fixed before calibration and form a finite, exhaustive partition of each admissible coarse history set. Their empirical distributions use the same coarse reward–next-state coordinates. Count every distinct coarse or auxiliary fine estimation tuple in G, even if its representation is not selectable. A sufficient residual premise is a supplied bound on the pairwise TV diameter within each finer set; convexity then bounds distance to any visit average by the same number. This premise must cover unvisited histories. Predictive sufficiency of the finer label in the nominal environment makes that residual zero. If residual bounds are estimated, their own simultaneous validity event must also be included in the confidence allocation.

Proof. Fix t, e, z, a with visited cell $C = C_t^e(z)$ and let $\rho = \rho_t^e(z, a)$. Write $\nu(h), \nu^M(h)$ for its nominal and deployment conditionals and $\bar{\nu}_{\mathrm{vis}}, \bar{\nu}_{\mathrm{vis}}(x)$ for the coarse and fine visit averages. First suppose all finer tuples are visited; unvisited tuples are treated below. On the common event of Lemma A.1,

$$\mathrm{TV}(\widehat{\nu}, \bar{\nu}_{\mathrm{vis}}) \leq u_0, \qquad \mathrm{TV}(\widehat{\nu}_x, \bar{\nu}_{\mathrm{vis}}(x)) \leq u_x, \qquad \sup_{h \in C_x} \mathrm{TV}(\nu(h), \bar{\nu}_{\mathrm{vis}}(x)) \leq \varsigma(x).$$

Lower bound. Convexity over the fine visits and $C_x \subseteq C$ give

$$\mathrm{TV}(\bar{\nu}_{\mathrm{vis}}(x), \bar{\nu}_{\mathrm{vis}}) \leq \sup_{h \in C_x} \mathrm{TV}(\nu(h), \bar{\nu}_{\mathrm{vis}}) \leq \alpha_{\mathrm{vis}},$$
$$\mathrm{TV}(\widehat{\nu}_x, \widehat{\nu}) \leq u_x + \alpha_{\mathrm{vis}} + u_0,$$
$$\max_{x \in \mathcal{X}_C} [\mathrm{TV}(\widehat{\nu}_x, \widehat{\nu}) - u_x - u_0]_+ \leq \alpha_{\mathrm{vis}}.$$

Upper bound and calibration. For every $h \in C_x$,

$$\begin{aligned} \mathrm{TV}(\nu(h), \widehat{\nu}) &\le \varsigma(x) + u_x + \mathrm{TV}(\widehat{\nu}_x, \widehat{\nu}), \\ \sup_{h \in C} \mathrm{TV}(\nu(h), \widehat{\nu}) &\le \varepsilon_{\mathrm{fine}}, \\ \alpha_{\mathrm{vis}} &\le \min\{1, \varepsilon_{\mathrm{fine}} + u_0\}, \\ \sup_{M \in \mathfrak{M},\, h \in C} \mathrm{TV}(\nu^M(h), \widehat{\nu}) &\le \min\{1, \rho + \varepsilon_{\mathrm{fine}}\}. \end{aligned}$$

Here the second line uses the exhaustive partition and $\mathrm{TV} \le 1$. Nominal sufficiency gives $\nu(h) = \bar{\nu}_{\mathrm{vis}}(x)$ on C_x, hence $\varsigma(x) = 0$.

For an unvisited nonsingleton fine tuple, $u_x = 1$ makes its lower term zero and $\varepsilon_{\mathrm{fine}} = 1$; no fine visit average is needed. For a known singleton, use its known point mass and $u_x = \varsigma(x) = 0$; the same deductions hold. An unvisited nonsingleton coarse tuple uses radius one, with α_{vis} undefined. All defined bounds hold on the same event. □

A.6 Proof of Theorem 5

Proof. Fix $\mathcal{D} \in \mathcal{C}$ and e, and write $v_t = \widehat{V}_t^e$, $\widehat{\pi} = \widehat{\pi}^e$. Define $v_{T+1}^+ = 0$ and

$$\begin{aligned} v_t^+(z) &= \max_{a \in \mathcal{A}} \sup_{\nu \in \widehat{\mathcal{U}}_t^e(z,a)} \nu[R + v_{t+1}^+(Z')], \\ U(e) &= \lambda(e)[v_1^+]. \end{aligned} \tag{A.4}$$

Backward induction from $v_{T+1} = v_{T+1}^+ = 0$ gives $0 \le v_t \le v_t^+ \le \tau_t$.

History-wise induction. For fixed M, every history policy satisfies

$$V_t^{\pi,M}(h) = \sum_a \pi_t(a \mid h) \int \Big[r + V_{t+1}^{\pi,M}(h \oplus (a, r, o')) \Big] \, P_t^M(dr, do' \mid h, a).$$

The induction claims are $v_t(e_t(h)) \le V_t^{\widehat{\pi},M}(h)$ and $V_t^{\pi,M}(h) \le v_t^+(e_t(h))$ for all π, h. Both hold at $T+1$. Assuming them at $t+1$, let $z = e_t(h)$ and $\widehat{a} = \widehat{\pi}_t(z)$. Forward closure and calibration imply

$$\begin{aligned} V_t^{\widehat{\pi},M}(h) &\ge \nu_t^{e,M}(\cdot \mid h, \widehat{a})[R + v_{t+1}(Z')] \ge \inf_{\nu \in \widehat{\mathcal{U}}_t^e(z,\widehat{a})} \nu[R + v_{t+1}(Z')] = v_t(z), \\ V_t^{\pi,M}(h) &\le \sum_a \pi_t(a \mid h) \nu_t^{e,M}(\cdot \mid h, a)[R + v_{t+1}^+(Z')] \\ &\le \sum_a \pi_t(a \mid h) \sup_{\nu \in \widehat{\mathcal{U}}_t^e(z,a)} \nu[R + v_{t+1}^+(Z')] \le v_t^+(z). \end{aligned}$$

Integrating against μ and taking the required infima and suprema yields

$$0 \le L(e) \le \inf_M J_M(\widehat{\pi}^e) \le \sup_{\pi \in \Pi(e)} \inf_M J_M(\pi) \le \sup_{\pi \in \Pi} \inf_M J_M(\pi) \le U(e).$$

No exchange of $\sup_\pi$ and $\inf_M$, Markov assumption, or rectangularity is used.

Width induction. Set $D_t = \|v_t^+ - v_t\|_\infty$ and $g_t(r, z') = r + v_{t+1}(z')$, so $\mathrm{span}(g_t) \le \tau_t$. For every (z, a), (A.1) gives

$$\begin{aligned} &\sup_{\nu \in \widehat{\mathcal{U}}_t^e(z,a)} \nu[R + v_{t+1}^+(Z')] - \inf_{\nu \in \widehat{\mathcal{U}}_t^e(z,a)} \nu[g_t] \\ &\quad \le D_{t+1} + \sup_{\nu, \nu' \in \widehat{\mathcal{U}}_t^e(z,a)} (\nu[g_t] - \nu'[g_t]) \le D_{t+1} + \tau_t \bar{d}_t^e, \end{aligned}$$

$$D_t \le D_{t+1} + \tau_t \bar{d}_t^e \le \sum_{s=t}^{T} \tau_s \bar{d}_s^e, \qquad D_{T+1} = 0.$$

Thus

$$U(e) - L(e) \le \min\{T, D_1\} \le \min\left\{ T, \sum_{t=1}^{T} \tau_t \bar{d}_t^e \right\} = B(e). \tag{A.5}$$

The argument holds for every e on $\mathcal{C}$, whose probability is at least $1 - \delta$. □

The width reflects estimation, history variation, and the local relaxation of deployment uncertainty; exact centers or predictive sufficiency alone need not remove it. A positive radius need not create a gap when supported continuation values coincide. A total-return cap H_{ret} sharpens the trivial width cap only if it holds throughout the relaxed recursion; FrozenLake's single goal-entry reward and zero terminal continuation give $H_{\mathrm{ret}} = 1$.

A.7 Proof of Theorem 6

Proof. On the common event $\mathcal{C}$, $\widehat{e} \in \arg\max_e L(e)$ and Theorem 5 give

$$J(\widehat{\pi}^{\widehat{e}}) \geq L(\widehat{e}) = \max_e L(e), \qquad J^\star \leq \min_e U(e),$$
$$\begin{aligned} J^\star - J(\widehat{\pi}^{\widehat{e}}) &\leq \min_e U(e) - \max_e L(e) \\ &\leq \min_e \{U(e) - L(e)\} \leq \min_e B(e). \end{aligned}$$

All inequalities hold simultaneously before the data-dependent choice; hence no additional confidence penalty is needed. □

A.8 Proof of Proposition 7

Proof. (i) Policy classes. For $\pi^{e_c} \in \Pi(e_c)$, define

$$\pi_t^{e_f}(a \mid z_f) = \pi_t^{e_c}(a \mid \kappa_t(z_f)) \implies \pi_t^{e_f}(a \mid e_{f,t}(h)) = \pi_t^{e_c}(a \mid e_{c,t}(h)).$$

The induced trajectory laws agree in every M. Therefore $\Pi(e_c) \subseteq \Pi(e_f)$ and $J^\star(e_c) \leq J^\star(e_f)$.

(ii) Conditional variation on common coordinates. Fix $M \in \mathfrak{M} \cup \{M_0\}$. Write $C_f = C_t^{e_f}(z_f) \subseteq C_c = C_t^{e_c}(\kappa_t(z_f))$. Since $\nu_t^{e_f,M}(\cdot \mid h, a) \circ \bar{\kappa}_t^{-1} = \nu_t^{e_c,M}(\cdot \mid h, a)$,

$$\begin{aligned} &\sup_{h,h' \in C_f} \mathrm{TV}\Big(\nu_t^{e_f,M}(\cdot \mid h, a) \circ \bar{\kappa}_t^{-1}, \nu_t^{e_f,M}(\cdot \mid h', a) \circ \bar{\kappa}_t^{-1}\Big) \\ &\quad \leq \sup_{h,h' \in C_c} \mathrm{TV}\Big(\nu_t^{e_c,M}(\cdot \mid h, a), \nu_t^{e_c,M}(\cdot \mid h', a)\Big). \end{aligned}$$

The same restriction argument applies to distance from any common coarse center; it does not compare separately centered radii.

(iii) Certificate induction. Let Z_f' denote the fine successor coordinate. Terminal values are zero. Assume $\widehat{V}_{t+1}^{e_c} \circ \kappa_{t+1} \leq \widehat{V}_{t+1}^{e_f}$. The stipulated set inclusion implies

$$\begin{aligned} \widehat{Q}_t^{e_c}(\kappa_t(z_f), a) &\leq \inf_{\nu \in \widehat{\mathcal{U}}_t^{e_f}(z_f,a)} \nu[R + \widehat{V}_{t+1}^{e_c}(\kappa_{t+1}(Z_f'))] \\ &\leq \inf_{\nu \in \widehat{\mathcal{U}}_t^{e_f}(z_f,a)} \nu[R + \widehat{V}_{t+1}^{e_f}(Z_f')] = \widehat{Q}_t^{e_f}(z_f, a), \\ \widehat{V}_t^{e_c} \circ \kappa_t &\leq \widehat{V}_t^{e_f}. \end{aligned}$$

Backward induction and $\lambda(e_c) = \lambda(e_f) \circ \kappa_1^{-1}$ give $L(e_c) \leq L(e_f)$.

(iv) Counterexample. Take $T = 1$, one action, reward $R = 1$, and any strict refinement. Let † denote the terminal represented state. Identifying r with $(r, \dagger)$, choose

$$\widehat{\mathcal{U}}^{e_c} = \{\delta_1\}, \qquad \widehat{\mathcal{U}}^{e_f} = \mathscr{P}(\{0,1\}) \quad \implies \quad L(e_c) = 1 > 0 = L(e_f).$$

Both sets contain the true distribution, so calibration alone does not order certificates. □

A.9 Proof of Theorem 8

Proof. Fix M and let $V_t = V_t^{\widetilde{\pi},M}$ and $E_t^{\mathrm{pol}} = \sum_{s=t}^T c_s$, with $E_{T+1}^{\mathrm{pol}} = 0$. We prove

$$\widetilde{V}_t(e_t(h)) \leq V_t(h) + E_t^{\mathrm{pol}} \qquad (\forall t, h)$$

by backward induction. At $T+1$, all three quantities are zero. Suppose the claim holds at $t+1$ and set $a = \widetilde{\pi}_t(e_t(h))$. The premise and forward closure give

$$\begin{aligned}\widetilde{V}_t(e_t(h)) &= \widetilde{Q}_t(e_t(h), a) \\ &\le \mathbb{E}_M[R_t + \widetilde{V}_{t+1}(e_{t+1}(H_{t+1})) \mid h, a] + c_t \\ &\le \mathbb{E}_M[R_t + V_{t+1}(H_{t+1}) \mid h, a] + E^{\mathrm{pol}}_{t+1} + c_t = V_t(h) + E^{\mathrm{pol}}_t.\end{aligned}$$

Consequently, for every M,

$$\lambda(e)[\widetilde{V}_1] - \sum_{t=1}^{T} c_t \le J_M(\widetilde{\pi}) \implies \lambda(e)[\widetilde{V}_1] - \sum_{t=1}^{T} c_t \le \inf_M J_M(\widetilde{\pi}).$$

On a joint event the deduction holds for all e, including any subsequently selected representation. □

B Supporting results used in the paper

B.1 Certification after policy learning

Corollary B.1 (Certification of an arbitrary learned policy)**.** *Under Theorem 5's conditions, let $\pi^e \in \Pi(e)$ be any data-dependent policy. Define its calibrated evaluation by $v^e_{T+1} = 0$ and*

$$v^e_t(z) = \sum_a \pi^e_t(a \mid z)(\mathcal{T}^e_t v^e_{t+1})(z, a), \qquad L^{\mathrm{eval}}(e, \pi^e) = \lambda(e)[v^e_1].$$

On $\mathcal{C}$, $L^{\mathrm{eval}}(e, \pi^e) \le J(\pi^e)$ simultaneously over the representations. This remains valid after selecting a pair by its evaluated lower bound.

Proof. Fix $\mathcal{D} \in \mathcal{C}$ and the learned policies. At $T+1$, $v^e_{T+1} = V^{\pi^e, M}_{T+1} = 0$. If $v^e_{t+1} \circ e_{t+1} \le V^{\pi^e, M}_{t+1}$, calibration gives, for $z = e_t(h)$,

$$\begin{aligned}v^e_t(z) &= \sum_a \pi^e_t(a \mid z) \inf_{\nu \in \widehat{\mathcal{U}}^e_t(z,a)} \nu[R + v^e_{t+1}(Z')] \\ &\le \sum_a \pi^e_t(a \mid z) \mathbb{E}_M[R_t + V^{\pi^e, M}_{t+1}(H_{t+1}) \mid h, a] = V^{\pi^e, M}_t(h).\end{aligned}$$

Induction and integration imply $L^{\mathrm{eval}}(e, \pi^e) \le \inf_M J_M(\pi^e) = J(\pi^e)$. The argument is deterministic on $\mathcal{C}$ and applies to every returned policy, so training and selection incur no further confidence penalty. □

B.2 A shared reference forces certificate dominance

Let f be a finite reference representation that is predictively sufficient in every nominal and deployment environment, and let every candidate be a coarsening $e_t = \kappa^e_t \circ f_t$. For $x = f_t(h)$, write $K^M_t(\cdot \mid x, a)$ for the conditional distribution of $(R_t, f_{t+1}(H_{t+1}))$, with nominal kernel K_t and probability estimate $\widehat{K}_t$. Let $u_t(x, a), \rho_t(x, a) \ge 0$ be supplied estimation and deployment bounds; suppress (x, a) in kernel arguments when unambiguous. The reference outcome support $\mathcal{Y}^f_t(x, a)$ supports $\widehat{K}_t$ and every specified true conditional distribution. Suppose $\mathrm{TV}(K^M_t, K_t) \le \rho_t(x, a)$ and the estimation event $\mathcal{C}_{\mathrm{ref}} = \{\mathrm{TV}(\widehat{K}_t, K_t) \le u_t(x, a),\ \forall t, x, a\}$ has probability at least $1 - \delta$.

let $F^e_t(r, x') = (r, \kappa^e_{t+1}(x'))$. Here $x \in \mathcal{Z}^f_t$ and $x \mapsto \kappa^e_t(x)$ maps reference states to coarse states. For each coarse cell, choose a center $\widehat{\nu}^e_t(\cdot \mid z, a)$ on a support containing $\bigcup_{x:\kappa^e_t(x)=z} F^e_t(\mathcal{Y}^f_t(x, a))$. Denote the TV balls restricted to these respective supports by $\widehat{\mathcal{U}}^f_t(x, a)$ and $\widehat{\mathcal{U}}^e_t(z, a)$. Define the reference ball with radius $\min\{1, u_t(x, a) + \rho_t(x, a)\}$ at $\widehat{K}_t$, and the coarse ball with radius

$$\begin{aligned}\varepsilon^e_{\mathrm{ref},t}(z, a) = \min\Big\{1, \max_{x:\kappa^e_t(x)=z} \big[u_t(x, a) + \rho_t(x, a) \\ + \mathrm{TV}\Big(\widehat{K}_t(\cdot \mid x, a) \circ (F^e_t)^{-1}, \widehat{\nu}^e_t(\cdot \mid z, a)\Big)\big]\Big\}. \qquad \text{(B.1)}\end{aligned}$$

Proposition B.2 (Shared-reference calibration and dominance)**.** *The reference and coarse balls are jointly calibrated on $\mathcal{C}_{\mathrm{ref}}$. Moreover, for every dataset, their sets satisfy*

$$\{q \circ (F_t^e)^{-1} : q \in \widehat{\mathcal{U}}_t^f(x,a)\} \subseteq \widehat{\mathcal{U}}_t^e(\kappa_t^e(x), a).$$

Under exact optimization and the common initial distribution, $\widehat{V}_t^f(x) \geq \widehat{V}_t^e(\kappa_t^e(x))$ and $L(f) \geq L(e)$. If f is selectable, no coarsening can have a strictly larger certificate; ties remain possible.

Proof. On $\mathcal{C}_{\mathrm{ref}}$, for every (t, x, a, M),

$$\mathrm{TV}(K_t^M, \widehat{K}_t) \leq \mathrm{TV}(K_t^M, K_t) + \mathrm{TV}(K_t, \widehat{K}_t) \leq \rho_t(x,a) + u_t(x,a), \qquad K_t^M \in \widehat{\mathcal{U}}_t^f(x,a).$$

For any dataset, $q \in \widehat{\mathcal{U}}_t^f(x,a)$, and $z = \kappa_t^e(x)$, suppress the fixed kernel arguments (x,a) and center arguments (z,a). Contraction gives

$$\begin{aligned}\mathrm{TV}(q \circ (F_t^e)^{-1}, \widehat{\nu}_t^e) &\leq \mathrm{TV}(q, \widehat{K}_t) + \mathrm{TV}(\widehat{K}_t \circ (F_t^e)^{-1}, \widehat{\nu}_t^e) \\ &\leq u_t(x,a) + \rho_t(x,a) + \mathrm{TV}(\widehat{K}_t \circ (F_t^e)^{-1}, \widehat{\nu}_t^e).\end{aligned}$$

The cap $\mathrm{TV} \leq 1$ and the support inclusion imply

$$q \circ (F_t^e)^{-1} \in \widehat{\mathcal{U}}_t^e(z,a).$$

Taking $q = K_t^M$ proves coarse calibration on $\mathcal{C}_{\mathrm{ref}}$. Applying Proposition 7(iii) to the set inclusion yields

$$\widehat{V}_t^f \geq \widehat{V}_t^e \circ \kappa_t^e, \qquad L(f) \geq L(e).$$

Thus strict coarse-to-reference certificate crossover is impossible under this shared-reference construction. □

B.3 RFQI AGREEMENT AND SEPARATION

Proposition B.3 (Markov agreement and separation under conditional variation)**.** *(i) For a predictively sufficient representation and known deterministic state–action rewards, COR-FQI and RFQI have identical operators, values, and policies when their nominal estimates, uncertainty sets, supports, action domains, optimizers, initialization, and tie-breaking agree. This holds for finite-horizon recursion and discounted iteration. (ii) There is a three-stage process with an insufficient representation in which the behavior-centered RFQI update without a conditional-variation correction overstates its policy's worst-case return, while calibrated COR-FQI returns a better policy with a valid certificate.*

Proof. (i) Operator identity. Let $r_t(z,a)$ be the known reward, $\mathcal{P}_t(z,a)$ the common transition set, and $\mathcal{T}_t^{\mathrm{RFQI}}$ the RFQI backup. For $F_{t,z,a}(z') = (r_t(z,a), z')$, take the joint-distribution set $\widehat{\mathcal{U}}_t^e(z,a) = \{p \circ F_{t,z,a}^{-1} : p \in \mathcal{P}_t(z,a)\}$. Then

$$(\mathcal{T}_t^e v)(z,a) = r_t(z,a) + \inf_{p \in \mathcal{P}_t(z,a)} p[v] = (\mathcal{T}_t^{\mathrm{RFQI}} v)(z,a).$$

Equal terminal values (or initialization), operators, optimizers, and tie-breaking imply equal values and policies by induction. For discount $\gamma \in (0,1)$, replacing v by γv proves the discounted case. For known next-state rewards, the same identity holds with integrand $r_t(z,a,z') + v(z')$ when both methods use that convention; this is not a claim about the released RFQI code.

(ii) Separation. Let $T = 3$ and $\mathcal{A} = \{0,1\}$; name these actions $+, -$ at stage one and risky/safe at stage two. Stage-one actions $+, -$ give rewards $0, r_{\mathrm{bonus}} = .05$; stage-two actions risky/safe give zero reward; stage three pays $Y \in \{0, c, 1\}$, $c = .6$, and terminates. The representation is constant at stages one and two and retains Y at stage three. Let h_+, h_- be the stage-two histories following the two stage-one actions, and let $\mathbb{P}_b$ denote the behavior-policy probability. The nominal laws of Y given the stage-two history and action are

	h_+	h_-
risky	δ_1	δ_0
safe	δ_c	δ_c

$$\mathbb{P}_b(H_2 = h_+ \mid A_2 = a) = p_{\mathrm{beh}} = .8 \quad (a = \mathrm{risky}, \mathrm{safe}).$$

The behavior uses the same positive risky probability at both histories. Deployment independently perturbs each stage-two law within TV radius $\rho_{\mathrm{sep}} = .1$ on $\{0, c, 1\}$; other kernels remain fixed. Write $Q^{\mathrm{RFQI}}, Q^{\mathrm{COR}}$ for the stage-two action values, V^{RFQI} for the reported initial value, and L^{COR} for the calibrated certificate; stage three has no effective action choice. With exact population centers, the uncorrected update satisfies

$$Q^{\mathrm{RFQI}}(\mathrm{risky}) = p_{\mathrm{beh}} - \rho_{\mathrm{sep}} = .7 > c(1 - \rho_{\mathrm{sep}}) = .54 = Q^{\mathrm{RFQI}}(\mathrm{safe}),$$
$$\pi^{\mathrm{RFQI}} = (-, \mathrm{risky}), \qquad V^{\mathrm{RFQI}} = .05 + .7 = .75, \qquad J(\pi^{\mathrm{RFQI}}) = .05.$$

In this fixed cell, abbreviate the action-specific radius and history-to-center variation by $\varepsilon(a), \alpha(a)$. For calibrated COR-FQI,

$$\alpha(\mathrm{risky}) = \max\{p_{\mathrm{beh}}, 1 - p_{\mathrm{beh}}\} = .8, \qquad \varepsilon(\mathrm{risky}) = .9, \qquad Q^{\mathrm{COR}}(\mathrm{risky}) = 0,$$
$$\alpha(\mathrm{safe}) = 0, \qquad \varepsilon(\mathrm{safe}) = .1, \qquad Q^{\mathrm{COR}}(\mathrm{safe}) = .54,$$
$$\pi^{\mathrm{COR}} = (-, \mathrm{safe}), \qquad L^{\mathrm{COR}} = J(\pi^{\mathrm{COR}}) = .59.$$

The risky ball contains δ_0. Using the pairwise bound $b(\mathrm{risky}) = 1$ instead of the exact history-to-center radius .8 gives the same policy and certificate. For the policy $\pi^{+,\mathrm{risky}}$ choosing $+$ and then risky,

$$J(\pi^{+,\mathrm{risky}}) = 1 - \rho_{\mathrm{sep}} = .9, \qquad .9 - L^{\mathrm{COR}} = .31,$$

so the example demonstrates validity and strict policy improvement, while retaining conservatism within the same representation class. □

B.4 Fitted iteration with explicit error terms

For deterministic e-policies $\widetilde{\pi}^e$, let $\widetilde{V}^e_t(z) = \widetilde{Q}^e_t(z, \widetilde{\pi}^e_t(z))$ and $\widetilde{V}^e_{T+1} = 0$. For infinite state or action spaces, assume standard Borel spaces, measurable kernels and selected policies, bounded measurable value functions, and robust backups and action suprema that preserve bounded measurability. Approximate action maximizers must be measurable. These are additional assumptions beyond the finite construction; uniform diameters use suprema.

Theorem B.4 (Fitted COR-FQI under uniform approximation-error conditions)**.** *Assume conditional calibration and the preceding measurability conditions. For each stage, let $y^e_t(z, a)$ be a lower inner target and let $c_{\mathrm{in},t}, c_{\mathrm{fit},t}, c_{\mathrm{act},t} \geq 0$ bound inner-solver, regression, and action-selection errors. Suppose, uniformly over admissible (z, a),*

$$\mathcal{T}^e_t \widetilde{V}^e_{t+1} - c_{\mathrm{in},t} \leq y^e_t \leq \mathcal{T}^e_t \widetilde{V}^e_{t+1}, \tag{B.2}$$
$$\|\widetilde{Q}^e_t - y^e_t\|_\infty \leq c_{\mathrm{fit},t}, \tag{B.3}$$
$$\widetilde{V}^e_t(z) \geq \sup_a \widetilde{Q}^e_t(z, a) - c_{\mathrm{act},t}, \tag{B.4}$$

Then

$$\widetilde{L}(e) := \lambda(e)[\widetilde{V}^e_1] - \sum_t c_{\mathrm{fit},t} \leq J(\widetilde{\pi}^e), \tag{B.5}$$
$$J^\star - \widetilde{L}(e) \leq B(e) + \sum_t (c_{\mathrm{in},t} + 2c_{\mathrm{fit},t} + c_{\mathrm{act},t}). \tag{B.6}$$

If the premises hold jointly over the family, selecting $\widetilde{e} \in \arg\max_{e \in \mathcal{E}} \widetilde{L}(e)$ with fixed tie-breaking bounds the selected policy's full-history optimality gap by the minimum of the right-hand sides of (B.6)*.*

Proof. Fix e and suppress it. The inner-target and regression conditions imply

$$\mathcal{T}_t \widetilde{V}_{t+1} - c_{\mathrm{in},t} - c_{\mathrm{fit},t} \leq \widetilde{Q}_t \leq \mathcal{T}_t \widetilde{V}_{t+1} + c_{\mathrm{fit},t}.$$

Policy certificate. Calibration gives, for every h, a, M,

$$\widetilde{Q}_t(e_t(h), a) \leq \nu^{e,M}_t(\cdot \mid h, a)[R + \widetilde{V}_{t+1}(Z')] + c_{\mathrm{fit},t}.$$

Theorem 8 with $c_t = c_{\mathrm{fit},t}$ yields (B.5).

Error induction. Define the exact control recursion $\widehat{V}_{T+1} = 0$, $\widehat{V}_t = \sup_a \mathcal{T}_t \widehat{V}_{t+1}$ and

$$E_t = \sum_{s=t}^{T} (c_{\mathrm{in},s} + c_{\mathrm{fit},s} + c_{\mathrm{act},s}), \qquad E_{T+1} = 0.$$

We prove $\widetilde{V}_t \geq \widehat{V}_t - E_t$. The terminal case is equality. If it holds at $t+1$, monotonicity and $\mathcal{T}_t(v + c) = \mathcal{T}_t v + c$ for constants $c \in \mathbb{R}$ imply

$$\begin{aligned}
\widetilde{V}_t &\geq \sup_a \widetilde{Q}_t - c_{\mathrm{act},t} \\
&\geq \sup_a \mathcal{T}_t \widetilde{V}_{t+1} - c_{\mathrm{in},t} - c_{\mathrm{fit},t} - c_{\mathrm{act},t} \\
&\geq \widehat{V}_t - E_{t+1} - c_{\mathrm{in},t} - c_{\mathrm{fit},t} - c_{\mathrm{act},t} = \widehat{V}_t - E_t.
\end{aligned}$$

The history and diameter comparisons in Theorem 5, using suprema where necessary, give $J^\star \leq \lambda(e)[\widehat{V}_1] + B(e)$. Hence

$$\begin{aligned}
J^\star - \widetilde{L}(e) &\leq B(e) + \lambda(e)[\widehat{V}_1 - \widetilde{V}_1] + \sum_t c_{\mathrm{fit},t} \\
&\leq B(e) + \sum_t (c_{\mathrm{in},t} + 2c_{\mathrm{fit},t} + c_{\mathrm{act},t}).
\end{aligned}$$

On the joint event, $J(\widetilde{\pi}^{\widetilde{e}}) \geq \max_e \widetilde{L}(e)$; subtracting from $J^\star$ and minimizing the preceding bound proves selection. □

These are uniform error premises, not consequences of empirical regression loss. Nonnegative returns permit the stronger certificate $\max\{0, \widetilde{L}(e)\}$. With restricted action search, the gap bound requires either an error bound against the full action supremum or a correspondingly restricted comparator; policy certification itself does not require action optimality.

B.5 Discounted stationary extension

Extend the history domains and kernels to all $t \geq 1$. Fix a stationary representation e with finite range $\mathcal{Z}^e$ and a finite action set; let rewards lie in $[0, 1]$ and $\gamma \in (0, 1)$. Suppressing e, for nonempty sets $\widehat{\mathcal{U}}(z, a) \subseteq \mathscr{P}([0, 1] \times \mathcal{Z}^e)$, define the action and control backups by

$$(\mathcal{T}_\gamma v)(z, a) = \inf_{\nu \in \widehat{\mathcal{U}}(z,a)} \nu[R + \gamma v(Z')], \qquad (\mathcal{T}_\gamma^{\mathrm{ctrl}} v)(z) = \max_a (\mathcal{T}_\gamma v)(z, a).$$

For a stationary deterministic e-policy π, let $(\mathcal{T}_\gamma^\pi v)(z) = (\mathcal{T}_\gamma v)(z, \pi(z))$.

Proposition B.5 (Discounted contraction and lower bound)**.** *The control operator $\mathcal{T}_\gamma^{\mathrm{ctrl}}$ is a γ-contraction in the uniform norm and has a unique bounded fixed point $\widehat{V}_\gamma$. Let π be its greedy policy with fixed tie-breaking. If $\mathcal{L}_M((R_t, e(H_{t+1})) \mid H_t = h, A_t = a) \in \widehat{\mathcal{U}}(e(h), a)$ for every $t \geq 1$, $h \in \mathsf{H}_t^\circ$, $a \in \mathcal{A}$, and $M \in \mathfrak{M}$, then*

$$\widehat{V}_\gamma(e(h)) \leq \inf_{M \in \mathfrak{M}} \mathbb{E}_{M,\pi} \left[\sum_{j=0}^{\infty} \gamma^j R_{t+j} \,\middle|\, H_t = h \right].$$

Proof. Contraction. For bounded v, w,

$$\begin{aligned}
|(\mathcal{T}_\gamma v)(z, a) - (\mathcal{T}_\gamma w)(z, a)| &\leq \sup_{\nu \in \widehat{\mathcal{U}}(z,a)} \gamma |\nu[v(Z') - w(Z')]| \leq \gamma \|v - w\|_\infty, \\
\|\mathcal{T}_\gamma^{\mathrm{ctrl}} v - \mathcal{T}_\gamma^{\mathrm{ctrl}} w\|_\infty &\leq \gamma \|v - w\|_\infty.
\end{aligned}$$

Banach's fixed-point theorem gives a unique $\widehat{V}_\gamma$; iteration from zero gives $0 \leq \widehat{V}_\gamma \leq (1 - \gamma)^{-1}$.

History-wise comparison. Fix M, h. Greediness and calibration imply

$$\widehat{V}_\gamma(e(h)) = (\mathcal{T}_\gamma^\pi \widehat{V}_\gamma)(e(h)) \leq \mathbb{E}_{M,\pi}[R_t + \gamma \widehat{V}_\gamma(e(H_{t+1})) \mid h].$$

Induction on n, using the tower property at each step, yields

$$\widehat{V}_\gamma(e(h)) \le \mathbb{E}_{M,\pi}\left[\sum_{j=0}^{n-1}\gamma^j R_{t+j} + \gamma^n \widehat{V}_\gamma(e(H_{t+n}))\,\middle|\, h\right].$$

Since $0 \le \gamma^n \widehat{V}_\gamma \le \gamma^n/(1-\gamma) \to 0$ and the discounted return is bounded, let $n \to \infty$ and then take $\inf_M$. □

Continuous-action scope. The same argument allows standard Borel state and action spaces and any bounded reward interval, replacing $[0,1]$ in the outcome space. Assume the backups and action supremum preserve bounded measurability, and a stationary deterministic measurable policy π satisfies, for a constant $c_{\mathrm{act}} \ge 0$,

$$\widehat{V}_\gamma \le \mathcal{T}_\gamma^\pi \widehat{V}_\gamma + c_{\mathrm{act}}.$$

Let $J_\gamma(\pi \mid h)$ denote the worst-case discounted conditional return in Proposition B.5. Conditional containment and the same n-step induction give

$$\widehat{V}_\gamma(e(h)) \le \mathbb{E}_{M,\pi}\left[\sum_{j=0}^{n-1}\gamma^j R_{t+j} + \gamma^n \widehat{V}_\gamma(e(H_{t+n}))\,\middle|\, h\right] + c_{\mathrm{act}}\frac{1-\gamma^n}{1-\gamma},$$

$$J_\gamma(\pi \mid h) \ge \widehat{V}_\gamma(e(h)) - \frac{c_{\mathrm{act}}}{1-\gamma}.$$

An exact measurable maximizer gives $c_{\mathrm{act}} = 0$; an actor step or finite action sample alone does not establish this premise.

B.6 Exact TV backups and the quantile update

Enumerate the entire known finite admissible support $\{(r_j, z'_j)\}_{j=1}^K$, including atoms with zero empirical mass. Let $\widehat{\mathbf{p}} \in \mathbb{R}^K$ be the probability center, $\varepsilon \in [0,1]$ the TV radius, and $\mathbf{g} = (g_j)_{j=1}^K$ with $g_j = r_j + v(z'_j)$ for a bounded continuation v. Write $\mathbf{1}_K$ for the all-ones vector in $\mathbb{R}^K$. The optimization variables $\mathbf{p}, \mathbf{s}^+, \mathbf{s}^- \in \mathbb{R}^K$ are nonnegative; vector inequalities are coordinatewise.

Proposition B.6 (Exact TV linear program and mass transfer)**.** *The TV-robust backup equals the linear program*

$$\min_{\mathbf{p},\mathbf{s}^+,\mathbf{s}^- \ge 0} \mathbf{g}^\top \mathbf{p} \quad \textit{subject to} \quad \mathbf{1}_K^\top \mathbf{p} = 1, \quad \mathbf{p} - \widehat{\mathbf{p}} = \mathbf{s}^+ - \mathbf{s}^-, \quad \mathbf{1}_K^\top(\mathbf{s}^+ + \mathbf{s}^-) \le 2\varepsilon, \tag{B.7}$$

An optimizer transfers at most ε probability from largest-cost outcomes to smallest-cost outcomes, subject to available mass and capacity.

Proof. For probability vectors $\mathbf{p}, \widehat{\mathbf{p}}$, let $\mathbf{d} = \mathbf{p} - \widehat{\mathbf{p}}$. Then

$$\mathbf{1}_K^\top \mathbf{d} = 0, \qquad \mathbf{1}_K^\top [\mathbf{d}]_+ = \mathbf{1}_K^\top [-\mathbf{d}]_+ = \tfrac{1}{2}\|\mathbf{d}\|_1 = \mathrm{TV}(\mathbf{p}, \widehat{\mathbf{p}}).$$

Thus $\mathrm{TV}(\mathbf{p}, \widehat{\mathbf{p}}) \le \varepsilon$ implies LP feasibility with $\mathbf{s}^+ = [\mathbf{d}]_+$ and $\mathbf{s}^- = [-\mathbf{d}]_+$. Conversely, LP feasibility implies

$$\|\mathbf{p} - \widehat{\mathbf{p}}\|_1 \le \mathbf{1}_K^\top(\mathbf{s}^+ + \mathbf{s}^-) \le 2\varepsilon.$$

The feasible probability vectors coincide. For added and removed mass vectors of common total at most ε,

$$\mathbf{g}^\top \mathbf{p} = \mathbf{g}^\top \widehat{\mathbf{p}} + \mathbf{g}^\top \mathbf{s}^+ - \mathbf{g}^\top \mathbf{s}^-.$$

Moving an addition of size $s_0 > 0$ from cost g_i to an available $g_j \le g_i$ changes the objective by $s_0(g_j - g_i) \le 0$. Moving a removal from g_i to an available $g_j \ge g_i$ changes it by $-s_0(g_j - g_i) \le 0$. These exchanges preserve feasibility and order additions from lowest cost and removals from highest cost. Transfer stops at budget exhaustion or zero further reduction. □

Lemma B.7 (TV dual and conditional weighted quantile)**.** *Let $\widehat{\nu} = \sum_{j=1}^{K} w_j \delta_{y_j}$ be a probability distribution on $\mathcal{Y}$, with $w_j \geq 0$ and $\sum_j w_j = 1$. Write $\mathbf{w} = (w_j)_{j=1}^K$ and $\boldsymbol{\beta} = (\beta_j)_{j=1}^K$ for the weight and retained-mass vectors. For bounded measurable g, let $g_j = g(y_j)$, $g_{\min} = \inf_{y\in\mathcal{Y}} g(y)$, and $g_{\max} = \sup_{y\in\mathcal{Y}} g(y)$. For $\varepsilon \in [0,1]$,*

$$\inf_{\nu\in\mathscr{P}(\mathcal{Y}):\,\mathrm{TV}(\nu,\widehat{\nu})\leq\varepsilon} \nu[g] = \max_{\eta\in[g_{\min},g_{\max}]} \left\{ (1-\varepsilon)\eta - \sum_j w_j[\eta - g_j]_+ \right\} + \varepsilon g_{\min}. \tag{B.8}$$

For $q \in (0,1)$, a weighted q-quantile is a value η satisfying $\sum_{g_j<\eta} w_j \leq q \leq \sum_{g_j\leq\eta} w_j$. For $0 < \varepsilon < 1$, any weighted $(1-\varepsilon)$-quantile of the value–weight pairs (g_j, w_j) is a maximizer. At $\varepsilon = 0$ the value is $\sum_j w_j g_j$; at $\varepsilon = 1$ it is $g_{\min}$. The infimum need not be attained if $g_{\min}$ is not attained.

Proof. Retained-mass formulation. A feasible ν shares a submeasure of mass $1 - \mathrm{TV}(\nu,\widehat{\nu}) \geq 1-\varepsilon$ with $\widehat{\nu}$. Retaining exactly $1-\varepsilon$ of this mass gives

$$\nu[g] \geq \min_{0\leq\beta_j\leq w_j,\ \sum_j \beta_j = 1-\varepsilon} \sum_j \beta_j g_j + \varepsilon g_{\min}.$$

Conversely, take a minimizer $\boldsymbol{\beta}^\star$ and $y_n^\star \in \mathcal{Y}$ with $g(y_n^\star) \downarrow g_{\min}$. Then

$$\nu_n = \sum_j \beta_j^\star \delta_{y_j} + \varepsilon\delta_{y_n^\star}, \qquad \mathrm{TV}(\nu_n,\widehat{\nu}) \leq \varepsilon, \qquad \nu_n[g] \longrightarrow \sum_j \beta_j^\star g_j + \varepsilon g_{\min}.$$

This proves equality of the two infima, including when the support infimum is unattained.

Dual and quantile. For the mass-constraint multiplier $\eta \in \mathbb{R}$, define $F(\eta) = (1-\varepsilon)\eta - \sum_j w_j[\eta - g_j]_+$. Finite linear-program duality gives

$$\begin{aligned}\min_{0\leq\boldsymbol{\beta}\leq\mathbf{w},\ \mathbf{1}_K^\top\boldsymbol{\beta}=1-\varepsilon} \sum_j \beta_j g_j &= \sup_{\eta\in\mathbb{R}} \inf_{0\leq\boldsymbol{\beta}\leq\mathbf{w}} \left[(1-\varepsilon)\eta + \sum_j \beta_j(g_j - \eta) \right] \\ &= \sup_{\eta\in\mathbb{R}} F(\eta).\end{aligned}$$

Let F'_- and F'_+ denote the left and right derivatives, respectively. For $0 < \varepsilon < 1$, the concave function F is maximized precisely when

$$F'_-(\eta) = 1 - \varepsilon - \sum_{g_j<\eta} w_j \geq 0, \qquad F'_+(\eta) = 1 - \varepsilon - \sum_{g_j\leq\eta} w_j \leq 0,$$
$$\sum_{g_j<\eta} w_j \leq 1-\varepsilon \leq \sum_{g_j\leq\eta} w_j.$$

A weighted $(1-\varepsilon)$-quantile satisfies this condition and lies in $[g_{\min}, g_{\max}]$. Adding $\varepsilon g_{\min}$ proves (B.8). At $\varepsilon = 0$, the retained-mass vector is $\mathbf{w}$; at $\varepsilon = 1$, it is zero, giving the stated endpoint values. □

The support infimum concerns all admissible outcomes, not only observed ones. Replacing it by a justified lower bound $\underline{g} \leq g_{\min}$ and allowing $\eta \in [\underline{g}, g_{\max}]$ gives a lower target whose shortfall is exactly $\varepsilon(g_{\min} - \underline{g})$. Evaluating the correct dual at a shared batch threshold also gives a lower target, but can add an uncontrolled optimization shortfall. Dropping $\varepsilon\underline{g}$ is justified only when that lower bound is zero; negative rewards or critic values do not permit this omission.

B.7 Calibration bounds for the command-history channel

Let x contain the tile and command registers. For each possible age j, let $a'_j(x,a)$ be the command of age j, with $a'_0(x,a) = a$. An independent random age ι selects the executed command. let $p_d^{\mathrm{age}} = \mathbb{P}(\iota = d)$, $\theta = (q, p_{\mathrm{rep}})$, where $q \in [0,1]$ is the intended-direction probability and $p_{\mathrm{rep}} \in [0,1]$ the uniform command-replacement probability. Let $\mathrm{dir}_q(a',\cdot)$ assign probability q to the intended

direction a' and $(1-q)/2$ to each perpendicular direction. Write $\mathrm{Unif}(\mathcal{A})$ for the uniform distribution on commands, let $\theta_0 = (.9, 0)$, and define

$$\mathrm{dir}_\theta(a', \cdot) = (1 - p_{\mathrm{rep}})\mathrm{dir}_q(a', \cdot) + p_{\mathrm{rep}}\,\mathrm{Unif}(\mathcal{A}),$$

For a finite deployment family Θ, define

$$\rho(\Theta) = \max_{\theta\in\Theta,\, a'\in\mathcal{A}} \mathrm{TV}(\mathrm{dir}_\theta(a', \cdot), \mathrm{dir}_{\theta_0}(a', \cdot)). \tag{B.9}$$

Lemma B.8 (Calibration bounds for the command-history channel)**.** *Let $K(x, a; \theta)$ be the next-tile law and let $e_m(x)$ retain the tile and last m commands. If $e_m(x) = e_m(x')$, then*

$$\mathrm{TV}(K(x, a; \theta_0), K(x', a; \theta_0)) \le b(m) := \mathbb{P}(\iota > m),$$
$$\mathrm{TV}(K(x, a; \theta), K(x, a; \theta_0)) \le \rho(\Theta), \qquad \theta \in \Theta.$$

The same bounds hold for reward–next-representation outcomes.

Proof. Couple the two nominal transitions with the same age and directional randomness; let $S'_x, S'_{x'}$ be their next tiles. Then

$$\iota \le m \implies a'_\iota(x, a) = a'_\iota(x', a) \implies S'_x = S'_{x'},$$

The coupling inequality gives

$$\mathrm{TV}(K(x, a; \theta_0), K(x', a; \theta_0)) \le \mathbb{P}(S'_x \ne S'_{x'}) \le \mathbb{P}(\iota > m).$$

For fixed x, a, convexity gives

$$\mathrm{TV}\left(\sum_d p_d^{\mathrm{age}} \mathrm{dir}_\theta(a'_d(x, a), \cdot), \sum_d p_d^{\mathrm{age}} \mathrm{dir}_{\theta_0}(a'_d(x, a), \cdot)\right)$$
$$\le \sum_d p_d^{\mathrm{age}}\, \mathrm{TV}(\mathrm{dir}_\theta(a'_d(x, a), \cdot), \mathrm{dir}_{\theta_0}(a'_d(x, a), \cdot)) \le \rho(\Theta).$$

The maps from direction to next tile and from next tile to reward and next represented state are common to the compared laws. Applying (A.2) proves both outcome bounds. □

C Experimental protocols and implementation scope

This appendix records the calibration protocol and the function-approximation implementation supporting Section 6.

C.1 FrozenLake calibration protocol

For Section 6.1, use map `SFFF/FHFF/FFFF/FFHG`, horizon $T = 10$, and commands left/down/right/up. The state contains the tile and two command registers; the registers shift deterministically. The age probabilities and deployment parameters are

$$(p_0^{\mathrm{age}}, p_1^{\mathrm{age}}, p_2^{\mathrm{age}}) = (.6, .3, .1), \quad (b(0), b(1), b(2)) = (.4, .1, 0),$$
$$\theta_0 = (.9, 0), \qquad \Theta = \{.9, .89\} \times \{0, .02\}, \qquad \rho(\Theta) = .0228.$$

The bounds follow from Lemma B.8; each environment keeps one $\theta \in \Theta$ throughout an episode. Boundary moves remain in place; first goal entry pays one, and holes and the goal have zero continuation. Logging uses uniform commands and resets uniformly over nonterminal tiles and registers; evaluation starts at the standard start tile with both registers set to left.

The 200 independent seeds and nested sizes specified use $G = 1344$ shared state–action tuples and $\delta = .05$. Supports include admissible outcomes with zero counts; unvisited nonsingleton tuples use radius one and a uniform supported center. Fine-log calibration uses e_2 with zero residual and the complete finer partition. Excluding e_2 from selection retains its calibration tuples and G. Returns are evaluated on the full state.

C.2 Function approximation: implementation and scope

Algorithm 2 specifies the discounted implementation in Section 6.2, using local conditional mean embedding (CME) estimates and a shared batch threshold. Here represented states and actions are

finite-dimensional real vectors, $\mathcal{A}$ is bounded, and $\gamma \in (0,1)$. Let the nonempty set $\mathcal{I}_{\mathrm{tr}}$ index training transitions extracted from $\mathcal{D}$. For each recorded transition i, h_i, a_i, r_i, h_i' denote the current history, action, observed reward, and successor history. The recorded terminal indicator $\mathrm{done}_i \in \{0,1\}$ equals one when the dataset marks transition i as terminal; $1 - \mathrm{done}_i$ is the stored `not_done` mask. A time-limit truncation need not be marked terminal. For a fixed stationary representation e, set $z_i = e(h_i)$ and $z_i' = e(h_i')$. Write $\widetilde{Q}^{(1)}, \widetilde{Q}^{(2)}$ for the current critics and use subscript tgt for target copies, suppressing their dependence on e. A behavior variational autoencoder (VAE) supplies action proposals.

Algorithm 2 Approximate COR-FQI for continuous actions

Input: offline data $\mathcal{D}$, representations $\mathcal{E}$, discount γ, supplied TV radius $\varepsilon \in (0,1)$, kernel and optimization parameters

for each $e \in \mathcal{E}$ **do**
  Encode the training transitions using e.
  CME estimate: construct local conditional distributions by (C.1)–(C.2).
  Initialize twin critics, an actor, their target copies, and a behavior VAE.
  for each training update **do**
    Sample a nonempty minibatch of training indices $\mathcal{B}$ (with repetitions); update the VAE.
    Evaluate: compute successor values by (C.3), applying the configured terminal mask.
    Robust target: compute the shared threshold $\eta_{\mathcal{B}}$ and targets y_i by (C.4)–(C.5).
    Fit: regress both critics to $\{y_i : i \in \mathcal{B}\}$.
    Improve: update the actor by ascending the current $\widetilde{Q}^{(1)}$.
    Update the target critics and target actor by Polyak averaging.
  end for
  Define $\widetilde{\pi}^e$ from the fitted networks by (C.6).
end for
return $\{\widetilde{\pi}^e : e \in \mathcal{E}\}$.

CME estimation and fitted targets. Let $\mathbf{u}_i = (z_i, a_i)$ and $\bar{\mathbf{u}}_i$ be its coordinate-wise standardized version using training statistics. Let $\mathcal{N}(i) \subseteq \mathcal{I}_{\mathrm{tr}}$ contain the k_{nn} nearest training indices, including i, with $1 \le k_{\mathrm{nn}} \le |\mathcal{I}_{\mathrm{tr}}|$. For bandwidth $\sigma_{\mathrm{RBF}} > 0$, define

$$\mathrm{k}(\mathbf{u}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{u}-\mathbf{v}\|_2^2}{2\sigma_{\mathrm{RBF}}^2}\right), \qquad \mathbf{K}_i = [\mathrm{k}(\bar{\mathbf{u}}_j, \bar{\mathbf{u}}_l)]_{j,l \in \mathcal{N}(i)}, \qquad \mathbf{k}_i = [\mathrm{k}(\bar{\mathbf{u}}_i, \bar{\mathbf{u}}_j)]_{j \in \mathcal{N}(i)}.$$

Let $\xi > 0$ be the ridge, $w_{\min} \in [0,1)$ the weight floor, and $\mathbf{I}$ the identity matrix of matching dimension. For a nonnegative vector, N_i divides by its sum if that sum exceeds 10^{-12}; otherwise it places unit weight on i. With $\mathbf{w}_i = (w_{ij})_{j \in \mathcal{N}(i)}$, define

$$\widetilde{\mathbf{w}}_i = \mathsf{N}_i\big([(\mathbf{K}_i + \xi \mathbf{I})^{-1}\mathbf{k}_i]_+\big), \qquad \mathbf{w}_i = \mathsf{N}_i(\widetilde{\mathbf{w}}_i \mathbf{1}\{\widetilde{\mathbf{w}}_i \ge w_{\min}\}), \tag{C.1}$$

$$\widehat{P}_i = \sum_{j \in \mathcal{N}(i)} w_{ij} \delta_{z_j'}, \qquad w_{ij} \ge 0, \qquad \sum_{j \in \mathcal{N}(i)} w_{ij} = 1. \tag{C.2}$$

Set $w_{ij} = 0$ outside $\mathcal{N}(i)$; thresholding and multiplication act coordinatewise. The regularized CME coefficients are corrected into probability weights. The measure $\widehat{P}_i$ estimates the successor distribution, while the observed reward r_i is used separately.

Let $\mathcal{A}_{\mathrm{tgt}}(z) \subseteq \mathcal{A}$ be a finite nonempty set of VAE action samples perturbed by the target actor. Parenthesized superscripts index the critics; their mixing coefficient is $\chi \in [0,1]$. The successor values are

$$v_j = \max_{a \in \mathcal{A}_{\mathrm{tgt}}(z_j')} \left\{ \chi \min_{c \in \{1,2\}} \widetilde{Q}_{\mathrm{tgt}}^{(c)}(z_j', a) + (1-\chi) \max_{c \in \{1,2\}} \widetilde{Q}_{\mathrm{tgt}}^{(c)}(z_j', a) \right\}. \tag{C.3}$$

When terminal masking is enabled, replace v_j by $(1 - \mathrm{done}_j)v_j$. Equivalently, use the marked-successor law $\widehat{P}_i^{\mathrm{term}} = \sum_{j \in \mathcal{N}(i)} w_{ij} \delta_{(z_j', \mathrm{done}_j)}$ and zero continuation on terminal marks. For a normalized weighted multiset $\{(x_j, w_j)\}$ and $q \in (0,1)$, define

$$\mathrm{WQuant}_q\{(x_j, w_j)\} = \inf\left\{ \eta : \sum_j w_j \mathbf{1}\{x_j \le \eta\} \ge q \right\}.$$

Thus ties use the smallest value reaching the required cumulative weight. Write $[x]_{[l,u]} = \min\{u, \max\{l, x\}\}$ for clipping, where $l \le u$. For a nonempty minibatch $\mathcal{B}$ of training indices counted with multiplicity, the implemented update is

$$\eta_{\mathcal{B}} = \left[\mathrm{WQuant}_{1-\varepsilon} \left\{ \left(v_j, \frac{w_{ij}}{|\mathcal{B}|} \right) : i \in \mathcal{B},\ j \in \mathcal{N}(i) \right\} \right]_{[0,\, 1/(\varepsilon(1-\gamma))]}, \tag{C.4}$$

$$y_i = r_i + \gamma \left[(1-\varepsilon)\eta_{\mathcal{B}} - \sum_{j \in \mathcal{N}(i)} w_{ij} [\eta_{\mathcal{B}} - v_j]_+ \right]. \tag{C.5}$$

Both critics fit these fixed targets by squared-error regression. The actor improves the current first critic's value at perturbed VAE samples. At deployment, the finite nonempty action set $\mathcal{A}_{\mathrm{sel}}(z) \subseteq \mathcal{A}$ uses the current actor and fresh VAE samples, giving

$$\widetilde{\pi}^e(z) \in \operatorname*{arg\,max}_{a \in \mathcal{A}_{\mathrm{sel}}(z)} \widetilde{Q}^{(1)}(z, a). \tag{C.6}$$

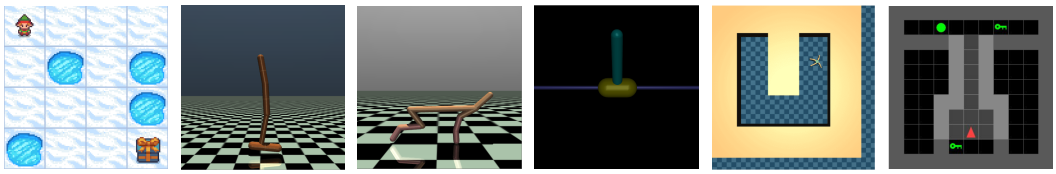
Ties use a fixed ordering. Here $\widetilde{\pi}^e(z)$ is the sampled action, and $\widetilde{\pi}^e(\cdot \mid z)$ denotes its distribution over candidate-sampler randomness.

Relation to the exact backup. For a fixed input i, let $\mathcal{S}_i$ be its entire specified admissible successor support, containing the local atoms and including terminal marks when masking is used. Let $\widetilde{V}$ be a bounded continuation on $\mathcal{S}_i$, with v_j its value at atom j. let $v_{\min,i} = \inf_{s \in \mathcal{S}_i} \widetilde{V}(s)$ and $v_{\max,i} = \sup_{s \in \mathcal{S}_i} \widetilde{V}(s)$. Lemma B.7 gives the exact per-input target

$$y_i^{\mathrm{TV}} = r_i + \gamma \left[\max_{\eta \in [v_{\min,i}, v_{\max,i}]} \left\{ (1-\varepsilon)\eta - \sum_{j \in \mathcal{N}(i)} w_{ij} [\eta - v_j]_+ \right\} + \varepsilon v_{\min,i} \right]. \tag{C.7}$$

The extrema concern the entire admissible support. When $v_{\min,i} = 0$, the shared threshold gives a lower dual value, with an unquantified shortfall from the per-input optimum. This condition is not enforced by unrestricted neural critics; for signed values the minimum term cannot generally be omitted. Calibration and uniform approximation-error bounds remain additional premises of Theorem B.4; Algorithm 2 alone supplies no policy certificate.

D ENVIRONMENT & IMPLEMENTATION DETAILS



(a) (b) (c) (d) (e) (f)

Figure 4: Environments used in the experiments: (a) *FrozenLake*, (b) *Hopper*, (c) *HalfCheetah*, (d) *Inverted Pendulum*, (e) *Umaze-diverse*, and (f) *MiniGrid Memory*.

Table 3: Details of the environments.

Environment	Action type	Action size	Observation	Timestep (s)
FrozenLake (4×4)	Discrete	4	16	—
Hopper	Continuous	3	11	0.008
HalfCheetah	Continuous	6	17	0.05
Inverted Pendulum	Continuous	1	4	0.04
Umaze-diverse	Continuous	8	29	0.1
MiniGrid Memory	Discrete	3	$84 \times 84 \times 3$	—

Table 4: Nominal parameters.

Environment	Model parameter	Nominal value
Hopper	Foot-joint stiffness	0
HalfCheetah	Back thigh-joint stiffness	240
	Back shin-joint stiffness	180
	Back foot-joint stiffness	120
Umaze-diverse	Sliding-friction coefficient	1.0

D.1 OFFLINE DATASETS

COR-FQI is fitted to fixed offline data. Changes to physical parameters and grid size are introduced during evaluation.

Hopper/ Halfcheetah. We use the mixed dataset from the RFQI protocol (Panaganti et al., 2022), containing 10^6 transitions collected with an SAC behavior policy. During collection, a uniformly sampled action replaces the policy action with probability 0.3.

Inverted Pendulum and Umaze-diverse. Inverted Pendulum uses 10^6 transitions collected with an SAC behavior policy. Its action-delay implementation relabels recorded trajectories while retaining rewards and terminal indicators. Umaze-diverse uses the D4RL offline dataset with 10^6 transitions, following the evaluation setting of DT-CORL (Zhan et al., 2026).

MiniGrid Memory. Following RATE (Cherepanov et al., 2026), we use $10{,}000$ trajectories collected on 41×41 grids with a PPO policy using a Transformer-XL backbone (Schulman et al., 2017; Dai et al., 2019). The task requires retaining an initial cue until a later branch choice. Evaluation on other grid sizes uses the fixed learned policy.

D.2 Implementation

Continuous-control COR-FQI. Our implementation follows RFQI's BCQ-based architecture (Panaganti et al., 2022; Fujimoto et al., 2019): a behavior variational autoencoder (VAE) proposes actions, a perturbation actor adjusts them, and twin critics estimate their values. The actor and each critic have ReLU hidden layers of widths $(400, 300)$. The VAE encoder and decoder each have widths $(750, 750)$, with latent dimension twice the action dimension. Each update fits the VAE, regresses both critics onto the robust targets, improves the actor using the first critic, and updates the target networks.

The conditional mean embedding (CME) estimates the successor distribution from eight nearest training transitions in standardized represented-state–action coordinates, including the query transition. A Gaussian kernel and ridge regression give weights that are clipped to be nonnegative, normalized, pruned below $.01$, and normalized again. The bandwidth is the median distance to the fourth other neighbor, with a floor of 10^{-6}. Unlike RFQI's single-successor estimate, this target uses a local weighted successor distribution. The weighted successor values determine one quantile threshold shared across the minibatch. The observed reward is added outside the robust continuation term. Algorithm 2 and (C.5) give the update precisely. This shared threshold approximates the per-input optimization; the implementation alone does not establish a return certificate.

E Relation to robust learning and representation selection

Robust learning. RFQI supplies the robust fitted-learning template; Proposition B.3 states the matched-input identity. Huang et al. already establish robust offline learning for non-Markovian processes, including low-rank predictive-state representations and general model classes. Our finite construction addresses policies deliberately restricted to possibly insufficient representations, and the history variation that their local estimates conceal. It does not supersede their model-learning or sample-complexity results.

Information retained by the policy. Approximate information states quantify control loss when a history summary only approximately preserves relevant predictions (Subramanian et al., 2022). The worst-case formulation of Dave et al. (2024) uses bounded disturbances without a probability model; its minimax objective differs from worst-case expected return over stochastic environments. Both make history compression a substantive approximation. Here, supplied or finer-log variation bounds enlarge local probability sets and enter a common confidence event. Such bounds are additional information, not consequences of accurate coarse mixture prediction alone.

Selection and scope. Batch abstraction selection already trades approximation against estimation error (Jiang et al., 2015). Our selector compares jointly certified worst-case returns across a fixed family and can be used when none is predictively sufficient. It chooses the strongest lower bound, not necessarily the largest true return. Learned encoders require a separate training split before calibration or a valid uniform guarantee over the learned class. Neural fitting, encoder quality, and predictive selection do not establish those guarantees by themselves.