UNIVERSAL BELLMAN COMPLETENESS FORCES LUMPABLE STATE PARTITIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Bellman completeness is a key assumption in fitted value methods and offline reinforcement learning because repeated Bellman updates must remain representable in a fixed value space. However, its universal form leaves the structure of a complete value space and a way to check it unclear. We study this converse for a finite Markov decision process (MDP) with at least two actions and a value space containing the constant functions. We prove that closure under every optimal Bellman update using rewards and continuation values from the space holds exactly when the value space consists of functions constant on the blocks of a unique state partition, and the transition kernel for every action is lumpable with respect to that partition. Independent action rewards expose closure under pointwise maxima and isolate each action's transition kernel. Together, these constructions yield an exact finite certificate based on adaptive Bellman membership queries. On a block-constant value space, they also give an exact relation between the closure defect and transition mismatch. Under separation or conditioning assumptions, this relation supports block recovery and yields value and policy loss bounds.

1 INTRODUCTION

Many fitted value methods in approximate dynamic programming and offline reinforcement learning repeatedly apply Bellman updates within a fixed function class, so stability of the representation requires the relevant updates to remain in the class (Antos et al., 2008; Parr et al., 2008; Song et al., 2016; Chang et al., 2022). Bellman completeness formalizes the needed stability by requiring the relevant Bellman updates to remain representable (Chen & Jiang, 2019; Chang et al., 2022). However, a representation can express the current value and still fail after one update, as shown in Figure 1. Although Bellman completeness excludes this failure, its definition does not reveal which state distinctions are needed to exclude all such failures. Therefore, if a given representation remains closed under every Bellman update, what must that representation and the underlying transition dynamics look like?

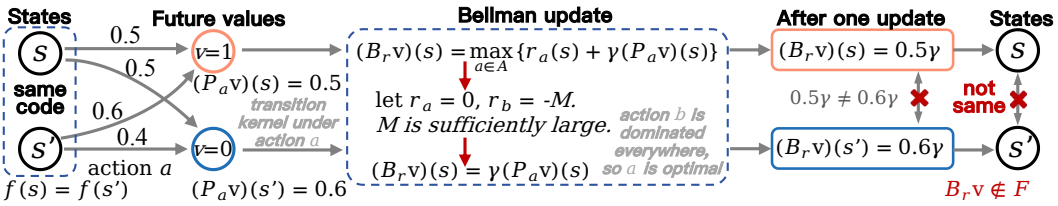


Figure 1: A representation assigns the same code to s and s' (every $f \in F$ satisfies $f(s) = f(s')$), but action a induces different expectations ($(P_a v)(s) = 0.5$ and $(P_a v)(s') = 0.6$) for the same represented continuation v. Choosing rewards so that a is optimal at both states makes the next Bellman update assign them different values (0.5γ and 0.6γ), so the update leaves the representation space ($B_r v \notin F$).

One established line of work addresses that question by specifying which states a representation should treat as equivalent. Classical state abstraction and stochastic bisimulation encode this choice as a state partition, impose reward consistency and transition lumpability, and show that Bellman

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.

planning can be carried out on the resulting quotient (Castro, 2020; van der Pol et al., 2020; Dean & Givan, 1997; Givan et al., 2003; Ferns et al., 2004; Li et al., 2006). Related work on value equivalence uses Bellman updates for selected policies and functions, while representation learning based on Bellman completeness uses approximate closure as a learning objective (Grimm et al., 2020; 2021; Chang et al., 2022). Across these lines, the analysis begins with a prescribed structure or criterion and derives the Bellman consequences it guarantees. The converse remains open. Can stability under every relevant Bellman update determine the structure that the representation must have?

In this paper, we answer this converse by using the Bellman maximum as a structural probe. The freedom to choose each action's reward function independently in the value space lets us turn Bellman updates into tests for two algebraic properties of the representation. The value space must be closed under pointwise maxima and invariant under each action's transition. In the finite setting with the constant functions in the value space, closure under pointwise maxima forces a unique state partition. The value space consists exactly of functions that are constant on the blocks of this partition. Transition invariance then forces the transition kernel for every action to be lumpable with respect to the same partition. We prove the reverse direction as well, so universal Bellman completeness is exactly the value space of a state partition whose transition kernels are lumpable for every action. This probe also yields finite certificates for this condition and gives approximate closure a direct transition and planning interpretation.

We make the following contributions:

(1) We prove the exact converse characterization. Universal Bellman completeness holds if and only if the value space consists of functions that are constant on the blocks of a unique state partition and every action's transition kernel is lumpable with respect to that partition. The partition is intrinsic to the represented subspace, and encoder coordinates remain free up to invertible reparameterization.

(2) We replace the universal closure condition with an exact finite certificate. For an encoder with $k \geq 2$ distinct rows, at most $3k-4$ adaptive probes construct the block indicators and $|A|k$ isolation probes verify the transition condition. With access to an exact Bellman membership oracle, all queries succeed if and only if F is universally Bellman complete, and each queried reward and continuation has sup norm at most one.

(3) We quantify approximate closure. On a block-constant space, $\mathcal{E}_R(F) = \gamma\lambda$ for $R \geq 1$, where λ is the largest total variation distance between block transition distributions of states in the same block. Under margin or conditioning assumptions, a small defect recovers an exact or nearby block-constant space, and the resulting transition mismatch yields value and policy loss bounds.

2 Preliminaries

2.1 Bellman Updates and Represented Value Spaces

To study representation stability under repeated Bellman updates, we fix a finite Markov decision process (MDP) (Puterman, 1994) and a supplied linear value space. Let S be a finite state set, with $n = |S|$, and let A be a finite action set with $|A| \geq 2$. For each action $a \in A$, let $P_a \in [0,1]^{S\times S}$ be a transition matrix whose rows are probability distributions. We fix a discount factor $\gamma \in (0,1)$ and identify every value function on S with a vector in $\mathbb{R}^S$. For $v \in \mathbb{R}^S$, $(P_a v)(s)$ is the expected value of v after taking action a at state s.

Let $r = (r_a)_{a\in A}$ collect the reward functions associated with the actions. Given a continuation value v, the discounted Bellman update is

$$(B_r v)(s) = \max_{a\in A} \{r_a(s) + \gamma(P_a v)(s)\}. \tag{1}$$

We study this update on a linear value space $F \leq \mathbb{R}^S$ containing the constant function $\mathbf{1}$, where $\mathbf{1}$ denotes the function equal to one at every state. Write $d = \dim F$. The represented continuations are the functions in F, and the reward function associated with each action can be chosen independently from F. We write F^A for the collection of reward tuples whose components all belong to F.

Definition 1 (Universal Bellman completeness)**.** The space F is universally Bellm an complete for $(P_a)_{a\in A}$ if

$$B_r v \in F \qquad \text{for every } r \in F^A \text{ and every } v \in F. \tag{2}$$

For a given MDP and value space F, every Bellman update generated from represented rewards and continuations must remain in F. The constant function provides uniform reward offsets, and having at least two actions lets the Bellman maximum compare independently chosen rewards. Thus, the condition asks whether one value space can absorb every optimal update generated from functions it already represents.

2.2 State Partitions and Lumpability

To describe the state distinctions retained by a value space, let $\mathcal{P} = \{C_1, \ldots, C_k\}$ be a partition of S. The associated space of functions constant on each block is $F_{\mathcal{P}} = \{v \in \mathbb{R}^S : v(s) = v(t) \text{ whenever } s, t \in C_i \text{ for some } i\} = \operatorname{span}\{\mathbf{1}_{C_1}, \ldots, \mathbf{1}_{C_k}\}$, where $\mathbf{1}_{C_i}$ denotes the indicator of C_i. Every function in $F_{\mathcal{P}}$ is determined by one value for each block.

For an action a, the transition matrix P_a is lumpable with respect to $\mathcal{P}$ when, for every source block C_i, states $s, t \in C_i$, and target block C_j, $\sum_{x\in C_j} P_a(s, x) = \sum_{x\in C_j} P_a(t, x)$. This condition compares the total probability sent to each target block. Probabilities assigned to individual states inside a target block may differ. We say that the transition kernels are actionwise lumpable with respect to $\mathcal{P}$ when the condition holds for every action (Givan et al., 2003; Ferns et al., 2004; Li et al., 2006).

These definitions give the basic mechanism behind planning on a partition. If $F = F_{\mathcal{P}}$ and the transition kernels are actionwise lumpable, every represented reward and continuation is constant within each block, and $P_a v$ is constant within each source block. Each action value in equation 1 is therefore constant on that block, and the pointwise maximum preserves the same equality.

When the supplied value space is represented by an encoder $\Phi \in \mathbb{R}^{S\times d}$ with linearly independent columns, we write $F = \operatorname{col}(\Phi)$. Its rows provide concrete feature codes for the states. Later certification and recovery results use these codes, while the structural statements concern the subspace F. With these objects fixed, the next section determines whether universal Bellman completeness forces the partition and lumpability conditions introduced above.

3 Structural Consequences of Universal Bellman Completeness

A Bellman update can leave the value space through its pointwise maximum, or through a transition that separates states identified by the representation. Universal Bellman completeness must control both effects. We expose them in this order. First, we remove the transition term and use the maximum to determine which states the value space identifies. We then isolate one action and use its transition to determine how those identifications must be preserved.

3.1 From Bellman Closure to Partition and Lumpability

Max Closure and the Induced Partition. The first question is which state distinctions are compatible with all Bellman updates generated by represented rewards. A linear space already contains f and g whenever $f, g \in F$, but it need not contain their pointwise maximum. The Bellman operator lets us test this operation directly.

Take $f, g \in F$, choose two distinct actions $a_0, a_1 \in A$, and set $r_{a_0} = f$, $r_{a_1} = g$, and $v = 0$. For every remaining action, set $r_b = -M\mathbf{1}$, where $-M < \min\{\min_{s\in S} f(s), \min_{s\in S} g(s)\}$. The remaining actions are dominated at every state, while the transition term vanishes for the selected actions. Therefore,

$$B_r 0 = f \vee g. \tag{3}$$

Since f and g are arbitrary, universal Bellman completeness implies that F is closed under pointwise maxima. Linearity then gives closure under pointwise minima and positive parts.

This gives an algebraic closure property. We next determine what state structure this property implies. For a linear space $L \leq \mathbb{R}^S$, define

$$s \sim_L t \quad \Longleftrightarrow \quad f(s) = f(t) \quad \text{for every } f \in L.$$

The following finite-space fact converts maximum closure into a partition (Aliprantis & Border, 2006).

Lemma 2 (Finite spaces with pointwise maximum closure induce a unique partition)**.** *Let $L \leq \mathbb{R}^S$ contain $\mathbf{1}$ and be closed under pointwise maxima. Then L is exactly the space of functions that are constant on the equivalence classes of $\sim_L$. If these classes form a partition $\mathcal{P}$, then $L = F_{\mathcal{P}}$, the number of classes is $\dim L$, and $\mathcal{P}$ is unique up to reordering.*

Proof idea: Every function in L is constant on the classes of $\sim_L$, which gives $L \subseteq F_{\mathcal{P}}$. For the reverse inclusion, choose a linear combination $h \in L$ that takes a different value on every class. Such a choice exists because each pair of classes excludes one proper hyperplane in the coefficient space, and there are finitely many pairs. Order the values as $\beta_1 < \cdots < \beta_k$. Positive-part ramps isolate the two extreme levels, and the minimum of two ramps isolates an interior level. For example, an interior class indicator is obtained from

$$\left(\frac{h - \beta_{i-1}\mathbf{1}}{\beta_i - \beta_{i-1}}\right)_+ \wedge \left(\frac{\beta_{i+1}\mathbf{1} - h}{\beta_{i+1} - \beta_i}\right)_+ .$$

Thus every class indicator belongs to L, and these indicators span $F_{\mathcal{P}}$. This proves the equality, the dimension statement, and uniqueness. The finite construction is given in Appendix A.1.

The lemma answers the first part of the structural question. Applying it to the max closure forced by equation 3 gives a unique partition $\mathcal{P} = \{C_1, \ldots, C_d\}$ such that $F = F_{\mathcal{P}}$. The partition records exactly which states the value space identifies. We now ask how the transition kernels must act on these identified states.

Action Isolation and Transition Lumpability. Once the partition is known, a transition can violate closure in one specific way. Two states in the same block receive the same value from every function in F, but an action may give them different expectations for a represented continuation. We therefore isolate one action and inspect its transition directly.

Fix $a \in A$ and $v \in F$. Set $r_a = 0$ and $r_b = -M\mathbf{1}$ for every $b \neq a$, where

$$M > \gamma \left(\max_{x \in S} v(x) - \min_{x \in S} v(x) \right). \tag{4}$$

Because every P_b is row-stochastic, each value of $P_b v$ lies between the minimum and maximum of v. Hence, for every state s and competitor $b \neq a$,

$$\begin{aligned} &\big[r_a(s) + \gamma(P_a v)(s)\big] - \big[r_b(s) + \gamma(P_b v)(s)\big] \\ &\quad \geq M - \gamma \left(\max_x v(x) - \min_x v(x) \right) > 0. \end{aligned} \tag{5}$$

Action a is the unique maximizer at every state, so $B_r v = \gamma P_a v$. Universal Bellman completeness applies to every $v \in F$. Since $\gamma > 0$, this gives $P_a F \subseteq F$ for every $a \in A$.

The invariance statement becomes concrete on the partition already recovered from max closure. For a target block C_j, applying P_a to its indicator gives $(P_a \mathbf{1}_{C_j})(s) = \sum_{x \in C_j} P_a(s, x)$.

The left-hand side belongs to $F_{\mathcal{P}}$, so it is constant over each source block C_i. Therefore, states $s, t \in C_i$ send the same total probability to every target block C_j. This is exactly the condition that P_a is lumpable with respect to $\mathcal{P}$.

The converse uses the same block indicators. If the target-block masses agree for states in the same source block, then every linear combination of the indicators is preserved by P_a. Hence $P_a F_{\mathcal{P}} \subseteq F_{\mathcal{P}}$. On a block-constant space, transition invariance and lumpability are therefore equivalent.

We have now translated the two ways in which a Bellman update can leave the value space into two structural requirements. Max closure determines the partition, and action isolation requires every action transition kernel to be lumpable with respect to that partition. The next result asks whether these requirements together are exactly sufficient.

3.2 Exact Characterization

We now verify the reverse direction.

Theorem 3 (Finite-state universal characterization)**.** *For the finite MDP and unital value space defined above, the following statements are equivalent.*

(i) F is universally Bellman complete.

(ii) There is a unique unordered partition $\mathcal{P} = \{C_1, \dots, C_d\}$ of S such that $F = F_{\mathcal{P}}$, and every action transition kernel P_a is lumpable with respect to $\mathcal{P}$.

Proof idea: For *(i)* $\Rightarrow$ *(ii)*, choose two action rewards and zero continuation. Equation equation 3 gives pointwise maximum closure, and the finite partition lemma yields the unique partition and $F = F_{\mathcal{P}}$. Next, use the negative reward offset in equation 4. It isolates any chosen action for any continuation in F, giving $P_a F \subseteq F$. Applying this invariant transition to each block indicator gives the lumpability condition.

For *(ii)* $\Rightarrow$ *(i)*, take arbitrary $r \in {F_{\mathcal{P}}}^A$ and $v \in F_{\mathcal{P}}$. Each action reward is constant on every block, and lumpability makes $P_a v$ constant on every block. Therefore every action value $r_a + \gamma P_a v$ is block-constant. The pointwise maximum over actions preserves this equality, so $B_r v \in F_{\mathcal{P}}$.

The theorem identifies the represented subspace itself. Its partition is determined by agreement of all functions in F, so it does not depend on a chosen basis. An invertible change of coordinates in a full-rank encoder leaves the value space and the partition unchanged.

The constant function is a convenient order unit for the argument (Aliprantis & Border, 2006). If F contains a strictly positive function u, diagonal rescaling by u reduces the problem to a unital value space. The same reasoning then gives a weighted block space and the corresponding weighted transition condition. The exact statement and proof are given in Appendix A.2.

4 Consequences for Planning and Certification

4.1 Exact Quotient Planning

Theorem 3 identifies the state distinctions retained by F. We now ask how this identification supports planning. Let $\mathcal{P} = \{C_1, \dots, C_d\}$ be the partition from the theorem, and let $\Psi \in \mathbb{R}^{S \times d}$ be its block-indicator matrix, with $\Psi_{s,j} = \mathbf{1}_{C_j}(s)$. Every represented value has a unique form $v = \Psi q$ for some $q \in \mathbb{R}^d$.

For each action a, define the quotient transition matrix by $\bar{P}_a(i,j) = \sum_{x \in C_j} P_a(s,x),\ s \in C_i$ (Givan et al., 2003; Li et al., 2006). Lumpability makes the right-hand side independent of the representative s. The quotient matrices are row-stochastic, and grouping next states by their blocks gives $P_a \Psi = \Psi \bar{P}_a$. Thus, applying P_a to a block-constant value is equivalent to applying $\bar{P}_a$ to its block values and lifting the result back to S.

A represented reward for action a has the form $r_a = \Psi \rho_a$. Define the quotient Bellman operator by $(\bar{B}_\rho q)_i = \max_{a \in A}\{\rho_a(i) + \gamma(\bar{P}_a q)_i\}$. The transition relation remains valid after taking the action maximum:

$$B_r(\Psi q) = \Psi \bar{B}_\rho q, \qquad q \in \mathbb{R}^d. \tag{6}$$

This identity applies to every represented continuation. Each ground Bellman update can therefore be computed on the d quotient values and lifted back to the original state space.

Proof idea: The relation $P_a \Psi = \Psi \bar{P}_a$ follows by grouping the sum over next states according to target blocks. Substituting this relation and $r_a = \Psi \rho_a$ into the ground Bellman update gives equation 6. Repeated application gives $B_r^H(\Psi q_0) = \Psi \bar{B}_\rho^H q_0$ for finite-horizon planning. Since the ground and quotient operators are γ-contractions, their fixed points satisfy $v^* = \Psi q^*$ (Puterman, 1994). Backward induction lifts finite-horizon maximizing actions, and a quotient-greedy action at q^* lifts blockwise to an optimal stationary policy in the ground MDP.

The partition is intrinsic to F, while encoder coordinates only choose a basis for the same space. If a full-rank encoder Φ spans F, then $\Phi = \Psi T$ for an invertible matrix T, and $P_a \Phi = \Phi(T^{-1} \bar{P}_a T)$. Encoder coordinates therefore give a reparameterization of the same quotient planning problem.

4.2 FINITE CERTIFICATION

The quotient representation identifies the finite objects that determine universal Bellman completeness. We now ask how to check them when equation 2 ranges over every represented reward and continuation.

Let $\Phi \in \mathbb{R}^{S\times d}$ have full column rank and $\mathrm{col}(\Phi) = F$. Its rows induce the relation

$$s \equiv_\Phi t \quad \Longleftrightarrow \quad \Phi_{s,:} = \Phi_{t,:}.$$

Let $\mathcal{P}_\Phi$ be the resulting row partition and let $k_\Phi = |\mathcal{P}_\Phi|$. We use an exact Bellman membership oracle that receives $r \in F^A$ and $v \in F$, and returns the bit indicating whether $B_r v \in F$. The certificate has two parts. The row partition must have exactly d classes, and every action transition must preserve the block structure of that partition.

Theorem 4 (Finite exact certificate)**.** *Let Φ have full column rank and let $\mathbf{1} \in \mathrm{col}(\Phi) = F$. Then F is universally Bellman complete if and only if $k_\Phi = d$ and, for every action a, every source class $C_i \in \mathcal{P}_\Phi$, every target class $C_j \in \mathcal{P}_\Phi$, and all $s, t \in C_i$, $\sum_{x\in C_j} P_a(s,x) = \sum_{x\in C_j} P_a(t,x)$.*

Proof idea: Equal encoder rows imply $F \subseteq F_{\mathcal{P}_\Phi}$, while evaluation on one representative from each row class gives $d \le k_\Phi$. If $k_\Phi = d$, the two spaces have the same dimension, so $F = F_{\mathcal{P}_\Phi}$ and its block indicators belong to F. For an action a and a target class C_j, use the indicator $\mathbf{1}_{C_j}$ as the continuation, assign zero reward to a, and assign $-\mathbf{1}$ to every competing action. The selected action is then isolated, and membership of the resulting update is equivalent to $P_a\mathbf{1}_{C_j} \in F$. Since the block indicators form a basis, these tests are equivalent to lumpability. Conversely, universal Bellman completeness gives the unique d-block partition from Theorem 3; equality of encoder rows coincides with its blocks, which yields $k_\Phi = d$ and the block-mass condition.

Proposition 5 (Finite bounded reward probes)**.** *Let $k = k_\Phi$ and suppose $\mathbf{1} \in F = \mathrm{col}(\Phi)$. For $k \ge 2$, one can construct an adaptive sequence of at most $3k - 4 + |A|k$ Bellman membership queries such that all queries succeed if and only if F is universally Bellman complete. Every queried reward and continuation has sup norm at most one. When $k = 1$, the value space is the constant space and no query is needed.*

Proof idea: Choose a represented function whose values are distinct on the k row classes. Normalize affine ramps of this function and use membership queries to construct positive parts. The two extreme class indicators require one query each, and each interior indicator requires two positive-part queries and one minimum query, giving $3k - 4$ queries. Once the indicators are available, one action-isolation query for each pair (a, C_j) tests $P_a\mathbf{1}_{C_j} \in F$, giving $|A|k$ additional queries. If all queries succeed, the indicators imply $k \le d$, while row evaluation gives $d \le k$; hence $k = d$, and Theorem 4 applies. The constructed Bellman outputs are known symbolically, so later queries can be formed from the earlier membership bits without observing output vectors.

When the transition matrices are explicitly available, the same certificate is checked by aggregating each transition row over the encoder row classes. Universal Bellman completeness is therefore reduced to finite tests of the represented state structure and the transition of each action on its block indicators.

4.3 APPROXIMATE CLOSURE AND PLANNING CONSEQUENCES

Finite certification handles exact closure. When the closure residual is nonzero, we ask what transition failure it measures and how that failure affects planning (Chen & Jiang, 2019; Chang et al., 2022; Farahmand et al., 2017; Farahmand, 2018; Ayoub et al., 2020). For $z \in \mathbb{R}^S$, let $\mathrm{dist}_\infty(z, F) = \inf_{f\in F} \|z - f\|_\infty$. For $R \ge 1$, define the normalized closure defect

$$\mathcal{E}_R(F) = \sup_{\substack{v\in F,\ \|v\|_\infty \le 1 \\ r\in F^A,\ \max_a \|r_a\|_\infty \le R}} \mathrm{dist}_\infty(B_r v, F). \tag{7}$$

The meaning of the closure defect. Let $\mathcal{P} = \{C_1, \dots, C_d\}$ be a state partition. For $s \in S$ and $a \in A$, define the transition distribution over its blocks by $p_a^s(j) = \sum_{x\in C_j} P_a(s,x)$, and define the block-transition mismatch by

$$\lambda = \max_{\substack{a,i \\ s,t\in C_i}} \mathrm{TV}(p_a^s, p_a^t), \qquad \mathrm{TV}(p,q) = \frac{1}{2}\sum_j |p(j) - q(j)|.$$

When the value space is block-constant on $\mathcal{P}$, i.e., $F = F_{\mathcal{P}}$, the closure defect $\mathcal{E}_R(F)$ measures the block-transition mismatch λ exactly.

Theorem 6 (Exact defect identity). *Let $F = F_{\mathcal{P}}$ and $R \geq 1$. Then $\mathcal{E}_R(F) = \gamma\lambda$.*

Proof idea: For the upper bound, represented rewards agree inside each block. Two states in the same block can therefore differ only through their transition expectations. Total variation bounds the difference for each action by $2\gamma\lambda$, and the action maximum preserves this range. The distance to a block-constant space is half of the largest range inside a block, which gives $\mathcal{E}_R(F) \leq \gamma\lambda$. For the lower bound, choose an action, two states, and a subset of target blocks that attain λ. Use the signed indicator of that subset as the continuation and use rewards $+\mathbf{1}$ and $-\mathbf{1}$ to isolate the selected action. The resulting Bellman output has range $2\gamma\lambda$ on the selected source block, which gives the reverse inequality.

Recovering the state partition. The identity gives an exact interpretation once the block space is known. We next ask whether a small closure defect can recover that structure. Let $\mathcal{P}_F$ be the partition in which s and t belong to the same class when $f(s) = f(t)$ for every $f \in F$. Write $k = |\mathcal{P}_F|$, choose one representative s_i from each class, and define

$$\Delta_F = \max_{\substack{h \in F \\ \|h\|_\infty \leq 1}} \min_{i \neq j} |h(s_i) - h(s_j)|.$$

This margin measures how strongly one represented function can separate the classes induced by F.

Theorem 7 (Margin-conditioned recovery). *Let $d \geq 2$, $R \geq 1$, and $\varepsilon = \mathcal{E}_R(F)$.*
If $\varepsilon < \min\left\{1, \frac{\Delta_F}{6k}\right\}$, then $k = d$, $F = F_{\mathcal{P}_F}$, and the transition mismatch of $\mathcal{P}_F$ satisfies $\lambda \leq \frac{\varepsilon}{\gamma}$.

Proof idea: The closure defect first gives approximate closure under pointwise maxima, positive parts, and minima. Applying these operations to a represented function that separates the k classes produces an approximation in F to every class indicator, with error at most $6\varepsilon/\Delta_F$. Evaluating these approximations on the class representatives gives a matrix whose distance from I_k is less than one in the induced sup norm. The matrix is invertible, so its rank gives $k \leq d$. Evaluation on the representatives gives $d \leq k$, hence $k = d$ and $F = F_{\mathcal{P}_F}$. The action-isolation construction applied to signed block indicators then gives the transition bound. Appendix D provides the complete proof and a recovery theorem conditioned on anchor interpolation that bounds the distance to a nearby block-constant space.

Planning under transition mismatch. The transition mismatch determines how accurately a quotient Markov decision process (MDP) represents the original planning problem (Taylor et al., 2008; Abel et al., 2016). For each block C_i, choose a representative s_i and define $\bar{P}_a(i,j) = p_a^{s_i}(j)$. For a block-constant reward $r_a = \Psi\rho_a$, let $\bar{B}_\rho$ be the quotient Bellman operator formed from these reference transitions.

Theorem 8 (Approximate quotient planning). *Let $\mathcal{P}$ have transition mismatch at most λ, and suppose $\max_a \|\rho_a\|_\infty \leq R_{\max}$. Then, for every $q \in \mathbb{R}^d$,*

$$\|B_r(\Psi q) - \Psi\bar{B}_\rho q\|_\infty \leq 2\gamma\lambda\|q\|_\infty. \tag{8}$$

If v^* and q^* are the optimal fixed points of the original and reference quotient Bellman operators, and π is the blockwise lift of a quotient-greedy policy at q^*, then

$$\|v^* - \Psi q^*\|_\infty \leq \frac{2\gamma\lambda R_{\max}}{(1-\gamma)^2}, \qquad \|v^* - v^\pi\|_\infty \leq \frac{4\gamma\lambda R_{\max}}{(1-\gamma)^2}. \tag{9}$$

Proof idea: For $s \in C_i$, the block transition distributions at s and the representative s_i differ by at most λ in total variation. Their expectations of q therefore differ by at most $2\lambda\|q\|_\infty$. The rewards agree inside each block, and Bellman nonexpansiveness gives equation 8. Since $\|q^*\|_\infty \leq R_{\max}/(1-\gamma)$, this inequality bounds the ground Bellman residual of Ψq^*. The contraction residual bound gives the first inequality in equation 9. Applying the same argument to the action selected by the quotient-greedy policy bounds $\|v^\pi - \Psi q^*\|_\infty$. The triangle inequality then gives the policy loss bound.

The results above depend on transition distributions over the recovered blocks. With transition samples, empirical block frequencies estimate λ, and confidence bounds for these estimates can be substituted into the planning inequalities. The corresponding concentration and finite-precision recovery results are given in Appendix C.

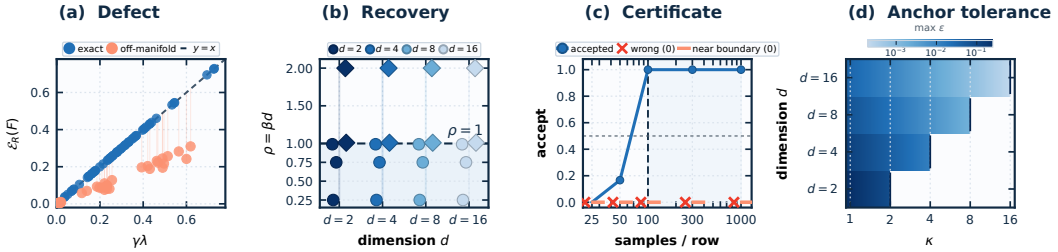


Figure 2: Finite structural audits. Exact block spaces follow the defect identity in (a), while the plotted spaces outside the block space family move away from the identity line. Panel (b) displays the sufficient recovery trigger through the relative conditioning $\rho = \beta d$. Panel (c) shows certificate acceptance as the sample budget for each row grows. Panel (d) visualizes the maximum admissible defect obtained from the theorem as a function of representation dimension and anchor conditioning.

5 Experiments

We use experiments to answer the following three questions: ***(i)*** Do the max and action isolation constructions, together with the finite certificate, identify the block structure and test lumpability? ***(ii)*** Under approximate closure, does the measured defect track transition mismatch and planning error? ***(iii)*** With shared transition samples, how does quotient pooling affect planning?

5.1 Experimental Setup

Our benchmark is a synthetic family of randomized finite Markov decision processes (MDPs) with block structure. Each trial chooses a state partition $\mathcal{P}$, its number of blocks and block sizes, and a finite action set, then uses the associated space $F_{\mathcal{P}}$ of functions constant on each block. On every exact instance, the transition kernel for each action is lumpable with respect to $\mathcal{P}$ (Givan et al., 2003). We construct a paired instance by keeping $\mathcal{P}$ and $F_{\mathcal{P}}$ fixed and redistributing probability in one source row so that its mass on a target block differs from the other states in that source block. The pair shares the representation and changes the transition condition tested by the certificate.

Structural certificate audits use $d \in \{2, \dots, 5\}$, two to four actions, and block sizes from two to four. They run exact Bellman-membership checks on explicit kernels and use multinomial row estimates in finite-sample variants. The planning benchmark contains 80 block MDPs with $d = 5$, block size 4, three actions, $\gamma = 0.92$, and budgets from 8 to 256 samples per state-action pair. We compare ground plug-in estimation, oracle and recovered quotient pooling, perturbed and mismatched assignments, and clustering based on approximate bisimulation, value equivalence, and random features. Every planning method receives the same sampled transition counts for each seed and budget. Full parameter ranges appear in Appendix F.3.

5.2 Structural Audits of Closure and Certification

The first question follows the two constructions in the proof. Max closure determines which states share a value code, and action isolation checks whether each action preserves those classes. We test this chain on paired finite instances. Exact instances use value spaces that are constant on blocks and transition kernels that are lumpable with respect to those blocks. Each corrupted partner changes one source row so that its mass on a target block no longer agrees with the other states in that source block. The paired instances are then passed through the finite certificate.

Panel (c) of Figure 2 shows the finite certificate as the row budget increases. In the displayed sweep, exact block instances reach full acceptance at 100 samples per row, while wrong assignments and assignments near the boundary remain rejected.

5.3 Approximate Recovery

The second question asks whether a small closure defect retains the quantitative meaning predicted by the theory. We vary the defect, transition mismatch, separation between row codes, and anchor conditioning, and report the resulting recovery diagnostics.

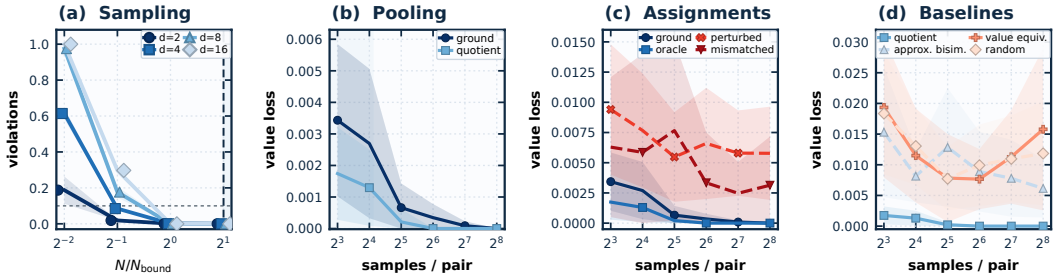


Figure 3: Finite planning audits under shared transition samples. Panel (a) reports transition violations against the normalized sample budget N/N_{bound}. Panel (b) compares policy value loss for ground and quotient pooling. Panel (c) changes the assignment while keeping the aggregation procedure fixed. Panel (d) compares quotient pooling with the clustering baselines. Shaded regions indicate variation across the displayed instances.

Panel (a) of Figure 2 compares the measured $\mathcal{E}_R(F)$ with $\gamma\lambda$. Exact block instances lie on the diagonal, matching the defect identity, while spaces outside the block space family deviate from it. Panels (b) and (d) examine the recovery condition. In panel (b), $\rho < 1$ marks the sufficient trigger $\beta < 1/d$. Recovery succeeds at every tested point, including points beyond this trigger for the displayed generator. Panel (d) displays the admissible defect $1/[d\kappa(1+\kappa)]$, making the dependence on dimension and anchor conditioning visible.

5.4 Quotient Planning under Shared Samples

The third question asks whether the recovered quotient changes planning when all methods receive the same transition data. The displayed budgets range from 8 to 256 samples per state action pair. We compare ground plug-in estimation, matched and oracle quotient pooling, perturbed and mismatched assignments, and clustering based on approximate bisimulation (Abel et al., 2016; Ferns et al., 2004; Castro, 2020; van der Pol et al., 2020), value equivalence (Grimm et al., 2020; 2021), and random features. Every method receives the same sampled transition counts for each seed and budget.

Panel (a) of Figure 3 calibrates the sample regime. Transition violations fall sharply near $N/N_{\text{bound}} = 1$, and larger representation dimensions have more violations below the sufficient budget. Panel (b) shows the planning consequence of pooling. At the smallest budget, matched quotient pooling has about half the policy value loss of the ground plug-in estimate. The gap narrows as more samples are available, and the ground estimate slightly catches up at the largest budget. Panel (c) keeps sample counts and the aggregation rule fixed while changing the assignment. The oracle quotient has the smallest loss, while perturbed and mismatched assignments retain substantially larger losses. Panel (d) places the same quotient curve beside the three clustering baselines. The quotient has the lowest displayed loss across the budgets in this benchmark.

The paired assignment controls connect the planning comparison to the structural condition. On these instances, the gain is observed for the quotient whose transitions satisfy that condition, and it disappears under the perturbed and mismatched assignments. The results provide numerical support for the planning consequences predicted by the theorem on the stated finite benchmark.

6 Conclusion

We studied the converse of Bellman completeness. In a finite MDP with at least two actions and the constant functions represented, closure under every Bellman update generated by represented rewards and continuations forces F to be exactly the space of functions constant on the blocks of a unique state partition, with every action transition kernel lumpable with respect to that partition. The max construction exposes the partition, and action isolation exposes lumpability; the reverse implication proves sufficiency. Universal Bellman completeness therefore has a concrete quotient interpretation, with finite certification and approximate planning bounds following from the same structure.

AI USE STATEMENT

In this work, we used generative AI tools for aiding and polishing the writing. We take responsibility for the final content of this work.

REFERENCES

David Abel, David Hershkowitz, and Michael L. Littman. Near optimal behavior via approximate state abstraction. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2915–2923, 2016.

Alekh Agarwal, Sham M. Kakade, Akshay Krishnamurthy, and Wen Sun. FLAMBE: Structural complexity and representation learning of low rank MDPs. In *Advances in Neural Information Processing Systems 33*, pp. 20095–20107, 2020.

Rishabh Agarwal, Marlos C. Machado, Pablo Samuel Castro, and Marc G. Bellemare. Contrastive behavioral similarity embeddings for generalization in reinforcement learning. In *International Conference on Learning Representations*, 2021.

Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, 3 edition, 2006.

András Antos, Csaba Szepesvári, and Rémi Munos. Fitted Q-iteration in continuous action-space MDPs. In *Advances in Neural Information Processing Systems 20*, 2008.

Kavosh Asadi, Evan Cater, Dipendra Misra, and Michael L. Littman. Equivalence between wasserstein and value-aware loss for model-based reinforcement learning. In *FAIM Workshop on Prediction and Generative Modeling in Reinforcement Learning*, 2018.

Alex Ayoub, Zeyu Jia, Csaba Szepesvári, Mengdi Wang, and Lin F. Yang. Model-based reinforcement learning with value-targeted regression. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 463–474, 2020.

Marc G. Bellemare, Will Dabney, Robert Dadashi, Adrien Ali Taiga, Pablo Samuel Castro, Nicolas Le Roux, Dale Schuurmans, Tor Lattimore, and Clare Lyle. A geometric perspective on optimal representations for reinforcement learning. In *Advances in Neural Information Processing Systems 32*, 2019.

Pablo Samuel Castro. Scalable methods for computing state similarity in deterministic markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10069–10076, 2020. doi: 10.1609/aaai.v34i06.6564.

Jonathan D. Chang, Kaiwen Wang, Nathan Kallus, and Wen Sun. Learning Bellman complete representations for offline policy evaluation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2938–2971, 2022.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1042–1051, 2019.

Will Dabney, André Barreto, Mark Rowland, Robert Dadashi, John Quan, Marc G. Bellemare, and David Silver. The value-improvement path: Towards better representations for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7160–7168, 2021.

Thomas Dean and Robert Givan. Model minimization in markov decision processes. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, pp. 106–111, 1997.

Amir-massoud Farahmand. Iterative value-aware model learning. In *Advances in Neural Information Processing Systems 31*, pp. 9090–9101, 2018.

Amir-massoud Farahmand, André Barreto, and Daniel Nikovski. Value-aware loss function for model-based reinforcement learning. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1486–1494, 2017.

Gregory Farquhar, Tim Rocktäschel, Maximilian Igl, and Shimon Whiteson. TreeQN and ATreeC: Differentiable tree-structured models for deep reinforcement learning. In *International Conference on Learning Representations*, 2018.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite markov decision processes. In *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence*, pp. 162–169, 2004.

Dylan J. Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 3489–3489, 2022.

Vincent François-Lavet, Yoshua Bengio, Doina Precup, and Joelle Pineau. Combined reinforcement learning via abstract representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3582–3589, 2019. doi: 10.1609/aaai.v33i01.33013582.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. DeepMDP: Learning continuous latent space models for representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2170–2179, 2019.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in markov decision processes. *Artificial Intelligence*, 147(1–2):163–223, 2003.

Christopher Grimm, André Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5541–5552, 2020.

Christopher Grimm, André Barreto, Gregory Farquhar, David Silver, and Satinder Singh. Proper value equivalence. In *Advances in Neural Information Processing Systems*, volume 34, pp. 7773–7786, 2021.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1704–1713, 2017.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2137–2143, 2020.

Joshua Joseph, Alborz Geramifard, John W. Roberts, Jonathan P. How, and Nicholas Roy. Reinforcement learning with misspecified model classes. In *2013 IEEE International Conference on Robotics and Automation*, pp. 939–946, 2013. doi: 10.1109/ICRA.2013.6630686.

Lihong Li, Thomas J. Walsh, and Michael L. Littman. Towards a unified theory of state abstraction for MDPs. In *Proceedings of the 9th International Symposium on Artificial Intelligence and Mathematics*, 2006.

Aditya Modi, Jinglin Chen, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. Model-free representation learning and exploration in low-rank MDPs. *Journal of Machine Learning Research*, 25(6):1–76, 2024.

Junhyuk Oh, Satinder Singh, and Honglak Lee. Value prediction network. In *Advances in Neural Information Processing Systems 30*, 2017.

Ronald Parr, Lihong Li, Gavin Taylor, Christopher Painter-Wakefield, and Michael L. Littman. An analysis of linear models, linear value-function approximation, and feature selection for reinforcement learning. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 752–759, 2008. doi: 10.1145/1390156.1390251.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588 (7839):604–609, 2020. doi: 10.1038/s41586-020-03051-4.

David Silver, Hado van Hasselt, Matteo Hessel, Tom Schaul, Arthur Guez, Tim Harley, Gabriel Dulac-Arnold, David Reichert, Neil Rabinowitz, and André Barreto. The predictron: End-to-end learning and planning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3191–3199, 2017.

Zhao Song, Ronald E. Parr, Xuejun Liao, and Lawrence Carin. Linear feature encoding for reinforcement learning. In *Advances in Neural Information Processing Systems 29*, 2016.

Aviv Tamar, Yi Wu, Garrett Thomas, Sergey Levine, and Pieter Abbeel. Value iteration networks. In *Advances in Neural Information Processing Systems 29*, 2016.

Jonathan Taylor, Doina Precup, and Prakash Panangaden. Bounding performance loss in approximate MDP homomorphisms. In *Advances in Neural Information Processing Systems 21*, pp. 1649–1656, 2008.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *The Tenth International Conference on Learning Representations*, 2022.

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline RL in low-rank MDPs. In *International Conference on Learning Representations*, 2022.

Elise van der Pol, Thomas Kipf, Frans A. Oliehoek, and Max Welling. Plannable approximations to MDP homomorphisms: Equivariance under actions. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1431–1439, 2020.

Manuel Watter, Jost Tobias Springenberg, Joschka Boedecker, and Martin Riedmiller. Embed to control: A locally linear latent dynamics model for control from raw images. In *Advances in Neural Information Processing Systems 28*, 2015.

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

Marvin Zhang, Sharad Vikram, Laura Smith, Pieter Abbeel, Matthew Johnson, and Sergey Levine. SOLAR: Deep structured representations for model-based reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7444–7453, 2019.

A Proofs of the Structural Results

A.1 Proof of the finite lattice lemma

We give the finite construction behind Lemma 2. It is the step that turns closure under pointwise maxima into explicit indicators of the states identified by the value space.

Proof of Lemma 2. Let $C_1, \dots, C_k$ be the equivalence classes of $\sim_L$, and choose $x_i \in C_i$. Every $f \in L$ is constant on each C_i. Hence, for

$$G = \{f \in \mathbb{R}^S : f \text{ is constant on every } C_i\},$$

we have $L \subseteq G$. We show that each indicator $\mathbf{1}_{C_i}$ is in L.

Choose a basis $f_1, \dots, f_m$ of L. For $i \neq j$, the coefficient vectors for which a linear combination has equal values at x_i and x_j form

$$H_{ij} = \left\{ c \in \mathbb{R}^m : \sum_{\ell=1}^{m} c_\ell \big(f_\ell(x_i) - f_\ell(x_j)\big) = 0 \right\}.$$

This is a proper hyperplane. Otherwise every member of L would agree at x_i and x_j, which would put the two representatives in the same equivalence class. A finite union of proper hyperplanes cannot cover $\mathbb{R}^m$, so we choose c outside their union and set

$$h = \sum_{\ell=1}^{m} c_\ell f_\ell \in L.$$

The values of h are distinct across the classes. Relabel the classes so that

$$h|_{C_i} = \beta_i, \qquad \beta_1 < \beta_2 < \cdots < \beta_k.$$

If $k = 1$, all members of L are constant and $\mathbf{1} \in L$, so $L = G = \operatorname{span}\{\mathbf{1}\}$. Assume $k \geq 2$. Max closure and linearity imply closure under positive parts and minima: $x_+ = x \vee 0$ and $x \wedge y = -((-x) \vee (-y))$. Define

$$e_1 = \left(\frac{\beta_2 \mathbf{1} - h}{\beta_2 - \beta_1} \right)_+, \qquad e_k = \left(\frac{h - \beta_{k-1}\mathbf{1}}{\beta_k - \beta_{k-1}} \right)_+,$$

and, for $1 < i < k$,

$$e_i = \left(\frac{h - \beta_{i-1}\mathbf{1}}{\beta_i - \beta_{i-1}} \right)_+ \wedge \left(\frac{\beta_{i+1}\mathbf{1} - h}{\beta_{i+1} - \beta_i} \right)_+.$$

All denominators are positive and all operations preserve L. Evaluating at the levels β_j shows that e_i equals one on C_i and zero on every other class. Thus $e_i = \mathbf{1}_{C_i}$.

The class indicators span G, so $G \subseteq L$, and hence $L = G$. They are linearly independent because their supports are disjoint and nonempty. Therefore $k = \dim L$. The equivalence relation is defined only by L, and any partition whose block-constant space is L has precisely these equivalence classes. This proves uniqueness up to reordering. □

A.2 Proof of the finite-state characterization

Proof of Theorem 3. Assume that F is universally Bellman complete. For arbitrary $f, g \in F$, choose two distinct actions, set their rewards to f and g, set the continuation to zero, and give every remaining action reward $-M\mathbf{1}$, where $-M < \min_s\{f(s), g(s)\}$. The update is $f \vee g$ by equation 3. Thus F is closed under pointwise maxima, and Lemma 2 gives a unique partition $\mathcal{P} = \{C_1, \dots, C_d\}$ with $F = F_{\mathcal{P}}$.

Fix $a \in A$ and $v \in F$. Let $\operatorname{osc}(v) = \max_s v(s) - \min_s v(s)$, choose $M > \gamma \operatorname{osc}(v)$, set $r_a = 0$, and set $r_b = -M\mathbf{1}$ for $b \neq a$. Row stochasticity gives

$$\min_x v(x) \leq (P_b v)(s) \leq \max_x v(x)$$

for every b and s. Consequently,

$$\big[r_a(s) + \gamma (P_a v)(s)\big] - \big[r_b(s) + \gamma (P_b v)(s)\big] \geq M - \gamma \operatorname{osc}(v) > 0.$$

Action a is selected at every state, so $B_r v = \gamma P_a v$. Universal completeness and $\gamma > 0$ imply $P_a v \in F$, and hence $P_a F \subseteq F$ for every action.

Now fix $s, t \in C_i$ and a target block C_j. Since $P_a \mathbf{1}_{C_j} \in F = F_{\mathcal{P}}$, it is constant on C_i. Therefore

$$\sum_{x \in C_j} P_a(s, x) = (P_a \mathbf{1}_{C_j})(s) = (P_a \mathbf{1}_{C_j})(t) = \sum_{x \in C_j} P_a(t, x),$$

which is lumpability of P_a with respect to $\mathcal{P}$.

Conversely, suppose $F = F_{\mathcal{P}}$ and every P_a is lumpable with respect to $\mathcal{P}$. Take arbitrary $r \in F^A$ and $v \in F$. Write $v = \sum_j q_j \mathbf{1}_{C_j}$. For $s \in C_i$,

$$(P_a v)(s) = \sum_j q_j \sum_{x \in C_j} P_a(s, x).$$

Every inner sum is constant as s ranges over C_i, so $P_a v$ is block-constant. The same is true of each r_a. Thus every action value in the Bellman maximum is constant on C_i, and taking the maximum over actions preserves this equality. Hence $B_r v \in F_{\mathcal{P}} = F$, proving universal Bellman completeness. □

The block-indicator calculation also proves the elementary equivalence

$$P_a F_{\mathcal{P}} \subseteq F_{\mathcal{P}} \quad \Longleftrightarrow \quad P_a \text{ is lumpable with respect to } \mathcal{P}.$$

The forward direction applies P_a to each $\mathbf{1}_{C_j}$; the reverse direction expands an arbitrary block-constant function in the indicator basis.

A.3 Positive order units

The following extension is useful when the represented space contains a strictly positive function in place of the constant function.

Theorem 9 (Positive-order-unit characterization)**.** *Let $F \leq \mathbb{R}^S$ contain u with $u(s) > 0$ for every state. If $|A| \geq 2$ and $\gamma > 0$, then universal closure for all $r \in F^A$ and $v \in F$ holds if and only if there is a unique unordered partition $\mathcal{P} = \{C_1, \ldots, C_d\}$ such that*

$$F = D_u F_{\mathcal{P}}, \qquad D_u = \mathrm{diag}(u),$$

and, for every a, source block C_i, target block C_j, and $s, t \in C_i$,

$$\frac{1}{u(s)} \sum_{x \in C_j} P_a(s, x) u(x) = \frac{1}{u(t)} \sum_{x \in C_j} P_a(t, x) u(x).$$

The partition does not depend on which strictly positive member of F is chosen as u.

Proof of Theorem 9. For $f, g \in F$, assign them to two action rewards, assign f to every remaining action, and take zero continuation. The resulting update is $f \vee g$, so F is a vector sublattice. Put $G = D_u^{-1} F$. Then $\mathbf{1} \in G$, and positivity of u gives

$$D_u^{-1}(f \vee g) = (D_u^{-1} f) \vee (D_u^{-1} g).$$

Lemma 2 yields a unique partition with $G = F_{\mathcal{P}}$, which is equivalent to $F = D_u F_{\mathcal{P}}$.

Fix a and $v \in F$, and let $m = \min_s u(s) > 0$. Choose

$$M > \frac{\gamma}{m} \max_{\substack{s \in S \\ b \neq a}} \left| (P_b v)(s) - (P_a v)(s) \right|.$$

Set $r_a = 0$ and $r_b = -Mu$ for $b \neq a$. The selected action is unique, so universal closure gives $P_a v \in F$. The functions $u\mathbf{1}_{C_j}$ form a basis of F. Applying P_a to one of them and dividing by $u(s) > 0$ gives the weighted block-mass identity.

For sufficiency, the weighted identities imply $P_a F \subseteq F$ by linearity on the basis $u\mathbf{1}_{C_j}$. Every action value $r_a + \gamma P_a v$ belongs to F. If $f = uq$ and $g = uq'$ with $q, q' \in F_{\mathcal{P}}$, then $f \vee g = u(q \vee q')$, so the maximum over actions also belongs to F. Finally, the classes of $D_u^{-1} F = F_{\mathcal{P}}$ are equivalently characterized by $f(s)u(t) = f(t)u(s)$ for every $f \in F$. If w is another positive member of F, then w/u is positive and block-constant, and multiplication by its reciprocal preserves $F_{\mathcal{P}}$. The recovered partition is therefore unchanged. □

B Proofs for Quotient Planning

B.1 Intertwining identity

Let Ψ be the block-indicator matrix and let $\bar{P}_a$ be the quotient transition matrix from Section 4.1. For $s \in C_i$,

$$(P_a \Psi q)(s) = \sum_{x \in S} P_a(s,x)(\Psi q)(x)$$
$$= \sum_{j=1}^{d} q_j \sum_{x \in C_j} P_a(s,x) = \sum_{j=1}^{d} \bar{P}_a(i,j) q_j = (\bar{P}_a q)_i.$$

Lifting the final expression gives $P_a \Psi = \Psi \bar{P}_a$. If $r_a = \Psi \rho_a$, the same calculation inside the action maximum gives

$$(B_r(\Psi q))(s) = \max_{a \in A} \{\rho_a(i) + \gamma(\bar{P}_a q)_i\} = (\bar{B}_\rho q)_i = (\Psi \bar{B}_\rho q)(s).$$

This proves equation 6.

B.2 Finite-horizon planning

For $q_0 \in \mathbb{R}^d$, the lifted identity is immediate at horizon zero. If

$$B_r^H(\Psi q_0) = \Psi \bar{B}_\rho^H q_0,$$

then applying the one-step identity gives

$$B_r^{H+1}(\Psi q_0) = B_r\big(\Psi \bar{B}_\rho^H q_0\big) = \Psi \bar{B}_\rho^{H+1} q_0.$$

Induction proves the finite-horizon equality. At each stage, an action maximizing the quotient expression at block i also maximizes the ground expression at every state in C_i. A blockwise lift therefore attains the ground finite-horizon value.

B.3 Discounted fixed points and policy lifting

For any $u, v \in \mathbb{R}^S$, row stochasticity gives

$$|(P_a u)(s) - (P_a v)(s)| \le \|u - v\|_\infty.$$

The inequality

$$\Big|\max_a x_a - \max_a y_a\Big| \le \max_a |x_a - y_a|$$

then makes the ground and quotient Bellman operators γ-contractions. They have unique fixed points. Applying the intertwining identity at the quotient fixed point gives

$$B_r(\Psi q^*) = \Psi \bar{B}_\rho q^* = \Psi q^*.$$

Uniqueness of the ground fixed point yields $v^* = \Psi q^*$.

Choose a quotient-greedy action at every block and lift it to a stationary policy π^*. The selected action satisfies the ground Bellman equation at every state in its block, so $T_{\pi^*} v^* = B_r v^* = v^*$. The policy operator is a contraction, and its unique fixed point is $v^{\pi^*} = v^*$. Thus the lifted policy is optimal.

B.4 Encoder coordinate changes

The columns of Ψ and a full-rank encoder Φ spanning F are two bases of the same space. There is therefore an invertible matrix T with

$$\Phi = \Psi T.$$

Using $P_a \Psi = \Psi \bar{P}_a$ gives

$$P_a \Phi = \Phi\big(T^{-1} \bar{P}_a T\big).$$

For $r_a = \Phi w_a$ and $v = \Phi\theta$, write $r_a = \Psi(T w_a)$ and $v = \Psi(T\theta)$. Substitution into the quotient identity yields

$$B_r(\Phi\theta) = \Phi\left[T^{-1} \bar{B}_{Tw}(T\theta)\right].$$

Thus the represented subspace and its partition are unchanged by an invertible change of encoder coordinates.

C Proofs for Finite Certification

C.1 Exact certificate

Proof of Theorem 4. Let $\mathcal{P}_\Phi$ be the partition induced by equal rows of Φ. For any coefficient vector w, the function Φw is constant on every row class, so

$$F = \mathrm{col}(\Phi) \subseteq F_{\mathcal{P}_\Phi}.$$

Evaluation on one representative of each row class is injective on F: a represented function that vanishes at all representatives vanishes on every state. Thus $d \le k_\Phi$. If $k_\Phi = d$, the two finite-dimensional spaces in the inclusion have the same dimension, and $F = F_{\mathcal{P}_\Phi}$.

Under $k_\Phi = d$, the block indicators $e_j = \mathbf{1}_{C_j}$ belong to F. Fix a and j, set $v = e_j$, $r_a = 0$, and $r_b = -M\mathbf{1}$ for $b \neq a$, with $M > \gamma$. Since $0 \le (P_b v)(s) \le 1$,

$$\gamma(P_a v)(s) - [-M + \gamma(P_b v)(s)] \ge M - \gamma > 0.$$

The selected action is unique and $B_r v = \gamma P_a e_j$. Membership of this update is therefore equivalent to $P_a e_j \in F$. Testing all a, j is equivalent to $P_a F \subseteq F$ for every action, because the e_j form a basis. On a block space this is exactly lumpability, so the two finite conditions imply universal Bellman completeness by Theorem 3.

Conversely, suppose F is universally Bellman complete. Theorem 3 gives a unique d-block partition $\mathcal{P}$ with $F = F_{\mathcal{P}}$. States in one block agree on every column of Φ, hence have equal rows. States in different blocks are separated by a block indicator in F, so their rows cannot be equal. Thus $\mathcal{P}_\Phi = \mathcal{P}$, $k_\Phi = d$, and the required mass equalities follow from the theorem. □

C.2 Bounded adaptive probes

Proof of Proposition 5. Let $k = k_\Phi$. If $k = 1$, injective evaluation gives $d \le 1$, while $\mathbf{1} \in F$ gives $d \ge 1$. Hence $F = \mathrm{span}\{\mathbf{1}\}$, and row stochasticity makes every admissible Bellman update constant. No query is needed.

Assume $k \ge 2$. Choose $h \in F$ whose values $\beta_1 < \cdots < \beta_k$ are distinct on the row classes. This is possible by the same finite-hyperplane argument used in the proof of Lemma 2. We describe how to query the positive part of each affine ramp using rewards whose sup norm is at most one.

For a known $x \in F$, let $L_x = \max\{1, \|x\|_\infty\}$. Assign x/L_x and zero to two actions, assign $-\mathbf{1}$ to every other action, and use zero continuation. The Bellman output is $(x/L_x)_+$, because the first two heads have maximum at least zero and all other heads equal -1. Membership of this output is equivalent to $x_+ \in F$, and all queried rewards have sup norm at most one.

Use this positive-part query for the two extreme ramps and for the two ramps in each interior tent. After the two positive parts p, q of an interior tent have passed, let $L_{pq} = \max\{1, \|p\|_\infty, \|q\|_\infty\}$, assign $-p/L_{pq}$ and $-q/L_{pq}$ to two actions, assign $-\mathbf{1}$ to the remaining actions, and use zero continuation. The output is $-(p \wedge q)/L_{pq}$, so its membership is equivalent to $p \wedge q \in F$. The two extreme indicators use one query each and each of the $k-2$ interior indicators uses three queries. The order-structure part therefore uses $2 + 3(k-2) = 3k - 4$ queries.

If all these queries pass, the tent formulas put all k row-class indicators in F. They are linearly independent, so $k \le d$. The row evaluation inequality already gives $d \le k$, hence $k = d$ and $F = F_{\mathcal{P}_\Phi}$. For every a, j, use $v = \mathbf{1}_{C_j}$, reward zero on a, reward $-\mathbf{1}$ on every competitor, and zero continuation. The selected action wins by at least $1-\gamma$, so the output is $\gamma P_a \mathbf{1}_{C_j}$. The $|A|k$ resulting membership bits are equivalent to transition invariance on the indicator basis, hence to lumpability. If all queries pass, Theorem 3 gives universal Bellman completeness. The converse follows because universal completeness makes every constructed output a member of F. All continuations are zero or indicators, and all reward heads were scaled to have sup norm at most one. □

C.3 Finite optimization and sampled defect

The next two results make the defect computable or statistically estimable when the transition rows are accessed directly.

Proposition 10 (Exact finite optimization). *Suppose Φ, all P_a, R, and γ are rational. Then $\mathcal{E}_R(F)$ is the maximum of a finite family of rational linear programs. Let*

$$K_F = \{y \in \mathbb{R}^S : \Phi^\top y = 0,\ \|y\|_1 \le 1\}.$$

The family enumerates a statewise maximizing-action map $\sigma : S \to A$ and a vertex of K_F. For $k_\Phi \ge 2$, Δ_F is the maximum of at most $k_\Phi!$ linear programs, one for each ordering of the row classes.

Proof of Proposition 10. Finite-dimensional ℓ_∞/ℓ_1 duality gives

$$\operatorname{dist}_\infty(z, F) = \max_{\substack{y \in \mathbb{R}^S \\ \Phi^\top y = 0,\ \|y\|_1 \le 1}} y^\top z.$$

Write $v = \Phi\theta$ and $r_a = \Phi w_a$. The constraints $\|v\|_\infty \le 1$ and $\|r_a\|_\infty \le R$ are finite linear inequalities in θ, w_a. For a map $\sigma : S \to A$, impose

$$(\Phi w_{\sigma(s)})(s) + \gamma(P_{\sigma(s)}\Phi\theta)(s) \ge (\Phi w_b)(s) + \gamma(P_b\Phi\theta)(s) \quad \text{for every } s, b.$$

On this polyhedral cell the Bellman output is affine. Fixing a vertex of K_F makes the dual objective linear, so maximizing it on the cell is a rational linear program. The cells cover the compact admissible domain, including tie boundaries, and the finite vertex set gives the exact supremum.

For Δ_F, choose one representative row ϕ_i for each class. For each permutation π, maximize t subject to

$$-\mathbf{1} \le \Phi\theta \le \mathbf{1}, \qquad (\phi_{\pi(i+1)} - \phi_{\pi(i)})\theta \ge t \quad (1 \le i < k_\Phi).$$

The smallest pairwise gap of an ordered scalar list is an adjacent gap. Every feasible function has an ordering, and every solution with $t > 0$ realizes that ordering. Maximizing over the finitely many permutations gives Δ_F. □

Proposition 11 (Sampled defect certificate). *Let $\widehat{P}_a(s, \cdot)$ be formed from N independent samples at each state-action pair, and let $\widehat{\mathcal{E}}_R(F)$ use these rows in place of P_a. If $\eta, \delta \in (0, 1)$ and*

$$N \ge \frac{2}{\eta^2}\left(n \log 2 + \log \frac{2|A|n}{\delta}\right),$$

then, with probability at least $1 - \delta$,

$$\max_{a,s} \|\widehat{P}_a(s, \cdot) - P_a(s, \cdot)\|_1 \le \eta, \qquad \left|\widehat{\mathcal{E}}_R(F) - \mathcal{E}_R(F)\right| \le \gamma\eta.$$

Proof of Proposition 11. For a fixed row and subset $J \subseteq S$, Hoeffding's inequality gives

$$\Pr\{|\widehat{P}_a(s, J) - P_a(s, J)| > \eta/2\} \le 2e^{-N\eta^2/2}.$$

Using $\|p - q\|_1 = 2\max_J |p(J) - q(J)|$, a union bound over 2^n subsets and $|A|n$ rows gives failure probability at most $|A|n2^{n+1}e^{-N\eta^2/2}$, which is at most δ under the stated sample size. On this event,

$$\max_a \|(\widehat{P}_a - P_a)v\|_\infty \le \eta\|v\|_\infty \le \eta$$

for every admissible continuation. The max inequality then gives $\|B_r^{\widehat{P}}v - B_r^P v\|_\infty \le \gamma\eta$. Distance to a fixed subspace is one-Lipschitz, so taking the two suprema proves the defect bound. □

C.4 Finite-precision row and anchor recovery

Proposition 12 (Separation-stable block recovery). *Let $\mathcal{P} = \{C_1, \dots, C_d\}$, let $t_1, \dots, t_d \in \mathbb{R}^d$ be distinct, and suppose*

$$\|\widetilde{\Phi}_{s,:} - t_i\|_\infty \le \nu \qquad (s \in C_i).$$

Set $m = \min_{i \ne j} \|t_i - t_j\|_\infty$. If $\nu < m/4$, then

$$s, t \text{ lie in the same block} \iff \|\widetilde{\Phi}_{s,:} - \widetilde{\Phi}_{t,:}\|_\infty \le 2\nu.$$

Pairwise thresholding costs $O(n^2 d)$ operations.

Proof of Proposition 12. If $s, t \in C_i$, the triangle inequality gives

$$\|\widetilde{\Phi}_{s,:} - \widetilde{\Phi}_{t,:}\|_\infty \le 2\nu.$$

If $s \in C_i, t \in C_j$ with $i \neq j$, the reverse triangle inequality gives

$$\|\widetilde{\Phi}_{s,:} - \widetilde{\Phi}_{t,:}\|_\infty \ge m - 2\nu > 2\nu.$$

Thus the threshold recovers the blocks. Computing all pairwise distances uses d coordinates for each of $O(n^2)$ pairs. □

Proposition 13 (Certified finite-precision anchors)**.** *Let Φ be an exact encoder and $\widetilde{\Phi}$ satisfy $\max_{s,j} |\widetilde{\Phi}_{sj} - \Phi_{sj}| \le \rho$. Choose a nonsingular observed anchor matrix $\widetilde{A} = \widetilde{\Phi}_{I,:}$, and set*

$$a = \|\widetilde{A}^{-1}\|_{\infty\to\infty}, \quad q = d\rho a, \quad \widetilde{Z} = \widetilde{\Phi}\widetilde{A}^{-1}, \quad \widetilde{\kappa} = \max_s \|\widetilde{Z}_{s,:}\|_1.$$

If $q < 1$, the exact anchor matrix $A = \Phi_{I,:}$ is nonsingular and, for $Z = \Phi A^{-1}$,

$$\|Z - \widetilde{Z}\|_{\infty\to\infty} \le \omega := \frac{d\rho a(1 + \|\widetilde{\Phi}\|_{\infty\to\infty} a)}{1 - q}, \qquad \kappa \le \widetilde{\kappa} + \omega,$$

where $\kappa = \max_s \|Z_{s,:}\|_1$. If a defect upper bound ε_U satisfies

$$\beta_U = \varepsilon_U(\widetilde{\kappa} + \omega)(1 + \widetilde{\kappa} + \omega) < 1/d, \qquad 2\omega < 1/d - \beta_U,$$

then the argmax partition computed from $\widetilde{Z}$ equals the partition obtained from Z.

Proof of Proposition 13. Let $E = \widetilde{\Phi} - \Phi$. Entrywise error gives $\|E\|_{\infty\to\infty} \le d\rho$, and the same bound holds for the anchor submatrix E_I. Since $A = \widetilde{A} - E_I$,

$$A = \widetilde{A}(I - \widetilde{A}^{-1}E_I).$$

The norm of the second factor's perturbation is at most $q < 1$, so the Neumann series gives invertibility and $\|A^{-1}\|_{\infty\to\infty} \le a/(1 - q)$. Also

$$A^{-1} - \widetilde{A}^{-1} = A^{-1}E_I\widetilde{A}^{-1}.$$

Adding and subtracting $\widetilde{\Phi}A^{-1}$, then applying the preceding bounds, gives the displayed ω estimate. Since $\kappa = \max_s \|Z_{s,:}\|_1$, the same estimate gives $\kappa \le \widetilde{\kappa} + \omega$.

The exact Lagrange coordinates have one winning coordinate at least $1/d$ and all other coordinates at most β_U, by the anchor recovery proof below. Their gap is at least $1/d - \beta_U$. Every observed coordinate is within ω, and 2ω is smaller than this gap, so all argmax decisions are preserved. □

C.5 A NOISY END-TO-END CERTIFICATE

Theorem 14 (Noisy recovery, certification, and planning)**.** *Assume the perturbation model of Proposition 12 with $\nu < m/4$. Recover the blocks by its threshold rule. From N samples at each ground state-action pair, aggregate next states into recovered blocks and let $\widehat{p}_a^s$ be the empirical block distribution. Define*

$$\widehat{\lambda} = \max_{\substack{a,i \\ s,t\in C_i}} \mathrm{TV}(\widehat{p}_a^s, \widehat{p}_a^t).$$

For $\xi \in (0, 2]$, $\delta \in (0, 1)$, if

$$N \ge \frac{2}{\xi^2}\left(d\log 2 + \log\frac{2|A|n}{\delta}\right),$$

then with probability at least $1 - \delta$,

$$|\widehat{\lambda} - \lambda| \le \xi, \qquad \max_{a,s} \|\widehat{p}_a^s - p_a^s\|_1 \le \xi.$$

For $0 \le \lambda_0 < \lambda_1 \le 1$ and $\xi < (\lambda_1 - \lambda_0)/2$, thresholding $\widehat{\lambda}$ at $(\lambda_0 + \lambda_1)/2$ separates the two regimes. If $\widehat{q}^*$ and $\widehat{\pi}$ are obtained by planning on empirical quotient rows from one representative per recovered block, then

$$\|v^* - \Psi\widehat{q}^*\|_\infty \le \frac{\gamma(2\widehat{\lambda} + 3\xi)R_{\max}}{(1 - \gamma)^2}, \qquad \|v^* - v^{\widehat{\pi}}\|_\infty \le \frac{2\gamma(2\widehat{\lambda} + 3\xi)R_{\max}}{(1 - \gamma)^2}.$$

Proof of Theorem 14. The threshold proposition recovers the true blocks deterministically. For a single ground row, the subset argument with d block categories gives

$$\Pr\{\|\widehat{p}_a^s - p_a^s\|_1 > \xi\} \le 2^{d+1} e^{-N\xi^2/2}.$$

A union bound over $|A|n$ rows proves the row event. On this event,

$$\left|\mathrm{TV}(\widehat{p}_a^s, \widehat{p}_a^t) - \mathrm{TV}(p_a^s, p_a^t)\right| \le \mathrm{TV}(\widehat{p}_a^s, p_a^s) + \mathrm{TV}(\widehat{p}_a^t, p_a^t) \le \xi,$$

which proves the bound on $\widehat{\lambda}$ and the testing statement.

Let $\bar{P}$ be the reference quotient formed from the true representative rows and let $\bar{q}^*$ be its optimal fixed point. Applying the approximate planning bound to the true quotient gives

$$\|v^* - \Psi\bar{q}^*\|_\infty \le \frac{2\gamma\lambda R_{\max}}{(1-\gamma)^2}.$$

The empirical and true quotient operators differ by at most $\gamma\xi R_{\max}/(1-\gamma)$ on the relevant fixed-point range, so contraction comparison gives

$$\|\bar{q}^* - \widehat{q}^*\|_\infty \le \frac{\gamma\xi R_{\max}}{(1-\gamma)^2}.$$

Using $\lambda \le \widehat{\lambda} + \xi$ and the triangle inequality proves the value bound. For the policy, compare the empirical policy with its true quotient value; the fixed-policy contraction argument contributes $2\gamma\xi R_{\max}/(1-\gamma)^2$, while lifting through the approximate quotient contributes $2\gamma\lambda R_{\max}/(1-\gamma)^2$. Adding the three terms and substituting $\lambda \le \widehat{\lambda} + \xi$ gives the stated policy bound. □

C.6 Quotient-row sampling

Theorem 15 (Quotient sampling and planning)**.** *Assume the exact conditions of Theorem 3, let $d = |\mathcal{P}|$, and sample N next states for each representative state-action pair. Let $\widehat{P}_a(i,j)$ be the empirical target-block frequency. If*

$$N \ge \frac{1}{2\tau^2}\log\frac{2|A|d^2}{\delta},$$

then with probability at least $1-\delta$,

$$\max_{a,i,j} |\widehat{P}_a(i,j) - \bar{P}_a(i,j)| \le \tau.$$

For block-readable rewards with $\max_a \|\rho_a\|_\infty \le R_{\max}$, the empirical and true quotient fixed points satisfy

$$\|\widehat{q}^* - q^*\|_\infty \le \frac{\gamma d\tau R_{\max}}{(1-\gamma)^2}, \qquad \|v^* - v^{\widehat{\pi}}\|_\infty \le \frac{2\gamma d\tau R_{\max}}{(1-\gamma)^2}.$$

Alternatively, if

$$N \ge \frac{2}{\eta^2}\left(d\log 2 + \log\frac{2|A|d}{\delta}\right),$$

then $\max_{a,i} \|\widehat{P}_a(i,\cdot) - \bar{P}_a(i,\cdot)\|_1 \le \eta$, and the two bounds above hold with $d\tau$ replaced by η.

Proof of Theorem 15. For fixed a, i, j, the block indicator of a sampled next state is Bernoulli with mean $\bar{P}_a(i,j)$. Hoeffding's inequality and a union bound over $|A|d^2$ triples give the entrywise bound. On this event,

$$\|(\widehat{P}_a - \bar{P}_a)q\|_\infty \le d\tau\|q\|_\infty.$$

The max inequality gives

$$\|\widehat{\bar{B}}_\rho q - \bar{B}_\rho q\|_\infty \le \gamma d\tau\|q\|_\infty.$$

Both operators are contractions and their fixed points have norm at most $R_{\max}/(1-\gamma)$. Evaluating the last inequality at an empirical fixed point and rearranging gives the value bound. Comparing the empirical greedy policy with its true quotient policy operator gives the same bound once more; the triangle inequality gives the policy bound.

For the row-ℓ_1 result, apply Hoeffding to every subset of the d target blocks. The identity $\|p-q\|_1 = 2\max_J |p(J) - q(J)|$ and a union bound over $2^d|A|d$ events give the stated sample size and row bound. Replacing $d\tau$ by η in the contraction argument proves the final two inequalities. □

D Quantitative Proofs

D.1 Approximate lattice operations

Lemma 16 (Approximate lattice operations)**.** *Let $R \geq 1$, $\varepsilon = \mathcal{E}_R(F)$, and $f, g \in F$. Then*

$$\operatorname{dist}_\infty(f \vee g, F) \leq \varepsilon \max\{\|f\|_\infty, \|g\|_\infty\}.$$

Consequently,

$$\operatorname{dist}_\infty(f_+, F) \leq \varepsilon \|f\|_\infty,$$

and the same bound holds for $f \wedge g$.

Proof of Lemma 16. For unit-bounded f, g, use them as two rewards, use zero continuation, and assign $-R\mathbf{1}$ to every remaining action. The Bellman output is $f \vee g$, so its distance is at most ε. For general f, g, divide by $L = \max\{\|f\|_\infty, \|g\|_\infty\}$ when $L > 0$, apply the unit bound, and use homogeneity of distance to a linear subspace. The positive-part statement takes $g = 0$, and the minimum statement follows from $f \wedge g = -((-f) \vee (-g))$. □

D.2 Exact defect identity

For $z \in \mathbb{R}^S$, the distance to a block-constant space satisfies

$$\operatorname{dist}_\infty(z, F_{\mathcal{P}}) = \frac{1}{2} \max_i \left\{ \max_{s \in C_i} z(s) - \min_{s \in C_i} z(s) \right\}.$$

The midpoint of the minimum and maximum on each block attains this distance, and no block constant can do better.

Proof of Theorem 6. Fix admissible r, v, and write $v = \Psi q$ with $\|q\|_\infty \leq 1$. For $s, t \in C_i$, rewards agree for every action. The total-variation inequality gives

$$\left| \sum_j (p_a^s(j) - p_a^t(j)) q_j \right| \leq 2 \operatorname{TV}(p_a^s, p_a^t) \|q\|_\infty \leq 2\lambda.$$

Thus the action values at s and t differ by at most $2\gamma\lambda$. The maximum over actions preserves this range, and the block-distance formula gives $\operatorname{dist}_\infty(B_r v, F) \leq \gamma\lambda$. Taking the supremum proves the upper bound.

For the reverse inequality, choose a, i, s, t and a set J of target blocks attaining λ, with the order of s, t chosen so that $p_a^s(J) - p_a^t(J) = \lambda$. Set

$$v = 2\mathbf{1}_{\bigcup_{j \in J} C_j} - \mathbf{1}, \qquad r_a = \mathbf{1}, \qquad r_b = -\mathbf{1} \quad (b \neq a).$$

The selected action wins by at least $2(1 - \gamma) > 0$, and the Bellman output is $\mathbf{1} + \gamma P_a v$. Its values at s, t differ by $2\gamma\lambda$. The block-distance formula gives distance at least $\gamma\lambda$, proving equality. □

D.3 Margin-conditioned recovery

Proof of Theorem 7. Let $C_1, \ldots, C_k$ be the classes of $\mathcal{P}_F$, with representatives s_i. A maximizer $h \in F$ of Δ_F exists by compactness of the unit ball in finite dimension. Relabel the classes so that

$$-1 \leq \beta_1 < \cdots < \beta_k \leq 1, \qquad h|_{C_i} = \beta_i.$$

Every adjacent gap is at least Δ_F. The affine ramps in the lattice construction have norm at most $2/\Delta_F$. Lemma 16 therefore gives an element of F within $2\varepsilon/\Delta_F$ of each extreme indicator.

For an interior indicator, approximate the two positive parts first. Their approximants have norm at most $2(1 + \varepsilon)/\Delta_F \leq 4/\Delta_F$, since $\varepsilon < 1$. The approximate minimum contributes at most $4\varepsilon/\Delta_F$, while the two positive-part errors contribute $2\varepsilon/\Delta_F$ in total. Hence for every class there is $g_i \in F$ with

$$\|g_i - \mathbf{1}_{C_i}\|_\infty \leq \frac{6\varepsilon}{\Delta_F}.$$

Let $G_{ij} = g_j(s_i)$. Then

$$\|G - I_k\|_{\infty\to\infty} \le \frac{6k\varepsilon}{\Delta_F} < 1.$$

The Neumann series makes G invertible. Its columns are evaluations of functions in the d-dimensional space F, so $k \le d$. Evaluation on the representatives is injective, so $d \le k$. Thus $k = d$, and the inclusion $F \subseteq F_{\mathcal{P}_F}$ becomes equality.

To bound transitions, fix a, i, j, write $e_j = \mathbf{1}_{C_j}$ and $v_j = 2e_j - \mathbf{1}$, and choose $r_a = \mathbf{1}$, $r_b = -\mathbf{1}$ for $b \neq a$. The selected action has margin at least $2(1-\gamma)$, so $B_r v_j = \mathbf{1} + \gamma P_a v_j$. Choose $u \in F$ within ε of this output. For $s, t \in C_i$, $u(s) = u(t)$, while $(P_a v_j)(s) = 2p_a^s(j) - 1$. Therefore

$$2\gamma|p_a^s(j) - p_a^t(j)| \le 2\varepsilon,$$

and each coordinate mismatch is at most ε/γ. Repeating the construction with $v_J = 2\mathbf{1}_{\bigcup_{j\in J} C_j} - \mathbf{1}$ for every J and using $\mathrm{TV}(p,q) = \max_J |p(J) - q(J)|$ gives $\lambda \le \varepsilon/\gamma$. □

D.4 Anchor-conditioned recovery (supplementary)

For completeness, we record the continuous recovery result used by the finite-precision audits. For a full-rank encoder, choose d anchor states $I = \{x_1, \dots, x_d\}$ such that $A = \Phi_{I,:}$ is nonsingular, and let

$$Z = \Phi A^{-1}, \qquad \kappa_I = \max_s \|Z_{s,:}\|_1, \qquad \kappa = \min_I \kappa_I.$$

Let $\ell_i \in F$ be the Lagrange functions satisfying $\ell_i(x_j) = \mathbf{1}_{\{i\}}(j)$. Define the symmetric gap

$$\mathrm{gap}_\infty(U, V) = \max\left\{ \sup_{\substack{u\in U\\ \|u\|_\infty\le 1}} \mathrm{dist}_\infty(u, V), \sup_{\substack{v\in V\\ \|v\|_\infty\le 1}} \mathrm{dist}_\infty(v, U) \right\}.$$

Proposition 17 (Universal anchor conditioning)**.** *Every unital d-dimensional subspace satisfies $1 \le \kappa \le d$. For every odd d, a unital subspace on $d+1$ states attains $\kappa = d$.*

Proof of Proposition 17. Choose a nonsingular d-row submatrix A of maximum absolute determinant. Cramer's rule shows that every coordinate of every row of $Z = \Phi A^{-1}$ has absolute value at most one, because replacing one row of A by any encoder row cannot increase the maximal determinant. Hence $\kappa \le d$, while $\kappa \ge 1$ follows by comparing the maximum of a nonzero represented function with its values on the anchors.

For odd d, take the rows $e_1^\top, \dots, e_d^\top, z^\top$, where $z \in \{-1, 1\}^d$ and $\sum_i z_i = 1$. The column space contains $\mathbf{1}$. For the anchor set of the first d rows, the omitted row has coefficient vector z and ℓ_1-norm d. If row $e_j^\top$ is omitted, then

$$e_j^\top = z_j \left(z^\top - \sum_{k\neq j} z_k e_k^\top \right),$$

whose coefficient vector also has ℓ_1-norm d. Every anchor set therefore has condition at least d, attaining the upper bound. □

Theorem 18 (Anchor-conditioned robust nearness)**.** *Let $d \ge 2$, $R \ge 1$, and $\varepsilon = \mathcal{E}_R(F)$. Set $\beta = \varepsilon\kappa(1+\kappa)$. If $\beta < 1/d$, assign each state to*

$$C_i = \{s : i = \arg\max_j \ell_j(s)\}.$$

Then the maximizer is unique, every C_i is nonempty, and, with $\zeta = (d-1)\beta$,

$$\max_i \|\ell_i - \mathbf{1}_{C_i}\|_\infty \le \zeta, \qquad \mathrm{gap}_\infty(F, F_{\mathcal{P}^\sharp}) \le 2\zeta.$$

Moreover, if $\lambda^\sharp$ is the total-variation mismatch of the recovered blocks, then

$$\lambda^\sharp \le \frac{\varepsilon}{\gamma} + \frac{2\zeta(1+\gamma+\varepsilon)}{\gamma} + \frac{4\zeta}{1+2\zeta}.$$

Proof of Theorem 18. At an anchor x_j, $\ell_i(x_j)$ is zero or one. By the approximate lattice lemma, choose $g_i \in F$ with $\|g_i - (\ell_i)_+\|_\infty \le \varepsilon\kappa$. The anchor values of $(\ell_i)_+$ equal those of ℓ_i, so interpolation gives

$$\|g_i - \ell_i\|_\infty \le \varepsilon\kappa^2.$$

Consequently

$$\|(\ell_i)_+ - \ell_i\|_\infty \le \varepsilon\kappa(1+\kappa) = \beta, \qquad \ell_i(s) \ge -\beta.$$

For $i \ne j$, approximate $\ell_i \wedge \ell_j$ by an element of F. Its anchor values are zero, so the interpolation bound gives

$$\|\ell_i \wedge \ell_j\|_\infty \le \beta.$$

The Lagrange coordinates sum to one, so their maximum is at least $1/d > \beta$. At most one coordinate can exceed β, and the maximizer is unique. Every anchor belongs to a different recovered block. If $s \in C_i$, the losing coordinates lie in $[-\beta, \beta]$, and

$$|\ell_i(s) - 1| = \left|\sum_{j \ne i} \ell_j(s)\right| \le (d-1)\beta = \zeta.$$

The zero-sum of the coordinate error also gives

$$\sum_j |\ell_j(s) - \mathbf{1}_{C_j}(s)| \le 2\zeta.$$

For $f \in F$, interpolation and the last inequality yield

$$\|f - \sum_i f(x_i)\mathbf{1}_{C_i}\|_\infty \le 2\zeta\|f\|_\infty.$$

The same estimate in the other direction, for a block-constant function, proves the gap bound.

For the transition estimate, let $e_J = \mathbf{1}_{\bigcup_{j\in J} C_j}$ and $h_J = \sum_{j\in J} \ell_j$. The zero-sum relation improves the preceding bound to $\|h_J - e_J\|_\infty \le \zeta$. Put

$$v_J = \frac{2h_J - \mathbf{1}}{1+2\zeta} \in F.$$

Then $\|v_J\|_\infty \le 1$ and

$$\|v_J - (2e_J - \mathbf{1})\|_\infty \le \frac{4\zeta}{1+2\zeta}.$$

Give the selected action reward $\mathbf{1}$ and each competitor reward $-\mathbf{1}$. The selected action is isolated and the output is $y = \mathbf{1} + \gamma P_a v_J$. Choose $z \in F$ with $\|y - z\|_\infty \le \varepsilon$. For s, t in the same recovered block, the interpolation estimate gives

$$|z(s) - z(t)| \le 4\zeta\|z\|_\infty \le 4\zeta(1+\gamma+\varepsilon).$$

Dividing the resulting expectation difference by γ, and adding the error from replacing v_J by $2e_J - \mathbf{1}$, gives

$$|p_a^s(J) - p_a^t(J)| \le \frac{\varepsilon}{\gamma} + \frac{2\zeta(1+\gamma+\varepsilon)}{\gamma} + \frac{4\zeta}{1+2\zeta}.$$

Maximizing over a, J, s, t proves the claim. □

D.5 A LOCAL CONVERSE

Proposition 19 (Gap to defect)**.** *Let $G = F_{\mathcal{P}}$ be universally Bellman complete and let $\delta = \mathrm{gap}_\infty(F, G)$. Then*

$$\mathcal{E}_R(F) \le (R+\gamma)(2+\delta)\delta.$$

Proof of Proposition 19. Fix admissible $v \in F$ and $r \in F^A$. Choose $v_0 \in G$ and $r_{0,a} \in G$ with

$$\|v - v_0\|_\infty \le \delta, \qquad \|r_a - r_{0,a}\|_\infty \le R\delta.$$

Then $\|v_0\|_\infty \le 1+\delta$ and $\|r_{0,a}\|_\infty \le R(1+\delta)$. Universal completeness of G gives $z_0 = B_{r_0} v_0 \in G$, with $\|z_0\|_\infty \le (R+\gamma)(1+\delta)$. Bellman nonexpansiveness and the reverse directed gap give

$$\mathrm{dist}_\infty(B_r v, F) \le (R+\gamma)\delta + (R+\gamma)(1+\delta)\delta = (R+\gamma)(2+\delta)\delta.$$

Taking the supremum proves the result. □

D.6 APPROXIMATE QUOTIENT PLANNING

Proof of Theorem 8. For $s \in C_i$, the definition of the representative quotient and the transition mismatch give

$$\left|(P_a\Psi q)(s) - (\bar{P}_a q)_i\right| = \left|\sum_j (p_a^s(j) - p_a^{s_i}(j))q_j\right| \le 2\lambda\|q\|_\infty.$$

Rewards agree within the block. Applying the max inequality over actions and multiplying by γ proves equation 8.

The quotient fixed point satisfies $\|q^*\|_\infty \le R_{\max}/(1-\gamma)$. Since $\bar{B}_\rho q^* = q^*$, the preceding inequality bounds the ground Bellman residual of Ψq^* by $2\gamma\lambda R_{\max}/(1-\gamma)$. If T is a γ-contraction with fixed point z^*, then

$$\|z - z^*\|_\infty \le \frac{\|Tz - z\|_\infty}{1-\gamma}.$$

Applying this to the ground operator proves the first bound in equation 9.

Let π be the blockwise lift of a quotient-greedy action at q^*. The same expectation comparison for the selected action bounds the residual of Ψq^* under the ground policy operator by $2\gamma\lambda R_{\max}/(1-\gamma)$. Policy contraction gives

$$\|v^\pi - \Psi q^*\|_\infty \le \frac{2\gamma\lambda R_{\max}}{(1-\gamma)^2}.$$

Combining this with the fixed-point bound by the triangle inequality yields the policy loss bound in equation 9. □

E ADDITIONAL TIGHTNESS RESULTS

Proposition 20 (Minimax quotient-row lower bound)**.** *Let $d \ge 2$ and $0 < \eta \le 1/16$. If an estimator based on N independent samples from an arbitrary $p \in \Delta_d$ satisfies*

$$\inf_{p\in\Delta_d} \Pr_p\{\|\widehat{p} - p\|_1 \le \eta\} \ge \frac{2}{3},$$

then

$$N \ge \frac{d}{40000\,\eta^2}.$$

Proof of Proposition 20. Set $m = \lfloor d/2 \rfloor$, so $m \ge d/3$. Let $\theta \in \{-1,1\}^m$, put $\alpha = 8\eta \le 1/2$, and define

$$p_\theta(2j-1) = \frac{1+\alpha\theta_j}{2m}, \qquad p_\theta(2j) = \frac{1-\alpha\theta_j}{2m} \quad (1 \le j \le m),$$

with probability zero on an unused final coordinate when d is odd. These are valid distributions.

If θ and $\theta^{(j)}$ differ only at coordinate j, then

$$D_{\mathrm{KL}}(p_\theta \| p_{\theta^{(j)}}) = \frac{\alpha}{m}\log\frac{1+\alpha}{1-\alpha} \le \frac{4\alpha^2}{m}.$$

Additivity of KL and Pinsker's inequality give

$$\mathrm{TV}(p_\theta^{\otimes N}, p_{\theta^{(j)}}^{\otimes N}) \le \sqrt{\frac{2N\alpha^2}{m}}.$$

For $N < d/(40000\eta^2)$, this is smaller than $1/10$. Decode each θ_j from the sign of $\widehat{p}(2j-1) - \widehat{p}(2j)$. The two-point testing bound gives error probability at least $9/20$ for each bit under the uniform prior, so if H is the number of decoding errors,

$$\mathbb{E}H \ge \frac{9m}{20}.$$

An incorrect ordering in one pair costs at least α/m in ℓ_1 distance, hence

$$\|\widehat{p} - p_\theta\|_1 \geq \frac{\alpha H}{m}.$$

Success at radius $\eta = \alpha/8$ implies $H \leq m/8$. Since $0 \leq H \leq m$,

$$\mathbb{E}H \leq \frac{m}{8} + \frac{7m}{8} \Pr\{H > m/8\}.$$

Combining the last two displays gives $\Pr\{H > m/8\} \geq 13/35 > 1/3$ on average over θ. Some p_θ therefore violates the claimed success probability, proving the lower bound. □

Proposition 21 (Attained discounted value sensitivity)**.** *Fix $0 < \gamma < 1$ and $R_{\max} > 0$. For $0 < \alpha \leq \min\{1, (1-\gamma)/\gamma\}$, there are two two-state, two-action MDPs with the same rewards and maximum transition-row ℓ_1 distance $\eta = 2\alpha$ such that*

$$\|v^*_\alpha - v^*_0\|_\infty = \frac{\gamma\alpha R_{\max}}{(1-\gamma)(1-\gamma+\gamma\alpha)} \geq \frac{\gamma\eta R_{\max}}{4(1-\gamma)^2}.$$

Proof of Proposition 21. Use two states, duplicate the transition matrix across both actions, and give both actions reward $(R_{\max}, 0)^\top$. Let

$$P_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad P_\alpha = \begin{pmatrix} 1 & 0 \\ \alpha & 1-\alpha \end{pmatrix}.$$

The largest rowwise ℓ_1 difference is 2α. In both models $v^*(1) = R_{\max}/(1-\gamma)$, while $v^*_0(2) = 0$. In the second model,

$$v^*_\alpha(2) = \gamma\left(\frac{\alpha R_{\max}}{1-\gamma} + (1-\alpha)v^*_\alpha(2)\right).$$

Solving gives the equality. The condition $\gamma\alpha \leq 1-\gamma$ bounds the denominator by $2(1-\gamma)$, proving the lower bound. □

F Boundary Example and Experimental Details

F.1 A four-state audit

Let $S = \{1,2,3,4\}$, $A = \{a,b\}$, and

$$C_1 = \{1,2\}, \qquad C_2 = \{3,4\}, \qquad F = \operatorname{span}\{\mathbf{1}_{C_1}, \mathbf{1}_{C_2}\}.$$

Consider

$$P_a = \begin{pmatrix} .3 & .2 & .1 & .4 \\ .1 & .4 & .3 & .2 \\ .2 & .1 & .4 & .3 \\ 0 & .3 & .2 & .5 \end{pmatrix}, \qquad P_b = \begin{pmatrix} .6 & .1 & .2 & .1 \\ .2 & .5 & 0 & .3 \\ .1 & .1 & .4 & .4 \\ .2 & 0 & .3 & .5 \end{pmatrix}.$$

For P_a, states $1,2$ have next-block masses $(.5,.5)$, and states $3,4$ have $(.3,.7)$. For P_b, these masses are $(.7,.3)$ on C_1 and $(.2,.8)$ on C_2. Both matrices are lumpable, with quotient transitions

$$\bar{P}_a = \begin{pmatrix} .5 & .5 \\ .3 & .7 \end{pmatrix}, \qquad \bar{P}_b = \begin{pmatrix} .7 & .3 \\ .2 & .8 \end{pmatrix}.$$

For $v = q_1\mathbf{1}_{C_1} + q_2\mathbf{1}_{C_2}$ and

$$r_a = \alpha_1\mathbf{1}_{C_1} + \alpha_2\mathbf{1}_{C_2}, \qquad r_b = \beta_1\mathbf{1}_{C_1} + \beta_2\mathbf{1}_{C_2},$$

the Bellman update has block values

$$(B_r v)|_{C_1} = \max\{\alpha_1 + \gamma(.5q_1 + .5q_2), \beta_1 + \gamma(.7q_1 + .3q_2)\},$$
$$(B_r v)|_{C_2} = \max\{\alpha_2 + \gamma(.3q_1 + .7q_2), \beta_2 + \gamma(.2q_1 + .8q_2)\}.$$

It is therefore block-constant for every represented reward and continuation.

F.2 Breaking one block-mass equality

Replace only the second row of P_a by

$$\widetilde{P}_a(2,\cdot) = (.2, .4, .2, .2).$$

The new row sends mass .6 to C_1, while state 1 sends .5. For $v = \mathbf{1}_{C_1}$,

$$(P_a v)(1) = .5, \qquad (\widetilde{P}_a v)(2) = .6,$$

so $\widetilde{P}_a v \notin F$. Set $r_a = 0$ and $r_b = -2\mathbf{1}$. Since $\mathrm{osc}(v) = 1$ and $2 > \gamma$, action a is selected at every state, and

$$B_r v = \gamma \widetilde{P}_a v \notin F.$$

This is an explicit represented continuation and reward tuple that violates universal Bellman completeness.

F.3 Experiment protocol details

The structural certificate experiments sample $d \in \{2, \ldots, 5\}$, between two and four actions, and block sizes from two to four. Each exact instance uses transition kernels lumpable on its partition. Its paired corruption changes one source row's target-block mass while keeping the representation fixed. The tolerance-probe group uses 160 seeds, compares exact membership with multinomial row estimates using 300 samples per state-action row, and includes wrong merges, wrong splits, and random assignments. The harder group uses per-row budgets $25, 50, 100, 300, 1000$, together with missing rows, wrong partitions, random spaces, near-boundary assignments, and perturbed learned rows. A missing row is recorded as non-certifiable.

For defect and recovery audits, randomized block spaces are used to measure $\max |\widehat{\mathcal{E}}_R - \gamma\lambda|$. Off-manifold checks perturb lumpability, use dense random feature assignments, and estimate kernels with finite samples. The nonpartition group draws dense feature spaces with $k > d$, including near-partition interpolation cases, and recomputes the residual after pointwise maximum closure. Recovery records the defect, conditioning parameter, row-code margin, anchor mode, and recovery status. The boundary sweep uses $d = 2, 4, 8, 16$, $\beta/(1/d) \in \{0.25, 0.75, 0.99, 1.01, 2.0\}$, and 30 seeds per cell.

The sampling sweep uses $d = 2, 4, 8, 16$, $\eta \in \{0.25, 0.35, 0.45\}$, and $N/N_{\mathrm{bound}} \in \{0.25, 0.5, 1, 2\}$, with 500 trials per cell for direct multinomial, coordinatewise, and Dirichlet-smoothed estimators. The finite lower-bound check compares multinomial, Dirichlet posterior-mean, and corner-grid estimators on simplex corners, near-uniform rows, and two-heavy rows. The planning benchmark samples 80 block MDPs with $d = 5$, block size 4, three actions, $\gamma = 0.92$, and budgets $8, 16, 32, 64, 128, 256$ per state-action pair. Every baseline receives the same sampled counts at each seed and budget. Baselines are ground plug-in estimation, oracle and recovered quotient pooling, approximate-bisimulation clustering on row signatures, value-equivalence clustering, random-feature clustering, perturbed recovered blocks, and mismatched blocks.

G Extend Background and Preliminaries

Dynamic programming evaluates a model through the rewards and transition kernels in its Bellman recursion (Puterman, 1994). This perspective has motivated decision-aware model learning for misspecified models, value-aware losses, iterative fitting, and value-targeted regression (Joseph et al., 2013; Farahmand et al., 2017; Farahmand, 2018; Ayoub et al., 2020; Asadi et al., 2018). Value equivalence makes policy and function tests explicit, and proper value equivalence extends these tests across repeated updates and planning-sufficient model classes (Grimm et al., 2020; 2021). Planning-oriented latent models also couple learned predictions directly to control (Watter et al., 2015; Tamar et al., 2016; Oh et al., 2017; Silver et al., 2017; Farquhar et al., 2018; Zhang et al., 2019; Schrittwieser et al., 2020).

State abstraction and model minimization construct reduced Markov decision processes from state relations or abstraction maps (Dean & Givan, 1997; Givan et al., 2003; Li et al., 2006). Markov decision process homomorphisms and stochastic bisimulation impose action-compatible aggregation,

while approximate variants control behavioral or performance error (Ferns et al., 2004; Taylor et al., 2008; Abel et al., 2016). Scalable state-similarity methods and action-equivariant approximations extend this line to larger or learned representations (Castro, 2020; van der Pol et al., 2020). These approaches start from an aggregation structure and derive the Bellman consequences it guarantees.

Representation-learning methods turn abstraction principles into objectives for rich observations. Continuous latent dynamics, reward-predictive features, discrete abstractions, and behavioral-similarity losses learn relaxations of model or bisimulation structure (Gelada et al., 2019; Zhang et al., 2021; Agarwal et al., 2021; François-Lavet et al., 2019); value geometry and improvement paths provide related criteria (Bellemare et al., 2019; Dabney et al., 2021). Function-space analyses study inherent Bellman error, linear value-function approximation, and feature encodings (Antos et al., 2008; Parr et al., 2008; Song et al., 2016). Bellman rank, linear Markov decision processes, and low-rank representation learning connect related structure to exploration and statistical efficiency (Jiang et al., 2017; Jin et al., 2020; Agarwal et al., 2020; Modi et al., 2024; Uehara et al., 2022). Offline reinforcement learning analyses identify coverage, misspecification, and partial coverage as constraints on value estimation (Chen & Jiang, 2019; Foster et al., 2022; Uehara & Sun, 2022). Our setting starts from a supplied state-value subspace and studies the structure forced by closure under all represented rewards and continuations with optimal control.