Bellman–Whitney Envelopes: Sharp Partial Identification in Offline Control under Support Holes

Manoj Saravanan [1] **Rohit Kumar Salla** [1]

Abstract

We study finite-horizon offline evaluation and control when a target policy enters state–action regions with zero behavior support, so the target value is not point-identified. We introduce a Bellman–Lipschitz compatibility class that constrains candidate Q-sequences only through Bellman equalities on the observed support and Lipschitz extensions off support. Under a rectangular Bellman–Lipschitz closure condition, we prove that the exact identified interval of the target-policy value is given by a backward Bellman–Whitney recursion, and that this recursion recovers the sharp smooth no-overlap interval exactly when $H = 1$. We further show that the same endpoints admit a no-gap dual characterization via one-sided Bellman relaxations, and we identify a dynamic support-hole geometry for the interval width that is sharp on explicit least-favorable sequential families. On the statistical side, we prove deterministic stability of the recursive endpoints under joint perturbations of the support sets and supported Bellman operators, derive stagewise additive finite-sample endpoint-estimation bounds, and establish an oracle minimax lower bound on a favorable zero-width subclass. Finally, under the control analogue of our closure assumption, we derive Bellman–Whitney action certificates that partition actions into certifiably good, certifiably bad, and intrinsically ambiguous sets.

1. Introduction and Positioning

A central assumption in offline decision-making is that the target rule is sufficiently supported by the logged data. When this assumption holds, one can often retain point identification through overlap, weak-overlap, or partial-coverage conditions and then analyze statistical efficiency or policy learning in the resulting point-identified problem (Uehara & Sun, 2022; Mehrabi & Wager, 2024; Liu et al., 2026). When the target rule enters a genuine support hole, however, the situation changes qualitatively: the relevant density ratio may fail to exist, and the object of interest is no longer a single latent value but an *identified set*. In the one-step setting, this shift has already led to sharp partial identification under smoothness assumptions (Khan et al., 2024). The long-horizon analogue, by contrast, is still missing.

This paper studies finite-horizon offline control and evaluation under genuine support holes. We fix an arbitrary target policy π and allow its occupancy to leave the support of the behavior occupancy at intermediate stages. Our structural assumption is not global point identification of rewards or transitions off support, but a *Bellman–Lipschitz compatibility* class: on the observed support, candidate Q-functions must satisfy the Bellman equation for the target continuation value; off support, they must extend those supported Bellman relations with prescribed Lipschitz radii. Under a recursive Bellman–Lipschitz closure condition, we show that offline control becomes an identified-set problem with an exact dynamic solution.

The central claim of the paper is that the sharp identified interval of the target-policy value is given by a backward Bellman–Whitney recursion built from the classical extremal Lipschitz extension formulas of McShane and Whitney (McShane, 1934; Whitney, 1934). Writing

$$\begin{aligned} g_h^-(c) &:= (B_h^\pi \underline{V}_{h+1}^\pi)(c), \\ g_h^+(c) &:= (B_h^\pi \overline{V}_{h+1}^\pi)(c), \\ & c \in C_h. \end{aligned}$$

the recursion takes the form

$$\begin{aligned} \underline{Q}_h^\pi &= \mathcal{W}_{h,L_h}^- \, g_h^-, \\ \overline{Q}_h^\pi &= \mathcal{W}_{h,L_h}^+ \, g_h^+, \\ \underline{V}_h^\pi &= \mathsf{A}_h^\pi \underline{Q}_h^\pi, \\ \overline{V}_h^\pi &= \mathsf{A}_h^\pi \overline{Q}_h^\pi. \end{aligned}$$

[1]Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA, USA. Correspondence to: Manoj Saravanan <manojsaravanan@vt.edu>, Rohit Kumar Salla <rohits25@vt.edu>.

Accepted at ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning.

with terminal condition $\underline{V}^{\pi}_{H+1} = \overline{V}^{\pi}_{H+1} \equiv 0$. Our first main theorem shows that

$$\mathcal{I}^{\pi} = \left[\underline{V}^{\pi}_{1}(s_1),\, \overline{V}^{\pi}_{1}(s_1)\right].$$

Thus the long-horizon support-hole problem admits a sharp dynamic-programming solution rather than only conservative lower and upper certificates.

Contributions. The paper makes five contributions. First, it introduces the Bellman–Lipschitz compatibility framework for arbitrary fixed target policies under genuine support holes and proves a *sharp Bellman–Whitney interval theorem*. Second, it derives a *no-gap dual characterization*: the same endpoints arise as exact optima of one-sided Bellman relaxations, so the Bellman–Whitney tails are simultaneously primal extreme points and dual certificates. Third, it identifies a *dynamic support-hole geometry* for the interval width and proves that this geometry is sharp on an explicit family of least-favorable sequential instances. Fourth, it separates *irreducible identified-set width* from *endpoint-estimation difficulty*: the recursive envelopes are stable under support and Bellman-target perturbations, and the resulting stagewise additive endpoint-estimation rate is shown to be minimax-sharp on a favorable zero-width oracle subclass. Fifth, it derives a *certifiable control* corollary: under the control analogue of Bellman–Lipschitz closure, the same envelope construction yields sets of certifiably good, certifiably bad, and intrinsically ambiguous actions.

Positioning relative to prior work. The closest one-step antecedent is the smooth no-overlap theory of Khan et al. (2024); our exact $H = 1$ reduction shows that the present framework recovers that theory without approximation, but the main novelty is a *sequential Bellman collapse* from long-horizon partial identification to a stagewise backward recursion. Our setting is also distinct from weak-overlap sequential OPE (Mehrabi & Wager, 2024), where point identification is retained and the main difficulty is heavy-tailed importance weighting rather than genuine support holes. It differs from interval methods for OPE under misspecification (Jiang & Huang, 2020), where intervals quantify bias in an otherwise point-identified problem rather than partial identification under support failure. It also differs from sequential lower/upper bounds under confounding (Zhang & Bareinboim, 2025): there the source of ambiguity is hidden bias, whereas here it is metric support extrapolation under a Bellman-smooth class. At a more abstract level, our dual programs can be viewed as a Bellman-specific specialization of conditional-LP partial identification (Ben-Michael, 2025); the distinctive contribution is that the Bellman structure yields an exact dynamic recursion, strong duality, and a sharp width geometry. Finally, unlike partial-coverage offline RL (Uehara & Sun, 2022; Liu et al., 2026), we do not compare against covered policies or complexity-controlled comparators: we characterize the sharp identified set of an arbitrary fixed target policy that may itself be uncovered.

Why this problem is structurally different. The Bellman–Whitney perspective reveals that support-hole offline control is not merely "offline RL with worse coverage." The primitive object is the identified interval itself, not a point estimate with a pessimism correction. The relevant hardness is not a density ratio but a dynamic metric distance to the observed support, propagated through the Bellman operator. And the appropriate dual object is not a generic occupancy-based certificate but a pair of one-sided Bellman relaxations whose optima coincide exactly with the Bellman–Whitney endpoints.

Organization. Section 2 defines the Bellman–Lipschitz compatibility class and the target identified set. Section 3 proves the sharp Bellman–Whitney interval theorem and gives the exact $H = 1$ reduction. Section 4 establishes the no-gap dual characterization. Section 5 develops the dynamic support-hole geometry and its sharpness. Section 6 proves deterministic stability, finite-sample endpoint-estimation guarantees, and a minimax lower bound. Section 7 gives the certifiable-control corollary. The appendices contain the full proofs, primitive sufficient conditions, counterexamples, and extended comparisons to neighboring literatures.

2. Setup and Bellman–Lipschitz Compatibility

We study a finite-horizon offline control problem over the stagewise compact metric state–action spaces introduced in Appendix A. For each stage $h \in [H]$, the offline dataset induces a behavior occupancy measure ρ^{b}_h on $\mathcal{X}_h$ and hence a closed support $C_h = \mathrm{supp}(\rho^{\mathrm{b}}_h)$. Our object of interest is the value of a *fixed* target policy π whose induced occupancy may enter points $x \notin C_h$. At such points, the one-step Bellman backup is not identified from the logged data alone, and the target-policy value must therefore be treated as an identified-set functional rather than as a point-identified estimand. This differs conceptually from comparator-based partial-coverage offline RL, which benchmarks against policies that are sufficiently covered by the offline data (Uehara & Sun, 2022); when $H = 1$, our problem reduces to the smoothness-based no-overlap policy-evaluation setting studied by Khan et al. (2024).

Fix a target policy $\pi = (\pi_h)_{h=1}^{H}$. For each $h \in [H]$, Appendix A defines the supported Bellman operator

$$(B^{\pi}_h v)(x) := \mathbb{E}[R_h + v(S_{h+1}) \mid X_h = x], \qquad x \in C_h,$$

for every bounded Borel function $v : \mathcal{S}_{h+1} \to \mathbb{R}$, under the

convention $V_{H+1} \equiv 0$. The collection

$$\mathfrak{D}^\pi := \{(C_h, B_h^\pi) : h \in [H]\}$$

is the population object identified by on-support data. All partial identification in the sequel arises from the nonuniqueness of extending these supported Bellman relations off C_h.

Assumption 2.1 (Bounded rewards)**.** There exists $R_{\max} < \infty$ such that

$$|R_h| \le R_{\max} \qquad \text{almost surely for every } h \in [H].$$

assumption 2.1 is used only to keep the Bellman-compatible class uniformly bounded on compact state–action spaces. With additional notation, it can be replaced by stagewise integrable envelope conditions.

For any policy ν and any bounded Borel function $q : \mathcal{X}_h \to \mathbb{R}$, we define the stagewise policy-averaging operator

$$(\mathsf{A}_h^\nu q)(s) := \sum_{a \in \mathcal{A}} \nu_h(a \mid s)\, q(s,a), \qquad s \in \mathcal{S}_h.$$

Since $\mathcal{A}$ is finite, $\mathsf{A}_h^\nu q$ is Borel measurable whenever q is Borel measurable.

We now define the central model class of the paper. The class is formulated at the level of Bellman-compatible function sequences rather than at the level of a globally identified primitive MDP. This is deliberate: under genuine support holes, the off-support continuation values are exactly the source of non-identification.

Definition 2.2 (Bellman–Lipschitz-compatible sequence)**.** Fix a target policy π and a radius vector $L = (L_1, \dots, L_H) \in [0,\infty)^H$. A sequence

$$(Q_h, V_h)_{h=1}^{H+1}$$

belongs to the compatibility class $\mathfrak{C}_L^\pi$ if the following hold:

1. $V_{H+1} \equiv 0$ on $\mathcal{S}_{H+1}$.
2. For every $h \in [H]$, the function $Q_h : \mathcal{X}_h \to \mathbb{R}$ is bounded, Borel measurable, and satisfies
$$\mathrm{Lip}_h(Q_h) \le L_h.$$
3. For every $h \in [H]$, the corresponding state-value function is given by target-policy averaging:
$$V_h = \mathsf{A}_h^\pi Q_h.$$
4. For every $h \in [H]$, the Bellman relation holds pointwise on the behavior support:
$$Q_h(x) = (B_h^\pi V_{h+1})(x), \qquad x \in C_h.$$

Remark 2.3 (Function-level formulation). Definition 2.2 requires only that the supported backup $x \mapsto (B_h^\pi V_{h+1})(x)$ admit an L_h-Lipschitz extension from C_h to the whole state–action space $\mathcal{X}_h$. It does *not* assume that rewards or transition kernels are themselves point-identified outside C_h. Distinct globally defined Q-sequences may therefore agree with exactly the same supported Bellman data $\mathfrak{D}^\pi$ and yet yield different target-policy values. Primitive reward/transition smoothness conditions may be imposed at the model level to ensure that a genuine controlled Markov model induces an element of $\mathfrak{C}_L^\pi$, but the present paper works directly with the resulting function-level compatibility condition.

Assumption 2.4 (Admissible radius vector)**.** The radius vector L is admissible for the supported Bellman data $\mathfrak{D}^\pi$ and target policy π, in the sense that

$$\mathfrak{C}_L^\pi \neq \varnothing.$$

assumption 2.4 is the sequential analogue of assuming that the on-support conditional mean admits a Lipschitz extension in the one-step no-overlap setting. It is an identifiability-class assumption, not an algorithmic assumption: throughout the paper, L is treated as part of the model class, not as a learned quantity.

The parameter of interest is the identified set of target-policy values compatible with the supported Bellman data and the Bellman–Lipschitz radius vector:

$$\mathcal{I}^\pi := \big\{V_1(s_1) : (Q_h, V_h)_{h=1}^{H+1} \in \mathfrak{C}_L^\pi\big\}. \tag{2.1}$$

Its dependence on $(\mathfrak{D}^\pi, L)$ is suppressed for notational readability. Under assumptions 2.1 and 2.4, $\mathcal{I}^\pi$ is nonempty and bounded. The central question of the paper is to characterize the *exact* endpoints of (2.1). Our main result will show that these endpoints are given by a backward Bellman–Whitney envelope recursion, thereby collapsing a priori infinite-dimensional off-support ambiguity into a sharp dynamic-programming representation.

3. Sharp Partial Identification via Bellman–Whitney Envelopes

We now identify the exact endpoints of the value set $\mathcal{I}^\pi$. The key structural observation is that support-hole ambiguity is resolved stagewise by combining the supported Bellman recursion with the extremal Lipschitz extension formulas of McShane and Whitney (McShane, 1934; Whitney, 1934). In the present sequential setting, however, exact interpolation on the support is not automatic: it must be guaranteed recursively for every admissible continuation value. This leads to the following closure condition.

For $h \in [H]$, define

$$\mathcal{V}_h^{\pi,L} := \big\{\mathsf{A}_h^\pi q : q \in \mathrm{Lip}_{L_h}(\mathcal{X}_h)\big\}, \qquad \mathcal{V}_{H+1}^{\pi,L} := \{0\}.$$

Thus $\mathcal{V}_h^{\pi,L}$ is the collection of state-value functions that can be generated at stage h by a L_h-Lipschitz state–action value under the fixed target policy.

Assumption 3.1 (Rectangular Bellman–Lipschitz closure)**.** For every stage $h \in [H]$ and every $v \in \mathcal{V}_{h+1}^{\pi,L}$, the supported Bellman target

$$c \longmapsto (B_h^\pi v)(c), \qquad c \in C_h,$$

is L_h-Lipschitz with respect to the restricted metric on C_h.

assumption 3.1 is the precise recursive compatibility condition under which the stagewise Bellman targets admit exact L_h-Lipschitz extensions from the support to the full state–action space. It is therefore stronger than the bare nonemptiness requirement in assumption 2.4, but it is exactly what is needed to convert one-sided support envelopes into sharp identified-set endpoints. At a primitive model level, this requirement can be enforced through reward and transition regularity conditions that make the supported Bellman targets Lipschitz.

With assumption 3.1 in force, define the lower and upper supported Bellman targets recursively by

$$\begin{aligned} g_h^-(c) &:= (B_h^\pi \underline{V}_{h+1}^\pi)(c), \\ g_h^+(c) &:= (B_h^\pi \overline{V}_{h+1}^\pi)(c), \\ &\qquad c \in C_h. \end{aligned} \tag{3.1}$$

starting from the terminal condition

$$\underline{V}_{H+1}^\pi = \overline{V}_{H+1}^\pi \equiv 0.$$

The corresponding Bellman–Whitney envelopes are

$$\begin{aligned} \underline{Q}_h^\pi &= \mathcal{W}_{h,L_h}^- \, g_h^-, \\ \overline{Q}_h^\pi &= \mathcal{W}_{h,L_h}^+ \, g_h^+, \\ \underline{V}_h^\pi &= \mathsf{A}_h^\pi \underline{Q}_h^\pi, \\ \overline{V}_h^\pi &= \mathsf{A}_h^\pi \overline{Q}_h^\pi. \end{aligned} \tag{3.2}$$

for $h = H, H-1, \ldots, 1$.

The notation in Equation (3.2) is deliberate. The operator $\mathcal{W}_{h,L_h}^-$ is the smallest L_h-Lipschitz function that dominates its input on C_h, while $\mathcal{W}_{h,L_h}^+$ is the largest L_h-Lipschitz function that is dominated by its input on C_h; see Appendix B. Under assumption 3.1, the supported targets in Equation (3.1) are themselves L_h-Lipschitz on the support, so the two operators become exact stagewise extension maps. The next theorem shows that the resulting recursion is not merely valid but sharp.

Theorem 3.2 (Sharp Bellman–Whitney interval)**.** *Assume assumptions 2.1 and 3.1. Let $(\underline{Q}_h^\pi, \overline{Q}_h^\pi, \underline{V}_h^\pi, \overline{V}_h^\pi)_{h=1}^{H+1}$ be defined by Equations (3.1) and (3.2). Then:*

1. *the lower and upper recursive tails are feasible, i.e.,*
$$(\underline{Q}_h^\pi, \underline{V}_h^\pi)_{h=1}^{H+1} \in \mathfrak{C}_L^\pi, \qquad (\overline{Q}_h^\pi, \overline{V}_h^\pi)_{h=1}^{H+1} \in \mathfrak{C}_L^\pi;$$
2. *every compatible sequence $(Q_h, V_h)_{h=1}^{H+1} \in \mathfrak{C}_L^\pi$ satisfies the stagewise sandwich inequalities*
$$\underline{Q}_h^\pi(x) \le Q_h(x) \le \overline{Q}_h^\pi(x), \qquad x \in \mathcal{X}_h,$$
and
$$\underline{V}_h^\pi(s) \le V_h(s) \le \overline{V}_h^\pi(s), \qquad s \in \mathcal{S}_h,$$
for all $h \in [H]$;
3. *the target-policy value identified set is the sharp interval*
$$\mathcal{I}^\pi = \Big[\underline{V}_1^\pi(s_1), \, \overline{V}_1^\pi(s_1)\Big].$$
In particular, both endpoints are attained by feasible Bellman–Lipschitz compatible tails.

Proof sketch; full proof in Appendix D. The appendix proves the stronger stagewise statement that, for every $h \in [H]$ and every $s \in \mathcal{S}_h$, the full tail value set generated by all compatible continuations from stage h onward is exactly $[\underline{V}_h^\pi(s), \overline{V}_h^\pi(s)]$.

The argument has three steps. First, under assumption 3.1, each supported Bellman target g_h^- and g_h^+ is L_h-Lipschitz on C_h, so the McShane–Whitney formulas yield exact feasible stagewise extensions. This shows that the recursive lower and upper tails themselves belong to the compatibility class. Second, for an arbitrary compatible tail, backward induction and monotonicity of conditional expectation imply

$$B_h^\pi \underline{V}_{h+1}^\pi \le Q_h \le B_h^\pi \overline{V}_{h+1}^\pi \qquad \text{on } C_h.$$

The extremal characterization of the Bellman–Whitney operators from Appendix B then yields $\underline{Q}_h^\pi \le Q_h \le \overline{Q}_h^\pi$ on all of $\mathcal{X}_h$, and hence $\underline{V}_h^\pi \le V_h \le \overline{V}_h^\pi$ after target-policy averaging. Third, the feasible tail class is convex, and the two recursive tails attain the two endpoints, so the entire closed interval between them is feasible. Taking $h = 1$ proves the claim. □

Corollary 3.3 (Exact one-step reduction)**.** *When $H = 1$, the Bellman–Whitney recursion collapses to a single-stage smoothness envelope. In particular, if the target policy is deterministic at the initial state, $\pi_1(\cdot \mid s_1) = \delta_{a^\star}$, then*

$$\mathcal{I}^\pi = \begin{bmatrix} \sup\limits_{c \in C_1} \Big\{ m(c) - L_1 d_1\big((s_1, a^\star), c\big) \Big\}, \\ \inf\limits_{c \in C_1} \Big\{ m(c) + L_1 d_1\big((s_1, a^\star), c\big) \Big\} \end{bmatrix}.$$

where

$$m(c) := \mathbb{E}[R_1 \mid X_1 = c], \qquad c \in C_1.$$

This is exactly the sharp smooth no-overlap interval in the one-step setting of Khan et al. (2024). Appendix F gives the formal reduction.

Remark 3.4 (Why the theorem is stronger than pessimistic bounding). Theorem 3.2 does not merely provide conservative lower and upper certificates. It identifies the *exact* range of values compatible with the supported Bellman data and the Bellman–Lipschitz class. This distinction is fundamental. In point-identified offline RL under partial coverage, the main question is how to estimate a single latent value; here, by contrast, the primitive object is the identified interval itself, and the Bellman–Whitney recursion computes its endpoints sharply.

4. Dual Characterization and Strong Duality

The Bellman–Whitney interval in theorem 3.2 is not merely the sharp primal identified set over the equality-constrained compatibility class. It also admits an exact dual characterization through one-sided Bellman relaxations. This is important conceptually: the support-hole problem can be viewed either as optimization over all Bellman–Lipschitz-compatible tails, or equivalently as optimization over all Lipschitz tails that satisfy Bellman consistency only from one side on the observed support.

For $h \in [H]$ and $s \in \mathcal{S}_h$, define the stagewise primal endpoint values

$$\mathsf{P}_h^+(s) := \sup\Big\{V_h(s) : (Q_t, V_t)_{t=h}^{H+1} \in \mathfrak{C}_{h:H}^{\pi,L}\Big\},$$
$$\mathsf{P}_h^-(s) := \inf\Big\{V_h(s) : (Q_t, V_t)_{t=h}^{H+1} \in \mathfrak{C}_{h:H}^{\pi,L}\Big\}.$$

Next define the upper dual class $\mathfrak{U}_h^{\pi,L}$ to consist of all sequences $U = (U_t)_{t=h}^H$ such that, for every $t \in \{h, \dots, H\}$,

$$U_t \in \mathrm{Lip}_{L_t}(\mathcal{X}_t), \qquad V_t^U := \mathsf{A}_t^\pi U_t, \qquad V_{H+1}^U \equiv 0,$$

and

$$U_t(c) \le (B_t^\pi V_{t+1}^U)(c) \qquad \forall c \in C_t.$$

Similarly, the lower dual class $\mathfrak{L}_h^{\pi,L}$ is defined by reversing the last inequality:

$$L_t(c) \ge (B_t^\pi V_{t+1}^L)(c) \qquad \forall c \in C_t,$$

where $V_t^L := \mathsf{A}_t^\pi L_t$ and $V_{H+1}^L \equiv 0$. The associated dual objective values are

$$\mathsf{D}_h^+(s) := \sup\Big\{V_h^U(s) : U \in \mathfrak{U}_h^{\pi,L}\Big\},$$
$$\mathsf{D}_h^-(s) := \inf\Big\{V_h^L(s) : L \in \mathfrak{L}_h^{\pi,L}\Big\}.$$

The upper dual is therefore a Bellman *minorant* relaxation, while the lower dual is a Bellman *majorant* relaxation. A primal-feasible tail satisfies Bellman equality on the support and is thus feasible for both duals. The nontrivial question is whether these one-sided relaxations are exact.

Theorem 4.1 (Strong duality and dual characterization)**.** *Assume assumptions 2.1 and 3.1. Then, for every stage $h \in [H]$ and every state $s \in \mathcal{S}_h$,*

$$\mathsf{D}_h^-(s) = \mathsf{P}_h^-(s) = \underline{V}_h^\pi(s), \qquad \mathsf{P}_h^+(s) = \mathsf{D}_h^+(s) = \overline{V}_h^\pi(s).$$

Consequently,

$$\mathcal{I}^\pi = \big[\mathsf{D}_1^-(s_1),\, \mathsf{D}_1^+(s_1)\big] = \Big[\underline{V}_1^\pi(s_1),\, \overline{V}_1^\pi(s_1)\Big].$$

Moreover, the recursive Bellman–Whitney tails $(\underline{Q}_t^\pi, \underline{V}_t^\pi)_{t=1}^{H+1}$ and $(\overline{Q}_t^\pi, \overline{V}_t^\pi)_{t=1}^{H+1}$ are simultaneously primal-feasible and dual-optimal.

Proof sketch; full proof in Appendix E. Weak duality is immediate: every primal-feasible tail satisfies both dual constraints with equality on the support and is therefore feasible for both one-sided relaxations. Hence

$$\mathsf{D}_h^-(s) \le \mathsf{P}_h^-(s) \le \mathsf{P}_h^+(s) \le \mathsf{D}_h^+(s).$$

The key step is pointwise extremality of the Bellman–Whitney recursion over the dual classes. Let $U \in \mathfrak{U}_h^{\pi,L}$ be arbitrary. By backward induction, if $V_{t+1}^U \le \overline{V}_{t+1}^\pi$, then on the support

$$U_t \le B_t^\pi V_{t+1}^U \le B_t^\pi \overline{V}_{t+1}^\pi = g_t^+,$$

where g_t^+ is the supported upper Bellman target from Equation (3.1). Since U_t is L_t-Lipschitz, Appendix B implies

$$U_t \le \mathcal{W}_{t,L_t}^+ g_t^+ = \overline{Q}_t^\pi,$$

and therefore $V_t^U \le \overline{V}_t^\pi$ after target-policy averaging. The symmetric argument for $L \in \mathfrak{L}_h^{\pi,L}$ yields

$$\underline{Q}_t^\pi \le L_t, \qquad \underline{V}_t^\pi \le V_t^L.$$

Thus $\overline{V}_h^\pi$ is the pointwise maximum over the upper dual class and $\underline{V}_h^\pi$ is the pointwise minimum over the lower dual class.

Finally, the recursive Bellman–Whitney tails are primal-feasible by theorem 3.2 and dual-feasible because they satisfy Bellman equality on the support. Hence they attain both primal and dual extrema, establishing the claimed no-gap identities. □

Remark 4.2 (Duality as Bellman collapse). Theorem 4.1 shows that the support-hole problem admits an exact duality theory without introducing trajectory-space multipliers or generic history-indexed linear programs. The Bellman structure collapses the dual relaxations to the same backward recursion that already yields the sharp primal identified interval. In this sense, the Bellman–Whitney envelopes are simultaneously primal extreme points and dual certificates.

5. Geometry of Dynamic Support Holes

The Bellman–Whitney interval from theorem 3.2 is an exact identified set, but its width still requires interpretation. We now show that the interval width is governed by a recursively propagated support-hole geometry. The resulting object is neither a density-ratio quantity nor a generic pessimism penalty: it is a dynamic metric extrapolation functional tied directly to the Bellman recursion.

For each stage $h \in [H]$, define the state–action width

$$w_h^\pi(x) := \overline{Q}_h^\pi(x) - \underline{Q}_h^\pi(x), \qquad x \in \mathcal{X}_h, \tag{5.1}$$

and the corresponding state-value width

$$\Delta_h(s) := \overline{V}_h^\pi(s) - \underline{V}_h^\pi(s), \qquad s \in \mathcal{S}_h,$$

with terminal convention $\Delta_{H+1} \equiv 0$. By linearity of target-policy averaging,

$$\Delta_h = \mathsf{A}_h^\pi w_h^\pi.$$

Next define the supported continuation operator

$$(\mathsf{T}_h f)(c) := \mathbb{E}[f(S_{h+1}) \mid X_h = c], \qquad c \in C_h,$$

for bounded Borel functions $f : \mathcal{S}_{h+1} \to \mathbb{R}$. The reward term plays no role in the width recursion, since it cancels between the upper and lower Bellman targets.

This motivates the following dynamic support-hole envelope:

$$\Gamma_h^\pi(x) := \inf_{c \in C_h} \Big\{ 2L_h d_h(x,c) + (\mathsf{T}_h \Delta_{h+1})(c) \Big\}, \quad x \in \mathcal{X}_h. \tag{5.2}$$

The first term is the local extrapolation price from the observed support to the query point x, while the second term is the continuation ambiguity already present at the support anchor c.

Theorem 5.1 (Dynamic support-hole upper bound)**.** *Assume assumptions 2.1 and 3.1. Then, for every $h \in [H]$ and every $x \in \mathcal{X}_h$,*

$$w_h^\pi(x) \le \Gamma_h^\pi(x).$$

Moreover, on the behavior support itself,

$$w_h^\pi(c) = \Gamma_h^\pi(c) = (\mathsf{T}_h \Delta_{h+1})(c), \qquad c \in C_h.$$

The theorem has a simple interpretation. On-support ambiguity is exactly the expected continuation width. Off-support ambiguity is obtained by transporting that continuation width away from the support at rate $2L_h$. Thus the Bellman–Whitney interval width decomposes into an immediate geometric cost of leaving the support and a future ambiguity cost propagated through the observed dynamics.

The next theorem shows that this geometry is sharp. Consider the explicit family in which each stage support collapses to a single anchor point and the supported transition deterministically propagates the target trajectory forward. Appendix G formalizes this as a *serial singleton-support family*.

Theorem 5.2 (Sharpness on serial singleton-support families)**.** *Fix any deterministic target policy and any serial singleton-support family in the sense of Appendix G. Then for every stage $h \in [H]$,*

$$w_h^\pi(x_h^\star) = \Gamma_h^\pi(x_h^\star),$$

where $x_h^\star$ is the designated target state–action point at stage h. Equivalently,

$$w_h^\pi(x_h^\star) = 2L_h d_h(x_h^\star, c_h) + w_{h+1}^\pi(x_{h+1}^\star), \quad h \in \{1, \dots, H-1\},$$

with terminal identity

$$w_H^\pi(x_H^\star) = 2L_H d_H(x_H^\star, c_H).$$

Hence

$$w_h^\pi(x_h^\star) = 2\sum_{t=h}^{H} L_t d_t(x_t^\star, c_t).$$

Proof sketch; full proofs in Appendix G. On the support, exact interpolation of the Bellman–Whitney envelopes implies

$$w_h^\pi(c) = (B_h^\pi \overline{V}_{h+1}^\pi)(c) - (B_h^\pi \underline{V}_{h+1}^\pi)(c) = (\mathsf{T}_h \Delta_{h+1})(c).$$

For an arbitrary off-support point x, the defining formulas for $\overline{Q}_h^\pi$ and $\underline{Q}_h^\pi$ yield, for every anchor $c \in C_h$,

$$w_h^\pi(x) \le 2L_h d_h(x,c) + w_h^\pi(c),$$

and taking the infimum over c proves the upper bound. Sharpness is established by an explicit least-favorable family in which each stage support is a singleton, so the Bellman–Whitney envelopes become affine cones around the anchor point and the recursive inequality is attained with equality at every step. □

Example 5.3 (Two-stage additive ambiguity)**.** Suppose $H = 2$, the action space is a singleton, the stagewise state spaces are $[0,1]$ with the Euclidean metric, and the support sets are singletons $C_1 = \{c_1\}$ and $C_2 = \{c_2\}$ with

$$c_1 = (0, a), \qquad c_2 = (0, a).$$

Let the target evaluation points be

$$x_1^\star = (\delta_1, a), \qquad x_2^\star = (\delta_2, a), \qquad \delta_1, \delta_2 \in [0,1],$$

and suppose the supported transition from c_1 deterministically reaches the second-stage state δ_2. Then Appendix G shows

$$w_2^\pi(x_2^\star) = 2L_2\delta_2, \qquad w_1^\pi(x_1^\star) = 2L_1\delta_1 + 2L_2\delta_2.$$

Thus the two-stage identified-set width is exactly additive across stages: first-stage off-support extrapolation contributes $2L_1\delta_1$, and second-stage continuation ambiguity contributes $2L_2\delta_2$.

Remark 5.4 (What the geometry theorem says). Theorems 5.1 and 5.2 identify the right scale of irreducible ambiguity under support holes. The relevant difficulty is not a heavy-tailed importance ratio but a dynamic metric distance to the observed support, propagated through the Bellman operator. In this sense, the Bellman–Whitney interval exposes a geometric notion of offline hardness that is invisible in purely point-identified analyses.

6. Endpoint Estimation and Minimax Lower Bounds

The Bellman–Whitney interval separates two distinct sources of uncertainty: its *width*, which is an irreducible identified-set object, and the statistical error incurred when estimating its *endpoints* from finite data. We now analyze the second component. The main message is that the backward Bellman–Whitney recursion is stable under joint perturbations of the support sets and the supported Bellman operators, and that the resulting endpoint-estimation rate is minimax-sharp even on a favorable zero-width subclass.

For each stage $h \in [H]$, let $\widehat{C}_h \subseteq \mathcal{X}_h$ be a nonempty compact support estimator and let $\widehat{B}_h^\pi$ be an empirical supported Bellman operator. Appendix H defines the empirical Bellman–Whitney recursion

$$\underline{\widehat{Q}}_h^\pi,\ \widehat{\overline{Q}}_h^\pi,\ \underline{\widehat{V}}_h^\pi,\ \widehat{\overline{V}}_h^\pi,$$

obtained by replacing (C_h, B_h^π) with $(\widehat{C}_h, \widehat{B}_h^\pi)$ at every stage.

The right deterministic perturbation measure is the stagewise support-value discrepancy

$$\varepsilon_h := \sup_{v \in \mathcal{V}_{h+1}^{\mathrm{env}}} \Delta_{C_h, \widehat{C}_h}^{(L_h)}\Big(c \mapsto (B_h^\pi v)(c),\ \hat{c} \mapsto (\widehat{B}_h^\pi v)(\hat{c})\Big), \tag{6.1}$$

where $\mathcal{V}_{h+1}^{\mathrm{env}}$ is the uniformly bounded continuation class from Appendix H. Thus ε_h simultaneously captures support-set error and supported Bellman-target error at stage h.

Theorem 6.1 (Deterministic endpoint stability)**.** *Assume assumptions 2.1, 3.1 and H.1. Then for every stage $h \in [H]$,*

$$\begin{aligned}
&\|\underline{\widehat{Q}}_h^\pi - \underline{Q}_h^\pi\|_\infty \\
&\vee \|\widehat{\overline{Q}}_h^\pi - \overline{Q}_h^\pi\|_\infty \\
&\vee \|\underline{\widehat{V}}_h^\pi - \underline{V}_h^\pi\|_\infty \\
&\vee \|\widehat{\overline{V}}_h^\pi - \overline{V}_h^\pi\|_\infty \\
&\le \sum_{t=h}^{H} \varepsilon_t.
\end{aligned}$$

In particular,

$$\left|\underline{\widehat{V}}_1^\pi(s_1) - \underline{V}_1^\pi(s_1)\right| \vee \left|\widehat{\overline{V}}_1^\pi(s_1) - \overline{V}_1^\pi(s_1)\right| \le \sum_{t=1}^{H} \varepsilon_t.$$

Proof sketch; full proof in Appendix H. The proof is a backward perturbation argument. At a fixed stage h, the pair stability lemma for Bellman–Whitney envelopes implies that the error in $(\underline{\widehat{Q}}_h^\pi, \widehat{\overline{Q}}_h^\pi)$ is bounded by the discrepancy between the true and empirical supported Bellman targets evaluated at the next-stage continuation values. This discrepancy splits into a stagewise approximation term ε_h plus the propagated continuation error from stage $h+1$. Because target-policy averaging is a contraction in sup norm, the same bound transfers to $(\underline{\widehat{V}}_h^\pi, \widehat{\overline{V}}_h^\pi)$. Iterating backward yields the additive recursion. □

The preceding theorem is deterministic. To obtain a statistical rate, it suffices to upper bound ε_h on a high-probability event. Appendix H gives one convenient sufficient condition: if each stage admits a support estimation error η_h in Hausdorff distance and a supported Bellman-target approximation error α_h, then

$$\varepsilon_h \le \alpha_h + 2L_h\eta_h.$$

Combining this with theorem 6.1 yields the following immediate finite-sample consequence.

Corollary 6.2 (Finite-sample endpoint estimation)**.** *Suppose there exists an event $\mathcal{E}_n$ with $\mathbb{P}(\mathcal{E}_n) \ge 1 - \delta_n$ such that on $\mathcal{E}_n$, for every $h \in [H]$,*

$$\begin{aligned}
d_{\mathrm{H}}\Big(C_h, \widehat{C}_h\Big) &\le \eta_{h,n}, \\
\sup_{v \in \mathcal{V}_{h+1}^{\mathrm{env}}} \sup_{\hat{c} \in \widehat{C}_h} \Big|(\widehat{B}_h^\pi v)(\hat{c}) - f_{h,v}(\hat{c})\Big| &\le \alpha_{h,n}.
\end{aligned}$$

where $f_{h,v}$ is an L_h-Lipschitz extension of $c \mapsto (B_h^\pi v)(c)$ from C_h to $\mathcal{X}_h$. Then on $\mathcal{E}_n$,

$$\begin{aligned}
&\left|\underline{\widehat{V}}_1^\pi(s_1) - \underline{V}_1^\pi(s_1)\right| \vee \left|\widehat{\overline{V}}_1^\pi(s_1) - \overline{V}_1^\pi(s_1)\right| \\
&\qquad \le \sum_{h=1}^{H} (\alpha_{h,n} + 2L_h\eta_{h,n}).
\end{aligned} \tag{6.2}$$

In particular, if

$$\alpha_{h,n} = \widetilde{O}\Big(n_h^{-1/(2+d_h)}\Big), \qquad \eta_{h,n} = \widetilde{O}\Big(n_h^{-1/d_h}\Big),$$

then

$$\begin{aligned}&\Big|\widehat{\underline{V}}_1^{\pi}(s_1) - \underline{V}_1^{\pi}(s_1)\Big| \\ &\vee \Big|\widehat{\overline{V}}_1^{\pi}(s_1) - \overline{V}_1^{\pi}(s_1)\Big| \\ &= \widetilde{O}_p\Bigg(\sum_{h=1}^{H} n_h^{-1/(2+d_h)}\Bigg).\end{aligned}$$

The final question is whether the stagewise additive endpoint-estimation rate is optimal. To answer this cleanly, Appendix I studies a statistically favorable oracle subclass in which the behavior support is the entire state–action space at every stage, the target policy is deterministic, and the Bellman–Whitney interval has zero width. Thus any lower bound there is a lower bound for endpoint estimation alone, uncontaminated by support-hole ambiguity. The proof uses a two-point minimax construction in the sense of Yu (1997); Tsybakov (2009).

Theorem 6.3 (Oracle minimax lower bound)**.** *Fix the model-class parameters $(H, L, d_1, \ldots, d_H, R_{\max})$. There exists a constant $c_\star > 0$ depending only on these parameters such that for every sample-size vector $(n_1, \ldots, n_H)$,*

$$\begin{aligned}\inf_{(\hat{\ell},\hat{u})} \sup_{M \in \mathfrak{O}^L_{(n_1,\ldots,n_H)}} \mathbb{E}_M\Big[&\Big|\hat{\ell} - \underline{V}_1^{\pi}(\mathbf{0})\Big| \\ &+ \Big|\hat{u} - \overline{V}_1^{\pi}(\mathbf{0})\Big|\Big] \\ &\geq c_\star \sum_{h=1}^{H} n_h^{-1/(d_h+2)}.\end{aligned}$$

Moreover, on this oracle class the Bellman–Whitney interval degenerates to a singleton, so the lower bound concerns endpoint estimation alone.

Proof sketch; full proof in Appendix I. The appendix constructs a zero-width two-point subclass in which, at each stage, the reward mean contains a localized L_h-Lipschitz bump of amplitude $n_h^{-1/(d_h+2)}$ near the evaluation state. The stagewise bumps add linearly in the target value, so the parameter separation is of order $\sum_{h=1}^{H} n_h^{-1/(d_h+2)}$. At the same time, the joint Kullback–Leibler divergence between the two induced data distributions remains bounded by a constant. A two-point minimax inequality then yields the stated lower bound. □

Remark 6.4 (Two distinct hardness sources). Theorems 6.1 and 6.3 should be read together with the geometry results from Section 5. The width of the Bellman–Whitney interval is an irreducible support-hole phenomenon, while the rate in theorem 6.3 is an endpoint-estimation phenomenon that persists even when the width is zero. The theory therefore cleanly separates *identified-set ambiguity* from *statistical estimation difficulty*.

7. Certifiable Control under Ambiguity

The fixed-policy Bellman–Whitney theory extends naturally to control once target-policy averaging is replaced by the Bellman optimality operator. However, this change destroys the convexity structure used in theorem 3.2: the map

$$Q \longmapsto \max_{a \in \mathcal{A}} Q(\cdot, a)$$

is nonlinear, so the control-compatible tail class is generally not convex. Accordingly, the correct control conclusion is not a sharp interval theorem for optimal values, but an action-certification theorem.

For each stage $h \in [H]$, define the control value class

$$\mathcal{V}_h^{\mathrm{ctl},L} := \left\{\begin{array}{l} v : \mathcal{S}_h \to \mathbb{R} : \\ \exists\, q \in \mathrm{Lip}_{L_h}(\mathcal{X}_h) \\ \text{with } v(s) = \max_{a \in \mathcal{A}} q(s,a) \\ \text{for all } s \end{array}\right\},$$

with terminal convention $\mathcal{V}_{H+1}^{\mathrm{ctl},L} = \{0\}$. We assume the control analogue of rectangular Bellman–Lipschitz closure: for every $h \in [H]$ and every $v \in \mathcal{V}_{h+1}^{\mathrm{ctl},L}$, the supported Bellman target

$$c \longmapsto (B_h^{\pi} v)(c), \qquad c \in C_h,$$

is L_h-Lipschitz on C_h. Under this condition, the control Bellman–Whitney recursion is

$$\begin{aligned}\underline{V}_{H+1}^{\mathrm{ctl}} &= \overline{V}_{H+1}^{\mathrm{ctl}} \equiv 0, \\ \underline{Q}_h^{\mathrm{ctl}} &= \mathcal{W}^-_{h,L_h}\Big(c \mapsto (B_h^{\pi} \underline{V}_{h+1}^{\mathrm{ctl}})(c)\Big), \\ \overline{Q}_h^{\mathrm{ctl}} &= \mathcal{W}^+_{h,L_h}\Big(c \mapsto (B_h^{\pi} \overline{V}_{h+1}^{\mathrm{ctl}})(c)\Big).\end{aligned}$$

and

$$\underline{V}_h^{\mathrm{ctl}}(s) = \max_{a \in \mathcal{A}} \underline{Q}_h^{\mathrm{ctl}}(s,a), \qquad \overline{V}_h^{\mathrm{ctl}}(s) = \max_{a \in \mathcal{A}} \overline{Q}_h^{\mathrm{ctl}}(s,a).$$

Appendix J proves that these envelopes sandwich every control-compatible tail, i.e., every bounded Lipschitz sequence satisfying $V_t = \max_a Q_t$ and Bellman equality on the observed support.

This leads to the following certification sets. For a state $s \in \mathcal{S}_h$ and tolerance $\delta \geq 0$, define

$$\mathcal{A}_h^{\mathrm{good}}(s;\delta) := \left\{\begin{array}{l} a \in \mathcal{A} : \\ \underline{Q}_h^{\mathrm{ctl}}(s,a) \geq \max_{a' \in \mathcal{A}} \overline{Q}_h^{\mathrm{ctl}}(s,a') - \delta \end{array}\right\},$$

$$\mathcal{A}_h^{\mathrm{bad}}(s;\delta) := \left\{ \begin{array}{l} a \in \mathcal{A} : \\ \overline{Q}_h^{\mathrm{ctl}}(s,a) < \\ \max_{a' \in \mathcal{A}} \underline{Q}_h^{\mathrm{ctl}}(s,a') - \delta \end{array} \right\},$$

and

$$\mathcal{A}_h^{\mathrm{amb}}(s;\delta) := \mathcal{A} \setminus \left(\mathcal{A}_h^{\mathrm{good}}(s;\delta) \cup \mathcal{A}_h^{\mathrm{bad}}(s;\delta) \right).$$

Corollary 7.1 (Certifiable control under ambiguity)**.** *Assume assumptions 2.1 and J.1. Fix $h \in [H]$, $s \in \mathcal{S}_h$, and $\delta \geq 0$. Then:*

1. *every action in $\mathcal{A}_h^{\mathrm{good}}(s;\delta)$ is uniformly δ-optimal over the entire control-compatible class;*
2. *every action in $\mathcal{A}_h^{\mathrm{bad}}(s;\delta)$ is uniformly δ-suboptimal over the entire control-compatible class;*
3. *actions in $\mathcal{A}_h^{\mathrm{amb}}(s;\delta)$ are not resolvable from the Bellman–Whitney envelopes alone.*

More precisely, if $(Q_t, V_t)_{t=h}^{H+1} \in \mathfrak{C}_{h:H}^{\mathrm{ctl},L}$ is any control-compatible tail, then for every $a \in \mathcal{A}_h^{\mathrm{good}}(s;\delta)$,

$$Q_h(s,a) \geq V_h(s) - \delta,$$

while for every $a \in \mathcal{A}_h^{\mathrm{bad}}(s;\delta)$,

$$Q_h(s,a) \leq V_h(s) - \delta.$$

Proof sketch; full proof in Appendix J. Appendix J first proves a control Bellman–Whitney sandwich theorem:

$$\underline{Q}_t^{\mathrm{ctl}} \leq Q_t \leq \overline{Q}_t^{\mathrm{ctl}}, \qquad \underline{V}_t^{\mathrm{ctl}} \leq V_t \leq \overline{V}_t^{\mathrm{ctl}},$$

for every control-compatible tail. The proof is the same backward-induction scheme as in the fixed-policy case, except that target-policy averaging is replaced by pointwise maximization over actions. Once these inequalities hold, the certification statements follow immediately. If $a \in \mathcal{A}_h^{\mathrm{good}}(s;\delta)$, then

$$\underline{Q}_h^{\mathrm{ctl}}(s,a) \geq \max_{a'} \overline{Q}_h^{\mathrm{ctl}}(s,a') - \delta \geq V_h(s) - \delta,$$

and therefore $Q_h(s,a) \geq V_h(s) - \delta$. The bad-action guarantee is symmetric. □

Remark 7.2 (What is and is not certified). The corollary certifies actions, not optimal values. This is the right level of generality under support holes. Actions in $\mathcal{A}_h^{\mathrm{good}}(s;\delta)$ are provably near-optimal uniformly over the entire Bellman–Lipschitz ambiguity class, actions in $\mathcal{A}_h^{\mathrm{bad}}(s;\delta)$ are provably suboptimal, and the remaining actions are genuinely unresolved by the available offline information. Thus the Bellman–Whitney framework yields a set-valued notion of optimal decision-making under partial identification.

Impact Statement

This paper studies the theoretical foundations of offline decision-making under genuine support holes, where a target policy enters state–action regions that are not represented in the logged data. The positive potential impact of this work is methodological rather than immediate deployment: it provides a rigorous framework for distinguishing what is point-identifiable from what is only partially identifiable, and for replacing unjustified point estimates with sharp intervals, dual certificates, and action-level ambiguity sets. In high-stakes domains such as healthcare, education, public policy, and scientific experimentation, this perspective can help prevent overconfident conclusions from observational or biased historical data.

At the same time, the framework could be misused if its guarantees are invoked without verifying the modeling assumptions. In particular, the Bellman–Lipschitz intervals and action certificates are only as meaningful as the underlying support characterization, the chosen Lipschitz radii, and the Bellman closure conditions. If these assumptions are misspecified, the resulting intervals may be either too narrow, giving a false appearance of certainty, or too wide to be operationally useful. Accordingly, this work should not be interpreted as a license for automated deployment in safety-critical settings without domain validation, sensitivity analysis, and application-specific oversight.

More broadly, we view the main ethical contribution of this paper as advocating for principled abstention. When the data do not identify a policy value or an action ranking, the correct output is not an artificially precise estimate but a transparent description of what remains ambiguous. Our control corollary makes this explicit by partitioning actions into certifiably good, certifiably bad, and intrinsically ambiguous sets. We hope this theoretical viewpoint supports a more cautious and scientifically honest use of offline decision-making methods in real-world systems.

References

Ben-Michael, E. Partial identification via conditional linear programs: Estimation and policy learning. *arXiv preprint arXiv:2506.12215*, 2025. doi: 10.48550/arXiv.2506.12215. URL https://arxiv.org/abs/2506.12215.

Iyengar, G. N. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005. doi: 10.1287/moor.1040.0129.

Jiang, N. and Huang, J. Minimax value interval for off-policy evaluation and policy optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2747–2758. Curran Associates, Inc., 2020. URL

https://proceedings.neurips.cc/paper/2020/hash/1cd138d0499a68f4bb72bee04bbec2d7-Abstract.html.

Khan, S., Saveski, M., and Ugander, J. Off-policy evaluation beyond overlap: Sharp partial identification under smoothness. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23734–23757. PMLR, July 2024. URL https://proceedings.mlr.press/v235/khan24b.html.

Liu, H., Snyder, B., and Wei, C.-Y. On the complexity of offline reinforcement learning with $Q^\star$-approximation and partial coverage. *arXiv preprint arXiv:2602.12107*, 2026. doi: 10.48550/arXiv.2602.12107. URL https://arxiv.org/abs/2602.12107.

McShane, E. J. Extension of range of functions. *Bulletin of the American Mathematical Society*, 40(12):837–842, 1934. doi: 10.1090/S0002-9904-1934-05978-0.

Mehrabi, M. and Wager, S. Off-policy evaluation in markov decision processes under weak distributional overlap. *arXiv preprint arXiv:2402.08201*, 2024. doi: 10.48550/arXiv.2402.08201. URL https://arxiv.org/abs/2402.08201.

Nilim, A. and Ghaoui, L. E. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005. doi: 10.1287/opre.1050.0216.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, NY, 2009. ISBN 978-0-387-79051-0. doi: 10.1007/b13794.

Uehara, M. and Sun, W. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=tyrJsbKAe6.

Villani, C. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009. ISBN 978-3-540-71049-3. doi: 10.1007/978-3-540-71050-9.

Whitney, H. Analytic extensions of differentiable functions defined in closed sets. *Transactions of the American Mathematical Society*, 36(1):63–89, 1934. doi: 10.1090/S0002-9947-1934-1501735-3.

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

Zhang, J. and Bareinboim, E. Causal eligibility traces for confounding robust off-policy evaluation. In Chiappa, S. and Magliacane, S. (eds.), *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 4933–4942. PMLR, 2025. URL https://proceedings.mlr.press/v286/zhang25d.html.

A. Additional Preliminaries and Notation

This appendix records the topological, measure-theoretic, and geometric conventions used throughout the proofs. The main text is formulated at the level of Bellman-compatible function classes rather than at the level of a fully specified primitive controlled Markov model. This distinction is essential under genuine support holes: on-support Bellman data need not determine a unique off-support extension, and the central object of the paper is precisely the identified set generated by those admissible extensions.

Throughout, we write $[H] := \{1, \dots, H\}$ for the horizon index set. Unless stated otherwise, all random variables are defined on a common complete probability space, all measurable spaces are equipped with their Borel σ-fields, and all conditional expectations are understood through fixed Borel versions. For clarity of exposition, the initial state is taken to be a deterministic point $s_1 \in \mathcal{S}_1$. Every statement admits the obvious extension to an initial distribution ν_1 by replacing $V_1(s_1)$ with $\int_{\mathcal{S}_1} V_1(s)\, \nu_1(ds)$.

A.1. Stagewise Spaces, Metrics, and Policies

Assumption A.1 (Standing topological conventions)**.** For each stage $h \in [H]$:

1. $\mathcal{S}_h$ is a nonempty compact metric space with metric d_h^{S}.
2. The action space $\mathcal{A}$ is a nonempty finite set.
3. The state–action space is the product $\mathcal{X}_h := \mathcal{S}_h \times \mathcal{A}$, endowed with the metric
$$d_h((s,a),(s',a')) := d_h^{\mathrm{S}}(s,s') + \mathbb{I}\{a \neq a'\}. \tag{A.1}$$
4. The stagewise reward R_h is integrable under every policy considered.

Because $\mathcal{A}$ is finite and $\mathcal{S}_h$ is compact, $\mathcal{X}_h$ is again a compact metric space. In particular, every $\mathcal{X}_h$ is Polish and therefore standard Borel. This guarantees the existence of regular conditional distributions and Borel versions of conditional expectations whenever they are needed later.

A (possibly history-independent) policy ν is identified with a sequence $\nu = (\nu_h)_{h=1}^H$, where each
$$\nu_h : \mathcal{S}_h \to \Delta(\mathcal{A})$$
is Borel measurable. Since $\mathcal{A}$ is finite, all policy averages are finite sums, and no measurable-selection issues arise from the action variable. We reserve π for the fixed target policy of interest and μ for the behavior policy generating the offline data.

Remark A.2 (Normalization of the action metric). The discrete contribution $\mathbb{I}\{a \neq a'\}$ in Equation (A.1) is only a normalization convention. Any equivalent product metric that separates distinct actions by a strictly positive amount would lead to the same theory after a deterministic rescaling of the Lipschitz radii. Fixing the action separation at one simply removes an irrelevant degree of freedom from later constants.

A.2. Occupancy Measures, Supports, and Distance-to-Support

For any policy ν and stage $h \in [H]$, let
$$X_h := (S_h, A_h) \in \mathcal{X}_h,$$
and define the stagewise occupancy measure
$$\rho_h^\nu(B) := \mathbb{P}^\nu(X_h \in B), \qquad B \subseteq \mathcal{X}_h \text{ Borel.}$$
The behavior occupancy at stage h is denoted ρ_h^{b}, and its topological support is
$$C_h := \mathrm{supp}(\rho_h^{\mathrm{b}}) = \big\{x \in \mathcal{X}_h : \rho_h^{\mathrm{b}}(U) > 0 \text{ for every open neighborhood } U \ni x\big\}.$$
Since ρ_h^{b} is a Borel probability measure on the compact metric space $\mathcal{X}_h$, the support C_h is nonempty, closed, and therefore compact.

For any nonempty set $A \subseteq \mathcal{X}_h$ and any $x \in \mathcal{X}_h$, we write

$$\mathrm{dist}_h(x, A) := \inf_{a \in A} d_h(x, a).$$

In particular, $\mathrm{dist}_h(x, C_h)$ measures the size of the support hole encountered by the point x at stage h.

Lemma A.3 (Distance-to-support regularity)**.** *For every stage $h \in [H]$, the map*

$$x \longmapsto \mathrm{dist}_h(x, C_h)$$

is 1*-Lipschitz on $\mathcal{X}_h$. Moreover,*

$$\mathrm{dist}_h(x, C_h) = 0 \quad \Longleftrightarrow \quad x \in C_h.$$

Proof. Fix $x, y \in \mathcal{X}_h$. For any $c \in C_h$, the triangle inequality gives

$$\mathrm{dist}_h(x, C_h) \le d_h(x, c) \le d_h(x, y) + d_h(y, c).$$

Taking the infimum over $c \in C_h$ yields

$$\mathrm{dist}_h(x, C_h) \le d_h(x, y) + \mathrm{dist}_h(y, C_h).$$

Swapping the roles of x and y gives

$$|\mathrm{dist}_h(x, C_h) - \mathrm{dist}_h(y, C_h)| \le d_h(x, y),$$

which proves the 1-Lipschitz claim.

If $x \in C_h$, then clearly $\mathrm{dist}_h(x, C_h) = 0$. Conversely, if $\mathrm{dist}_h(x, C_h) = 0$, then there exists a sequence $(c_n)_{n \ge 1} \subseteq C_h$ with $d_h(x, c_n) \to 0$. Since C_h is closed, this implies $x \in C_h$. □

A.3. Lipschitz Seminorms and Bounded Lipschitz Balls

For a function $f : \mathcal{X}_h \to \mathbb{R}$, define the stagewise Lipschitz seminorm

$$\mathrm{Lip}_h(f) := \sup_{x \ne y} \frac{|f(x) - f(y)|}{d_h(x, y)},$$

with the convention that the supremum over the empty set equals zero. Similarly, for a function $v : \mathcal{S}_h \to \mathbb{R}$, define

$$\mathrm{Lip}_h^{\mathrm{S}}(v) := \sup_{s \ne s'} \frac{|v(s) - v(s')|}{d_h^{\mathrm{S}}(s, s')}.$$

For $L \ge 0$, we write

$$\mathrm{Lip}_L(\mathcal{X}_h) := \{f : \mathcal{X}_h \to \mathbb{R} : \mathrm{Lip}_h(f) \le L\}.$$

Whenever a uniform sup-norm bound is also needed, we use the bounded Lipschitz ball

$$\mathrm{BL}_{M,L}(\mathcal{X}_h) := \{f : \mathcal{X}_h \to \mathbb{R} : \|f\|_\infty \le M,\ \mathrm{Lip}_h(f) \le L\}.$$

Lemma A.4 (Compactness of bounded Lipschitz balls)**.** *Fix $h \in [H]$, $M \ge 0$, and $L \ge 0$. Then $\mathrm{BL}_{M,L}(\mathcal{X}_h)$ is compact in $(C(\mathcal{X}_h), \|\cdot\|_\infty)$.*

Proof. Every function in $\mathrm{BL}_{M,L}(\mathcal{X}_h)$ is continuous, uniformly bounded by M, and the class is equicontinuous with common modulus $Ld_h(\cdot, \cdot)$. Since $\mathcal{X}_h$ is compact metric, Arzelà–Ascoli implies that every sequence in $\mathrm{BL}_{M,L}(\mathcal{X}_h)$ has a uniformly convergent subsequence. It therefore suffices to show that the class is closed under uniform limits.

Let $(f_n)_{n \ge 1} \subseteq \mathrm{BL}_{M,L}(\mathcal{X}_h)$ satisfy $\|f_n - f\|_\infty \to 0$. Then $\|f\|_\infty \le M$ follows by passing to the limit in $\|f_n\|_\infty \le M$. For any $x \ne y$,

$$\frac{|f(x) - f(y)|}{d_h(x, y)} \le \frac{|f(x) - f_n(x)|}{d_h(x, y)} + \frac{|f_n(x) - f_n(y)|}{d_h(x, y)} + \frac{|f_n(y) - f(y)|}{d_h(x, y)}.$$

Taking $n \to \infty$ yields

$$\frac{|f(x) - f(y)|}{d_h(x, y)} \le L.$$

Hence $\mathrm{Lip}_h(f) \le L$, so $f \in \mathrm{BL}_{M,L}(\mathcal{X}_h)$ and the class is closed. Therefore it is compact. □

A.4. Supported Bellman Data

For each $h \in [H-1]$ and each bounded Borel function $v : \mathcal{S}_{h+1} \to \mathbb{R}$, we write

$$(B_h^\pi v)(x) := \mathbb{E}[R_h + v(S_{h+1}) \mid X_h = x], \qquad x \in C_h, \tag{A.2}$$

where the right-hand side denotes a fixed Borel version of the corresponding conditional mean on C_h. At the terminal stage, we set $V_{H+1} \equiv 0$ by convention.

For any policy ν and any measurable $q : \mathcal{X}_h \to \mathbb{R}$, define the policy-averaging operator

$$(\mathsf{A}_h^\nu q)(s) := \sum_{a \in \mathcal{A}} \nu_h(a \mid s)\, q(s,a), \qquad s \in \mathcal{S}_h.$$

Thus, when a stagewise Q-function Q_h is given, its corresponding state-value function under the target policy is

$$V_h = \mathsf{A}_h^\pi Q_h.$$

The partial-identification problem studied in the main text conditions on the collection of supported Bellman data

$$\mathfrak{D}^\pi := \{(C_h, B_h^\pi) : h \in [H]\}.$$

Conceptually, $\mathfrak{D}^\pi$ is the population object pinned down by on-support data, while the off-support behavior of a candidate Q-sequence is governed by the Bellman–Lipschitz compatibility constraints introduced in Section 2. This is exactly where partial identification enters.

Remark A.5 (On the Bellman–Whitney terminology). The envelope operators introduced later in Section 3 combine a Bellman recursion with the extremal Lipschitz extension formulas of McShane and Whitney (McShane, 1934; Whitney, 1934). In the one-step case $H = 1$, the resulting identified interval coincides with the sharp smoothness-based no-overlap construction of Khan et al. (2024).

A.5. Hausdorff Geometry

For a nonempty set $A \subseteq \mathcal{X}_h$ and $\varepsilon \geq 0$, define the closed ε-enlargement

$$A^\varepsilon := \{x \in \mathcal{X}_h : \mathrm{dist}_h(x, A) \leq \varepsilon\}.$$

For two nonempty compact sets $A, B \subseteq \mathcal{X}_h$, their Hausdorff distance is

$$d_{\mathrm{H}}(A,B) := \max\Big[\sup_{a \in A} \mathrm{dist}_h(a,B),\ \sup_{b \in B} \mathrm{dist}_h(b,A)\Big].$$

Equivalently,

$$d_{\mathrm{H}}(A,B) \leq \varepsilon \quad \Longleftrightarrow \quad A \subseteq B^\varepsilon \ \text{ and } \ B \subseteq A^\varepsilon.$$

This is the natural geometry for perturbing estimated support sets in the statistical analysis.

The next identity will be used repeatedly when propagating support-estimation error through the Bellman–Whitney recursion.

Lemma A.6 (Hausdorff distance as a sup-norm distance between distance functions)**.** *Let $A, B \subseteq \mathcal{X}_h$ be nonempty compact sets. Then*

$$d_{\mathrm{H}}(A,B) = \sup_{x \in \mathcal{X}_h} |\mathrm{dist}_h(x,A) - \mathrm{dist}_h(x,B)|.$$

In particular,

$$\sup_{x \in \mathcal{X}_h} |\mathrm{dist}_h(x,A) - \mathrm{dist}_h(x,B)| \leq \varepsilon \quad \Longleftrightarrow \quad d_{\mathrm{H}}(A,B) \leq \varepsilon.$$

Proof. Set

$$\Delta(A,B) := \sup_{x \in \mathcal{X}_h} |\mathrm{dist}_h(x,A) - \mathrm{dist}_h(x,B)|.$$

We first show $\Delta(A,B) \le d_{\mathrm{H}}(A,B)$. Let $\eta := d_{\mathrm{H}}(A,B)$. By definition of Hausdorff distance, every point $a \in A$ satisfies $\mathrm{dist}_h(a,B) \le \eta$ and every point $b \in B$ satisfies $\mathrm{dist}_h(b,A) \le \eta$. Fix $x \in \mathcal{X}_h$ and $\delta > 0$. Choose $a_\delta \in A$ such that

$$d_h(x,a_\delta) \le \mathrm{dist}_h(x,A) + \delta.$$

Then

$$\mathrm{dist}_h(x,B) \le d_h(x,a_\delta) + \mathrm{dist}_h(a_\delta,B) \le \mathrm{dist}_h(x,A) + \delta + \eta.$$

Letting $\delta \downarrow 0$ yields

$$\mathrm{dist}_h(x,B) - \mathrm{dist}_h(x,A) \le \eta.$$

Exchanging the roles of A and B gives

$$|\mathrm{dist}_h(x,A) - \mathrm{dist}_h(x,B)| \le \eta.$$

Taking the supremum over x proves $\Delta(A,B) \le d_{\mathrm{H}}(A,B)$.

For the reverse inequality, fix $\delta > 0$. By definition of Hausdorff distance, either

$$\sup_{a\in A} \mathrm{dist}_h(a,B) \ge d_{\mathrm{H}}(A,B) - \delta$$

or

$$\sup_{b\in B} \mathrm{dist}_h(b,A) \ge d_{\mathrm{H}}(A,B) - \delta.$$

In the first case, choose $a_\delta \in A$ such that

$$\mathrm{dist}_h(a_\delta,B) \ge d_{\mathrm{H}}(A,B) - \delta.$$

Since $\mathrm{dist}_h(a_\delta,A) = 0$, we obtain

$$\Delta(A,B) \ge |\mathrm{dist}_h(a_\delta,A) - \mathrm{dist}_h(a_\delta,B)| = \mathrm{dist}_h(a_\delta,B) \ge d_{\mathrm{H}}(A,B) - \delta.$$

The second case is identical with A and B interchanged. Since $\delta > 0$ was arbitrary, this proves $\Delta(A,B) \ge d_{\mathrm{H}}(A,B)$.

Combining the two inequalities yields the claimed identity. □

A.6. How Appendix A Is Used Later

The roles of the present appendix in the rest of the paper are as follows.

- lemma A.3 provides the elementary support-hole geometry used later in the ambiguity analysis.
- lemma A.4 supplies the compactness input for the infinite-dimensional duality argument.
- lemma A.6 is the basic metric identity behind the stability theory for estimated support sets.

The remaining structural properties of the Bellman–Whitney envelope operators themselves are deferred to Appendix B, where they are proved once and then used throughout the sequel.

B. Fundamental Envelope Lemmas

The operators used in the main text are stagewise versions of the classical McShane–Whitney extremal Lipschitz constructions (McShane, 1934; Whitney, 1934). In our setting, however, the input supported Bellman data need *not* themselves be Lipschitz on the behavior support. The relevant objects are therefore not merely exact extensions, but rather one-sided extremal envelopes under support constraints. This appendix isolates the basic properties of those envelopes once and for all.

B.1. Generic Envelope Operators

Fix for the duration of this appendix a compact metric space (X, d), a nonempty compact subset $A \subseteq X$, and a radius $L \geq 0$. For any bounded Borel function $g : A \to \mathbb{R}$, define

$$\left(\mathcal{W}^-_{A,L} g\right)(x) := \sup_{a \in A} \{g(a) - L\, d(x, a)\}, \qquad x \in X, \tag{B.1}$$

and

$$\left(\mathcal{W}^+_{A,L} g\right)(x) := \inf_{a \in A} \{g(a) + L\, d(x, a)\}, \qquad x \in X. \tag{B.2}$$

In the stagewise setting of the main text, we apply these formulas with $X = \mathcal{X}_h$, $A = C_h$, and $L = L_h$, in which case

$$\mathcal{W}^-_{h,L_h} = \mathcal{W}^-_{C_h,L_h}, \qquad \mathcal{W}^+_{h,L_h} = \mathcal{W}^+_{C_h,L_h}.$$

Remark B.1 (One-sided envelopes versus exact interpolation). If g happens to be L-Lipschitz on A, then $\mathcal{W}^-_{A,L} g$ and $\mathcal{W}^+_{A,L} g$ coincide with the usual lower and upper McShane–Whitney extensions on A. In the partially identified regime studied in this paper, however, the supported Bellman data need not be L-Lipschitz. In that case, $\mathcal{W}^-_{A,L} g$ and $\mathcal{W}^+_{A,L} g$ no longer interpolate g exactly on A; instead, they represent the smallest feasible L-Lipschitz *majorant* and the largest feasible L-Lipschitz *minorant* compatible with the support information.

B.2. Basic Regularity and Order Properties

Lemma B.2 (Regularity, support bounds, and monotonicity)**.** *Let $g, \tilde{g} : A \to \mathbb{R}$ be bounded Borel functions. Then the following hold.*

1. *The functions $\mathcal{W}^-_{A,L} g$ and $\mathcal{W}^+_{A,L} g$ are bounded and L-Lipschitz on X.*
2. *For every $a \in A$,*
$$\left(\mathcal{W}^-_{A,L} g\right)(a) \geq g(a), \qquad \left(\mathcal{W}^+_{A,L} g\right)(a) \leq g(a). \tag{B.3}$$
3. *If $g \leq \tilde{g}$ pointwise on A, then*
$$\mathcal{W}^-_{A,L} g \leq \mathcal{W}^-_{A,L} \tilde{g}, \qquad \mathcal{W}^+_{A,L} g \leq \mathcal{W}^+_{A,L} \tilde{g} \quad \text{pointwise on } X. \tag{B.4}$$

Proof. We first prove boundedness of $\mathcal{W}^-_{A,L} g$. Since A is nonempty and compact, $\operatorname{diam}(A) \leq \operatorname{diam}(X) < \infty$. For any fixed $x \in X$,

$$\sup_{a \in A} g(a) - L \operatorname{diam}(X) \leq \sup_{a \in A} \{g(a) - L\, d(x, a)\} \leq \sup_{a \in A} g(a),$$

so $\mathcal{W}^-_{A,L} g$ is finite everywhere. The same argument yields

$$\inf_{a \in A} g(a) \leq \inf_{a \in A} \{g(a) + L\, d(x, a)\} \leq \inf_{a \in A} g(a) + L \operatorname{diam}(X),$$

hence $\mathcal{W}^+_{A,L} g$ is finite everywhere.

We now prove the Lipschitz claim for $\mathcal{W}^-_{A,L} g$. Fix $x, y \in X$. For every $a \in A$,

$$g(a) - L\, d(x, a) \leq g(a) - L\, d(y, a) + L\, d(x, y),$$

by the triangle inequality. Taking the supremum over $a \in A$ gives

$$\left(\mathcal{W}^-_{A,L} g\right)(x) \leq \left(\mathcal{W}^-_{A,L} g\right)(y) + L\, d(x, y).$$

Exchanging the roles of x and y yields

$$\left| \left(\mathcal{W}^-_{A,L} g\right)(x) - \left(\mathcal{W}^-_{A,L} g\right)(y) \right| \leq L\, d(x, y).$$

Thus $\mathcal{W}^-_{A,L}g$ is L-Lipschitz. The proof for $\mathcal{W}^+_{A,L}g$ is identical: for every $a \in A$,

$$g(a) + L\,d(x,a) \le g(a) + L\,d(y,a) + L\,d(x,y),$$

and taking infima gives the same Lipschitz bound.

For (B.3), fix $a \in A$. Since the point a is itself admissible in (B.1) and (B.2),

$$\left(\mathcal{W}^-_{A,L}g\right)(a) \ge g(a) - L\,d(a,a) = g(a),$$

and

$$\left(\mathcal{W}^+_{A,L}g\right)(a) \le g(a) + L\,d(a,a) = g(a).$$

Finally, if $g \le \tilde{g}$ on A, then for every $x \in X$ and every $a \in A$,

$$g(a) - L\,d(x,a) \le \tilde{g}(a) - L\,d(x,a), \qquad g(a) + L\,d(x,a) \le \tilde{g}(a) + L\,d(x,a).$$

Taking suprema in the first display and infima in the second proves (B.4). □

B.3. Extremal Characterization

The next result is the key structural fact used throughout the paper.

Lemma B.3 (Extremal one-sided characterization)**.** *Let $g : A \to \mathbb{R}$ be bounded Borel, and define the two feasible families*

$$\mathfrak{M}_{A,L}(g) := \{f \in \mathrm{Lip}_L(X) : f(a) \ge g(a) \text{ for all } a \in A\},$$

and

$$\mathfrak{m}_{A,L}(g) := \{f \in \mathrm{Lip}_L(X) : f(a) \le g(a) \text{ for all } a \in A\}.$$

Then $\mathfrak{M}_{A,L}(g)$ and $\mathfrak{m}_{A,L}(g)$ are nonempty, and for every $x \in X$,

$$\left(\mathcal{W}^-_{A,L}g\right)(x) = \inf_{f \in \mathfrak{M}_{A,L}(g)} f(x), \tag{B.5}$$

and

$$\left(\mathcal{W}^+_{A,L}g\right)(x) = \sup_{f \in \mathfrak{m}_{A,L}(g)} f(x). \tag{B.6}$$

Equivalently, $\mathcal{W}^-_{A,L}g$ is the pointwise smallest L-Lipschitz function on X that dominates g on A, while $\mathcal{W}^+_{A,L}g$ is the pointwise largest L-Lipschitz function on X that is dominated by g on A.

Proof. The classes are nonempty because constant functions are Lipschitz: if $M := \sup_{a\in A} g(a)$ and $m := \inf_{a\in A} g(a)$, then the constant function $f \equiv M$ belongs to $\mathfrak{M}_{A,L}(g)$ and the constant function $f \equiv m$ belongs to $\mathfrak{m}_{A,L}(g)$.

We prove (B.5). By lemma B.2, $\mathcal{W}^-_{A,L}g$ is itself L-Lipschitz on X and satisfies $\mathcal{W}^-_{A,L}g \ge g$ on A. Hence $\mathcal{W}^-_{A,L}g \in \mathfrak{M}_{A,L}(g)$, so

$$\inf_{f \in \mathfrak{M}_{A,L}(g)} f(x) \le \left(\mathcal{W}^-_{A,L}g\right)(x).$$

Conversely, let $f \in \mathfrak{M}_{A,L}(g)$ be arbitrary. Since f is L-Lipschitz, for every $a \in A$,

$$f(x) \ge f(a) - L\,d(x,a) \ge g(a) - L\,d(x,a).$$

Taking the supremum over $a \in A$ yields

$$f(x) \ge \left(\mathcal{W}^-_{A,L}g\right)(x).$$

Since this holds for every $f \in \mathfrak{M}_{A,L}(g)$,

$$\inf_{f \in \mathfrak{M}_{A,L}(g)} f(x) \ge \left(\mathcal{W}^-_{A,L}g\right)(x).$$

Together with the previous display, this proves (B.5).

The proof of (B.6) is symmetric. By lemma B.2, $\mathcal{W}^+_{A,L}g$ is L-Lipschitz and satisfies $\mathcal{W}^+_{A,L}g \le g$ on A, hence $\mathcal{W}^+_{A,L}g \in \mathfrak{m}_{A,L}(g)$. Therefore

$$\sup_{f\in\mathfrak{m}_{A,L}(g)} f(x) \ge \Big(\mathcal{W}^+_{A,L}g\Big)(x).$$

If $f \in \mathfrak{m}_{A,L}(g)$, then for every $a \in A$,

$$f(x) \le f(a) + L\,d(x,a) \le g(a) + L\,d(x,a).$$

Taking the infimum over $a \in A$ yields

$$f(x) \le \Big(\mathcal{W}^+_{A,L}g\Big)(x).$$

Since this holds for every $f \in \mathfrak{m}_{A,L}(g)$,

$$\sup_{f\in\mathfrak{m}_{A,L}(g)} f(x) \le \Big(\mathcal{W}^+_{A,L}g\Big)(x),$$

which proves (B.6). □

Corollary B.4 (Classical exact extension regime)**.** *Let $g : A \to \mathbb{R}$ be L-Lipschitz with respect to the restricted metric on A. Then the following hold.*

1. *$\mathcal{W}^-_{A,L}g$ and $\mathcal{W}^+_{A,L}g$ agree with g on A:*

$$\Big(\mathcal{W}^-_{A,L}g\Big)(a) = g(a) = \Big(\mathcal{W}^+_{A,L}g\Big)(a), \qquad a \in A.$$

2. *Every L-Lipschitz extension $f : X \to \mathbb{R}$ of g satisfies*

$$\mathcal{W}^-_{A,L}g \le f \le \mathcal{W}^+_{A,L}g \qquad \textit{pointwise on } X.$$

3. *In particular,*

$$\mathcal{W}^-_{A,L}g \le \mathcal{W}^+_{A,L}g \qquad \textit{pointwise on } X.$$

Proof. Fix $a_0 \in A$. Since g is L-Lipschitz on A, for every $a \in A$,

$$g(a) - L\,d(a_0,a) \le g(a_0) \le g(a) + L\,d(a_0,a).$$

Taking the supremum over $a \in A$ in the left inequality and the infimum over $a \in A$ in the right inequality yields

$$\Big(\mathcal{W}^-_{A,L}g\Big)(a_0) \le g(a_0) \le \Big(\mathcal{W}^+_{A,L}g\Big)(a_0).$$

Combined with the support bounds from lemma B.2, this gives equality on A.

If f is an L-Lipschitz extension of g, then $f \in \mathfrak{M}_{A,L}(g)$ and $f \in \mathfrak{m}_{A,L}(g)$ simultaneously. The sandwich inequality therefore follows immediately from lemma B.3. The final claim is obtained by taking $f = \mathcal{W}^+_{A,L}g$ in the left inequality, or equivalently $f = \mathcal{W}^-_{A,L}g$ in the right inequality. □

B.4. Stability with Respect to Supported Values

Lemma B.5 (Supremum-norm stability on a fixed support)**.** *Let $g, \tilde{g} : A \to \mathbb{R}$ be bounded Borel functions. Then*

$$\|\mathcal{W}^-_{A,L}g - \mathcal{W}^-_{A,L}\tilde{g}\|_\infty \le \|g - \tilde{g}\|_{\infty,A}, \tag{B.7}$$

and

$$\|\mathcal{W}^+_{A,L}g - \mathcal{W}^+_{A,L}\tilde{g}\|_\infty \le \|g - \tilde{g}\|_{\infty,A}. \tag{B.8}$$

Proof. Let

$$\delta := \|g - \tilde{g}\|_{\infty,A}.$$

Then $g \le \tilde{g} + \delta$ and $\tilde{g} \le g + \delta$ pointwise on A. Using lemma B.2 and the defining formulas,

$$\mathcal{W}^-_{A,L} g \le \mathcal{W}^-_{A,L}(\tilde{g} + \delta) = \mathcal{W}^-_{A,L}\tilde{g} + \delta,$$

and similarly

$$\mathcal{W}^-_{A,L}\tilde{g} \le \mathcal{W}^-_{A,L} g + \delta.$$

Hence

$$\|\mathcal{W}^-_{A,L} g - \mathcal{W}^-_{A,L}\tilde{g}\|_\infty \le \delta.$$

The proof for $\mathcal{W}^+_{A,L}$ is identical:

$$\mathcal{W}^+_{A,L} g \le \mathcal{W}^+_{A,L}(\tilde{g} + \delta) = \mathcal{W}^+_{A,L}\tilde{g} + \delta,$$

and vice versa, implying (B.8). □

B.5. Joint Stability with Respect to Support and Value Perturbations

The next lemma controls the effect of simultaneously perturbing both the support set and the supported function values. It is the correct generic tool for the statistical analysis, where estimated support sets and estimated supported Bellman targets are both present.

For nonempty compact sets $A, B \subseteq X$ and bounded Borel functions $g : A \to \mathbb{R}$, $\tilde{g} : B \to \mathbb{R}$, define the directed discrepancy

$$\Delta^{(L)}_{A\to B}(g, \tilde{g}) := \sup_{a\in A} \inf_{b\in B} \{|g(a) - \tilde{g}(b)| + L\, d(a, b)\}.$$

Define the symmetrized discrepancy

$$\Delta^{(L)}_{A,B}(g, \tilde{g}) := \max\Big[\Delta^{(L)}_{A\to B}(g, \tilde{g}),\ \Delta^{(L)}_{B\to A}(\tilde{g}, g)\Big]. \tag{B.9}$$

Lemma B.6 (Support-value pair stability)**.** *Let $A, B \subseteq X$ be nonempty compact sets, and let $g : A \to \mathbb{R}$, $\tilde{g} : B \to \mathbb{R}$ be bounded Borel functions. Then*

$$\|\mathcal{W}^-_{A,L} g - \mathcal{W}^-_{B,L}\tilde{g}\|_\infty \le \Delta^{(L)}_{A,B}(g, \tilde{g}), \tag{B.10}$$

and

$$\|\mathcal{W}^+_{A,L} g - \mathcal{W}^+_{B,L}\tilde{g}\|_\infty \le \Delta^{(L)}_{A,B}(g, \tilde{g}). \tag{B.11}$$

Proof. We first prove (B.10). Fix $x \in X$ and $\varepsilon > 0$. For each $a \in A$, choose $b_{a,\varepsilon} \in B$ such that

$$|g(a) - \tilde{g}(b_{a,\varepsilon})| + L\, d(a, b_{a,\varepsilon}) \le \Delta^{(L)}_{A\to B}(g, \tilde{g}) + \varepsilon.$$

Then

$$\begin{aligned} g(a) - L\, d(x, a) &\le \tilde{g}(b_{a,\varepsilon}) + |g(a) - \tilde{g}(b_{a,\varepsilon})| - L\, d(x, a) \\ &\le \tilde{g}(b_{a,\varepsilon}) - L\, d(x, b_{a,\varepsilon}) + |g(a) - \tilde{g}(b_{a,\varepsilon})| + L\, d(a, b_{a,\varepsilon}) \\ &\le \mathcal{W}^-_{B,L}\tilde{g}(x) + \Delta^{(L)}_{A\to B}(g, \tilde{g}) + \varepsilon. \end{aligned}$$

Taking the supremum over $a \in A$ yields

$$\mathcal{W}^-_{A,L} g(x) \le \mathcal{W}^-_{B,L}\tilde{g}(x) + \Delta^{(L)}_{A\to B}(g, \tilde{g}) + \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary,

$$\mathcal{W}^-_{A,L} g(x) \le \mathcal{W}^-_{B,L}\tilde{g}(x) + \Delta^{(L)}_{A\to B}(g, \tilde{g}).$$

Repeating the argument with (A, g) and $(B, \tilde{g})$ interchanged gives

$$\mathcal{W}^-_{B,L}\tilde{g}(x) \le \mathcal{W}^-_{A,L}g(x) + \Delta^{(L)}_{B\to A}(\tilde{g}, g).$$

Combining the two bounds and taking the supremum over $x \in X$ proves (B.10).

We next prove (B.11). Fix $x \in X$ and $\varepsilon > 0$. For each $b \in B$, choose $a_{b,\varepsilon} \in A$ such that

$$|g(a_{b,\varepsilon}) - \tilde{g}(b)| + L\, d(a_{b,\varepsilon}, b) \le \Delta^{(L)}_{B\to A}(\tilde{g}, g) + \varepsilon.$$

Then

$$\begin{aligned}\tilde{g}(b) + L\, d(x, b) &\le g(a_{b,\varepsilon}) + |g(a_{b,\varepsilon}) - \tilde{g}(b)| + L\, d(x, b)\\ &\le g(a_{b,\varepsilon}) + L\, d(x, a_{b,\varepsilon}) + |g(a_{b,\varepsilon}) - \tilde{g}(b)| + L\, d(a_{b,\varepsilon}, b)\\ &\le \mathcal{W}^+_{A,L}g(x) + \Delta^{(L)}_{B\to A}(\tilde{g}, g) + \varepsilon.\end{aligned}$$

Taking the infimum over $b \in B$ yields

$$\mathcal{W}^+_{B,L}\tilde{g}(x) \le \mathcal{W}^+_{A,L}g(x) + \Delta^{(L)}_{B\to A}(\tilde{g}, g) + \varepsilon.$$

Letting $\varepsilon \downarrow 0$ and then swapping the roles of (A, g) and $(B, \tilde{g})$ gives

$$|\mathcal{W}^+_{A,L}g(x) - \mathcal{W}^+_{B,L}\tilde{g}(x)| \le \Delta^{(L)}_{A,B}(g, \tilde{g}).$$

Taking the supremum over $x \in X$ proves (B.11). □

Corollary B.7 (Hausdorff perturbation under a common ambient Lipschitz envelope)**.** *Let $A, B \subseteq X$ be nonempty compact sets, and let $f : X \to \mathbb{R}$ be M-Lipschitz. Set*

$$g := f|_A, \qquad \tilde{g} := f|_B.$$

Then

$$\|\mathcal{W}^-_{A,L}g - \mathcal{W}^-_{B,L}\tilde{g}\|_\infty \le (M + L)\, d_{\mathrm{H}}(A, B), \tag{B.12}$$

and

$$\|\mathcal{W}^+_{A,L}g - \mathcal{W}^+_{B,L}\tilde{g}\|_\infty \le (M + L)\, d_{\mathrm{H}}(A, B). \tag{B.13}$$

In particular, if f is L-Lipschitz on X, then both bounds simplify to

$$\|\mathcal{W}^-_{A,L}g - \mathcal{W}^-_{B,L}\tilde{g}\|_\infty \vee \|\mathcal{W}^+_{A,L}g - \mathcal{W}^+_{B,L}\tilde{g}\|_\infty \le 2L\, d_{\mathrm{H}}(A, B).$$

Proof. Fix $a \in A$. By definition of Hausdorff distance, there exists $b \in B$ such that $d(a, b) \le d_{\mathrm{H}}(A, B)$. Since f is M-Lipschitz,

$$|g(a) - \tilde{g}(b)| + L\, d(a, b) = |f(a) - f(b)| + L\, d(a, b) \le (M + L)\, d_{\mathrm{H}}(A, B).$$

Taking the infimum over such b and then the supremum over $a \in A$ yields

$$\Delta^{(L)}_{A\to B}(g, \tilde{g}) \le (M + L)\, d_{\mathrm{H}}(A, B).$$

The same argument with A and B interchanged gives

$$\Delta^{(L)}_{B\to A}(\tilde{g}, g) \le (M + L)\, d_{\mathrm{H}}(A, B).$$

Hence

$$\Delta^{(L)}_{A,B}(g, \tilde{g}) \le (M + L)\, d_{\mathrm{H}}(A, B).$$

The stated inequalities now follow immediately from lemma B.6. The final display is the special case $M = L$. □

B.6. Role of the Envelope Lemmas in the Main Proofs

The preceding results are used in the sequel as follows.

- lemma B.3 is the core ingredient in the sharp interval theorem: it identifies the stagewise Bellman–Whitney recursion as the pointwise extremal feasible sequence under one-sided support constraints.
- corollary B.4 recovers the classical exact-extension regime and will be used to justify the $H = 1$ reduction.
- lemmas B.5 and B.6 are the basic perturbation tools needed later for the stability and statistical endpoint-estimation analysis.

C. Primitive Smoothness as a Sufficient Condition

The main text is formulated at the level of Bellman-compatible function classes. This is the mathematically cleanest way to study support-hole partial identification, because it avoids committing to a unique primitive off-support model. Nevertheless, it is important to understand when the abstract closure assumptions used in the main theorems follow from concrete reward and transition regularity. This appendix gives such primitive sufficient conditions.

The key estimate is a Bellman Lipschitz-propagation bound under Wasserstein-Lipschitz transitions. Its proof uses the Kantorovich–Rubinstein dual characterization of the one-Wasserstein distance; see, e.g., Villani (2009).

C.1. Primitive Assumptions

For this appendix only, suppose that for each stage $h \in [H-1]$ the primitive dynamics are specified by a Markov kernel

$$P_h(\cdot \mid x) \quad \text{on } \mathcal{S}_{h+1}, \qquad x \in \mathcal{X}_h,$$

and that for each stage $h \in [H]$ the conditional reward mean

$$r_h(x) := \mathbb{E}[R_h \mid X_h = x], \qquad x \in \mathcal{X}_h,$$

is well defined.

Assumption C.1 (Primitive reward and transition smoothness)**.** There exist nonnegative constants ρ_h for $h \in [H]$ and κ_h for $h \in [H-1]$ such that:

1. for every $h \in [H]$,
$$|r_h(x) - r_h(y)| \le \rho_h \, d_h(x, y) \qquad \forall x, y \in \mathcal{X}_h;$$
2. for every $h \in [H-1]$,
$$W_1(P_h(\cdot \mid x), P_h(\cdot \mid y)) \le \kappa_h \, d_h(x, y) \qquad \forall x, y \in \mathcal{X}_h,$$
where W_1 is the one-Wasserstein distance on probability measures over $\mathcal{S}_{h+1}$ induced by the state metric d^{S}_{h+1}.

The fixed-policy closure assumption from the main text requires an additional regularity condition on how the target policy varies with the state.

Assumption C.2 (Stagewise ℓ_1-Lipschitz target policy)**.** There exist nonnegative constants Λ_h for $h \in [H]$ such that

$$\|\pi_h(\cdot \mid s) - \pi_h(\cdot \mid s')\|_1 \le \Lambda_h \, d^{\mathrm{S}}_h(s, s') \qquad \forall s, s' \in \mathcal{S}_h.$$

Remark C.3 (Why assumption C.2 appears only in the fixed-policy result). For fixed-policy closure, the continuation class is

$$\mathcal{V}^{\pi,L}_{h+1} = \Big\{ \mathsf{A}^{\pi}_{h+1} q : q \in \mathrm{Lip}_{L_{h+1}}(\mathcal{X}_{h+1}) \Big\},$$

so one must control how policy averaging maps Lipschitz Q-functions into Lipschitz value functions. This depends on the state-regularity of π_{h+1}. By contrast, the control closure assumption uses

$$v(s) = \max_a q(s, a),$$

and taking a maximum over finitely many Lipschitz actions preserves Lipschitz regularity without any assumption on a policy map. This is why the control result below does not need assumption C.2.

C.2. Two Primitive Lipschitz-Propagation Lemmas

The first lemma controls policy averaging. The discrete action metric in Appendix A is crucial: it makes the cross-action oscillation of an L-Lipschitz Q-function automatically of order L.

Lemma C.4 (Policy averaging preserves Lipschitz regularity)**.** *Assume assumption C.2. Fix $h \in [H]$, and let $q : \mathcal{X}_h \to \mathbb{R}$ be bounded and L-Lipschitz with respect to d_h. Define*

$$v(s) := (\mathsf{A}_h^\pi q)(s) = \sum_{a\in\mathcal{A}} \pi_h(a \mid s)\, q(s,a), \qquad s \in \mathcal{S}_h.$$

Then v is d_h^{S}-Lipschitz and satisfies

$$\mathrm{Lip}_h^{\mathrm{S}}(v) \le \left(1 + \frac{\Lambda_h}{2}\right) L. \tag{C.1}$$

Proof. Fix $s, s' \in \mathcal{S}_h$. Write

$$\pi := \pi_h(\cdot \mid s), \qquad \pi' := \pi_h(\cdot \mid s').$$

Then

$$v(s) - v(s') = \sum_{a\in\mathcal{A}} \pi(a)\big(q(s,a) - q(s',a)\big) + \sum_{a\in\mathcal{A}} \big(\pi(a) - \pi'(a)\big) q(s',a).$$

Hence

$$\begin{aligned} |v(s) - v(s')| &\le \sum_{a\in\mathcal{A}} \pi(a)|q(s,a) - q(s',a)| + \left|\sum_{a\in\mathcal{A}} \big(\pi(a) - \pi'(a)\big) q(s',a)\right| \\ &\le L\, d_h^{\mathrm{S}}(s,s') + \left|\sum_{a\in\mathcal{A}} \big(\pi(a) - \pi'(a)\big) q(s',a)\right|. \end{aligned} \tag{C.2}$$

We now bound the second term. Let

$$m_{s'} := \frac{1}{2}\left(\max_{a\in\mathcal{A}} q(s',a) + \min_{a\in\mathcal{A}} q(s',a)\right).$$

Because $\sum_a (\pi(a) - \pi'(a)) = 0$, we may subtract the constant $m_{s'}$:

$$\sum_{a\in\mathcal{A}} \big(\pi(a) - \pi'(a)\big) q(s',a) = \sum_{a\in\mathcal{A}} \big(\pi(a) - \pi'(a)\big)\big(q(s',a) - m_{s'}\big).$$

Therefore

$$\left|\sum_{a\in\mathcal{A}} \big(\pi(a) - \pi'(a)\big) q(s',a)\right| \le \max_{a\in\mathcal{A}} |q(s',a) - m_{s'}| \cdot \|\pi - \pi'\|_1. \tag{C.3}$$

By construction of $m_{s'}$,

$$\max_{a\in\mathcal{A}} |q(s',a) - m_{s'}| = \frac{1}{2}\Big(\max_a q(s',a) - \min_a q(s',a)\Big).$$

Since the action metric contributes 1 whenever $a \neq a'$, any two actions at the same state are distance at most 1 apart in $\mathcal{X}_h$, and thus

$$\max_a q(s',a) - \min_a q(s',a) \le L.$$

Combining this with assumption C.2 and (C.3) gives

$$\left|\sum_{a\in\mathcal{A}} \big(\pi(a) - \pi'(a)\big) q(s',a)\right| \le \frac{L}{2}\, \Lambda_h\, d_h^{\mathrm{S}}(s,s').$$

Substituting into (C.2) yields

$$|v(s) - v(s')| \le \left(1 + \frac{\Lambda_h}{2}\right) L\, d_h^{\mathrm{S}}(s,s'),$$

which proves (C.1). □

The second lemma is the primitive Bellman propagation inequality.

Lemma C.5 (Primitive Bellman Lipschitz propagation)**.** *Assume assumption C.1. Fix $h \in [H-1]$, and let $v : \mathcal{S}_{h+1} \to \mathbb{R}$ be bounded and K-Lipschitz with respect to d^{S}_{h+1}. Then the function*

$$q^v_h(x) := r_h(x) + \int_{\mathcal{S}_{h+1}} v(s')\, P_h(ds' \mid x), \qquad x \in \mathcal{X}_h,$$

is $(\rho_h + \kappa_h K)$-Lipschitz on $\mathcal{X}_h$:

$$\mathrm{Lip}_h(q^v_h) \le \rho_h + \kappa_h K. \tag{C.4}$$

At the terminal stage,

$$q_H(x) := r_H(x)$$

is ρ_H-Lipschitz on $\mathcal{X}_H$.

Proof. Fix $x, y \in \mathcal{X}_h$ with $h \le H-1$. Then

$$|q^v_h(x) - q^v_h(y)| \le |r_h(x) - r_h(y)| + \left| \int v(s')\, P_h(ds' \mid x) - \int v(s')\, P_h(ds' \mid y) \right|.$$

By assumption C.1,

$$|r_h(x) - r_h(y)| \le \rho_h\, d_h(x, y).$$

For the second term, if $K = 0$ the bound is trivial. If $K > 0$, then v/K is 1-Lipschitz on $\mathcal{S}_{h+1}$, so the Kantorovich–Rubinstein dual representation of W_1 yields

$$\left| \int v\, dP_h(\cdot \mid x) - \int v\, dP_h(\cdot \mid y) \right| \le K\, W_1(P_h(\cdot \mid x), P_h(\cdot \mid y)) \le K\, \kappa_h\, d_h(x, y),$$

where the last inequality again uses assumption C.1; see Villani (2009). Therefore

$$|q^v_h(x) - q^v_h(y)| \le (\rho_h + \kappa_h K) d_h(x, y),$$

which proves (C.4). The terminal-stage statement is exactly the reward smoothness assumption. □

C.3. Primitive Sufficient Conditions for Fixed-Policy Closure

Define the effective policy-averaged radii

$$\bar{L}_h := \left(1 + \frac{\Lambda_h}{2}\right) L_h, \qquad h \in [H],$$

and set

$$\bar{L}_{H+1} := 0.$$

Proposition C.6 (Primitive sufficient condition for fixed-policy closure)**.** *Assume assumptions C.1 and C.2. Suppose the radius vector L satisfies*

$$L_H \ge \rho_H \qquad \textit{and} \qquad L_h \ge \rho_h + \kappa_h \bar{L}_{h+1} \quad \forall h \in \{1, \dots, H-1\}. \tag{C.5}$$

Then assumption 3.1 holds.

Proof. Fix $h \in [H]$. If $h = H$, then $\mathcal{V}^{\pi,L}_{H+1} = \{0\}$, so the supported Bellman target is just

$$c \longmapsto (B^\pi_H 0)(c) = r_H(c), \qquad c \in C_H,$$

which is ρ_H-Lipschitz on $\mathcal{X}_H$ and therefore L_H-Lipschitz on C_H by (C.5).

Now let $h \in \{1, \dots, H-1\}$ and take any $v \in \mathcal{V}^{\pi,L}_{h+1}$. By definition of the tail value class, there exists a bounded L_{h+1}-Lipschitz function $q : \mathcal{X}_{h+1} \to \mathbb{R}$ such that

$$v = \mathsf{A}^\pi_{h+1} q.$$

By lemma C.4,

$$\mathrm{Lip}^{\mathrm{S}}_{h+1}(v) \le \bar{L}_{h+1}.$$

Applying lemma C.5 with $K = \bar{L}_{h+1}$ shows that the function

$$x \longmapsto (B^{\pi}_h v)(x) = r_h(x) + \int v(s')\, P_h(ds' \mid x)$$

is $(\rho_h + \kappa_h \bar{L}_{h+1})$-Lipschitz on all of $\mathcal{X}_h$. By (C.5), this Lipschitz constant is at most L_h, so the restriction of $x \mapsto (B^{\pi}_h v)(x)$ to C_h is L_h-Lipschitz. This is exactly assumption 3.1. □

The next proposition shows that, under the same primitive conditions, the *actual* target-policy value sequence of the primitive model belongs to the Bellman–Lipschitz compatibility class.

Proposition C.7 (Primitive smoothness implies Bellman compatibility)**.** *Assume assumptions C.1 and C.2 and* (C.5)*. Define recursively the primitive target-policy value sequence by*

$$V^{\pi}_{H+1} \equiv 0,$$

$$Q^{\pi}_h(x) := r_h(x) + \int_{\mathcal{S}_{h+1}} V^{\pi}_{h+1}(s')\, P_h(ds' \mid x), \qquad x \in \mathcal{X}_h,$$

and

$$V^{\pi}_h := \mathsf{A}^{\pi}_h Q^{\pi}_h, \qquad h \in [H].$$

Then

$$\big(Q^{\pi}_h, V^{\pi}_h\big)_{h=1}^{H+1} \in \mathfrak{C}^{\pi}_L.$$

Proof. We prove by backward induction on h that

$$\mathrm{Lip}_h(Q^{\pi}_h) \le L_h, \qquad h \in [H].$$

At stage H, we have

$$Q^{\pi}_H(x) = r_H(x),$$

so

$$\mathrm{Lip}_H(Q^{\pi}_H) \le \rho_H \le L_H$$

by (C.5).

Now fix $h \in \{1, \dots, H-1\}$ and assume

$$\mathrm{Lip}_{h+1}\big(Q^{\pi}_{h+1}\big) \le L_{h+1}.$$

Then

$$V^{\pi}_{h+1} = \mathsf{A}^{\pi}_{h+1} Q^{\pi}_{h+1},$$

so lemma C.4 yields

$$\mathrm{Lip}^{\mathrm{S}}_{h+1}\big(V^{\pi}_{h+1}\big) \le \bar{L}_{h+1}.$$

Applying lemma C.5 to the continuation value V^{π}_{h+1} gives

$$\mathrm{Lip}_h(Q^{\pi}_h) \le \rho_h + \kappa_h \bar{L}_{h+1} \le L_h,$$

again by (C.5). This proves the inductive claim.

The Bellman compatibility conditions now follow directly from the definitions: $V^{\pi}_{H+1} \equiv 0$, $V^{\pi}_h = \mathsf{A}^{\pi}_h Q^{\pi}_h$, and

$$Q^{\pi}_h(x) = (B^{\pi}_h V^{\pi}_{h+1})(x) \qquad \forall x \in \mathcal{X}_h,$$

hence in particular on C_h. Since each Q^{π}_h is bounded, Borel measurable, and L_h-Lipschitz, the sequence belongs to $\mathfrak{C}^{\pi}_L$. □

C.4. Primitive Sufficient Conditions for Control Closure

The control case is simpler because the maximum of finitely many L_{h+1}-Lipschitz action-value slices is again L_{h+1}-Lipschitz on the state space, with no policy-regularity term.

Proposition C.8 (Primitive sufficient condition for control closure)**.** *Assume assumption C.1. Suppose the radius vector L satisfies*

$$L_H \geq \rho_H \qquad \textit{and} \qquad L_h \geq \rho_h + \kappa_h L_{h+1} \quad \forall h \in \{1, \dots, H-1\}. \tag{C.6}$$

Then assumption J.1 holds.

Proof. Fix $h \in [H]$. If $h = H$, then as before the only continuation value is zero, and the supported Bellman target is r_H, which is L_H-Lipschitz on C_H by (C.6).

Now let $h \in \{1, \dots, H-1\}$ and take any $v \in \mathcal{V}_{h+1}^{\mathrm{ctl},L}$. By definition, there exists $q \in \mathrm{Lip}_{L_{h+1}}(\mathcal{X}_{h+1})$ such that

$$v(s) = \max_{a \in \mathcal{A}} q(s,a), \qquad s \in \mathcal{S}_{h+1}.$$

For any $s, s' \in \mathcal{S}_{h+1}$,

$$v(s) - v(s') = \max_a q(s,a) - \max_{a'} q(s',a') \leq \max_a \big(q(s,a) - q(s',a)\big) \leq L_{h+1}\, d_{h+1}^{\mathrm{S}}(s,s').$$

Exchanging the roles of s and s' gives

$$\mathrm{Lip}_{h+1}^{\mathrm{S}}(v) \leq L_{h+1}.$$

Applying lemma C.5 with $K = L_{h+1}$ yields

$$\mathrm{Lip}_h(x \mapsto (B_h^\pi v)(x)) \leq \rho_h + \kappa_h L_{h+1} \leq L_h,$$

by (C.6). Restricting to C_h gives assumption J.1. □

C.5. Why Policy Regularity Is Genuinely Needed for Fixed-Policy Closure

The previous propositions show that primitive reward and transition smoothness plus policy regularity imply the fixed-policy closure condition. The next counterexample shows that the policy regularity assumption cannot simply be removed.

Proposition C.9 (Primitive smoothness alone does not imply fixed-policy closure)**.** *There exists a two-stage primitive model satisfying assumption C.1 with full support at both stages and a radius vector L such that assumption 3.1 fails.*

Proof. Take $H = 2$, and let

$$\mathcal{S}_1 = \mathcal{S}_2 = [0,1]$$

with the Euclidean metric. Let the action space be

$$\mathcal{A} = \{0,1\},$$

with the product metrics from Appendix A. Let both support sets be full:

$$C_1 = \mathcal{X}_1, \qquad C_2 = \mathcal{X}_2.$$

Set the reward means identically to zero:

$$r_1 \equiv 0, \qquad r_2 \equiv 0.$$

At stage 1, let the transition be deterministic identity on the state:

$$P_1(\cdot \mid (s,a)) = \delta_s \qquad \forall (s,a) \in \mathcal{X}_1.$$

Then assumption C.1 holds with

$$\rho_1 = \rho_2 = 0, \qquad \kappa_1 = 1,$$

since

$$W_1(\delta_s, \delta_{s'}) = |s - s'| \le d_1((s,a),(s',a')).$$

Now choose a discontinuous target policy at stage 2:

$$\pi_2(1 \mid s) = \mathbb{I}\{s \ge 1/2\}, \qquad \pi_2(0 \mid s) = 1 - \pi_2(1 \mid s).$$

Define

$$q(s,a) := a, \qquad (s,a) \in \mathcal{X}_2.$$

Because the action metric contributes 1 when $a \neq a'$, the function q is 1-Lipschitz on $\mathcal{X}_2$. Hence, if $L_2 \ge 1$, then $q \in \mathrm{Lip}_{L_2}(\mathcal{X}_2)$ and the corresponding target-averaged continuation value

$$v(s) := \mathsf{A}_2^\pi q(s) = \sum_{a \in \mathcal{A}} \pi_2(a \mid s)\, q(s,a) = \mathbb{I}\{s \ge 1/2\}$$

belongs to $\mathcal{V}_2^{\pi,L}$.

For this continuation value,

$$(B_1^\pi v)(s,a) = \int v(s')\, P_1(ds' \mid (s,a)) = v(s) = \mathbb{I}\{s \ge 1/2\}.$$

Since the support at stage 1 is the whole space $\mathcal{X}_1$, the supported Bellman target is discontinuous on C_1 and therefore not L_1-Lipschitz for any finite L_1. Thus assumption 3.1 fails.

The failure is not caused by support holes or by irregular primitive dynamics: the support is full and the transition is perfectly smooth. The obstruction is entirely due to the discontinuity of policy averaging. □

C.6. Summary

The three preceding results show that the abstract closure assumptions used in the main text are not ad hoc.

- Fixed-policy closure follows from primitive reward smoothness, Wasserstein-Lipschitz transitions, and stagewise ℓ_1-Lipschitz regularity of the target policy.
- The true primitive target-policy value sequence then belongs to the Bellman–Lipschitz compatibility class.
- Control closure follows from the same primitive reward/transition smoothness assumptions, but does not require policy regularity.

At the same time, proposition C.9 shows that policy regularity is a genuine structural requirement for the fixed-policy rectangular closure theorem, not a proof artifact.

D. Proof of the Sharp Interval Theorem

This appendix proves the sharp partial-identification theorem via a stronger stagewise statement. The key point is that the Bellman–Whitney recursion defined in the main text provides *a priori* one-sided bounds under support constraints. To upgrade those bounds into exact identified-set endpoints, one must ensure that the recursively generated supported Bellman targets remain admissible for exact L_h-Lipschitz extension at every stage. We make that closure requirement explicit below as the function-level condition under which the sharpness proof is carried out.

D.1. Tail Classes and Recursive Envelopes

For each stage $h \in [H]$, define the stagewise value class

$$\mathcal{V}_h^{\pi,L} := \{\mathsf{A}_h^\pi q : q \in \mathrm{Lip}_{L_h}(\mathcal{X}_h)\}.$$

By convention,

$$\mathcal{V}_{H+1}^{\pi,L} := \{0\}.$$

Assumption D.1 (Rectangular Bellman–Lipschitz closure)**.** For every stage $h \in [H]$ and every $v \in \mathcal{V}_{h+1}^{\pi,L}$, the supported Bellman target

$$c \longmapsto (B_h^\pi v)(c), \qquad c \in C_h,$$

is L_h-Lipschitz with respect to the restricted metric on C_h, i.e.,

$$|(B_h^\pi v)(c) - (B_h^\pi v)(c')| \le L_h\, d_h(c, c') \qquad \forall c, c' \in C_h.$$

Remark D.2 (Why assumption 3.1 is the right closure condition). The Bellman–Whitney recursion is stagewise and rectangular: the feasible continuation class at stage h depends only on the next-stage target-value class $\mathcal{V}_{h+1}^{\pi,L}$ and not on the earlier trajectory. Assumption 3.1 is therefore the exact compatibility condition needed to make the backward recursion *feasible*, rather than merely pointwise valid as a lower/upper bound construction. In particular, once $v \in \mathcal{V}_{h+1}^{\pi,L}$, the supported Bellman target $B_h^\pi v$ has an exact L_h-Lipschitz extension on $\mathcal{X}_h$ by corollary B.4.

For proof purposes, it is convenient to work with tail classes.

Definition D.3 (Tail compatibility class)**.** Fix $h \in [H]$. The tail compatibility class $\mathfrak{C}_{h:H}^{\pi,L}$ consists of all sequences

$$(Q_t, V_t)_{t=h}^{H+1}$$

such that:

1. $V_{H+1} \equiv 0$;
2. for every $t \in \{h, \dots, H\}$, the function $Q_t : \mathcal{X}_t \to \mathbb{R}$ is bounded, Borel measurable, and L_t-Lipschitz;
3. for every $t \in \{h, \dots, H\}$,
$$V_t = \mathsf{A}_t^\pi Q_t;$$
4. for every $t \in \{h, \dots, H\}$,
$$Q_t(x) = (B_t^\pi V_{t+1})(x), \qquad x \in C_t.$$

Thus $\mathfrak{C}_{1:H}^{\pi,L} = \mathfrak{C}_L^\pi$ from Section 2 and definition 2.2. For each $h \in [H]$ and each $s \in \mathcal{S}_h$, define the stagewise identified-value set

$$\mathcal{I}_h^\pi(s) := \Big\{ V_h(s) : (Q_t, V_t)_{t=h}^{H+1} \in \mathfrak{C}_{h:H}^{\pi,L} \Big\}.$$

We now define the Bellman–Whitney recursion used throughout the theorem.

Set

$$\underline{V}_{H+1}^\pi = \overline{V}_{H+1}^\pi \equiv 0. \tag{D.1}$$

For $h = H, H-1, \dots, 1$, define the supported lower and upper Bellman targets

$$g_h^-(c) := (B_h^\pi \underline{V}_{h+1}^\pi)(c), \qquad g_h^+(c) := (B_h^\pi \overline{V}_{h+1}^\pi)(c), \qquad c \in C_h, \tag{D.2}$$

and then define the stagewise Bellman–Whitney envelopes

$$\underline{Q}_h^\pi := \mathcal{W}_{h,L_h}^- g_h^-, \qquad \overline{Q}_h^\pi := \mathcal{W}_{h,L_h}^+ g_h^+, \tag{D.3}$$

together with the induced value functions

$$\underline{V}_h^\pi := \mathsf{A}_h^\pi \underline{Q}_h^\pi, \qquad \overline{V}_h^\pi := \mathsf{A}_h^\pi \overline{Q}_h^\pi. \tag{D.4}$$

D.2. Convexity of the Tail Classes

Lemma D.4 (Convexity of tail compatibility classes)**.** *For every $h \in [H]$, the class $\mathfrak{C}^{\pi,L}_{h:H}$ is convex. Consequently, for every fixed $s \in \mathcal{S}_h$, the scalar set $\mathcal{I}^{\pi}_h(s)$ is a convex subset of $\mathbb{R}$.*

Proof. Fix $h \in [H]$, and let

$$(Q^{(0)}_t, V^{(0)}_t)_{t=h}^{H+1}, \qquad (Q^{(1)}_t, V^{(1)}_t)_{t=h}^{H+1}$$

belong to $\mathfrak{C}^{\pi,L}_{h:H}$. For $\lambda \in [0,1]$, define

$$Q^{(\lambda)}_t := (1-\lambda)Q^{(0)}_t + \lambda Q^{(1)}_t, \qquad V^{(\lambda)}_t := (1-\lambda)V^{(0)}_t + \lambda V^{(1)}_t.$$

Boundedness and Borel measurability are preserved by convex combination. Moreover, if $Q^{(0)}_t$ and $Q^{(1)}_t$ are both L_t-Lipschitz, then so is $Q^{(\lambda)}_t$, because

$$|Q^{(\lambda)}_t(x) - Q^{(\lambda)}_t(y)| \le (1-\lambda)L_t d_t(x,y) + \lambda L_t d_t(x,y) = L_t d_t(x,y).$$

Linearity of A^{π}_t gives

$$V^{(\lambda)}_t = \mathsf{A}^{\pi}_t Q^{(\lambda)}_t.$$

Finally, on C_t,

$$\begin{aligned} Q^{(\lambda)}_t &= (1-\lambda)Q^{(0)}_t + \lambda Q^{(1)}_t \\ &= (1-\lambda)B^{\pi}_t V^{(0)}_{t+1} + \lambda B^{\pi}_t V^{(1)}_{t+1} \\ &= B^{\pi}_t\Big((1-\lambda)V^{(0)}_{t+1} + \lambda V^{(1)}_{t+1}\Big) \\ &= B^{\pi}_t V^{(\lambda)}_{t+1}, \end{aligned}$$

where the third line uses linearity of conditional expectation. Hence $(Q^{(\lambda)}_t, V^{(\lambda)}_t)_{t=h}^{H+1} \in \mathfrak{C}^{\pi,L}_{h:H}$.

The final statement follows because the map

$$(Q_t, V_t)_{t=h}^{H+1} \longmapsto V_h(s)$$

is affine on $\mathfrak{C}^{\pi,L}_{h:H}$. □

D.3. Feasibility of the Recursive Extremal Tails

Lemma D.5 (Feasibility of the recursive lower and upper tails)**.** *Assume assumption 3.1. Then for every $h \in [H]$,*

$$(\underline{Q}^{\pi}_t, \underline{V}^{\pi}_t)_{t=h}^{H+1} \in \mathfrak{C}^{\pi,L}_{h:H}, \qquad (\overline{Q}^{\pi}_t, \overline{V}^{\pi}_t)_{t=h}^{H+1} \in \mathfrak{C}^{\pi,L}_{h:H}.$$

In particular, $\mathfrak{C}^{\pi,L}_{h:H} \neq \varnothing$ for every $h \in [H]$.

Proof. We prove the statement by backward induction on h.

Base case: $h = H$. By (D.1), $\underline{V}^{\pi}_{H+1} = \overline{V}^{\pi}_{H+1} \equiv 0$, so $\underline{V}^{\pi}_{H+1}, \overline{V}^{\pi}_{H+1} \in \mathcal{V}^{\pi,L}_{H+1}$. Therefore assumption 3.1 implies that

$$g^-_H(c) = (B^{\pi}_H \underline{V}^{\pi}_{H+1})(c), \qquad g^+_H(c) = (B^{\pi}_H \overline{V}^{\pi}_{H+1})(c), \qquad c \in C_H,$$

are L_H-Lipschitz on C_H.

By corollary B.4, the envelope functions

$$\underline{Q}^{\pi}_H = \mathcal{W}^-_{H,L_H} g^-_H, \qquad \overline{Q}^{\pi}_H = \mathcal{W}^+_{H,L_H} g^+_H$$

are bounded, Borel measurable, L_H-Lipschitz on $\mathcal{X}_H$, and satisfy the exact interpolation identities

$$\underline{Q}^{\pi}_H(x) = g^-_H(x), \qquad \overline{Q}^{\pi}_H(x) = g^+_H(x), \qquad x \in C_H.$$

By definition, $\underline{V}^{\pi}_H = \mathsf{A}^{\pi}_H \underline{Q}^{\pi}_H$ and $\overline{V}^{\pi}_H = \mathsf{A}^{\pi}_H \overline{Q}^{\pi}_H$. Hence

$$(\underline{Q}^{\pi}_H, \underline{V}^{\pi}_H, V_{H+1} \equiv 0) \in \mathfrak{C}^{\pi,L}_{H:H}, \qquad (\overline{Q}^{\pi}_H, \overline{V}^{\pi}_H, V_{H+1} \equiv 0) \in \mathfrak{C}^{\pi,L}_{H:H}.$$

Induction step. Fix $h \in [H-1]$ and assume the claim holds for stage $h+1$, i.e.,

$$(\underline{Q}_t^\pi, \underline{V}_t^\pi)_{t=h+1}^{H+1} \in \mathfrak{C}_{h+1:H}^{\pi,L}, \qquad (\overline{Q}_t^\pi, \overline{V}_t^\pi)_{t=h+1}^{H+1} \in \mathfrak{C}_{h+1:H}^{\pi,L}.$$

Then $\underline{Q}_{h+1}^\pi$ and $\overline{Q}_{h+1}^\pi$ are respectively L_{h+1}-Lipschitz on $\mathcal{X}_{h+1}$. Therefore, by definition of $\mathcal{V}_{h+1}^{\pi,L}$,

$$\underline{V}_{h+1}^\pi = \mathsf{A}_{h+1}^\pi \underline{Q}_{h+1}^\pi \in \mathcal{V}_{h+1}^{\pi,L}, \qquad \overline{V}_{h+1}^\pi = \mathsf{A}_{h+1}^\pi \overline{Q}_{h+1}^\pi \in \mathcal{V}_{h+1}^{\pi,L}.$$

Applying assumption 3.1 with these two values shows that the supported Bellman targets

$$g_h^-(c) = (B_h^\pi \underline{V}_{h+1}^\pi)(c), \qquad g_h^+(c) = (B_h^\pi \overline{V}_{h+1}^\pi)(c), \qquad c \in C_h,$$

are L_h-Lipschitz on C_h.

Invoking corollary B.4 once more, the functions

$$\underline{Q}_h^\pi = \mathcal{W}_{h,L_h}^- g_h^-, \qquad \overline{Q}_h^\pi = \mathcal{W}_{h,L_h}^+ g_h^+$$

are bounded, Borel measurable, L_h-Lipschitz on $\mathcal{X}_h$, and satisfy the exact support identities

$$\underline{Q}_h^\pi(x) = g_h^-(x), \qquad \overline{Q}_h^\pi(x) = g_h^+(x), \qquad x \in C_h.$$

Defining $\underline{V}_h^\pi = \mathsf{A}_h^\pi \underline{Q}_h^\pi$ and $\overline{V}_h^\pi = \mathsf{A}_h^\pi \overline{Q}_h^\pi$ therefore shows

$$(\underline{Q}_t^\pi, \underline{V}_t^\pi)_{t=h}^{H+1} \in \mathfrak{C}_{h:H}^{\pi,L}, \qquad (\overline{Q}_t^\pi, \overline{V}_t^\pi)_{t=h}^{H+1} \in \mathfrak{C}_{h:H}^{\pi,L}.$$

This completes the induction. □

D.4. Domination of Arbitrary Compatible Tails

Lemma D.6 (Every compatible tail is sandwiched by the Bellman–Whitney tails)**.** *Let $h \in [H]$, and let*

$$(Q_t, V_t)_{t=h}^{H+1} \in \mathfrak{C}_{h:H}^{\pi,L}.$$

Assume assumption 3.1. Then, for every $t \in \{h, \dots, H\}$,

$$\underline{Q}_t^\pi(x) \le Q_t(x) \le \overline{Q}_t^\pi(x) \qquad \forall x \in \mathcal{X}_t, \tag{D.5}$$

and hence

$$\underline{V}_t^\pi(s) \le V_t(s) \le \overline{V}_t^\pi(s) \qquad \forall s \in \mathcal{S}_t. \tag{D.6}$$

Proof. We proceed by backward induction on t.

Base case: $t = H$. Since $(Q_H, V_H, V_{H+1} \equiv 0) \in \mathfrak{C}_{H:H}^{\pi,L}$, we have on C_H that

$$Q_H(x) = (B_H^\pi 0)(x) = g_H^-(x) = g_H^+(x).$$

Moreover, Q_H is L_H-Lipschitz on $\mathcal{X}_H$. Hence Q_H is an exact L_H-Lipschitz extension of the common supported function $g_H^- = g_H^+$. By corollary B.4,

$$\underline{Q}_H^\pi \le Q_H \le \overline{Q}_H^\pi \qquad \text{pointwise on } \mathcal{X}_H.$$

Applying A_H^π to this inequality yields

$$\underline{V}_H^\pi \le V_H \le \overline{V}_H^\pi \qquad \text{pointwise on } \mathcal{S}_H.$$

Induction step. Fix $t \in \{h, \dots, H-1\}$ and suppose the claim is true at stage $t+1$ for the given compatible tail, i.e.,

$$\underline{V}^{\pi}_{t+1} \le V_{t+1} \le \overline{V}^{\pi}_{t+1} \qquad \text{pointwise on } \mathcal{S}_{t+1}.$$

By monotonicity of conditional expectation,

$$(B^{\pi}_t \underline{V}^{\pi}_{t+1})(x) \le (B^{\pi}_t V_{t+1})(x) \le (B^{\pi}_t \overline{V}^{\pi}_{t+1})(x) \qquad \forall x \in C_t.$$

Since (Q_t, V_t) is compatible, $Q_t(x) = (B^{\pi}_t V_{t+1})(x)$ on C_t. Therefore

$$g^-_t(x) \le Q_t(x) \le g^+_t(x) \qquad \forall x \in C_t, \tag{D.7}$$

where $g^{\pm}_t$ are defined in (D.2).

Because Q_t is L_t-Lipschitz on $\mathcal{X}_t$ and satisfies the lower support constraint $Q_t \ge g^-_t$ on C_t, lemma B.3 implies

$$Q_t \ge \mathcal{W}^-_{t,L_t} g^-_t = \underline{Q}^{\pi}_t \qquad \text{pointwise on } \mathcal{X}_t.$$

Similarly, since $Q_t \le g^+_t$ on C_t, lemma B.3 implies

$$Q_t \le \mathcal{W}^+_{t,L_t} g^+_t = \overline{Q}^{\pi}_t \qquad \text{pointwise on } \mathcal{X}_t.$$

This proves (D.5) at stage t. Applying A^{π}_t to the pointwise inequality and using linearity of target-policy averaging gives (D.6) at stage t.

The induction is complete. □

D.5. Stagewise Strengthening of the Sharp Interval Theorem

We now prove a stagewise strengthening from which the main theorem follows immediately.

Theorem D.7 (Stagewise sharp Bellman–Whitney intervals)**.** *Assume assumptions 2.1 and 3.1. Then, for every stage $h \in [H]$ and every state $s \in \mathcal{S}_h$,*

$$\mathcal{I}^{\pi}_h(s) = \Big[\underline{V}^{\pi}_h(s),\, \overline{V}^{\pi}_h(s)\Big]. \tag{D.8}$$

Moreover, the recursive lower and upper tails

$$(\underline{Q}^{\pi}_t, \underline{V}^{\pi}_t)_{t=h}^{H+1}, \qquad (\overline{Q}^{\pi}_t, \overline{V}^{\pi}_t)_{t=h}^{H+1},$$

belong to $\mathfrak{C}^{\pi,L}_{h:H}$ and attain the two endpoints of (D.8).

Proof. Fix $h \in [H]$ and $s \in \mathcal{S}_h$.

Step 1: outer inclusion. Let $(Q_t, V_t)_{t=h}^{H+1} \in \mathfrak{C}^{\pi,L}_{h:H}$ be arbitrary. By lemma D.6,

$$\underline{V}^{\pi}_h(s) \le V_h(s) \le \overline{V}^{\pi}_h(s).$$

Since the compatible tail was arbitrary, this proves

$$\mathcal{I}^{\pi}_h(s) \subseteq \Big[\underline{V}^{\pi}_h(s),\, \overline{V}^{\pi}_h(s)\Big].$$

Step 2: endpoint attainability. By lemma D.5, both recursive tails

$$(\underline{Q}^{\pi}_t, \underline{V}^{\pi}_t)_{t=h}^{H+1}, \qquad (\overline{Q}^{\pi}_t, \overline{V}^{\pi}_t)_{t=h}^{H+1},$$

belong to $\mathfrak{C}^{\pi,L}_{h:H}$. Therefore

$$\underline{V}^{\pi}_h(s) \in \mathcal{I}^{\pi}_h(s), \qquad \overline{V}^{\pi}_h(s) \in \mathcal{I}^{\pi}_h(s).$$

Step 3: interval filling by convexity. By lemma D.4, the class $\mathfrak{C}^{\pi,L}_{h:H}$ is convex, hence the scalar image $\mathcal{I}^{\pi}_{h}(s)$ is a convex subset of $\mathbb{R}$. A convex subset of $\mathbb{R}$ containing both endpoints $\underline{V}^{\pi}_{h}(s)$ and $\overline{V}^{\pi}_{h}(s)$ must contain the whole closed interval between them. Thus

$$\left[\underline{V}^{\pi}_{h}(s),\, \overline{V}^{\pi}_{h}(s)\right] \subseteq \mathcal{I}^{\pi}_{h}(s).$$

Combining the last display with the outer inclusion from Step 1 proves (D.8). The endpoint-attainment statement has already been established in Step 2. □

Corollary D.8 (Main sharp interval theorem)**.** *Assume assumptions 2.1 and 3.1. Then*

$$\mathcal{I}^{\pi} = \left[\underline{V}^{\pi}_{1}(s_1),\, \overline{V}^{\pi}_{1}(s_1)\right].$$

In particular, this yields the main theorem stated in the text.

Proof. By definition,

$$\mathcal{I}^{\pi} = \mathcal{I}^{\pi}_{1}(s_1).$$

Applying theorem D.7 with $h = 1$ and $s = s_1$ gives the claim. □

D.6. Interpretation of the Proof Structure

The proof decomposes into three logically distinct ingredients.

1. **Exact feasibility of the recursive extremes.** Under assumption 3.1, the recursive lower and upper Bellman targets remain admissible for exact Lipschitz extension at every stage. This is what upgrades the Bellman–Whitney formulas from one-sided envelope bounds to genuine feasible tails.
2. **One-sided domination of every compatible tail.** Even without exact interpolation, the extremal characterization from Appendix B forces any compatible Q_t to dominate the lower Bellman–Whitney envelope and to be dominated by the upper one.
3. **Convexity of the feasible tail class.** Once the two endpoints are shown to be feasible, convexity fills the entire interval and converts endpoint bounds into a sharp identified-set result.

This separation is conceptually useful: the envelope lemmas provide the stagewise geometry, the rectangular closure assumption provides recursive feasibility, and convexity turns extremal values into a full scalar identified interval.

E. Dual Characterization and Strong Duality

This appendix gives an exact dual characterization of the Bellman–Whitney endpoints. The dual objects are *one-sided Bellman relaxations*: the upper value endpoint is characterized by a maximal Bellman-consistent Lipschitz *minorant* program, while the lower value endpoint is characterized by a minimal Bellman-consistent Lipschitz *majorant* program. Because the decision variables are entire function sequences, these are infinite-dimensional linear programs over Lipschitz cones. The content of the theorem below is that there is nevertheless no gap: the Bellman–Whitney recursion solves both the primal equality-constrained problem and the appropriate one-sided dual relaxation exactly.

E.1. Stagewise Primal Programs

Fix a stage $h \in [H]$ and a state $s \in \mathcal{S}_h$. Recall from Appendix D that the tail compatibility class $\mathfrak{C}^{\pi,L}_{h:H}$ consists of all Bellman–Lipschitz-compatible continuations from stage h onward. Define the upper and lower primal values

$$\mathsf{P}^{+}_{h}(s) := \sup\left\{V_h(s) : (Q_t, V_t)_{t=h}^{H+1} \in \mathfrak{C}^{\pi,L}_{h:H}\right\}, \tag{E.1}$$

and

$$\mathsf{P}^{-}_{h}(s) := \inf\left\{V_h(s) : (Q_t, V_t)_{t=h}^{H+1} \in \mathfrak{C}^{\pi,L}_{h:H}\right\}. \tag{E.2}$$

Equivalently, these are infinite-dimensional linear optimization problems over the affine equality-constrained class $\mathfrak{C}^{\pi,L}_{h:H}$.

E.2. One-Sided Dual Relaxations

The upper dual keeps the same Lipschitz geometry but relaxes Bellman equality to a Bellman *minorant* constraint on the behavior support.

Definition E.1 (Upper dual class)**.** Fix $h \in [H]$. The upper dual class $\mathfrak{U}_h^{\pi,L}$ consists of all sequences $U = (U_t)_{t=h}^H$ such that:

1. for every $t \in \{h, \dots, H\}$, the function $U_t : \mathcal{X}_t \to \mathbb{R}$ is bounded, Borel measurable, and L_t-Lipschitz;
2. if
$$V_t^U := \mathsf{A}_t^\pi U_t, \qquad t \in \{h, \dots, H\},$$
and $V_{H+1}^U \equiv 0$, then the one-sided Bellman constraint
$$U_t(c) \le (B_t^\pi V_{t+1}^U)(c), \qquad c \in C_t, \tag{E.3}$$
holds for every $t \in \{h, \dots, H\}$.

The lower dual is the symmetric Bellman *majorant* relaxation.

Definition E.2 (Lower dual class)**.** Fix $h \in [H]$. The lower dual class $\mathfrak{L}_h^{\pi,L}$ consists of all sequences $L = (L_t)_{t=h}^H$ such that:

1. for every $t \in \{h, \dots, H\}$, the function $L_t : \mathcal{X}_t \to \mathbb{R}$ is bounded, Borel measurable, and L_t-Lipschitz;
2. if
$$V_t^L := \mathsf{A}_t^\pi L_t, \qquad t \in \{h, \dots, H\},$$
and $V_{H+1}^L \equiv 0$, then the one-sided Bellman constraint
$$L_t(c) \ge (B_t^\pi V_{t+1}^L)(c), \qquad c \in C_t, \tag{E.4}$$
holds for every $t \in \{h, \dots, H\}$.

We then define the upper and lower dual objective values by

$$\mathsf{D}_h^+(s) := \sup\Big\{V_h^U(s) : U \in \mathfrak{U}_h^{\pi,L}\Big\}, \tag{E.5}$$

and

$$\mathsf{D}_h^-(s) := \inf\Big\{V_h^L(s) : L \in \mathfrak{L}_h^{\pi,L}\Big\}. \tag{E.6}$$

Remark E.3 (Why these are the right duals). A primal feasible tail satisfies Bellman equality on the support and is therefore feasible for both one-sided relaxations. The upper dual enlarges the primal feasible set by replacing equality with a Bellman minorant constraint, and the lower dual enlarges it by replacing equality with a Bellman majorant constraint. The Bellman–Whitney theorem from Appendix D shows that these relaxed programs are nevertheless exact.

E.3. Weak Duality

Lemma E.4 (Weak duality)**.** *Assume assumptions 2.1 and 3.1. Then, for every $h \in [H]$ and every $s \in \mathcal{S}_h$,*

$$\mathsf{D}_h^-(s) \le \mathsf{P}_h^-(s) \le \mathsf{P}_h^+(s) \le \mathsf{D}_h^+(s). \tag{E.7}$$

Proof. Fix $h \in [H]$. Let $(Q_t, V_t)_{t=h}^{H+1} \in \mathfrak{C}_{h:H}^{\pi,L}$ be arbitrary. Define $U_t := Q_t$ for $t = h, \dots, H$. Since the tail is primal feasible,

$$Q_t(c) = (B_t^\pi V_{t+1})(c), \qquad c \in C_t,$$

and $V_t = \mathsf{A}_t^\pi Q_t$. Therefore the sequence $U = (U_t)_{t=h}^H$ belongs to $\mathfrak{U}_h^{\pi,L}$, because the upper dual constraints (E.3) hold with equality. Hence

$$V_h(s) = V_h^U(s) \le \mathsf{D}_h^+(s).$$

Taking the supremum over all primal feasible tails gives

$$\mathsf{P}_h^+(s) \le \mathsf{D}_h^+(s).$$

Applying the same argument to the lower dual class yields $L = (Q_t)_{t=h}^H \in \mathfrak{L}_h^{\pi,L}$, because (E.4) also holds with equality. Therefore

$$\mathsf{D}_h^-(s) \le V_h(s).$$

Taking the infimum over all primal feasible tails gives

$$\mathsf{D}_h^-(s) \le \mathsf{P}_h^-(s).$$

Finally, the middle inequality $\mathsf{P}_h^-(s) \le \mathsf{P}_h^+(s)$ is immediate from the definitions (E.1)–(E.2). □

E.4. Pointwise Extremality of the Dual Programs

The next two lemmas show that the Bellman–Whitney tails are not merely optimal for the dual objectives at the initial state: they are pointwise extremal over the entire dual feasible classes.

Lemma E.5 (Pointwise maximality of the upper dual)**.** *Assume assumptions 2.1 and 3.1. Fix $h \in [H]$. Then:*

1. *the recursive upper Bellman–Whitney tail $(\overline{Q}_t^\pi, \overline{V}_t^\pi)_{t=h}^{H+1}$ from Section D is feasible for $\mathfrak{U}_h^{\pi,L}$;*
2. *for every $U \in \mathfrak{U}_h^{\pi,L}$ and every $t \in \{h, \dots, H\}$,*

$$U_t(x) \le \overline{Q}_t^\pi(x) \qquad \forall x \in \mathcal{X}_t, \tag{E.8}$$

and consequently

$$V_t^U(s) \le \overline{V}_t^\pi(s) \qquad \forall s \in \mathcal{S}_t. \tag{E.9}$$

In particular, for every $t \in \{h, \dots, H\}$ and every $x \in \mathcal{X}_t$,

$$\overline{Q}_t^\pi(x) = \sup\Big\{U_t(x) : U \in \mathfrak{U}_h^{\pi,L}\Big\},$$

and for every $s \in \mathcal{S}_t$,

$$\overline{V}_t^\pi(s) = \sup\Big\{V_t^U(s) : U \in \mathfrak{U}_h^{\pi,L}\Big\}.$$

Proof. Feasibility of the recursive upper tail follows immediately from lemma D.5: the equality-constrained tail $(\overline{Q}_t^\pi, \overline{V}_t^\pi)_{t=h}^{H+1} \in \mathfrak{C}_{h:H}^{\pi,L}$ satisfies the upper dual constraints with equality and is therefore in $\mathfrak{U}_h^{\pi,L}$.

We prove (E.8) by backward induction on t.

Base case: $t = H$. Let $U \in \mathfrak{U}_h^{\pi,L}$. Since $V_{H+1}^U \equiv 0$, the upper dual constraint gives

$$U_H(c) \le (B_H^\pi 0)(c) = g_H^+(c), \qquad c \in C_H,$$

where g_H^+ is the supported upper Bellman target defined in Section D. Because U_H is L_H-Lipschitz on $\mathcal{X}_H$, the extremal characterization lemma B.3 implies

$$U_H \le \mathcal{W}_{H,L_H}^+ g_H^+ = \overline{Q}_H^\pi \qquad \text{pointwise on } \mathcal{X}_H.$$

Applying A_H^π yields

$$V_H^U \le \overline{V}_H^\pi.$$

Induction step. Fix $t \in \{h, \dots, H-1\}$ and suppose the conclusion has been proved at stage $t+1$, i.e.,

$$V_{t+1}^U \le \overline{V}_{t+1}^\pi \qquad \text{pointwise on } \mathcal{S}_{t+1}.$$

By monotonicity of conditional expectation,

$$(B_t^\pi V_{t+1}^U)(c) \le (B_t^\pi \overline{V}_{t+1}^\pi)(c) = g_t^+(c), \qquad c \in C_t.$$

Since U is upper dual feasible,

$$U_t(c) \le (B_t^\pi V_{t+1}^U)(c) \le g_t^+(c), \qquad c \in C_t.$$

Because U_t is L_t-Lipschitz on $\mathcal{X}_t$, lemma B.3 yields

$$U_t \le \mathcal{W}_{t,L_t}^+ g_t^+ = \overline{Q}_t^\pi \qquad \text{pointwise on } \mathcal{X}_t.$$

Applying A_t^π gives

$$V_t^U \le \overline{V}_t^\pi.$$

This proves the induction.

Since the recursive upper tail itself is feasible and attains the right-hand sides, the final supremum identities follow immediately. □

Lemma E.6 (Pointwise minimality of the lower dual)**.** *Assume assumptions 2.1 and 3.1. Fix $h \in [H]$. Then:*

1. *the recursive lower Bellman–Whitney tail $(\underline{Q}_t^\pi, \underline{V}_t^\pi)_{t=h}^{H+1}$ from Section D is feasible for $\mathfrak{L}_h^{\pi,L}$;*
2. *for every $L \in \mathfrak{L}_h^{\pi,L}$ and every $t \in \{h, \dots, H\}$,*

$$\underline{Q}_t^\pi(x) \le L_t(x) \qquad \forall x \in \mathcal{X}_t, \tag{E.10}$$

and consequently

$$\underline{V}_t^\pi(s) \le V_t^L(s) \qquad \forall s \in \mathcal{S}_t. \tag{E.11}$$

In particular, for every $t \in \{h, \dots, H\}$ and every $x \in \mathcal{X}_t$,

$$\underline{Q}_t^\pi(x) = \inf\Big\{L_t(x) : L \in \mathfrak{L}_h^{\pi,L}\Big\},$$

and for every $s \in \mathcal{S}_t$,

$$\underline{V}_t^\pi(s) = \inf\Big\{V_t^L(s) : L \in \mathfrak{L}_h^{\pi,L}\Big\}.$$

Proof. Feasibility of the recursive lower tail again follows from lemma D.5, since the equality-constrained tail belongs to $\mathfrak{C}_{h:H}^{\pi,L}$ and therefore satisfies the lower dual constraints with equality.

We prove (E.10) by backward induction on t.

Base case: $t = H$. Let $L \in \mathfrak{L}_h^{\pi,L}$. Since $V_{H+1}^L \equiv 0$, the lower dual constraint gives

$$L_H(c) \ge (B_H^\pi 0)(c) = g_H^-(c), \qquad c \in C_H,$$

where g_H^- is the supported lower Bellman target from Section D. Because L_H is L_H-Lipschitz on $\mathcal{X}_H$, the extremal characterization lemma B.3 yields

$$\mathcal{W}_{H,L_H}^- g_H^- = \underline{Q}_H^\pi \le L_H \qquad \text{pointwise on } \mathcal{X}_H.$$

Applying A_H^π gives

$$\underline{V}_H^\pi \le V_H^L.$$

Induction step. Fix $t \in \{h, \dots, H-1\}$ and suppose

$$\underline{V}^{\pi}_{t+1} \le V^{L}_{t+1} \qquad \text{pointwise on } \mathcal{S}_{t+1}.$$

By monotonicity of conditional expectation,

$$g^{-}_{t}(c) = (B^{\pi}_{t}\underline{V}^{\pi}_{t+1})(c) \le (B^{\pi}_{t}V^{L}_{t+1})(c), \qquad c \in C_t.$$

Since L is lower dual feasible,

$$L_t(c) \ge (B^{\pi}_{t}V^{L}_{t+1})(c) \ge g^{-}_{t}(c), \qquad c \in C_t.$$

Because L_t is L_t-Lipschitz on $\mathcal{X}_t$, lemma B.3 implies

$$\underline{Q}^{\pi}_{t} = \mathcal{W}^{-}_{t,L_t} g^{-}_{t} \le L_t \qquad \text{pointwise on } \mathcal{X}_t.$$

Applying A^{π}_{t} yields

$$\underline{V}^{\pi}_{t} \le V^{L}_{t}.$$

This completes the induction.

The infimum identities follow because the recursive lower tail is itself dual-feasible and attains the left-hand sides. □

E.5. No-Gap Duality

We can now identify the primal and dual values exactly.

Theorem E.7 (Strong duality for the Bellman–Whitney programs)**.** *Assume assumptions 2.1 and 3.1. Then for every stage $h \in [H]$ and every state $s \in \mathcal{S}_h$,*

$$\mathsf{D}^{-}_{h}(s) = \mathsf{P}^{-}_{h}(s) = \underline{V}^{\pi}_{h}(s), \tag{E.12}$$

and

$$\mathsf{P}^{+}_{h}(s) = \mathsf{D}^{+}_{h}(s) = \overline{V}^{\pi}_{h}(s). \tag{E.13}$$

Moreover, the recursive Bellman–Whitney tails are simultaneously primal-feasible and dual-optimal:

$$(\underline{Q}^{\pi}_{t}, \underline{V}^{\pi}_{t})_{t=h}^{H+1} \quad \textit{solves both } \mathsf{P}^{-}_{h}(s) \textit{ and } \mathsf{D}^{-}_{h}(s),$$

$$(\overline{Q}^{\pi}_{t}, \overline{V}^{\pi}_{t})_{t=h}^{H+1} \quad \textit{solves both } \mathsf{P}^{+}_{h}(s) \textit{ and } \mathsf{D}^{+}_{h}(s).$$

Proof. Fix $h \in [H]$ and $s \in \mathcal{S}_h$.

Upper value. By weak duality (lemma E.4),

$$\mathsf{P}^{+}_{h}(s) \le \mathsf{D}^{+}_{h}(s).$$

By lemma E.5, every upper dual feasible sequence U satisfies

$$V^{U}_{h}(s) \le \overline{V}^{\pi}_{h}(s),$$

and the recursive upper Bellman–Whitney tail attains equality. Hence

$$\mathsf{D}^{+}_{h}(s) = \overline{V}^{\pi}_{h}(s).$$

On the other hand, lemma D.5 shows that the recursive upper tail is primal-feasible, so

$$\mathsf{P}^{+}_{h}(s) \ge \overline{V}^{\pi}_{h}(s).$$

Combining the last three displays yields

$$\mathsf{P}^{+}_{h}(s) = \mathsf{D}^{+}_{h}(s) = \overline{V}^{\pi}_{h}(s).$$

Lower value. Again by weak duality,

$$\mathsf{D}_h^-(s) \le \mathsf{P}_h^-(s).$$

By lemma E.6, every lower dual feasible sequence L satisfies

$$\underline{V}_h^\pi(s) \le V_h^L(s),$$

and the recursive lower Bellman–Whitney tail attains equality. Therefore

$$\mathsf{D}_h^-(s) = \underline{V}_h^\pi(s).$$

Since the recursive lower tail is primal-feasible by lemma D.5, we also have

$$\mathsf{P}_h^-(s) \le \underline{V}_h^\pi(s).$$

Combining the last three displays yields

$$\mathsf{D}_h^-(s) = \mathsf{P}_h^-(s) = \underline{V}_h^\pi(s).$$

The optimality statements for the recursive lower and upper tails have already been established in the preceding argument. □

Corollary E.8 (No-gap dual characterization of the main identified interval)**.** *Assume assumptions 2.1 and 3.1. Then*

$$\mathcal{I}^\pi = \left[\mathsf{D}_1^-(s_1),\, \mathsf{D}_1^+(s_1)\right] = \left[\underline{V}_1^\pi(s_1),\, \overline{V}_1^\pi(s_1)\right].$$

Thus the Bellman–Whitney interval is simultaneously:

1. *the sharp primal identified interval over the equality-constrained Bellman–Lipschitz class;*
2. *the value interval generated by the no-gap upper minorant and lower majorant dual relaxations.*

Proof. The first equality follows from the theorem with $h = 1$, together with the definition of $\mathcal{I}^\pi$ as the primal tail value set at stage 1. The second equality is theorem 3.2. □

E.6. Interpretation

The upper and lower dual programs admit a simple geometric interpretation. The upper dual searches over all Lipschitz tails that are Bellman-consistent *from below* on the observed support, and extracts the largest achievable target value. The lower dual does the symmetric search over all Lipschitz tails that are Bellman-consistent *from above*. Theorem 4.1 shows that neither relaxation leaves slack: the Bellman–Whitney recursion already computes the exact extremal points of both relaxed classes and hence the exact sharp identified interval.

F. Exact $H = 1$ Reduction

This appendix proves that the sequential Bellman–Whitney theory collapses exactly to the classical one-step smooth no-overlap problem when $H = 1$. The reduction is not merely heuristic: the primal feasibility class, the one-sided dual relaxations, and the sharp identified interval all simplify to the usual McShane–Whitney extremal formulas on the single observed support set. In the deterministic-target case, this yields exactly the sharp interval of Khan et al. (2024).

Throughout this appendix, assume $H = 1$. Then $V_2 \equiv 0$, and the only supported Bellman target is the one-step conditional mean

$$m_1(c) := (B_1^\pi 0)(c) = \mathbb{E}[R_1 \mid X_1 = c], \qquad c \in C_1. \tag{F.1}$$

Under assumption 3.1, the function m_1 is L_1-Lipschitz on the restricted metric space C_1.

F.1. Collapse of the Primal Compatibility Class

Proposition F.1 (One-step collapse of the primal class)**.** *Assume assumptions 2.1 and 3.1 and $H = 1$. Then the Bellman–Lipschitz compatibility class from definition 2.2 reduces to*

$$\mathfrak{C}_L^\pi = \big\{(Q_1, V_1, V_2 \equiv 0) : Q_1 \in \mathrm{Lip}_{L_1}(\mathcal{X}_1),\ Q_1(c) = m_1(c)\ \forall c \in C_1,\ V_1 = \mathsf{A}_1^\pi Q_1\big\}.$$

Moreover, the Bellman–Whitney recursion from Equations (3.1) and (3.2) becomes

$$\underline{Q}_1^\pi = \mathcal{W}_{1,L_1}^- m_1, \qquad \overline{Q}_1^\pi = \mathcal{W}_{1,L_1}^+ m_1, \qquad \underline{V}_1^\pi = \mathsf{A}_1^\pi \underline{Q}_1^\pi, \qquad \overline{V}_1^\pi = \mathsf{A}_1^\pi \overline{Q}_1^\pi. \tag{F.2}$$

Proof. When $H = 1$, the compatibility definition imposes the terminal condition $V_2 \equiv 0$ and the support Bellman relation

$$Q_1(c) = (B_1^\pi V_2)(c) = (B_1^\pi 0)(c) = m_1(c), \qquad c \in C_1.$$

All remaining conditions in definition 2.2 are exactly: Q_1 is bounded, Borel measurable, and L_1-Lipschitz on $\mathcal{X}_1$, and $V_1 = \mathsf{A}_1^\pi Q_1$. This proves the stated representation of $\mathfrak{C}_L^\pi$.

The recursive formulas follow immediately from Equations (3.1) and (3.2). Indeed, since $\underline{V}_2^\pi = \overline{V}_2^\pi \equiv 0$,

$$g_1^-(c) = (B_1^\pi \underline{V}_2^\pi)(c) = m_1(c), \qquad g_1^+(c) = (B_1^\pi \overline{V}_2^\pi)(c) = m_1(c),$$

so the lower and upper Bellman–Whitney envelopes coincide with the lower and upper Whitney extensions of the same supported function m_1, which yields (F.2). □

F.2. Pointwise Sharp Identification in the One-Step Case

The next proposition shows that, when $H = 1$, the Bellman–Whitney formulas are not only sharp for the policy value at the initial state, but also sharp pointwise for the entire state–action value function.

Proposition F.2 (Exact one-step pointwise intervals)**.** *Assume assumptions 2.1 and 3.1 and $H = 1$. Then:*

1. *for every $x \in \mathcal{X}_1$,*
$$\{Q_1(x) : (Q_1, V_1, V_2 \equiv 0) \in \mathfrak{C}_L^\pi\} = \Big[\underline{Q}_1^\pi(x),\ \overline{Q}_1^\pi(x)\Big]; \tag{F.3}$$
2. *for every $s \in \mathcal{S}_1$,*
$$\{V_1(s) : (Q_1, V_1, V_2 \equiv 0) \in \mathfrak{C}_L^\pi\} = \Big[\underline{V}_1^\pi(s),\ \overline{V}_1^\pi(s)\Big]. \tag{F.4}$$

Proof. We first prove (F.3). Let $(Q_1, V_1, V_2 \equiv 0) \in \mathfrak{C}_L^\pi$ be arbitrary. By proposition F.1, the function Q_1 is L_1-Lipschitz on $\mathcal{X}_1$ and satisfies

$$Q_1(c) = m_1(c), \qquad c \in C_1.$$

Since m_1 is L_1-Lipschitz on C_1 by assumption 3.1, corollary B.4 implies

$$\underline{Q}_1^\pi = \mathcal{W}_{1,L_1}^- m_1 \le Q_1 \le \mathcal{W}_{1,L_1}^+ m_1 = \overline{Q}_1^\pi \qquad \text{pointwise on } \mathcal{X}_1.$$

Hence every feasible point value $Q_1(x)$ lies in $[\underline{Q}_1^\pi(x), \overline{Q}_1^\pi(x)]$.

To prove attainability of the entire interval, note first that $\underline{Q}_1^\pi$ and $\overline{Q}_1^\pi$ are themselves feasible by corollary B.4: they are L_1-Lipschitz on $\mathcal{X}_1$ and agree with m_1 on C_1. Therefore they belong to the primal class in proposition F.1. For any $\lambda \in [0, 1]$, define

$$Q_1^{(\lambda)} := (1 - \lambda)\underline{Q}_1^\pi + \lambda \overline{Q}_1^\pi.$$

Because convex combinations preserve boundedness, Borel measurability, and L_1-Lipschitzness, and because both envelopes agree with m_1 on C_1, the function $Q_1^{(\lambda)}$ is also primal feasible. Moreover, for every fixed $x \in \mathcal{X}_1$,

$$Q_1^{(\lambda)}(x) = (1 - \lambda)\underline{Q}_1^\pi(x) + \lambda \overline{Q}_1^\pi(x).$$

Thus every scalar value in $[\underline{Q}_1^\pi(x), \overline{Q}_1^\pi(x)]$ is achieved by some feasible $Q_1^{(\lambda)}$, which proves (F.3).

We next prove (F.4). Since every feasible V_1 is of the form $V_1 = \mathsf{A}_1^\pi Q_1$ for a feasible Q_1, the pointwise bounds from (F.3) imply

$$\underline{V}_1^\pi(s) = \sum_{a\in\mathcal{A}} \pi_1(a \mid s)\, \underline{Q}_1^\pi(s,a) \le V_1(s) \le \sum_{a\in\mathcal{A}} \pi_1(a \mid s)\, \overline{Q}_1^\pi(s,a) = \overline{V}_1^\pi(s).$$

Conversely, for each $\lambda \in [0,1]$, the feasible convex combination $Q_1^{(\lambda)}$ above induces the feasible value function

$$V_1^{(\lambda)} = \mathsf{A}_1^\pi Q_1^{(\lambda)} = (1-\lambda)\underline{V}_1^\pi + \lambda \overline{V}_1^\pi.$$

Hence every scalar value in $[\underline{V}_1^\pi(s), \overline{V}_1^\pi(s)]$ is attained at the state s, proving (F.4). □

F.3. Collapse of the Dual Programs

The one-step dual programs also admit an explicit closed form.

Proposition F.3 (One-step dual collapse)**.** *Assume assumptions 2.1 and 3.1 and $H = 1$. Then, for every $s \in \mathcal{S}_1$,*

$$\mathsf{D}_1^+(s) = \sup\left\{ \sum_{a\in\mathcal{A}} \pi_1(a \mid s)\, u(s,a) : u \in \mathrm{Lip}_{L_1}(\mathcal{X}_1),\ u(c) \le m_1(c)\ \forall c \in C_1 \right\},$$

and

$$\mathsf{D}_1^-(s) = \inf\left\{ \sum_{a\in\mathcal{A}} \pi_1(a \mid s)\, \ell(s,a) : \ell \in \mathrm{Lip}_{L_1}(\mathcal{X}_1),\ \ell(c) \ge m_1(c)\ \forall c \in C_1 \right\}.$$

Moreover,

$$\mathsf{D}_1^+(s) = \overline{V}_1^\pi(s), \qquad \mathsf{D}_1^-(s) = \underline{V}_1^\pi(s).$$

Proof. When $H = 1$, the upper dual feasibility condition from definition E.1 becomes

$$U_1(c) \le (B_1^\pi V_2^U)(c) = (B_1^\pi 0)(c) = m_1(c), \qquad c \in C_1,$$

with U_1 required to be L_1-Lipschitz on $\mathcal{X}_1$ and $V_1^U = \mathsf{A}_1^\pi U_1$. This proves the first display. The lower dual representation is identical with the inequality reversed.

We now identify the optimum values. Since $\overline{Q}_1^\pi$ is the pointwise largest L_1-Lipschitz function dominated by m_1 on C_1 by lemma B.3, every upper dual feasible u satisfies $u \le \overline{Q}_1^\pi$ pointwise on $\mathcal{X}_1$. Therefore

$$\sum_{a\in\mathcal{A}} \pi_1(a \mid s)\, u(s,a) \le \sum_{a\in\mathcal{A}} \pi_1(a \mid s)\, \overline{Q}_1^\pi(s,a) = \overline{V}_1^\pi(s).$$

Since $\overline{Q}_1^\pi$ itself is upper dual feasible, equality is attained and $\mathsf{D}_1^+(s) = \overline{V}_1^\pi(s)$.

The lower dual identity is symmetric. By lemma B.3, $\underline{Q}_1^\pi$ is the pointwise smallest L_1-Lipschitz function that dominates m_1 on the support. Hence every lower dual feasible ℓ satisfies $\underline{Q}_1^\pi \le \ell$, which implies

$$\underline{V}_1^\pi(s) = \sum_{a\in\mathcal{A}} \pi_1(a \mid s)\, \underline{Q}_1^\pi(s,a) \le \sum_{a\in\mathcal{A}} \pi_1(a \mid s)\, \ell(s,a).$$

Because $\underline{Q}_1^\pi$ is itself lower dual feasible, equality is attained and $\mathsf{D}_1^-(s) = \underline{V}_1^\pi(s)$. □

F.4. Proof of the Main-Text One-Step Corollary

We now prove the exact deterministic-target formula stated in corollary 3.3.

Corollary F.4 (Deterministic one-step target)**.** *Assume assumptions 2.1 and 3.1 and $H = 1$. Suppose the target policy is deterministic at the evaluation state, i.e.,*

$$\pi_1(\cdot \mid s_1) = \delta_{a^\star}$$

for some $a^\star \in \mathcal{A}$, and let

$$x^\star := (s_1, a^\star) \in \mathcal{X}_1.$$

Then

$$\mathcal{I}^\pi = \left[\sup_{c \in C_1} \{m_1(c) - L_1 d_1(x^\star, c)\},\ \inf_{c \in C_1} \{m_1(c) + L_1 d_1(x^\star, c)\}\right]. \tag{F.5}$$

Consequently, (F.5) is exactly the sharp smoothness-based no-overlap interval for one-step evaluation derived by Khan et al. (2024), after identifying $x^\star$ as the target state–action point, C_1 as the observed support, and m_1 as the supported conditional mean.

Proof. By proposition F.2, the identified set is

$$\mathcal{I}^\pi = \left[\underline{V}_1^\pi(s_1),\ \overline{V}_1^\pi(s_1)\right].$$

Because the target policy is deterministic at s_1,

$$\underline{V}_1^\pi(s_1) = \underline{Q}_1^\pi(s_1, a^\star) = \underline{Q}_1^\pi(x^\star), \qquad \overline{V}_1^\pi(s_1) = \overline{Q}_1^\pi(s_1, a^\star) = \overline{Q}_1^\pi(x^\star).$$

Using the definitions of the lower and upper Bellman–Whitney envelopes in the one-step case,

$$\underline{Q}_1^\pi(x^\star) = \mathcal{W}^-_{1,L_1} m_1(x^\star) = \sup_{c \in C_1} \{m_1(c) - L_1 d_1(x^\star, c)\},$$

and

$$\overline{Q}_1^\pi(x^\star) = \mathcal{W}^+_{1,L_1} m_1(x^\star) = \inf_{c \in C_1} \{m_1(c) + L_1 d_1(x^\star, c)\}.$$

Substituting these identities into the previous display proves (F.5). The final statement is immediate from the explicit formula. □

Proof of corollary 3.3. This is exactly corollary F.4. □

F.5. Interpretation of the Reduction

The preceding results show that the sequential theory is a genuine extension of the one-step sharp smoothness framework rather than a merely analogous construction. When $H = 1$:

1. the primal feasibility class reduces to the class of all L_1-Lipschitz extensions of the supported conditional mean m_1;
2. the Bellman–Whitney recursion reduces to the classical lower and upper McShane–Whitney envelopes of that supported function;
3. the upper and lower dual programs reduce to one-sided minorant and majorant extension programs;
4. the deterministic-target value interval is exactly the one-step sharp smooth no-overlap interval.

Thus the sequential Bellman–Whitney theory recovers the correct base case without loss or approximation.

G. Geometry of Dynamic Support Holes

This appendix isolates the geometric content of the Bellman–Whitney interval. The central quantity is the stagewise width

$$w_h^\pi(x) := \overline{Q}_h^\pi(x) - \underline{Q}_h^\pi(x), \qquad x \in \mathcal{X}_h,$$

which measures the irreducible ambiguity of the state–action value at the point x under the Bellman–Lipschitz class. The corresponding stagewise value width is

$$\Delta_h(s) := \overline{V}_h^\pi(s) - \underline{V}_h^\pi(s), \qquad s \in \mathcal{S}_h, \tag{G.1}$$

with the terminal convention

$$\Delta_{H+1} \equiv 0.$$

The purpose of this appendix is twofold. First, we prove an exact one-step upper bound showing that the stagewise width at x is controlled by the distance from x to the behavior support plus the propagated continuation ambiguity on the support. Second, we prove that this bound is sharp by constructing an explicit least-favorable family on which it is attained with equality at every stage.

G.1. Width Identities and the Supported Continuation Operator

For every $h \in [H]$, define the supported continuation operator

$$(\mathsf{T}_h f)(c) := \mathbb{E}[f(S_{h+1}) \mid X_h = c], \qquad c \in C_h,$$

for bounded Borel $f : \mathcal{S}_{h+1} \to \mathbb{R}$. Thus the supported Bellman operator can be decomposed as

$$(B_h^\pi f)(c) = r_h(c) + (\mathsf{T}_h f)(c), \qquad c \in C_h,$$

where

$$r_h(c) := \mathbb{E}[R_h \mid X_h = c].$$

For the geometry of widths, the reward term cancels, and only the continuation operator matters.

The next identity expresses stagewise value ambiguity as target-policy averaging of state–action ambiguity.

Lemma G.1 (Value width identity)**.** *For every $h \in [H]$ and every $s \in \mathcal{S}_h$,*

$$\Delta_h(s) = \sum_{a \in \mathcal{A}} \pi_h(a \mid s)\, w_h^\pi(s,a) = (\mathsf{A}_h^\pi w_h^\pi)(s). \tag{G.2}$$

Proof. By definition,

$$\Delta_h(s) = \overline{V}_h^\pi(s) - \underline{V}_h^\pi(s) = \mathsf{A}_h^\pi \overline{Q}_h^\pi(s) - \mathsf{A}_h^\pi \underline{Q}_h^\pi(s).$$

Since A_h^π is linear,

$$\Delta_h(s) = \mathsf{A}_h^\pi\Big(\overline{Q}_h^\pi - \underline{Q}_h^\pi\Big)(s) = \mathsf{A}_h^\pi w_h^\pi(s),$$

which is exactly (G.2). □

The next lemma is the basic support identity behind the entire geometry argument.

Lemma G.2 (Support width identity)**.** *Assume assumptions 2.1 and 3.1. Then for every $h \in [H]$ and every $c \in C_h$,*

$$w_h^\pi(c) = (\mathsf{T}_h \Delta_{h+1})(c). \tag{G.3}$$

Proof. By exact interpolation on the support, proved in Appendix D,

$$\underline{Q}_h^\pi(c) = g_h^-(c), \qquad \overline{Q}_h^\pi(c) = g_h^+(c), \qquad c \in C_h,$$

where

$$g_h^-(c) = (B_h^\pi \underline{V}_{h+1}^\pi)(c), \qquad g_h^+(c) = (B_h^\pi \overline{V}_{h+1}^\pi)(c).$$

Therefore

$$\begin{aligned}
w_h^\pi(c) &= \overline{Q}_h^\pi(c) - \underline{Q}_h^\pi(c) \\
&= (B_h^\pi \overline{V}_{h+1}^\pi)(c) - (B_h^\pi \underline{V}_{h+1}^\pi)(c) \\
&= \mathbb{E}\Big[R_h + \overline{V}_{h+1}^\pi(S_{h+1}) \mid X_h = c\Big] - \mathbb{E}\big[R_h + \underline{V}_{h+1}^\pi(S_{h+1}) \mid X_h = c\big] \\
&= \mathbb{E}\Big[\overline{V}_{h+1}^\pi(S_{h+1}) - \underline{V}_{h+1}^\pi(S_{h+1}) \mid X_h = c\Big] \\
&= (\mathsf{T}_h \Delta_{h+1})(c),
\end{aligned}$$

which proves (G.3). □

G.2. The Dynamic Support-Hole Envelope

We now define the geometric envelope that governs off-support ambiguity.

Definition G.3 (Dynamic support-hole envelope)**.** For every $h \in [H]$ and every $x \in \mathcal{X}_h$, define

$$\Gamma_h^\pi(x) := \inf_{c \in C_h} \{2L_h d_h(x,c) + (\mathsf{T}_h \Delta_{h+1})(c)\}. \tag{G.4}$$

The quantity $\Gamma_h^\pi(x)$ has a transparent interpretation. The term $2L_h d_h(x,c)$ is the local price of extending compatible values from an observed anchor point $c \in C_h$ to the query point x, while $(\mathsf{T}_h \Delta_{h+1})(c)$ is the ambiguity already present in the continuation problem at that anchor. The infimum selects the most favorable support anchor from which to extrapolate.

Theorem G.4 (Dynamic support-hole upper bound)**.** *Assume assumptions 2.1 and 3.1. Then for every $h \in [H]$ and every $x \in \mathcal{X}_h$,*

$$w_h^\pi(x) \le \Gamma_h^\pi(x). \tag{G.5}$$

Moreover, on the support itself,

$$w_h^\pi(c) = \Gamma_h^\pi(c) = (\mathsf{T}_h \Delta_{h+1})(c), \qquad c \in C_h. \tag{G.6}$$

Proof. Fix $h \in [H]$ and $x \in \mathcal{X}_h$. Let $c \in C_h$ be arbitrary. From the Bellman–Whitney definitions,

$$\overline{Q}_h^\pi(x) = \inf_{c' \in C_h} \left\{ g_h^+(c') + L_h d_h(x,c') \right\} \le g_h^+(c) + L_h d_h(x,c),$$

and

$$\underline{Q}_h^\pi(x) = \sup_{c' \in C_h} \left\{ g_h^-(c') - L_h d_h(x,c') \right\} \ge g_h^-(c) - L_h d_h(x,c).$$

Subtracting the two displays gives

$$\begin{aligned} w_h^\pi(x) &= \overline{Q}_h^\pi(x) - \underline{Q}_h^\pi(x) \\ &\le g_h^+(c) - g_h^-(c) + 2L_h d_h(x,c). \end{aligned}$$

By lemma G.2,

$$g_h^+(c) - g_h^-(c) = w_h^\pi(c) = (\mathsf{T}_h \Delta_{h+1})(c).$$

Therefore

$$w_h^\pi(x) \le 2L_h d_h(x,c) + (\mathsf{T}_h \Delta_{h+1})(c).$$

Since $c \in C_h$ was arbitrary, taking the infimum over c proves (G.5).

If $x = c \in C_h$, then (G.3) gives

$$w_h^\pi(c) = (\mathsf{T}_h \Delta_{h+1})(c).$$

On the other hand, the definition of $\Gamma_h^\pi(c)$ implies

$$\Gamma_h^\pi(c) \le 2L_h d_h(c,c) + (\mathsf{T}_h \Delta_{h+1})(c) = (\mathsf{T}_h \Delta_{h+1})(c).$$

Combining this with (G.5) yields

$$w_h^\pi(c) \le \Gamma_h^\pi(c) \le w_h^\pi(c),$$

hence equality holds throughout, proving (G.6). □

Remark G.5 (Geometry beyond weak overlap). The bound in theorem G.4 is qualitatively different from weak-overlap or density-ratio analyses. It is driven by a metric notion of support holes and by continuation ambiguity propagated through the Bellman operator, not by moments of importance weights. In particular, it remains meaningful when the target policy enters regions of state–action space that have genuinely zero behavior support.

G.3. A Least-Favorable Serial Family

We next show that the upper bound in theorem G.4 is sharp. The sharpness result is not merely existential; it is realized by an explicit recursive family in which the support at every stage collapses to a single anchor point and all future ambiguity propagates deterministically along one target-policy trajectory.

Definition G.6 (Serial singleton-support family)**.** Fix a horizon H, a deterministic target policy π, a sequence of evaluation points

$$x_t^\star = (s_t^\star, a_t^\star) \in \mathcal{X}_t, \qquad t \in [H],$$

with

$$\pi_t(\cdot \mid s_t^\star) = \delta_{a_t^\star},$$

and a sequence of anchor points

$$c_t \in \mathcal{X}_t, \qquad t \in [H].$$

A serial singleton-support family is the Bellman data specified by:

1. the support set at stage t is the singleton
$$C_t = \{c_t\};$$
2. the terminal supported Bellman target is zero:
$$(B_H^\pi 0)(c_H) = 0;$$
3. for every $t \in \{1, \dots, H-1\}$ and every bounded Borel $v : \mathcal{S}_{t+1} \to \mathbb{R}$,
$$(B_t^\pi v)(c_t) = v(s_{t+1}^\star).$$

Equivalently, the supported continuation operator is deterministic:

$$(\mathsf{T}_t f)(c_t) = f(s_{t+1}^\star), \qquad t \in \{1, \dots, H-1\}.$$

Because each support set is a singleton, assumption 3.1 holds automatically for every radius vector L: every function on a singleton support is trivially L_t-Lipschitz.

The next lemma computes the recursive widths exactly on this family.

Lemma G.7 (Least-favorable width recursion)**.** *Consider a serial singleton-support family in the sense of definition G.6. Then, for every $t \in [H]$,*

$$w_t^\pi(x_t^\star) = 2L_t d_t(x_t^\star, c_t) + \Delta_{t+1}(s_{t+1}^\star), \tag{G.7}$$

with the terminal convention $\Delta_{H+1} \equiv 0$. Since the target policy is deterministic at $s_{t+1}^\star$, this can equivalently be written as

$$w_t^\pi(x_t^\star) = 2L_t d_t(x_t^\star, c_t) + w_{t+1}^\pi(x_{t+1}^\star), \qquad t \in \{1, \dots, H-1\}. \tag{G.8}$$

Proof. Fix $t \in [H]$. Since $C_t = \{c_t\}$ is a singleton, the lower and upper Bellman–Whitney envelopes reduce to

$$\underline{Q}_t^\pi(x) = g_t^-(c_t) - L_t d_t(x, c_t), \qquad \overline{Q}_t^\pi(x) = g_t^+(c_t) + L_t d_t(x, c_t),$$

for all $x \in \mathcal{X}_t$. Therefore

$$w_t^\pi(x) = g_t^+(c_t) - g_t^-(c_t) + 2L_t d_t(x, c_t). \tag{G.9}$$

Evaluating at $x = x_t^\star$ gives

$$w_t^\pi(x_t^\star) = g_t^+(c_t) - g_t^-(c_t) + 2L_t d_t(x_t^\star, c_t).$$

By definition of the serial family, for $t < H$,

$$g_t^-(c_t) = (B_t^\pi \underline{V}_{t+1}^\pi)(c_t) = \underline{V}_{t+1}^\pi(s_{t+1}^\star),$$

and

$$g_t^+(c_t) = (B_t^\pi \overline{V}_{t+1}^\pi)(c_t) = \overline{V}_{t+1}^\pi(s_{t+1}^\star).$$

Hence

$$g_t^+(c_t) - g_t^-(c_t) = \Delta_{t+1}(s_{t+1}^\star),$$

which proves (G.7). When $t = H$, the same formula holds because $g_H^-(c_H) = g_H^+(c_H) = 0$ and $\Delta_{H+1} \equiv 0$ by convention.

Finally, since the target policy is deterministic at $s_{t+1}^\star$,

$$\Delta_{t+1}(s_{t+1}^\star) = \sum_{a \in \mathcal{A}} \pi_{t+1}(a \mid s_{t+1}^\star)\, w_{t+1}^\pi(s_{t+1}^\star, a) = w_{t+1}^\pi(x_{t+1}^\star),$$

which proves (G.8). □

We can now state the sharpness theorem.

Theorem G.8 (Sharpness on serial singleton-support families)**.** *Fix any deterministic target policy π and any serial singleton-support family in the sense of definition G.6. Then, for every $t \in [H]$,*

$$w_t^\pi(x_t^\star) = \Gamma_t^\pi(x_t^\star). \tag{G.10}$$

Equivalently,

$$w_t^\pi(x_t^\star) = 2L_t d_t(x_t^\star, c_t) + w_{t+1}^\pi(x_{t+1}^\star), \qquad t \in \{1, \dots, H-1\}, \tag{G.11}$$

with terminal identity

$$w_H^\pi(x_H^\star) = 2L_H d_H(x_H^\star, c_H).$$

In particular,

$$w_h^\pi(x_h^\star) = 2\sum_{t=h}^{H} L_t d_t(x_t^\star, c_t), \qquad h \in [H]. \tag{G.12}$$

Proof. Because $C_t = \{c_t\}$ is a singleton, the definition of Γ_t^π simplifies to

$$\Gamma_t^\pi(x_t^\star) = 2L_t d_t(x_t^\star, c_t) + (\mathsf{T}_t \Delta_{t+1})(c_t).$$

For $t < H$, the serial family definition implies

$$(\mathsf{T}_t \Delta_{t+1})(c_t) = \Delta_{t+1}(s_{t+1}^\star) = w_{t+1}^\pi(x_{t+1}^\star),$$

where the last identity uses determinism of the target policy at $s_{t+1}^\star$. Therefore

$$\Gamma_t^\pi(x_t^\star) = 2L_t d_t(x_t^\star, c_t) + w_{t+1}^\pi(x_{t+1}^\star).$$

By lemma G.7, the right-hand side is exactly $w_t^\pi(x_t^\star)$, proving (G.10) and (G.11). The terminal identity is the case $t = H$ of the same formula.

The closed-form expression (G.12) follows by iterating (G.11) backward from the terminal stage. □

G.4. A Two-Stage Closed-Form Example

The preceding theorem is abstract but completely explicit. The next proposition gives the simplest nontrivial illustration.

Proposition G.9 (Two-stage line-metric example)**.** *Assume $H = 2$, let the action space be a singleton, and let*

$$\mathcal{S}_1 = \mathcal{S}_2 = [0, 1]$$

with the Euclidean metric. Fix radii $L_1, L_2 > 0$, support anchors

$$c_1 = (0, a), \qquad c_2 = (0, a),$$

evaluation points

$$x_1^\star = (\delta_1, a), \qquad x_2^\star = (\delta_2, a), \qquad \delta_1, \delta_2 \in [0, 1],$$

and define a serial singleton-support family by setting

$$C_1 = \{c_1\}, \qquad C_2 = \{c_2\},$$

with terminal supported Bellman target zero and deterministic continuation from c_1 to state δ_2. Then

$$w_2^\pi(x_2^\star) = 2L_2\delta_2,$$

and

$$w_1^\pi(x_1^\star) = 2L_1\delta_1 + 2L_2\delta_2.$$

Hence the two-stage ambiguity is exactly additive across stages.

Proof. Because the action space is a singleton, target-policy averaging is trivial. At stage 2, the support is the singleton $c_2 = (0, a)$ and the supported target is zero, so

$$\underline{Q}_2^\pi(x) = -L_2 d_2(x, c_2), \qquad \overline{Q}_2^\pi(x) = L_2 d_2(x, c_2).$$

Evaluating at $x_2^\star = (\delta_2, a)$ gives

$$w_2^\pi(x_2^\star) = \overline{Q}_2^\pi(x_2^\star) - \underline{Q}_2^\pi(x_2^\star) = 2L_2\delta_2.$$

At stage 1, the deterministic continuation from c_1 to state δ_2 implies that the support width is exactly

$$\Delta_2(\delta_2) = w_2^\pi(x_2^\star) = 2L_2\delta_2.$$

Since $C_1 = \{c_1\}$ is again a singleton,

$$w_1^\pi(x_1^\star) = 2L_1 d_1(x_1^\star, c_1) + \Delta_2(\delta_2) = 2L_1\delta_1 + 2L_2\delta_2,$$

which proves the claim. □

G.5. Interpretation

The dynamic support-hole envelope from theorem G.4 is therefore exact in two distinct senses. First, on the support itself, it coincides with the recursively propagated continuation ambiguity. Second, on the explicit least-favorable family of theorem G.8, it is attained with equality at every stage and along every point of the designated target trajectory. Thus the Bellman–Whitney interval width is not merely bounded by the geometry of support holes; it is, in a precise minimax sense, generated by that geometry.

H. Stability and Endpoint Estimation

This appendix develops the perturbation theory of the Bellman–Whitney recursion. The guiding principle is to separate two sources of uncertainty:

1. *irreducible identification width*, quantified by the exact Bellman–Whitney interval itself;
2. *estimable endpoint error*, caused by replacing the population support sets and supported Bellman operators with empirical analogues.

The results below concern only the second component. They show that the recursive envelopes are stable under joint perturbations of the support sets and the supported Bellman targets, with additive error propagation across the horizon.

H.1. Empirical Bellman–Whitney Recursion

Let $\widehat{C}_h \subseteq \mathcal{X}_h$ be a nonempty compact estimator of the support C_h, and let

$$\widehat{B}_h^\pi : \{\text{bounded Borel } v : \mathcal{S}_{h+1} \to \mathbb{R}\} \to \{\text{bounded Borel functions on } \widehat{C}_h\}$$

be an empirical supported Bellman operator.

Assumption H.1 (Empirical Bellman regularity)**.** For every stage $h \in [H]$:

1. for every bounded Borel $v : \mathcal{S}_{h+1} \to \mathbb{R}$, the function $\widehat{B}_h^\pi v$ is bounded and Borel measurable on $\widehat{C}_h$;
2. (*sup-norm nonexpansiveness*) for all bounded Borel $v, w : \mathcal{S}_{h+1} \to \mathbb{R}$,
$$\sup_{\hat{c} \in \widehat{C}_h} \left| (\widehat{B}_h^\pi v)(\hat{c}) - (\widehat{B}_h^\pi w)(\hat{c}) \right| \le \|v - w\|_\infty;$$
3. (*boundedness*) for all bounded Borel $v : \mathcal{S}_{h+1} \to \mathbb{R}$,
$$\sup_{\hat{c} \in \widehat{C}_h} \left| (\widehat{B}_h^\pi v)(\hat{c}) \right| \le R_{\max} + \|v\|_\infty.$$

Remark H.2 (On assumption H.1). Assumption H.1 is satisfied by any plug-in Bellman operator of the form

$$(\widehat{B}_h^\pi v)(\hat{c}) = \hat{r}_h(\hat{c}) + \int v(s')\, \hat{P}_h(ds' \mid \hat{c}),$$

where $\hat{r}_h$ is bounded and $\hat{P}_h(\cdot \mid \hat{c})$ is a probability kernel. The nonexpansiveness property is then immediate from the fact that a Markov kernel is a contraction in sup norm.

Define the deterministic stagewise diameter

$$D_h := \operatorname{diam}(\mathcal{X}_h), \qquad h \in [H],$$

and the recursive envelope bounds

$$B_{H+1} := 0, \qquad B_h := R_{\max} + B_{h+1} + L_h D_h, \qquad h \in [H].$$

For each stage h, define the bounded value class

$$\mathcal{V}_h^{\text{env}} := \left\{ \mathsf{A}_h^\pi q : q \in \operatorname{Lip}_{L_h}(\mathcal{X}_h),\ \|q\|_\infty \le B_h \right\}, \qquad \mathcal{V}_{H+1}^{\text{env}} := \{0\}.$$

Definition H.3 (Empirical Bellman–Whitney recursion)**.** Set

$$\underline{\widehat{V}}_{H+1}^\pi = \overline{\widehat{V}}_{H+1}^\pi \equiv 0.$$

For $h = H, H-1, \ldots, 1$, define the empirical supported targets

$$\hat{g}_h^-(\hat{c}) := (\widehat{B}_h^\pi \underline{\widehat{V}}_{h+1}^\pi)(\hat{c}), \qquad \hat{g}_h^+(\hat{c}) := (\widehat{B}_h^\pi \overline{\widehat{V}}_{h+1}^\pi)(\hat{c}), \qquad \hat{c} \in \widehat{C}_h,$$

and the empirical Bellman–Whitney envelopes

$$\underline{\widehat{Q}}_h^\pi := \mathcal{W}^-_{\widehat{C}_h, L_h}\, \hat{g}_h^-, \qquad \overline{\widehat{Q}}_h^\pi := \mathcal{W}^+_{\widehat{C}_h, L_h}\, \hat{g}_h^+,$$

together with the associated value functions

$$\underline{\widehat{V}}_h^\pi := \mathsf{A}_h^\pi \underline{\widehat{Q}}_h^\pi, \qquad \overline{\widehat{V}}_h^\pi := \mathsf{A}_h^\pi \overline{\widehat{Q}}_h^\pi.$$

Lemma H.4 (Uniform recursive boundedness)**.** *Assume assumptions 2.1, 3.1 and H.1. Then for every $h \in [H]$,*

$$\|\underline{Q}_h^\pi\|_\infty \vee \|\overline{Q}_h^\pi\|_\infty \vee \|\underline{\widehat{Q}}_h^\pi\|_\infty \vee \|\overline{\widehat{Q}}_h^\pi\|_\infty \le B_h,$$

and hence

$$\underline{V}_h^\pi, \overline{V}_h^\pi, \underline{\widehat{V}}_h^\pi, \overline{\widehat{V}}_h^\pi \in \mathcal{V}_h^{\mathrm{env}}.$$

Proof. We argue by backward induction on h.

At stage $H+1$, all four value functions are identically zero, so the claim is trivial.

Now fix $h \in [H]$ and assume the statement holds at stage $h+1$. Since $\underline{V}_{h+1}^\pi, \overline{V}_{h+1}^\pi, \underline{\widehat{V}}_{h+1}^\pi, \overline{\widehat{V}}_{h+1}^\pi \in \mathcal{V}_{h+1}^{\mathrm{env}}$, their sup norms are bounded by B_{h+1}.

For the population upper envelope, the supported Bellman target satisfies

$$\sup_{c \in C_h} \left|g_h^+(c)\right| = \sup_{c \in C_h} \left|(B_h^\pi \overline{V}_{h+1}^\pi)(c)\right| \le R_{\max} + \|\overline{V}_{h+1}^\pi\|_\infty \le R_{\max} + B_{h+1},$$

and similarly for g_h^-. By the generic envelope bound from Appendix B,

$$\|\overline{Q}_h^\pi\|_\infty \le \sup_{c \in C_h} \left|g_h^+(c)\right| + L_h D_h \le R_{\max} + B_{h+1} + L_h D_h = B_h.$$

The same argument gives $\|\underline{Q}_h^\pi\|_\infty \le B_h$.

For the empirical upper envelope, assumption H.1 yields

$$\sup_{\hat{c} \in \widehat{C}_h} \left|\hat{g}_h^+(\hat{c})\right| = \sup_{\hat{c} \in \widehat{C}_h} \left|(\widehat{B}_h^\pi \overline{\widehat{V}}_{h+1}^\pi)(\hat{c})\right| \le R_{\max} + \|\overline{\widehat{V}}_{h+1}^\pi\|_\infty \le R_{\max} + B_{h+1}.$$

Applying the same envelope bound on $\widehat{C}_h$ gives

$$\|\overline{\widehat{Q}}_h^\pi\|_\infty \le B_h.$$

Likewise $\|\underline{\widehat{Q}}_h^\pi\|_\infty \le B_h$.

Finally, because target-policy averaging is a convex combination over the finite action space,

$$\|\underline{V}_h^\pi\|_\infty \vee \|\overline{V}_h^\pi\|_\infty \vee \|\underline{\widehat{V}}_h^\pi\|_\infty \vee \|\overline{\widehat{V}}_h^\pi\|_\infty \le B_h.$$

Moreover, each of the four Q-functions is L_h-Lipschitz by construction, so the associated value functions belong to $\mathcal{V}_h^{\mathrm{env}}$. □

H.2. A Stagewise Perturbation Modulus

We now quantify the pure stagewise discrepancy between the true and empirical supported Bellman operators when evaluated at the *same* continuation function. This isolates the estimation error that is not yet due to recursive propagation from later stages.

For each stage $h \in [H]$, define

$$\varepsilon_h := \sup_{v \in \mathcal{V}_{h+1}^{\mathrm{env}}} \Delta_{C_h, \widehat{C}_h}^{(L_h)} \Big(c \mapsto (B_h^\pi v)(c),\ \hat{c} \mapsto (\widehat{B}_h^\pi v)(\hat{c})\Big), \tag{H.1}$$

where $\Delta_{A,B}^{(L)}$ is the symmetrized support-value discrepancy from Section B. The quantity ε_h measures how accurately the empirical support-value pair $(\widehat{C}_h, \widehat{B}_h^\pi)$ reproduces the true pair (C_h, B_h^π) at stage h, uniformly over all continuation values that can arise from the Bellman–Whitney recursion.

Theorem H.5 (Deterministic stability of the Bellman–Whitney endpoints)**.** *Assume assumptions 2.1, 3.1 and H.1. Then, for every stage $h \in [H]$,*

$$\|\widehat{\underline{Q}}_h^\pi - \underline{Q}_h^\pi\|_\infty \vee \|\widehat{\overline{Q}}_h^\pi - \overline{Q}_h^\pi\|_\infty \vee \|\widehat{\underline{V}}_h^\pi - \underline{V}_h^\pi\|_\infty \vee \|\widehat{\overline{V}}_h^\pi - \overline{V}_h^\pi\|_\infty \le \sum_{t=h}^{H} \varepsilon_t. \tag{H.2}$$

In particular,

$$\left|\widehat{\underline{V}}_1^\pi(s_1) - \underline{V}_1^\pi(s_1)\right| \vee \left|\widehat{\overline{V}}_1^\pi(s_1) - \overline{V}_1^\pi(s_1)\right| \le \sum_{t=1}^{H} \varepsilon_t. \tag{H.3}$$

Proof. For each stage h, define the aggregate error

$$E_h := \|\widehat{\underline{Q}}_h^\pi - \underline{Q}_h^\pi\|_\infty \vee \|\widehat{\overline{Q}}_h^\pi - \overline{Q}_h^\pi\|_\infty \vee \|\widehat{\underline{V}}_h^\pi - \underline{V}_h^\pi\|_\infty \vee \|\widehat{\overline{V}}_h^\pi - \overline{V}_h^\pi\|_\infty,$$

and set $E_{H+1} := 0$.

We prove by backward induction that

$$E_h \le \varepsilon_h + E_{h+1}, \qquad h \in [H].$$

Iterating this recursion immediately yields (H.2), and (H.3) follows because point evaluation is dominated by the sup norm.

Fix $h \in [H]$. We first control the upper envelope. By lemma H.4, both $\overline{V}_{h+1}^\pi$ and $\widehat{\overline{V}}_{h+1}^\pi$ belong to $\mathcal{V}_{h+1}^{\mathrm{env}}$. By the pair-stability lemma from Appendix B,

$$\|\widehat{\overline{Q}}_h^\pi - \overline{Q}_h^\pi\|_\infty \le \Delta_{C_h,\widehat{C}_h}^{(L_h)}\big(g_h^+, \hat{g}_h^+\big),$$

where

$$g_h^+(c) = (B_h^\pi \overline{V}_{h+1}^\pi)(c), \qquad \hat{g}_h^+(\hat{c}) = (\widehat{B}_h^\pi \widehat{\overline{V}}_{h+1}^\pi)(\hat{c}).$$

We now compare $\hat{g}_h^+$ to the empirical target evaluated at the true continuation value. Let

$$\tilde{g}_h^+(\hat{c}) := (\widehat{B}_h^\pi \overline{V}_{h+1}^\pi)(\hat{c}), \qquad \hat{c} \in \widehat{C}_h.$$

By the definition of ε_h, since $\overline{V}_{h+1}^\pi \in \mathcal{V}_{h+1}^{\mathrm{env}}$,

$$\Delta_{C_h,\widehat{C}_h}^{(L_h)}\big(g_h^+, \tilde{g}_h^+\big) \le \varepsilon_h.$$

Moreover, by the definition of $\Delta_{A,B}^{(L)}$ and assumption H.1,

$$\begin{aligned}
&\Delta_{C_h,\widehat{C}_h}^{(L_h)}\big(g_h^+, \hat{g}_h^+\big) \\
&\le \Delta_{C_h,\widehat{C}_h}^{(L_h)}\big(g_h^+, \tilde{g}_h^+\big) + \sup_{\hat{c}\in\widehat{C}_h} \left|(\widehat{B}_h^\pi \overline{V}_{h+1}^\pi)(\hat{c}) - (\widehat{B}_h^\pi \widehat{\overline{V}}_{h+1}^\pi)(\hat{c})\right| \\
&\le \varepsilon_h + \|\overline{V}_{h+1}^\pi - \widehat{\overline{V}}_{h+1}^\pi\|_\infty \\
&\le \varepsilon_h + E_{h+1}.
\end{aligned}$$

Hence

$$\|\widehat{\overline{Q}}_h^\pi - \overline{Q}_h^\pi\|_\infty \le \varepsilon_h + E_{h+1}. \tag{H.4}$$

The same argument applies to the lower envelopes. Define

$$g_h^-(c) = (B_h^\pi \underline{V}_{h+1}^\pi)(c), \qquad \hat{g}_h^-(\hat{c}) = (\widehat{B}_h^\pi \widehat{\underline{V}}_{h+1}^\pi)(\hat{c}), \qquad \tilde{g}_h^-(\hat{c}) = (\widehat{B}_h^\pi \underline{V}_{h+1}^\pi)(\hat{c}).$$

Then

$$\|\widehat{\underline{Q}}_h^\pi - \underline{Q}_h^\pi\|_\infty \le \Delta_{C_h,\widehat{C}_h}^{(L_h)}\big(g_h^-, \hat{g}_h^-\big) \le \varepsilon_h + E_{h+1},$$

so

$$\|\widehat{\underline{Q}}_h^\pi - \underline{Q}_h^\pi\|_\infty \le \varepsilon_h + E_{h+1}. \tag{H.5}$$

Finally, because target-policy averaging is a contraction in sup norm,

$$\|\widehat{\underline{V}}_h^\pi - \underline{V}_h^\pi\|_\infty \le \|\widehat{\underline{Q}}_h^\pi - \underline{Q}_h^\pi\|_\infty, \qquad \|\widehat{\overline{V}}_h^\pi - \overline{V}_h^\pi\|_\infty \le \|\widehat{\overline{Q}}_h^\pi - \overline{Q}_h^\pi\|_\infty.$$

Combining these inequalities with (H.4)–(H.5) yields

$$E_h \le \varepsilon_h + E_{h+1},$$

which completes the induction. □

H.3. Separating Support Error from Bellman-Target Error

The preceding theorem is exact but uses the abstract perturbation modulus ε_h. The next corollary gives a concrete and interpretable sufficient condition that separates support-set error from Bellman-target error.

Corollary H.6 (Support and Bellman-target error decomposition)**.** *Assume assumptions 2.1, 3.1 and H.1. Suppose that for each stage $h \in [H]$ there exist nonnegative numbers α_h, η_h such that:*

1. *(*support estimation error*)*
$$d_{\mathrm{H}}\Big(C_h, \widehat{C}_h\Big) \le \eta_h;$$
2. *(*ambient Bellman-target approximation*) for every $v \in \mathcal{V}_{h+1}^{\mathrm{env}}$ there exists an L_h-Lipschitz function $f_{h,v} : \mathcal{X}_h \to \mathbb{R}$ satisfying*
$$f_{h,v}(c) = (B_h^\pi v)(c), \qquad c \in C_h,$$
and
$$\sup_{\hat{c} \in \widehat{C}_h} \Big|(\widehat{B}_h^\pi v)(\hat{c}) - f_{h,v}(\hat{c})\Big| \le \alpha_h.$$

Then

$$\varepsilon_h \le \alpha_h + 2L_h\eta_h, \qquad h \in [H],$$

and therefore

$$\|\widehat{\underline{Q}}_h^\pi - \underline{Q}_h^\pi\|_\infty \vee \|\widehat{\overline{Q}}_h^\pi - \overline{Q}_h^\pi\|_\infty \vee \|\widehat{\underline{V}}_h^\pi - \underline{V}_h^\pi\|_\infty \vee \|\widehat{\overline{V}}_h^\pi - \overline{V}_h^\pi\|_\infty \le \sum_{t=h}^{H}(\alpha_t + 2L_t\eta_t). \tag{H.6}$$

In particular,

$$\Big|\widehat{\underline{V}}_1^\pi(s_1) - \underline{V}_1^\pi(s_1)\Big| \vee \Big|\widehat{\overline{V}}_1^\pi(s_1) - \overline{V}_1^\pi(s_1)\Big| \le \sum_{t=1}^{H}(\alpha_t + 2L_t\eta_t). \tag{H.7}$$

Proof. Fix $h \in [H]$ and $v \in \mathcal{V}_{h+1}^{\mathrm{env}}$. Set

$$g(c) := (B_h^\pi v)(c), \qquad c \in C_h,$$

and

$$\hat{g}(\hat{c}) := (\widehat{B}_h^\pi v)(\hat{c}), \qquad \hat{c} \in \widehat{C}_h.$$

We claim that

$$\Delta_{C_h, \widehat{C}_h}^{(L_h)}(g, \hat{g}) \le \alpha_h + 2L_h\eta_h.$$

Indeed, let $c \in C_h$. Since $d_{\mathrm{H}}(C_h, \widehat{C}_h) \le \eta_h$, there exists $\hat{c} \in \widehat{C}_h$ with

$$d_h(c, \hat{c}) \le \eta_h.$$

Because $f_{h,v}$ is L_h-Lipschitz and extends g from C_h,

$$\begin{aligned}|g(c) - \hat{g}(\hat{c})| + L_h d_h(c, \hat{c}) &\le |f_{h,v}(c) - f_{h,v}(\hat{c})| + |f_{h,v}(\hat{c}) - \hat{g}(\hat{c})| + L_h d_h(c, \hat{c}) \\ &\le L_h d_h(c, \hat{c}) + \alpha_h + L_h d_h(c, \hat{c}) \\ &\le \alpha_h + 2L_h \eta_h.\end{aligned}$$

Taking the infimum over admissible $\hat{c}$ and then the supremum over $c \in C_h$ gives

$$\Delta^{(L_h)}_{C_h \to \widehat{C}_h}(g, \hat{g}) \le \alpha_h + 2L_h \eta_h.$$

The reverse directed bound

$$\Delta^{(L_h)}_{\widehat{C}_h \to C_h}(\hat{g}, g) \le \alpha_h + 2L_h \eta_h$$

is identical, using for each $\hat{c} \in \widehat{C}_h$ a point $c \in C_h$ within Hausdorff distance η_h. Hence

$$\Delta^{(L_h)}_{C_h, \widehat{C}_h}(g, \hat{g}) \le \alpha_h + 2L_h \eta_h.$$

Taking the supremum over $v \in \mathcal{V}^{\text{env}}_{h+1}$ proves $\varepsilon_h \le \alpha_h + 2L_h \eta_h$.

Substituting this bound into theorem H.5 yields (H.6), and (H.7) follows by point evaluation. □

H.4. A Conditional Finite-Sample Corollary

The previous corollary is deterministic. To extract a statistical rate, it is enough to place high-probability bounds on the two primitive quantities α_h and η_h.

Corollary H.7 (Conditional finite-sample endpoint rates)**.** *Let $\delta_n \in [0, 1]$. Suppose that for each sample size n there exists an event $\mathcal{E}_n$ with*

$$\mathbb{P}(\mathcal{E}_n) \ge 1 - \delta_n$$

such that on $\mathcal{E}_n$ the assumptions of corollary H.6 hold with stagewise errors $\alpha_{h,n}$ and $\eta_{h,n}$. Then on $\mathcal{E}_n$,

$$\left|\underline{\widehat{V}}^{\pi}_1(s_1) - \underline{V}^{\pi}_1(s_1)\right| \vee \left|\overline{\widehat{V}}^{\pi}_1(s_1) - \overline{V}^{\pi}_1(s_1)\right| \le \sum_{h=1}^{H} (\alpha_{h,n} + 2L_h \eta_{h,n}). \tag{H.8}$$

Consequently,

$$\mathbb{P}\left(\left|\underline{\widehat{V}}^{\pi}_1(s_1) - \underline{V}^{\pi}_1(s_1)\right| \vee \left|\overline{\widehat{V}}^{\pi}_1(s_1) - \overline{V}^{\pi}_1(s_1)\right| \le \sum_{h=1}^{H} (\alpha_{h,n} + 2L_h \eta_{h,n})\right) \ge 1 - \delta_n.$$

In particular, if for some stagewise dimensions $d_h > 0$ and effective sample sizes n_h,

$$\alpha_{h,n} = \widetilde{O}\Big(n_h^{-1/(2+d_h)}\Big), \qquad \eta_{h,n} = \widetilde{O}\Big(n_h^{-1/d_h}\Big),$$

then

$$\left|\underline{\widehat{V}}^{\pi}_1(s_1) - \underline{V}^{\pi}_1(s_1)\right| \vee \left|\overline{\widehat{V}}^{\pi}_1(s_1) - \overline{V}^{\pi}_1(s_1)\right| = \widetilde{O}_p\left(\sum_{h=1}^{H} n_h^{-1/(2+d_h)}\right).$$

Proof. The high-probability inequality (H.8) is exactly corollary H.6 applied on the event $\mathcal{E}_n$. The probability statement follows immediately.

For the final rate display, note that

$$n_h^{-1/d_h} = o\Big(n_h^{-1/(2+d_h)}\Big) \qquad \text{as } n_h \to \infty,$$

because $1/d_h > 1/(2 + d_h)$. Hence the support-estimation term is asymptotically no larger than the Bellman-target estimation term, and the overall rate is dominated by $\sum_{h=1}^{H} n_h^{-1/(2+d_h)}$ up to logarithmic factors. □

H.5. Interpretation

The stability theorem reveals a useful decomposition:

$$\text{endpoint estimation error} \lesssim \sum_{h=1}^{H} \Big(\text{Bellman-target estimation error at stage } h + \text{support-set estimation error at stage } h\Big).$$

Thus the recursive Bellman–Whitney endpoints behave like a stable dynamic program over perturbed support-value pairs. The irreducible interval width is a separate object, governed by the geometry of support holes (Appendix G); the results here show that the statistical task of estimating the interval *endpoints* is additive over stages and does not introduce an additional horizon-dependent instability beyond this backward accumulation.

I. Minimax Lower Bound

This appendix proves a minimax lower bound for *endpoint estimation*. The result is deliberately stated under a simplified oracle observation model in which the stagewise supported Bellman targets are observed through independent nonparametric Gaussian regression problems. This is a statistically favorable setting: the support is known exactly, there are no support holes, the target policy is deterministic, and the lower and upper Bellman–Whitney endpoints coincide. Proving a lower bound even in this regime isolates the intrinsic difficulty of *estimating the interval endpoints*, as distinct from the separate issue of irreducible identified-set width. The proof uses a standard two-point minimax construction in the spirit of Le Cam; see, e.g., Yu (1997); Tsybakov (2009).

I.1. Oracle Stagewise Regression Model

We work with a statistically favorable subclass in which all support-hole ambiguity vanishes and only endpoint-estimation difficulty remains.

Definition I.1 (Oracle stagewise regression model)**.** Fix horizon H, dimensions $(d_h)_{h=1}^{H}$ with $d_h \in \mathbb{N}$, and sample sizes $\mathbf{n} = (n_1, \dots, n_H) \in \mathbb{N}^H$. The oracle class $\mathfrak{O}_{\mathbf{n}}^{L}$ consists of all models satisfying the following properties:

1. For each stage $h \in [H]$,
$$\mathcal{S}_h = [0,1]^{d_h}$$
endowed with the ℓ_∞ metric, and the action space is a singleton
$$\mathcal{A} = \{a_0\}.$$
Consequently,
$$\mathcal{X}_h = [0,1]^{d_h} \times \{a_0\}, \qquad C_h = \mathcal{X}_h.$$
2. The target policy is deterministic and trivial:
$$\pi_h(\cdot \mid s) = \delta_{a_0} \qquad \forall s \in \mathcal{S}_h,\ \forall h \in [H].$$
3. For each stage $h \in \{1, \dots, H-1\}$, the controlled transition is deterministic to the origin:
$$S_{h+1} = \mathbf{0} \qquad \text{almost surely given } X_h.$$
4. The stage-h reward mean is a bounded function $r_h : [0,1]^{d_h} \to \mathbb{R}$ satisfying
$$\mathrm{Lip}_h(r_h) \le L_h, \qquad \|r_h\|_\infty \le R_{\max}.$$
5. The observed data are independent across stages and consist of independent stagewise Gaussian regression samples
$$\mathcal{D}_h = \{(Z_{h,i}, Y_{h,i})\}_{i=1}^{n_h}, \qquad h \in [H],$$
where
$$Z_{h,i} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([0,1]^{d_h}), \qquad Y_{h,i} = r_h(Z_{h,i}) + \xi_{h,i}, \qquad \xi_{h,i} \overset{\text{i.i.d.}}{\sim} N(0,1),$$
and all randomness is mutually independent across h and i.

Under definition I.1, the support equals the whole state–action space at every stage, so the Bellman–Whitney interval has zero width. The next lemma makes this reduction explicit.

Lemma I.2 (Endpoint collapse on the oracle class)**.** *Let $M \in \mathfrak{O}_{\mathbf{n}}^{L}$, and let*

$$\vartheta(M) := \sum_{h=1}^{H} r_h(\mathbf{0}).$$

Then:

1. *assumption 3.1 holds automatically;*
2. *the Bellman–Whitney interval degenerates:*
$$\mathcal{I}^{\pi} = \{\vartheta(M)\};$$
equivalently,
$$\underline{V}_1^{\pi}(\mathbf{0}) = \overline{V}_1^{\pi}(\mathbf{0}) = \vartheta(M).$$

Proof. Because $C_h = \mathcal{X}_h$ for every stage h, there is no support hole. Exact interpolation is therefore automatic, and the Bellman–Whitney envelopes coincide with the Bellman targets themselves. Since the target policy is deterministic and the transition is deterministic to $\mathbf{0}$,

$$V_h(\mathbf{0}) = r_h(\mathbf{0}) + V_{h+1}(\mathbf{0}), \qquad V_{H+1} \equiv 0.$$

Backward substitution yields

$$V_1(\mathbf{0}) = \sum_{h=1}^{H} r_h(\mathbf{0}) = \vartheta(M).$$

Because the support is full, the lower and upper envelopes agree with the same value function, proving

$$\underline{V}_1^{\pi}(\mathbf{0}) = \overline{V}_1^{\pi}(\mathbf{0}) = \vartheta(M).$$

□

I.2. A Two-Point Lower-Bound Principle

We use the following standard two-point absolute-error inequality.

Lemma I.3 (Two-point lower bound)**.** *Let P_+ and P_- be two probability measures on a common measurable space, and let $\theta_+, \theta_- \in \mathbb{R}$ with $\theta_+ \neq \theta_-$. Then for any measurable estimator $\hat{\theta}$,*

$$\max\Big\{\mathbb{E}_{P_+}\Big|\hat{\theta} - \theta_+\Big|, \mathbb{E}_{P_-}\Big|\hat{\theta} - \theta_-\Big|\Big\} \geq \frac{|\theta_+ - \theta_-|}{4}\,(1 - \mathrm{TV}(P_+, P_-)). \tag{I.1}$$

Consequently, by Pinsker's inequality,

$$\max\Big\{\mathbb{E}_{P_+}\Big|\hat{\theta} - \theta_+\Big|, \mathbb{E}_{P_-}\Big|\hat{\theta} - \theta_-\Big|\Big\} \geq \frac{|\theta_+ - \theta_-|}{4}\left(1 - \sqrt{\mathrm{KL}(P_+, P_-)/2}\right). \tag{I.2}$$

Proof. Set

$$\Delta := |\theta_+ - \theta_-| > 0, \qquad m := \frac{\theta_+ + \theta_-}{2}.$$

For any estimator $\hat{\theta}$ and any outcome ω, at least one of the two quantities

$$\Big|\hat{\theta}(\omega) - \theta_+\Big|, \qquad \Big|\hat{\theta}(\omega) - \theta_-\Big|$$

is at least $\Delta/2$. More precisely,

$$\Big|\hat{\theta} - \theta_+\Big| + \Big|\hat{\theta} - \theta_-\Big| \geq \Delta \qquad \text{pointwise.}$$

Let

$$A := \Big\{\omega : \hat{\theta}(\omega) \geq m\Big\}.$$

On A^c, we have $\hat{\theta} < m$, so

$$\Big|\hat{\theta} - \theta_+\Big| \geq \Delta/2.$$

On A, we have $\hat{\theta} \geq m$, so

$$\Big|\hat{\theta} - \theta_-\Big| \geq \Delta/2.$$

Therefore

$$\mathbb{E}_{P_+}\Big|\hat{\theta} - \theta_+\Big| \geq \frac{\Delta}{2} P_+(A^c), \qquad \mathbb{E}_{P_-}\Big|\hat{\theta} - \theta_-\Big| \geq \frac{\Delta}{2} P_-(A).$$

Taking the maximum of the two expectations and then averaging the two right-hand sides gives

$$\max\Big\{\mathbb{E}_{P_+}\Big|\hat{\theta} - \theta_+\Big|, \mathbb{E}_{P_-}\Big|\hat{\theta} - \theta_-\Big|\Big\} \geq \frac{\Delta}{4}(P_+(A^c) + P_-(A)).$$

Since

$$P_+(A^c) + P_-(A) = 1 - (P_+(A) - P_-(A)),$$

and

$$P_+(A) - P_-(A) \leq \mathrm{TV}(P_+, P_-),$$

we obtain (I.1). The second bound (I.2) follows from Pinsker's inequality

$$\mathrm{TV}(P_+, P_-) \leq \sqrt{\mathrm{KL}(P_+, P_-)/2}.$$

□

I.3. A Least-Favorable Two-Point Subclass

For each dimension $d \in \mathbb{N}$, define the bump profile

$$\psi_d(u) := (1 - \|u\|_\infty)_+, \qquad u \in [0,1]^d.$$

Then $\psi_d(0) = 1$, ψ_d is 1-Lipschitz on $[0,1]^d$ under the ℓ_∞ metric, and

$$\beta_d := \int_{[0,1]^d} \psi_d(u)^2 \, du = \frac{2}{(d+1)(d+2)}. \tag{I.3}$$

Indeed, writing $R = \|U\|_\infty$ for $U \sim \mathrm{Unif}([0,1]^d)$ gives $\mathbb{P}(R \leq r) = r^d$, and hence

$$\beta_d = d \int_0^1 (1-r)^2 r^{d-1} \, dr = \frac{2}{(d+1)(d+2)}.$$

For each stage $h \in [H]$, define

$$\gamma_h := \min\left\{\frac{1}{4}, \frac{R_{\max}}{2L_h}, \left(\frac{1}{16 H L_h^2 \beta_{d_h}}\right)^{\frac{1}{d_h+2}}\right\}, \qquad \delta_h := \gamma_h \, n_h^{-1/(d_h+2)}. \tag{I.4}$$

Define the scaled stagewise bump

$$\varphi_h(z) := \delta_h \, \psi_{d_h}(z/\delta_h), \qquad z \in [0,1]^{d_h},$$

with the convention that $\psi_{d_h}(z/\delta_h) = 0$ whenever some coordinate of z/δ_h exceeds one. Then:

1. φ_h is supported on $[0, \delta_h]^{d_h}$;
2. $\varphi_h(\mathbf{0}) = \delta_h$;

3. $\mathrm{Lip}_h(\varphi_h) \le 1$;

4. $$\int_{[0,1]^{d_h}} \varphi_h(z)^2 \, dz = \beta_{d_h} \, \delta_h^{d_h+2}. \tag{I.5}$$

We now define the least-favorable pair of oracle models. Let M^+ and M^- denote the two models in $\mathfrak{O}_{\mathbf{n}}^L$ whose stagewise reward means are

$$r_h^{\pm}(z) := \pm L_h \varphi_h(z), \qquad z \in [0,1]^{d_h}, \qquad h \in [H]. \tag{I.6}$$

By construction,

$$\mathrm{Lip}_h(r_h^{\pm}) \le L_h, \qquad \|r_h^{\pm}\|_\infty \le L_h \delta_h \le R_{\max}/2,$$

so $M^{\pm} \in \mathfrak{O}_{\mathbf{n}}^L$.

Lemma I.4 (Parameter separation and Kullback–Leibler control)**.** *For the least-favorable pair (M^+, M^-) defined above:*

1. *the endpoint parameter values satisfy*

$$\vartheta(M^+) - \vartheta(M^-) = 2 \sum_{h=1}^{H} L_h \delta_h; \tag{I.7}$$

2. *the joint data distributions satisfy*

$$\mathrm{KL}(P_{M^+}, P_{M^-}) \le \frac{1}{8}. \tag{I.8}$$

Proof. By lemma I.2,

$$\vartheta(M^{\pm}) = \sum_{h=1}^{H} r_h^{\pm}(\mathbf{0}).$$

Since $\varphi_h(\mathbf{0}) = \delta_h$,

$$r_h^+(\mathbf{0}) - r_h^-(\mathbf{0}) = 2L_h\delta_h.$$

Summing over h proves (I.7).

For the Kullback–Leibler divergence, note that the data are independent across stages, so

$$\mathrm{KL}(P_{M^+}, P_{M^-}) = \sum_{h=1}^{H} \mathrm{KL}(P_{h,+}, P_{h,-}),$$

where $P_{h,\pm}$ denotes the law of the stage-h sample $\mathcal{D}_h$ under $M^{\pm}$.

Conditioned on the design points $(Z_{h,i})_{i=1}^{n_h}$, the responses are independent Gaussians with unit variance and means $r_h^{\pm}(Z_{h,i})$. Therefore

$$\mathrm{KL}(P_{h,+}, P_{h,-} \mid Z_{h,1:n_h}) = \frac{1}{2} \sum_{i=1}^{n_h} \big(r_h^+(Z_{h,i}) - r_h^-(Z_{h,i})\big)^2.$$

Taking expectation over the design and using i.i.d. sampling yields

$$\mathrm{KL}(P_{h,+}, P_{h,-}) = \frac{n_h}{2} \int_{[0,1]^{d_h}} \big(r_h^+(z) - r_h^-(z)\big)^2 \, dz.$$

By (I.6),

$$r_h^+(z) - r_h^-(z) = 2L_h\varphi_h(z),$$

hence

$$\mathrm{KL}(P_{h,+}, P_{h,-}) = 2n_h L_h^2 \int_{[0,1]^{d_h}} \varphi_h(z)^2 \, dz.$$

Using (I.5),

$$\mathrm{KL}(P_{h,+}, P_{h,-}) = 2L_h^2 \beta_{d_h}\, n_h \delta_h^{d_h+2}.$$

By the definition of δ_h in (I.4),

$$n_h \delta_h^{d_h+2} = \gamma_h^{d_h+2} \le \frac{1}{16 H L_h^2 \beta_{d_h}}.$$

Therefore

$$\mathrm{KL}(P_{h,+}, P_{h,-}) \le \frac{1}{8H}.$$

Summing over $h = 1, \dots, H$ proves (I.8). □

I.4. Oracle Minimax Lower Bound

We can now prove the endpoint-estimation lower bound.

Theorem I.5 (Oracle minimax lower bound for endpoint estimation)**.** *Assume assumptions 2.1 and 3.1. Consider the oracle regression class $\mathfrak{O}_{\mathbf{n}}^L$ from definition I.1. Then there exists a subclass*

$$\mathcal{M}_{\mathrm{tp}}(\mathbf{n}) \subseteq \mathfrak{O}_{\mathbf{n}}^L$$

consisting of the two models $\{M^+, M^-\}$ above, such that for every measurable interval estimator $(\hat{\ell}, \hat{u})$ based on the full oracle dataset,

$$\sup_{M \in \mathcal{M}_{\mathrm{tp}}(\mathbf{n})} \mathbb{E}_M \Big[\Big| \hat{\ell} - \underline{V}_1^\pi(\mathbf{0}) \Big| + \Big| \hat{u} - \overline{V}_1^\pi(\mathbf{0}) \Big| \Big] \ge c_0 \sum_{h=1}^{H} L_h\, n_h^{-1/(d_h+2)}, \tag{I.9}$$

where

$$c_0 := \frac{3}{4} \min_{h \in [H]} \gamma_h$$

and $(\gamma_h)_{h=1}^H$ is given by (I.4). In particular, since H, L, and $(d_h)_{h=1}^H$ are fixed model-class parameters, there exists a constant

$$c_\star = c_\star(H, L, d_1, \dots, d_H, R_{\max}) > 0$$

such that

$$\inf_{(\hat{\ell}, \hat{u})} \sup_{M \in \mathfrak{O}_{\mathbf{n}}^L} \mathbb{E}_M \Big[\Big| \hat{\ell} - \underline{V}_1^\pi(\mathbf{0}) \Big| + \Big| \hat{u} - \overline{V}_1^\pi(\mathbf{0}) \Big| \Big] \ge c_\star \sum_{h=1}^{H} n_h^{-1/(d_h+2)}.$$

Proof. Fix an arbitrary interval estimator $(\hat{\ell}, \hat{u})$ and define its midpoint estimator

$$\hat{\theta} := \frac{\hat{\ell} + \hat{u}}{2}.$$

For any real number θ,

$$\Big| \hat{\ell} - \theta \Big| + |\hat{u} - \theta| \ge 2 \Big| \hat{\theta} - \theta \Big|.$$

Applying this pointwise with

$$\theta = \underline{V}_1^\pi(\mathbf{0}) = \overline{V}_1^\pi(\mathbf{0}) = \vartheta(M)$$

for $M \in \{M^+, M^-\}$ and then taking expectations gives

$$\sup_{M \in \{M^+, M^-\}} \mathbb{E}_M \Big[\Big| \hat{\ell} - \underline{V}_1^\pi(\mathbf{0}) \Big| + \Big| \hat{u} - \overline{V}_1^\pi(\mathbf{0}) \Big| \Big] \ge 2 \sup_{M \in \{M^+, M^-\}} \mathbb{E}_M \Big| \hat{\theta} - \vartheta(M) \Big|. \tag{I.10}$$

Now apply lemma I.3 to the scalar estimator $\hat{\theta}$ under the two models M^+ and M^-. By lemma I.4,

$$\vartheta(M^+) - \vartheta(M^-) = 2 \sum_{h=1}^{H} L_h \delta_h,$$

and

$$\mathrm{KL}(P_{M^+}, P_{M^-}) \le \frac{1}{8}.$$

Hence Pinsker's inequality gives

$$\mathrm{TV}(P_{M^+}, P_{M^-}) \le \sqrt{\frac{1}{16}} = \frac{1}{4}.$$

Substituting into (I.1),

$$\sup_{M \in \{M^+, M^-\}} \mathbb{E}_M \left| \hat{\theta} - \vartheta(M) \right| \ge \frac{1}{4}\left(1 - \frac{1}{4}\right) \cdot 2 \sum_{h=1}^{H} L_h \delta_h = \frac{3}{8} \sum_{h=1}^{H} L_h \delta_h.$$

Combining this with (I.10) yields

$$\sup_{M \in \{M^+, M^-\}} \mathbb{E}_M \left[\left| \hat{\ell} - \underline{V}_1^{\pi}(\mathbf{0}) \right| + \left| \hat{u} - \overline{V}_1^{\pi}(\mathbf{0}) \right| \right] \ge \frac{3}{4} \sum_{h=1}^{H} L_h \delta_h.$$

Since $\delta_h = \gamma_h n_h^{-1/(d_h+2)}$, this is exactly (I.9).

The final minimax lower bound over the entire oracle class follows because $\mathcal{M}_{\mathrm{tp}}(\mathbf{n}) = \{M^+, M^-\}$ is a subclass of $\mathfrak{O}_{\mathbf{n}}^L$. □

I.5. Separation Between Width and Estimation Difficulty

The minimax lower bound above concerns endpoint estimation on a zero-width subclass. It is therefore complementary to the sharpness theorem of Appendix G, which constructs explicit families with nontrivial identified-set width.

Corollary I.6 (Separation of estimable and irreducible components)**.** *The Bellman–Whitney framework contains two qualitatively distinct sources of difficulty:*

1. *a zero-width subclass $\mathfrak{O}_{\mathbf{n}}^L$ on which endpoint estimation alone incurs minimax error at least*

$$c_\star \sum_{h=1}^{H} n_h^{-1/(d_h+2)};$$

2. *serial singleton-support families on which the identified-set width is exactly*

$$w_h^{\pi}(x_h^\star) = 2 \sum_{t=h}^{H} L_t d_t(x_t^\star, c_t)$$

by theorem G.8.

Hence endpoint-estimation error and irreducible support-hole ambiguity are genuinely distinct phenomena, and neither can in general be removed by the other.

Proof. The first statement is exactly theorem I.5. The second is theorem G.8. Since the first family has zero width but nontrivial endpoint-estimation difficulty, whereas the second family can be constructed with exact population knowledge of the Bellman data but nontrivial identified-set width, the two sources of difficulty are distinct. □

J. Certifiable Control under Ambiguity

This appendix derives a control consequence of the Bellman–Whitney theory. Unlike the fixed-policy analysis in the main text, control requires comparing actions at a state and hence replacing target-policy averaging by the Bellman optimality operator. This changes one important structural feature: the resulting feasible tail class is generally *not* convex, so we do not claim a sharp interval characterization for optimal values analogous to theorem 3.2. What we *can* prove is exactly what is

needed for certification: a backward Bellman–Whitney recursion that yields lower and upper action-value envelopes, and pointwise sandwich inequalities showing that every control-compatible tail lies between them.

The strengthened assumption below is the control analogue of the fixed-policy rectangular closure condition. At a conceptual level, it is the function-level counterpart of the rectangularity assumptions that underlie time-consistent dynamic programming in classical robust MDP theory (Iyengar, 2005; Nilim & Ghaoui, 2005).

J.1. Control-Compatible Tails

For each stage $h \in [H]$, define the control value class

$$\mathcal{V}_h^{\mathrm{ctl},L} := \left\{ v : \mathcal{S}_h \to \mathbb{R} : \exists\, q \in \mathrm{Lip}_{L_h}(\mathcal{X}_h) \text{ such that } v(s) = \max_{a\in\mathcal{A}} q(s,a)\ \forall s \in \mathcal{S}_h \right\},$$

with terminal convention

$$\mathcal{V}_{H+1}^{\mathrm{ctl},L} := \{0\}.$$

Since $\mathcal{A}$ is finite and the maximum of finitely many L_h-Lipschitz functions is again L_h-Lipschitz on $\mathcal{S}_h$, every $v \in \mathcal{V}_h^{\mathrm{ctl},L}$ is bounded and L_h-Lipschitz.

Assumption J.1 (Rectangular control closure)**.** For every stage $h \in [H]$ and every continuation value $v \in \mathcal{V}_{h+1}^{\mathrm{ctl},L}$, the supported Bellman target

$$c \longmapsto (B_h^\pi v)(c), \qquad c \in C_h,$$

is L_h-Lipschitz with respect to the restricted metric on C_h.

Because B_h^π conditions on the full state–action pair $X_h = (S_h, A_h)$, it does not depend on the future action-selection rule. Accordingly, the same supported Bellman operator that appeared in the fixed-policy theory can be reused here.

Definition J.2 (Control-compatible tail)**.** Fix $h \in [H]$. The control-compatible tail class $\mathfrak{C}_{h:H}^{\mathrm{ctl},L}$ consists of all sequences

$$(Q_t, V_t)_{t=h}^{H+1}$$

such that:

1. $V_{H+1} \equiv 0$;
2. for every $t \in \{h, \dots, H\}$, the function $Q_t : \mathcal{X}_t \to \mathbb{R}$ is bounded, Borel measurable, and L_t-Lipschitz;
3. for every $t \in \{h, \dots, H\}$,
$$V_t(s) = \max_{a\in\mathcal{A}} Q_t(s,a), \qquad s \in \mathcal{S}_t;$$
4. for every $t \in \{h, \dots, H\}$,
$$Q_t(x) = (B_t^\pi V_{t+1})(x), \qquad x \in C_t.$$

J.2. Control Bellman–Whitney Recursion

Define recursively

$$\underline{V}_{H+1}^{\mathrm{ctl}} = \overline{V}_{H+1}^{\mathrm{ctl}} \equiv 0.$$

For $h = H, H-1, \dots, 1$, define the supported lower and upper control Bellman targets

$$g_{h,\mathrm{ctl}}^-(c) := (B_h^\pi \underline{V}_{h+1}^{\mathrm{ctl}})(c), \qquad g_{h,\mathrm{ctl}}^+(c) := (B_h^\pi \overline{V}_{h+1}^{\mathrm{ctl}})(c), \qquad c \in C_h,$$

and then the lower and upper control Bellman–Whitney envelopes

$$\underline{Q}_h^{\mathrm{ctl}} := \mathcal{W}_{h,L_h}^- g_{h,\mathrm{ctl}}^-, \qquad \overline{Q}_h^{\mathrm{ctl}} := \mathcal{W}_{h,L_h}^+ g_{h,\mathrm{ctl}}^+, \tag{J.1}$$

together with the induced lower and upper control values

$$\underline{V}_h^{\mathrm{ctl}}(s) := \max_{a\in\mathcal{A}} \underline{Q}_h^{\mathrm{ctl}}(s,a), \qquad \overline{V}_h^{\mathrm{ctl}}(s) := \max_{a\in\mathcal{A}} \overline{Q}_h^{\mathrm{ctl}}(s,a), \qquad s \in \mathcal{S}_h. \tag{J.2}$$

The lower envelope $\underline{Q}_h^{\mathrm{ctl}}$ should be interpreted as a *uniformly certified action lower bound*, while $\overline{Q}_h^{\mathrm{ctl}}$ is the corresponding uniform action upper bound. The theorem below shows that every control-compatible tail lies between them.

Theorem J.3 (Control Bellman–Whitney sandwich)**.** *Assume assumptions 2.1 and J.1. Then:*

1. *the recursive lower and upper control tails are feasible:*

$$(\underline{Q}_t^{\mathrm{ctl}}, \underline{V}_t^{\mathrm{ctl}})_{t=h}^{H+1} \in \mathfrak{C}_{h:H}^{\mathrm{ctl},L}, \qquad (\overline{Q}_t^{\mathrm{ctl}}, \overline{V}_t^{\mathrm{ctl}})_{t=h}^{H+1} \in \mathfrak{C}_{h:H}^{\mathrm{ctl},L}, \qquad h \in [H];$$

2. *for every $h \in [H]$ and every $(Q_t, V_t)_{t=h}^{H+1} \in \mathfrak{C}_{h:H}^{\mathrm{ctl},L}$,*

$$\underline{Q}_t^{\mathrm{ctl}}(x) \le Q_t(x) \le \overline{Q}_t^{\mathrm{ctl}}(x) \qquad \forall x \in \mathcal{X}_t,\ \forall t \in \{h, \dots, H\}, \tag{J.3}$$

and therefore

$$\underline{V}_t^{\mathrm{ctl}}(s) \le V_t(s) \le \overline{V}_t^{\mathrm{ctl}}(s) \qquad \forall s \in \mathcal{S}_t,\ \forall t \in \{h, \dots, H\}. \tag{J.4}$$

Proof. We begin with feasibility. The proof is by backward induction on h.

Base case: $h = H$**.** Since $\underline{V}_{H+1}^{\mathrm{ctl}} = \overline{V}_{H+1}^{\mathrm{ctl}} \equiv 0$, the zero function belongs to $\mathcal{V}_{H+1}^{\mathrm{ctl},L}$. By assumption J.1, both supported targets

$$g_{H,\mathrm{ctl}}^{-}(c) = (B_H^{\pi} 0)(c), \qquad g_{H,\mathrm{ctl}}^{+}(c) = (B_H^{\pi} 0)(c), \qquad c \in C_H,$$

are L_H-Lipschitz on C_H. Therefore, by corollary B.4, the envelopes $\underline{Q}_H^{\mathrm{ctl}}$ and $\overline{Q}_H^{\mathrm{ctl}}$ are bounded, Borel measurable, L_H-Lipschitz on $\mathcal{X}_H$, and satisfy

$$\underline{Q}_H^{\mathrm{ctl}}(x) = g_{H,\mathrm{ctl}}^{-}(x), \qquad \overline{Q}_H^{\mathrm{ctl}}(x) = g_{H,\mathrm{ctl}}^{+}(x), \qquad x \in C_H.$$

By definition,

$$\underline{V}_H^{\mathrm{ctl}}(s) = \max_a \underline{Q}_H^{\mathrm{ctl}}(s,a), \qquad \overline{V}_H^{\mathrm{ctl}}(s) = \max_a \overline{Q}_H^{\mathrm{ctl}}(s,a),$$

so both terminal control tails are feasible.

Induction step. Fix $h \in \{1, \dots, H-1\}$ and assume the recursive lower and upper control tails are feasible from stage $h+1$ onward. Then $\underline{Q}_{h+1}^{\mathrm{ctl}}$ and $\overline{Q}_{h+1}^{\mathrm{ctl}}$ are L_{h+1}-Lipschitz, so by definition

$$\underline{V}_{h+1}^{\mathrm{ctl}},\ \overline{V}_{h+1}^{\mathrm{ctl}} \in \mathcal{V}_{h+1}^{\mathrm{ctl},L}.$$

Applying assumption J.1 shows that the supported targets $g_{h,\mathrm{ctl}}^{-}$ and $g_{h,\mathrm{ctl}}^{+}$ are L_h-Lipschitz on C_h. By corollary B.4, the envelopes $\underline{Q}_h^{\mathrm{ctl}}$ and $\overline{Q}_h^{\mathrm{ctl}}$ are exact L_h-Lipschitz extensions of these supported targets. Defining $\underline{V}_h^{\mathrm{ctl}}$ and $\overline{V}_h^{\mathrm{ctl}}$ via (J.2) therefore yields feasible control tails from stage h onward.

This proves the first claim.

We next prove the sandwich inequalities, again by backward induction on t.

Base case: $t = H$**.** Let $(Q_H, V_H, V_{H+1} \equiv 0) \in \mathfrak{C}_{H:H}^{\mathrm{ctl},L}$. On the support,

$$Q_H(c) = (B_H^{\pi} 0)(c) = g_{H,\mathrm{ctl}}^{-}(c) = g_{H,\mathrm{ctl}}^{+}(c), \qquad c \in C_H.$$

Since Q_H is L_H-Lipschitz on $\mathcal{X}_H$, the extremal characterization in lemma B.3 gives

$$\underline{Q}_H^{\mathrm{ctl}} \le Q_H \le \overline{Q}_H^{\mathrm{ctl}} \qquad \text{pointwise on } \mathcal{X}_H.$$

Taking maxima over actions yields

$$\underline{V}_H^{\mathrm{ctl}} \le V_H \le \overline{V}_H^{\mathrm{ctl}} \qquad \text{pointwise on } \mathcal{S}_H.$$

Induction step. Fix $t \in \{h, \dots, H-1\}$ and assume

$$\underline{V}^{\mathrm{ctl}}_{t+1} \le V_{t+1} \le \overline{V}^{\mathrm{ctl}}_{t+1} \qquad \text{pointwise on } \mathcal{S}_{t+1}.$$

By monotonicity of conditional expectation,

$$(B^\pi_t \underline{V}^{\mathrm{ctl}}_{t+1})(c) \le (B^\pi_t V_{t+1})(c) \le (B^\pi_t \overline{V}^{\mathrm{ctl}}_{t+1})(c), \qquad c \in C_t.$$

Since (Q_t, V_t) is control-compatible,

$$Q_t(c) = (B^\pi_t V_{t+1})(c), \qquad c \in C_t,$$

and therefore

$$g^-_{t,\mathrm{ctl}}(c) \le Q_t(c) \le g^+_{t,\mathrm{ctl}}(c), \qquad c \in C_t.$$

Because Q_t is L_t-Lipschitz on $\mathcal{X}_t$, lemma B.3 implies

$$\underline{Q}^{\mathrm{ctl}}_t = \mathcal{W}^-_{t,L_t} g^-_{t,\mathrm{ctl}} \le Q_t \le \mathcal{W}^+_{t,L_t} g^+_{t,\mathrm{ctl}} = \overline{Q}^{\mathrm{ctl}}_t \qquad \text{pointwise on } \mathcal{X}_t.$$

Taking maxima over actions yields

$$\underline{V}^{\mathrm{ctl}}_t(s) = \max_a \underline{Q}^{\mathrm{ctl}}_t(s,a) \le \max_a Q_t(s,a) = V_t(s) \le \max_a \overline{Q}^{\mathrm{ctl}}_t(s,a) = \overline{V}^{\mathrm{ctl}}_t(s),$$

for all $s \in \mathcal{S}_t$.

This completes the induction. □

Remark J.4 (No sharp interval claim for control). The fixed-policy sharp interval theorem relied on convexity of the feasible tail class. In control, the map $Q \mapsto \max_a Q(\cdot, a)$ is nonlinear, and $\mathfrak{C}^{\mathrm{ctl},L}_{h:H}$ is generally not convex. Accordingly, theorem J.3 is stated as a uniform feasibility and sandwich result rather than as an exact interval characterization for optimal values. This is the mathematically correct level of generality for certification.

J.3. Certifiably Good, Bad, and Ambiguous Actions

We now define the action-level certification sets induced by the control Bellman–Whitney envelopes.

Definition J.5 (Uniform action certificates)**.** Fix a stage $h \in [H]$, a state $s \in \mathcal{S}_h$, and a tolerance $\delta \ge 0$. Define

$$\mathcal{A}^{\mathrm{good}}_h(s;\delta) := \left\{ a \in \mathcal{A} : \underline{Q}^{\mathrm{ctl}}_h(s,a) \ge \max_{a' \in \mathcal{A}} \overline{Q}^{\mathrm{ctl}}_h(s,a') - \delta \right\},$$

$$\mathcal{A}^{\mathrm{bad}}_h(s;\delta) := \left\{ a \in \mathcal{A} : \overline{Q}^{\mathrm{ctl}}_h(s,a) < \max_{a' \in \mathcal{A}} \underline{Q}^{\mathrm{ctl}}_h(s,a') - \delta \right\},$$

and

$$\mathcal{A}^{\mathrm{amb}}_h(s;\delta) := \mathcal{A} \setminus \left(\mathcal{A}^{\mathrm{good}}_h(s;\delta) \cup \mathcal{A}^{\mathrm{bad}}_h(s;\delta) \right).$$

The next lemma records the exact certification meaning of these sets.

Lemma J.6 (Uniform good/bad action guarantees)**.** *Assume assumptions 2.1 and J.1, and fix $h \in [H]$, $s \in \mathcal{S}_h$, and $\delta \ge 0$.*

1. *If $a \in \mathcal{A}^{\mathrm{good}}_h(s;\delta)$, then for every $(Q_t, V_t)_{t=h}^{H+1} \in \mathfrak{C}^{\mathrm{ctl},L}_{h:H}$,*

$$Q_h(s,a) \ge V_h(s) - \delta. \tag{J.5}$$

   Thus a is uniformly δ-optimal across the entire control-compatible class.

2. *If $a \in \mathcal{A}^{\mathrm{bad}}_h(s;\delta)$, then for every $(Q_t, V_t)_{t=h}^{H+1} \in \mathfrak{C}^{\mathrm{ctl},L}_{h:H}$,*

$$Q_h(s,a) \le V_h(s) - \delta. \tag{J.6}$$

More strongly, if

$$a^\dagger \in \arg\max_{a' \in \mathcal{A}} \underline{Q}_h^{\mathrm{ctl}}(s, a'),$$

then

$$Q_h(s, a^\dagger) \geq Q_h(s, a) + \delta \tag{J.7}$$

for every control-compatible tail.

Proof. Fix a control-compatible tail

$$(Q_t, V_t)_{t=h}^{H+1} \in \mathfrak{C}_{h:H}^{\mathrm{ctl},L}.$$

By theorem J.3,

$$\underline{Q}_h^{\mathrm{ctl}}(s, a') \leq Q_h(s, a') \leq \overline{Q}_h^{\mathrm{ctl}}(s, a') \qquad \forall a' \in \mathcal{A},$$

and

$$\underline{V}_h^{\mathrm{ctl}}(s) \leq V_h(s) \leq \overline{V}_h^{\mathrm{ctl}}(s).$$

Also,

$$V_h(s) = \max_{a' \in \mathcal{A}} Q_h(s, a').$$

Suppose first that $a \in \mathcal{A}_h^{\mathrm{good}}(s; \delta)$. Then

$$\underline{Q}_h^{\mathrm{ctl}}(s, a) \geq \max_{a'} \overline{Q}_h^{\mathrm{ctl}}(s, a') - \delta.$$

Hence

$$V_h(s) = \max_{a'} Q_h(s, a') \leq \max_{a'} \overline{Q}_h^{\mathrm{ctl}}(s, a') \leq \underline{Q}_h^{\mathrm{ctl}}(s, a) + \delta \leq Q_h(s, a) + \delta,$$

which proves (J.5).

Now suppose that $a \in \mathcal{A}_h^{\mathrm{bad}}(s; \delta)$. Let

$$a^\dagger \in \arg\max_{a' \in \mathcal{A}} \underline{Q}_h^{\mathrm{ctl}}(s, a').$$

By definition of $\mathcal{A}_h^{\mathrm{bad}}(s; \delta)$,

$$\overline{Q}_h^{\mathrm{ctl}}(s, a) < \underline{Q}_h^{\mathrm{ctl}}(s, a^\dagger) - \delta.$$

Therefore

$$Q_h(s, a) \leq \overline{Q}_h^{\mathrm{ctl}}(s, a) < \underline{Q}_h^{\mathrm{ctl}}(s, a^\dagger) - \delta \leq Q_h(s, a^\dagger) - \delta,$$

which proves (J.7). Since $V_h(s) \geq Q_h(s, a^\dagger)$, we also have

$$Q_h(s, a) \leq Q_h(s, a^\dagger) - \delta \leq V_h(s) - \delta,$$

proving (J.6). □

Corollary J.7 (Certifiable control under ambiguity)**.** *Assume assumptions 2.1 and J.1. Fix $h \in [H]$, $s \in \mathcal{S}_h$, and $\delta \geq 0$. Then:*

1. *every action in $\mathcal{A}_h^{\mathrm{good}}(s; \delta)$ is uniformly δ-optimal over the entire control-compatible class $\mathfrak{C}_{h:H}^{\mathrm{ctl},L}$;*
2. *every action in $\mathcal{A}_h^{\mathrm{bad}}(s; \delta)$ is uniformly δ-suboptimal over the entire control-compatible class $\mathfrak{C}_{h:H}^{\mathrm{ctl},L}$;*
3. *every action in $\mathcal{A}_h^{\mathrm{amb}}(s; \delta)$ is neither certifiably δ-optimal nor certifiably δ-suboptimal on the basis of the Bellman–Whitney envelopes alone.*

Proof. The first two claims are exactly lemma J.6. The third is immediate from the definition of $\mathcal{A}_h^{\mathrm{amb}}(s; \delta)$ as the complement of the certifiably good and certifiably bad sets. □

J.4. Interpretation

The corollary above provides the precise control consequence needed in the main text. The lower and upper control Bellman–Whitney envelopes do not merely bound the value of a fixed target policy; they induce action-level certificates. If an action survives the lower-vs-upper comparison in $\mathcal{A}_h^{\mathrm{good}}(s;\delta)$, then it is uniformly near-optimal across all control-compatible tails. If it falls into $\mathcal{A}_h^{\mathrm{bad}}(s;\delta)$, then it is uniformly suboptimal. The remaining actions are not artifacts of loose analysis: they are exactly those whose status cannot be resolved from the Bellman–Lipschitz ambiguity class alone.

K. Counterexamples and Failure Modes

This appendix records four short constructions that clarify the precise scope of the Bellman–Whitney theory. The first shows that genuine support holes place the problem outside density-ratio point-identification frameworks even when the Bellman–Whitney interval is finite and sharp. The second shows that comparator-style partial coverage does not identify arbitrary uncovered target policies (Uehara & Sun, 2022). The third explains why the control section of the paper is stated as an action-certification theorem rather than a sharp interval theorem. The fourth shows that Bellman collapse is not merely a convenient proof device: without it, a direct history-space reduction can be exponentially larger in the horizon, even when the underlying Markov state space is constant, in contrast to one-step contextual smoothness formulations (Khan et al., 2024).

K.1. Genuine Support Holes Destroy Density Ratios

Proposition K.1 (No finite density ratio under a genuine support hole)**.** *Consider the one-step problem $H = 1$ with:*

1. *a single state $s_\circ$;*
2. *action space $\mathcal{A} = \{0, 1\}$;*
3. *behavior support*
$$C_1 = \{(s_\circ, 0)\};$$
4. *supported reward mean*
$$\mathbb{E}[R_1 \mid X_1 = (s_\circ, 0)] = 0;$$
5. *Bellman–Lipschitz radius $L_1 = L > 0$.*

Let the target policy choose the uncovered action:

$$\pi_1(\cdot \mid s_\circ) = \delta_1.$$

Then:

1. *the target occupancy measure is not absolutely continuous with respect to the behavior occupancy measure, so no density ratio $d\rho_1^\pi / d\rho_1^{\mathrm{b}}$ exists;*
2. *the Bellman–Whitney identified interval is nevertheless finite and sharp:*
$$\mathcal{I}^\pi = [-L, L].$$

Proof. Because the behavior support is the singleton $(s_\circ, 0)$, the stagewise behavior occupancy is

$$\rho_1^{\mathrm{b}} = \delta_{(s_\circ,0)},$$

whereas the target occupancy is

$$\rho_1^\pi = \delta_{(s_\circ,1)}.$$

Suppose, for contradiction, that there were a measurable density ratio $w = d\rho_1^\pi / d\rho_1^{\mathrm{b}}$. Then

$$1 = \rho_1^\pi(\{(s_\circ, 1)\}) = \int \mathbb{I}\{(s_\circ, 1)\}(x)\, w(x)\, \rho_1^{\mathrm{b}}(dx) = 0,$$

since ρ_1^{b} puts no mass on $(s_\circ, 1)$. Thus $\rho_1^\pi \not\ll \rho_1^{\mathrm{b}}$ and no density ratio exists.

We now compute the Bellman–Whitney interval. Since $H = 1$, the supported mean is the scalar function

$$m_1(c) \equiv 0 \qquad \text{for } c = (s_\circ, 0) \in C_1.$$

Because the product metric from Appendix A separates the two actions by one unit,

$$d_1((s_\circ, 1), (s_\circ, 0)) = 1.$$

Therefore the lower and upper Bellman–Whitney envelopes at the uncovered target point are

$$\underline{Q}_1^\pi(s_\circ, 1) = \sup_{c \in C_1} \{m_1(c) - L\, d_1((s_\circ, 1), c)\} = 0 - L = -L,$$

and

$$\overline{Q}_1^\pi(s_\circ, 1) = \inf_{c \in C_1} \{m_1(c) + L\, d_1((s_\circ, 1), c)\} = 0 + L = L.$$

Since the target policy is deterministic at $s_\circ$,

$$\mathcal{I}^\pi = \left[\underline{V}_1^\pi(s_\circ), \overline{V}_1^\pi(s_\circ)\right] = \left[\underline{Q}_1^\pi(s_\circ, 1), \overline{Q}_1^\pi(s_\circ, 1)\right] = [-L, L].$$

This interval is sharp by theorem 3.2. □

Remark K.2. Proposition K.1 isolates the precise regime in which the present paper operates. The target value is not point-identified by overlap-based methods because the relevant density ratio does not exist, yet the Bellman–Whitney interval remains finite and informative.

K.2. Covered Comparators Do Not Identify Uncovered Targets

Proposition K.3 (Comparator coverage does not imply target-policy identification)**.** *Consider the two-stage problem $H = 2$ with:*

1. *singleton state spaces*
$$\mathcal{S}_1 = \{s_1\}, \qquad \mathcal{S}_2 = \{s_2\};$$
2. *action space $\mathcal{A} = \{0, 1\}$;*
3. *behavior support*
$$C_1 = \{(s_1, 0)\}, \qquad C_2 = \{(s_2, 0)\};$$
4. *zero stage-one reward and deterministic transition from $(s_1, 0)$ to s_2;*
5. *supported stage-two reward mean*
$$\mathbb{E}[R_2 \mid X_2 = (s_2, 0)] = 0;$$
6. *Bellman–Lipschitz radii $L_1, L_2 \geq 1$.*

Let π^{cov} denote the covered comparator policy that chooses action 0 at both stages, and let π^{tar} denote the target policy that also chooses action 0 at stage 1 but chooses the uncovered action 1 at stage 2. Then:

1. *the covered comparator is point-identified with value*
$$V_1^{\pi^{\mathrm{cov}}}(s_1) = 0;$$
2. *the uncovered target policy is not point-identified and satisfies*
$$\mathcal{I}^{\pi^{\mathrm{tar}}} = [-L_2, L_2].$$

Proof. Under π^{cov}, the only visited state–action points are the observed support points $(s_1, 0)$ and $(s_2, 0)$. Since both stagewise support Bellman targets are exactly known and equal to zero, the value of the covered comparator is

$$V_1^{\pi^{\mathrm{cov}}}(s_1) = 0.$$

For the uncovered target π^{tar}, stage 1 is still covered, but stage 2 visits the unsupported action $(s_2, 1)$. At stage 2, the same calculation as in proposition K.1 gives

$$\underline{Q}_2^{\pi}(s_2, 1) = -L_2, \qquad \overline{Q}_2^{\pi}(s_2, 1) = L_2.$$

Hence

$$\underline{V}_2^{\pi}(s_2) = -L_2, \qquad \overline{V}_2^{\pi}(s_2) = L_2,$$

because the target policy is deterministic at stage 2.

At stage 1, the target takes the covered action $(s_1, 0)$, the stage-one reward is zero, and the transition to s_2 is deterministic. Therefore the supported Bellman target at stage 1 is exactly the continuation value at s_2, so

$$\underline{V}_1^{\pi}(s_1) = \underline{V}_2^{\pi}(s_2) = -L_2, \qquad \overline{V}_1^{\pi}(s_1) = \overline{V}_2^{\pi}(s_2) = L_2.$$

By theorem 3.2,

$$\mathcal{I}^{\pi^{\mathrm{tar}}} = \left[\underline{V}_1^{\pi}(s_1), \overline{V}_1^{\pi}(s_1)\right] = [-L_2, L_2].$$

□

Remark K.4. Proposition K.3 shows that guarantees relative to covered comparators do not identify the value of an arbitrary target policy that enters a support hole, even if the target shares the entire earlier trajectory with a covered policy and deviates only at a single final decision.

K.3. The Control-Compatible Class Is Not Convex

The next example explains why the control section proves a sandwich/certification theorem rather than a sharp interval theorem for optimal values.

Proposition K.5 (Nonconvexity of the control-compatible tail class)**.** *Consider the one-step control problem $H = 1$ with:*

1. *a single state $s_\circ$;*
2. *action space $\mathcal{A} = \{0, 1\}$;*
3. *support*
$$C_1 = \{(s_\circ, 0)\};$$
4. *supported reward mean*
$$\mathbb{E}[R_1 \mid X_1 = (s_\circ, 0)] = 0;$$
5. *Bellman–Lipschitz radius $L_1 = 2$.*

Then the control-compatible class $\mathfrak{C}_{1:H}^{\mathrm{ctl},L}$ from Appendix J is not convex.

Proof. Define two state–action value functions

$$Q^+(s_\circ, 0) = 0, \qquad Q^+(s_\circ, 1) = 2,$$

and

$$Q^-(s_\circ, 0) = 0, \qquad Q^-(s_\circ, 1) = -2.$$

Because the action separation in the product metric is one,

$$\mathrm{Lip}_1(Q^+) = 2, \qquad \mathrm{Lip}_1(Q^-) = 2.$$

Both functions are therefore L_1-Lipschitz. On the observed support, both satisfy the Bellman equality

$$Q^+(s_\circ, 0) = 0 = Q^-(s_\circ, 0),$$

which matches the supported reward mean.

Now define the corresponding control values by pointwise maximization:

$$V^+(s_\circ) = \max\{Q^+(s_\circ, 0), Q^+(s_\circ, 1)\} = 2,$$

and

$$V^-(s_\circ) = \max\{Q^-(s_\circ, 0), Q^-(s_\circ, 1)\} = 0.$$

Hence both pairs

$$(Q^+, V^+, V_2 \equiv 0), \qquad (Q^-, V^-, V_2 \equiv 0)$$

belong to $\mathfrak{C}^{\mathrm{ctl},L}_{1:H}$.

Consider now the midpoint pair

$$\bar{Q} := \frac{1}{2}(Q^+ + Q^-), \qquad \bar{V} := \frac{1}{2}(V^+ + V^-).$$

Then

$$\bar{Q}(s_\circ, 0) = 0, \qquad \bar{Q}(s_\circ, 1) = 0, \qquad \bar{V}(s_\circ) = 1.$$

The function $\bar{Q}$ is still 2-Lipschitz and satisfies Bellman equality on the support, but

$$\max_{a \in \mathcal{A}} \bar{Q}(s_\circ, a) = 0 \neq 1 = \bar{V}(s_\circ).$$

Thus the averaged pair $(\bar{Q}, \bar{V}, V_2 \equiv 0)$ is not control-compatible. Therefore $\mathfrak{C}^{\mathrm{ctl},L}_{1:H}$ is not convex. □

Remark K.6. The obstruction in proposition K.5 is exactly the nonlinearity of the maximization operator $Q \mapsto \max_a Q(\cdot, a)$. This is why the control theory in Section 7 is framed in terms of certification sets rather than sharp scalar intervals.

K.4. History-Space Reduction Can Be Exponentially Larger

The final example shows that Bellman collapse is not merely a proof convenience. Without exploiting Markov structure, a direct history-space contextual reduction can grow exponentially with the horizon even when the Markov state space is constant.

Proposition K.7 (Exponential history-space blow-up without Bellman collapse)**.** *Fix any horizon $H \geq 1$. Consider the Markov family with:*

1. *a single state at each stage,*
$$\mathcal{S}_h = \{s_h\}, \qquad h \in [H];$$
2. *action space $\mathcal{A} = \{0, 1\}$;*
3. *behavior policy that randomizes independently and uniformly at every stage:*
$$\mu_h(\cdot \mid s_h) = \mathrm{Unif}\{0, 1\};$$
4. *deterministic state transition from s_h to s_{h+1} for every action.*

Then:

1. *the stagewise Markov state–action space has constant size*
$$|\mathcal{X}_h| = 2 \qquad \forall h \in [H];$$

2. *the stage-h history space*
$$\mathcal{H}_h := \{(a_1, \ldots, a_{h-1}, s_h) : a_i \in \{0, 1\}\}$$
has cardinality
$$|\mathcal{H}_h| = 2^{h-1};$$

3. *consequently, any direct one-step contextual reduction that treats distinct histories as distinct contexts at stage h (Khan et al., 2024) must represent at least*
$$|\mathcal{H}_h \times \mathcal{A}| = 2^h$$
history–action pairs, whereas the Bellman–Whitney recursion represents the same stage using only the two Markov state–action points in $\mathcal{X}_h$.

Proof. The first claim is immediate: each stage has one state and two actions, so
$$|\mathcal{X}_h| = |\{s_h\} \times \{0, 1\}| = 2.$$

For the second claim, a stage-h history consists of the fixed current state s_h together with the sequence of past actions $(a_1, \ldots, a_{h-1})$. Since each past action can be chosen independently from $\{0, 1\}$, the number of distinct histories is exactly
$$2^{h-1}.$$

The third claim follows because a one-step contextual reduction over histories must index the stage-h decision problem by the context-history pair $(H_h, a_h) \in \mathcal{H}_h \times \mathcal{A}$, which has cardinality
$$|\mathcal{H}_h \times \mathcal{A}| = |\mathcal{H}_h|\,|\mathcal{A}| = 2^{h-1} \cdot 2 = 2^h.$$

By contrast, the Bellman–Whitney recursion uses only the Markov state–action domain $\mathcal{X}_h$, whose size is 2. □

Remark K.8. Proposition K.7 shows that Bellman collapse is not cosmetic. Even on a constant-state MDP, a direct history-space treatment distinguishes exponentially many contexts, while the Bellman–Whitney recursion remains stagewise Markov and therefore linear in the horizon.

L. Extended Related Work

This appendix positions the Bellman–Whitney framework relative to the closest neighboring theory lines. The most relevant comparisons are not generic "offline RL" references, but works that are close along one of the following axes:

1. partial identification without overlap;
2. sequential off-policy evaluation under weak but nonzero overlap;
3. interval methods for bias or misspecification in point-identified OPE;
4. sequential lower/upper bounds under confounding or hidden bias;
5. generic conditional-linear-program formulations of partial identification;
6. offline RL under partial coverage;
7. robust dynamic programming under exogenous ambiguity.

The paper is closest to these lines, but not reducible to any of them.

One-step no-overlap partial identification under smoothness. The nearest one-step antecedent is Khan et al. (2024), who study off-policy evaluation without overlap in the contextual setting under nonparametric smoothness assumptions, with a particular focus on Lipschitz smoothness. Their main contribution is a sharp upper/lower characterization of the one-step value under no overlap, together with optimal estimators of those bounds. Our corollary 3.3 shows that the Bellman–Whitney framework recovers exactly that one-step geometry when $H = 1$. The distinction is that our main object is not a contextual LP or a single-stage smoothness program, but a *sequential Bellman collapse*: under Bellman–Lipschitz closure, the full long-horizon identified set reduces to a backward dynamic envelope recursion. This Bellman collapse is the structural novelty; Appendix K shows that a direct history-space reduction can be exponentially larger in the horizon.

Weak-overlap sequential OPE. A different line studies sequential OPE when overlap is weakened but not destroyed. The closest such paper is Mehrabi & Wager (2024), who study off-policy evaluation in MDPs under *weak distributional overlap*. Their analysis remains in a point-identified regime and assumes that the target/data distribution ratio satisfies a square-integrability or tail-type condition, leading to truncated doubly robust estimators with slower-than-$n^{-1/2}$ rates when the overlap deteriorates. By contrast, our target policy may enter regions of *genuine zero support*, so no density ratio need exist at all (Appendix K). The fundamental object is therefore not a single latent value but an identified interval. In short, Mehrabi & Wager (2024) weaken overlap while preserving point identification; we drop point identification itself and characterize the sharp identified set.

Intervals from approximation error versus intervals from support holes. Jiang & Huang (2020) introduce minimax value intervals for off-policy evaluation and policy optimization that unify value-learning and weight-learning views and quantify misspecification bias in function-approximation-based OPE. Those intervals are extremely important conceptually, but they serve a different purpose from ours. Their interval quantifies *bias under approximate models in a point-identified problem*; our interval quantifies *partial identification under genuine support holes*. In the Jiang–Huang framework, if the relevant function classes are correct, the interval can collapse even when coverage is imperfect. In our framework, even with exact population knowledge of the supported Bellman data, the interval may remain nondegenerate because the target policy leaves the observed support. This distinction is exactly the one formalized by the geometry and sharpness theorems in Sections 5 and G.

Sequential lower/upper bounds under confounding or hidden bias. Another nearby line studies sequential off-policy evaluation when the data are not only unsupported but also confounded. The closest paper here is Zhang & Bareinboim (2025), who derive Bellman-style lower and upper value bounds for confounding-robust off-policy evaluation using causal eligibility traces. The conceptual similarity is obvious: both papers replace point identification by lower/upper Bellman recursions. The distinction, however, is equally important. Their setting is a *causal identifiability* problem driven by unobserved confounding and observational bias; ours is a *support-hole partial-identification* problem in a fully observed Bellman-smooth model class. Their bounds arise from causal restrictions; ours arise from metric Bellman–Lipschitz extension geometry. Correspondingly, our main theorem is a *sharp Bellman–Whitney identified interval* together with a no-gap duality theorem and a dynamic support-hole geometry theorem, whereas their analysis is organized around causal Bellman bounds and eligibility-trace estimators.

Generic conditional LP formulations of partial identification. Ben-Michael (2025) develop a general framework for estimation, inference, and policy learning when the parameter of interest is partially identified by a family of conditional linear programs. That paper is perhaps the closest abstract methodological relative of our work: it treats partially identified decision problems as optimization problems over conditionally indexed linear constraints. The relationship is that our Bellman–Whitney programs can be viewed as a very structured sequential specialization of that philosophy. The distinction is that our paper exploits the *Bellman structure* to collapse a potentially trajectory-indexed partial-identification problem into a stagewise dynamic program with exact recursive envelopes, a strong duality theorem, and a width geometry that admits explicit least-favorable sequential instances. In other words, Ben-Michael (2025) provide a generic conditional-LP language, whereas our contribution is a Bellman-specific structural collapse theorem.

Offline RL under partial coverage. A separate and highly relevant line studies *partial coverage* in offline RL. The closest foundational paper here is Uehara & Sun (2022), who develop pessimistic model-based offline RL under partial coverage. More recently, Liu et al. (2026) study offline RL under $Q^\star$-approximation and partial coverage, introduce an intrinsic complexity notion for partial-coverage offline RL, and derive new complexity and sample-efficiency results together with a second-order performance-difference lemma. These papers are close in spirit because they explicitly abandon full

coverage. Nevertheless, they study a different formal object from ours. Their goal is to learn a good policy or characterize sample complexity in a *point-identified* offline RL problem under structural coverage conditions—often relative to suitably covered comparators or complexity classes. Our goal is to characterize the sharp identified set of an *arbitrary fixed target policy* that may itself enter uncovered regions. Appendix K makes this distinction explicit: covered comparators can be point-identified while an uncovered target policy remains only partially identified.

Robust MDPs and rectangular ambiguity. The control corollary in Section 7 is also related in spirit to the robust dynamic-programming literature, especially the rectangular uncertainty frameworks of Iyengar (2005) and Nilim & Ghaoui (2005). Both that literature and our control appendix rely on stagewise rectangularity to obtain time-consistent backward recursions. The difference is in the source and interpretation of the uncertainty set. In robust MDPs, the analyst posits an exogenous ambiguity set over transition laws and optimizes a worst-case value. In our setting, the uncertainty set is *endogenously induced* by observed support constraints together with a Bellman–Lipschitz extension class. The resulting control output is therefore not a distributionally robust policy in the usual sense, but a set of *certifiably good*, *certifiably bad*, and *intrinsically ambiguous* actions under partial identification.

What is genuinely specific to the present paper. Taken together, the distinctions above explain the exact contribution of the Bellman–Whitney framework:

1. it treats arbitrary target policies that may enter genuine support holes, rather than weak-overlap or comparator-only regimes;
2. it gives a *sharp identified interval*, not merely valid lower and upper certificates;
3. it derives a *no-gap dual characterization* of the interval endpoints through one-sided Bellman relaxations;
4. it identifies a *dynamic support-hole geometry* and proves its sharpness on explicit least-favorable sequential families;
5. it separates *irreducible identified-set width* from *endpoint-estimation difficulty*;
6. it yields *action certificates* for control without overclaiming a sharp optimal-control interval in a nonconvex class.

These are the exact senses in which the present work is adjacent to, but not subsumed by, the neighboring literatures above.