BELLMAN-CERTIFIED ROUNDING FOR SPARSE POLICY DEPLOYMENT IN MDPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Continuous policy optimization may spread an update across many states, even when deployment permits only a few complete state-level changes. We study how much discounted return can be retained when continuous row mixtures are rounded to sparse binary policies in finite MDPs. Policy-dependent visitation couples the row edits, while long horizons make global curvature bounds conservative. From $2d + 2$ Bellman solves, we derive reusable envelopes that support uniform and candidate-specific guarantees before rounding. A rank-two rational representation of each exchange further permits weighted curvature integration along the realized trajectory. We prove that linear dimension dependence is unavoidable when the budget scales, and that exact global curvature thresholding is hard. Candidate-specific bounds raise pre-rounding certification coverage from 48.2% to 74.1% on the structured suite. At $\gamma = 0.95$, local integration lowers the median bound-to-loss ratio from 402.3 to 2.08 on coupled instances.

1 INTRODUCTION

Continuous policy optimization naturally distributes small updates across many states. Deployment may instead require each modified state-level decision to be implemented as a complete row change, with only k such changes affordable. We study the return lost when a continuous policy is mapped to this sparse deployment class.

We consider d editable rows in a finite discounted Markov decision process. Each row retains its baseline action or adopts one proposed action. A binary vector y specifies these choices, giving the deployment problem $\max_{y\in\{0,1\}^d,\, \mathbf{1}^\top y\le k} J(y)$. Continuous row mixtures belong to $\mathcal{C}_k = \{x \in [0,1]^d : \mathbf{1}^\top x \le k\}$, and $\mathcal{C}_k^{\mathrm{bin}} = \mathcal{C}_k \cap \{0,1\}^d$ is the deployable class. Their mismatch is

$$\mathrm{Gap}(\mathcal{C}_k) = \max_{x\in\mathcal{C}_k} J(x) - \max_{y\in\mathcal{C}_k^{\mathrm{bin}}} J(y). \tag{1}$$

Our goal is to convert a continuous candidate into a feasible binary policy with a return guarantee. Changing one row alters later visits to other rows, so the values of the edits do not decouple. At long horizons, the largest curvature over the policy cube may also lie far from the rounding trajectory and substantially overstate its loss.

Every fractional row mixture is a stationary randomized policy in an auxiliary two-action MDP. Extremal Bellman values bracket both policy values and row occupancies. The resulting sensitivity envelopes require $2d + 2$ solves and give $J(y) \ge J(x) - B_{\mathrm{unif}}$ for every $x \in \mathcal{C}_k$, with $c_{d,k} \le k/4$. Once a candidate is known, its geometry sharpens the bound without another Bellman solve. For the remaining long-horizon slack, we use the rank-two rational form of each exchange and integrate curvature with the interpolation weight that determines endpoint loss.

Advance certification and deployment. A uniform loss bound controls the relaxation gap over the policy family. For a particular candidate, the condition $J(x) - B > J(0)$ certifies improvement over the baseline before the discrete policy is constructed. The same Bellman envelopes can be reused across candidates and budgets that share the baseline and proposed rows. On the structured suite, candidate-specific bounds increase pre-rounding certification coverage from 48.2% to 74.1%. Rounding uses a median of six return evaluations, compared with 26 for greedy, while their mean normalized optimality gaps are 0.185% and 0.178%. At $S = d = 16$, validated arithmetic makes

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

certificate construction $14.8\times$ faster than the exact-rational pipeline. The realized return can then be checked directly once rounding is complete.

1.1 Illustrative example

Consider m zero-reward states and one absorbing target with positive reward. At each zero-reward state, the baseline action self-loops and the proposed action moves to the target. The deployment budget is $k = m/2$, with one dummy editable coordinate at the target giving $d = m + 1$. The return separates across the m relevant rows, $J(x) = \frac{1}{m}\sum_{i=1}^{m} \frac{\gamma x_i}{(1-\gamma)(1-\gamma+\gamma x_i)}$. Each summand is increasing and concave. The continuous optimum assigns $x_i = 1/2$ to every relevant row, whereas a deployable policy selects $m/2$ rows and leaves the others unchanged. The continuous-to-binary loss is therefore explicit. Proposition 4.1 evaluates both the gap and its certificate in closed form.

1.2 Related work

Pipage rounding originates in the classical exchange framework of Ageev & Sviridenko (2004). Recent work extends continuous relaxations and rounding beyond standard submodularity, including one-sided-smooth objectives and nonmonotone submodular maximization (Ghadiri et al., 2025; Buchbinder & Feldman, 2025). Here the objective is the discounted return of an MDP, and its exchange curvature is induced by the full policy-dependent dynamics.

Recent safe-policy-improvement work studies finite-data guarantees, partial observability, scalable planning, reliable decision points, and learned models (Wienhöft et al., 2023; Simão et al., 2023; Castellini et al., 2023; Bianchi et al., 2024; 2025; Sharma et al., 2025; Hamel-De le Court et al., 2025; Delgrange et al., 2026). These methods control improvement relative to a baseline under data or model uncertainty. We address the return lost at the next stage of the pipeline, when a continuous model-based update must satisfy a k-row deployment constraint. Related work also studies stochastic optimization when the deployed solution must be deterministic (Montenegro et al., 2024). The constraint is also related to sparse and discrete optimization, where cardinality or binary structure is handled by hard thresholding, block-coordinate methods, or continuous reformulations (Yuan et al., 2020; Yuan, 2024; Yuan & Ghanem, 2016). Our starting point is an optimized continuous policy whose conversion to sparse binary deployment must be certified.

Recent formal and robust-MDP work studies parameter synthesis, structural constraints, uncertainty-aware computation, and robust policy optimization (Junges et al., 2024; Rickard et al., 2024; Heck et al., 2025; Azeem et al., 2025; Heck et al., 2026; Hartmanns et al., 2026; Schnitzer et al., 2026; Suilen et al., 2024; Lin et al., 2024; Meggendorfer et al., 2025; Wang et al., 2025; Li et al., 2026; Abate et al., 2026; Asadi et al., 2026; Kumar et al., 2023; Zhang et al., 2024; Chen & Huang, 2024; Sun et al., 2024). Complementary second-order policy-optimization work analyzes Hessian information directly (Maniyar et al., 2024; Mu & Klabjan, 2024; Khorasani et al., 2025; Zhang et al., 2026). Our Bellman envelopes instead provide uniform first- and second-order sensitivity bounds over the policy cube, which we specialize to sparse deployment.

1.3 Contributions

Certified deployment. Bellman envelopes and budget-aware rounding geometry yield uniform and candidate-specific return guarantees. Algorithm 1 constructs a binary policy with at most $2(d-1)$ endpoint evaluations.

Curvature and trajectory bounds. A lower-bound family identifies the necessary dimension dependence, while an NP-hardness result limits exact global curvature computation. For an executed exchange, its rank-two structure gives a weighted trajectory bound that avoids paying for curvature away from the path.

Empirical behaviour. The experiments separate pre-rounding certification, final policy quality, long-horizon tightness, and computation cost. They quantify when candidate geometry suffices, when trajectory integration is needed, and how validated arithmetic compares with exact rational computation.

2 PRELIMINARIES

We use $\mathbb{R}_+$ for the nonnegative reals and $\mathbf{1}$ for the all-ones vector. The vector $e_i \in \mathbb{R}^S$ selects a state row, while $\varepsilon_i \in \mathbb{R}^d$ selects a policy coordinate. For $a \in \mathbb{R}$, let $[a]_+ = \max\{a, 0\}$. We write $v = v^+ - v^-$ for the componentwise positive and negative parts of a vector. The operator $\mathrm{diag}(x)$ forms a diagonal matrix, and $\odot$ denotes outward-rounded interval multiplication. The cost of one discounted Bellman solve on S states is denoted by $T_{\mathrm{Bell}}(S)$.

2.1 PROBLEM SETTING AND ASSUMPTIONS

Let $\mathcal{S} = \{1, \dots, S\}$ be a finite state space. Let $\gamma \in (0, 1)$ be the discount, and let $\mu \in \mathbb{R}_+^S$ satisfy $\mathbf{1}^\top \mu = 1$. After relabeling, rows $1, \dots, d$ are editable, where $d \in \{1, \dots, S\}$. Each editable row i has a baseline action and one proposed action. The weight $x_i \in [0, 1]$ interpolates between them, giving

$$P(x) = P^0 + \sum_{i=1}^{d} x_i \, e_i d_i^\top, \qquad r(x) = r^0 + \sum_{i=1}^{d} x_i \, e_i \Delta r_i, \qquad x \in [0, 1]^d. \tag{2}$$

Here $d_i^\top$ is the difference between the two transition rows, and Δr_i is the reward difference.

Assumption 2.1 (Stochastic two-action rows)**.** P^0 is stochastic. For every editable row i, both $P^0_{i\cdot}$ and $P^0_{i\cdot} + d_i^\top$ are probability distributions over $\mathcal{S}$.

Assumption 2.1 makes $P(x)$ stochastic for every point in the cube. Therefore

$$R(x) = (I - \gamma P(x))^{-1}, \qquad V(x) = R(x) r(x), \qquad J(x) = \mu^\top V(x) \tag{3}$$

are well defined (Puterman, 1994). The function J is real analytic on a neighbourhood of the cube. We use the budget sets in (1), with $k \in \{0, \dots, d\}$. Their relaxation gap is nonnegative and vanishes when the continuous and binary optima coincide.

A second assumption concerns the arithmetic used to construct the bounds.

Assumption 2.2 (Exact or validated arithmetic)**.** Bellman solves use exact rational arithmetic or directed outward rounding (Moore et al., 2009). In the latter case, the stochastic input, Bellman inequalities, residuals, and all downstream interval operations are enclosed. Return comparisons in Algorithm 1 are exact, or the selected endpoint is certified to have return at least that of the alternative.

Ordinary floating-point solves with residual padding do not satisfy this assumption. The distinction is consequential in Section 6: double precision reports many zero bounds on the same cases for which exact arithmetic reports none.

2.2 SENSITIVITY IDENTITIES AND ROUNDING QUANTITIES

The rounding procedure uses coordinate increases and pairwise exchanges. Differentiating (3) introduces three quantities,

$$q_i(x) = \mu^\top R(x) e_i, \qquad a_i(x) = \Delta r_i + \gamma d_i^\top V(x), \qquad b_{ij}(x) = \gamma d_i^\top R(x) e_j, \tag{4}$$

and the identities

$$\partial_i J(x) = q_i(x) a_i(x), \tag{5}$$

$$H_{ii}(x) = 2 q_i(x) b_{ii}(x) a_i(x), \tag{6}$$

$$H_{ij}(x) = q_j(x) b_{ji}(x) a_i(x) + q_i(x) b_{ij}(x) a_j(x), \quad i \neq j. \tag{7}$$

Here $H(x) = \nabla^2 J(x)$. These are classical Markov-chain sensitivity identities (Cao, 1998; Furmston et al., 2016). Their factored form is important because every factor is evaluated at the same policy x.

Definition 2.3 (Quantities controlling the rounding loss)**.** Define the largest negative coordinate slope and the largest negative pairwise exchange curvature as

$$\tau = \max_i \sup_{x \in [0,1]^d} \left[-\partial_i J(x) \right]_+, \tag{8}$$

$$\beta = \max_{i<j} \sup_{x \in [0,1]^d} \left[-(\varepsilon_i - \varepsilon_j)^\top H(x) (\varepsilon_i - \varepsilon_j) \right]_+. \tag{9}$$

When $d = 1$, the pairwise maximum is empty and $\beta = 0$.

Thus every coordinate derivative is at least $-\tau$, and every exchange line has second derivative at least $-\beta$. The next section derives tractable upper bounds. Sections 4 and 5 then study their dimension dependence and the cost of evaluating β exactly.

3 BELLMAN BOUNDS FOR ROUNDING

This section turns Bellman value bounds into a constructive rounding certificate. Lemma 3.1 first bounds every policy in the continuous family. Proposition 3.2 propagates these envelopes to the required derivatives, and Theorem 3.3 supplies the rounding guarantee.

3.1 BELLMAN ENVELOPES

Consider an auxiliary discounted control problem with the two endpoint actions at each editable row. Every unedited row retains its fixed action. Let $\underline{V}$ and $\overline{V}$ be the fixed points of the min- and max-Bellman operators, so $\underline{V}_s \le V_s^\pi \le \overline{V}_s$ for every stationary randomized policy π and every state s. For each editable state j, let $\underline{h}_j$ and $\overline{h}_j$ be the corresponding min/max value functions when the stage reward is e_j.

Lemma 3.1 (Bellman envelope). *Under Assumption 2.1, every $x \in [0,1]^d$ satisfies*

$$\underline{V} \le V(x) \le \overline{V}, \qquad \underline{h}_j \le R(x)e_j \le \overline{h}_j \quad \text{for every } j \in \{1,\dots,d\}. \tag{10}$$

All inequalities are componentwise.

Proof of Lemma 3.1. Every $x \in [0,1]^d$ defines a stationary randomized policy over the two endpoint actions at each editable row. Standard discounted Bellman optimality therefore bounds its value componentwise, $\underline{V} \le V(x) \le \overline{V}$. Apply the same argument to the auxiliary control problem with stage reward e_j. The value of policy x in this problem is $R(x)e_j$, which gives $\underline{h}_j \le R(x)e_j \le \overline{h}_j$. Both statements hold simultaneously for every x and editable j. □

Splitting $d_i = d_i^+ - d_i^-$ propagates Lemma 3.1 to intervals for the sensitivity quantities in (4),

$$Q_i = [\mu^\top \underline{h}_i,\ \mu^\top \overline{h}_i], \tag{11}$$

$$A_i = \Big[\Delta r_i + \gamma\{(d_i^+)^\top \underline{V} - (d_i^-)^\top \overline{V}\},\ \Delta r_i + \gamma\{(d_i^+)^\top \overline{V} - (d_i^-)^\top \underline{V}\}\Big], \tag{12}$$

$$B_{ij} = \gamma\Big[(d_i^+)^\top \underline{h}_j - (d_i^-)^\top \overline{h}_j,\ (d_i^+)^\top \overline{h}_j - (d_i^-)^\top \underline{h}_j\Big]. \tag{13}$$

3.2 SENSITIVITY CERTIFICATE

Using Minkowski interval addition and outward-rounded multiplication $\odot$, define

$$G_i = Q_i \odot A_i, \qquad \widehat{\tau} = \max_i \left[-\inf G_i\right]_+, \tag{14}$$

$$\mathcal{H}_{ii} = 2\,Q_i \odot B_{ii} \odot A_i, \qquad \mathcal{H}_{ij} = (Q_j \odot B_{ji} \odot A_i) + (Q_i \odot B_{ij} \odot A_j), \tag{15}$$

$$\underline{S}_{ij} = \inf \mathcal{H}_{ii} + \inf \mathcal{H}_{jj} - 2\sup \mathcal{H}_{ij}, \qquad \widehat{\beta} = \max_{i<j} \left[-\underline{S}_{ij}\right]_+, \tag{16}$$

with $\widehat{\beta} = 0$ when $d = 1$.

Proposition 3.2 (Sensitivity certificate). *Let Assumptions 2.1 and 2.2 hold. For every $x \in [0,1]^d$ and all editable i, j, $q_i(x) \in Q_i, \quad a_i(x) \in A_i, \quad b_{ij}(x) \in B_{ij}$. Consequently, $\tau \le \widehat{\tau}, \qquad \beta \le \widehat{\beta}$.*

Proof of Proposition 3.2. Since $\mu \ge 0$, Lemma 3.1 gives $q_i(x) = \mu^\top R(x)e_i \in Q_i$. For any componentwise box $\underline{z} \le z \le \overline{z}$, the exact extrema of $d_i^\top z = (d_i^+)^\top z - (d_i^-)^\top z$ are $(d_i^+)^\top \underline{z} - (d_i^-)^\top \overline{z}$ and $(d_i^+)^\top \overline{z} - (d_i^-)^\top \underline{z}$. Taking $z = V(x)$ and $z = R(x)e_j$ proves $a_i(x) \in A_i$ and $b_{ij}(x) \in B_{ij}$.

Equation (5) then gives $\partial_i J(x) \in G_i$. Hence $[-\partial_i J(x)]_+ \le [-\inf G_i]_+ \le \widehat{\tau}$. Taking the maximum over i and the supremum over x yields $\tau \le \widehat{\tau}$. Equations (6) and (7) give $H_{ii}(x) \in \mathcal{H}_{ii}$ and $H_{ij}(x) \in \mathcal{H}_{ij}$. Therefore

$$(\varepsilon_i - \varepsilon_j)^\top H(x)(\varepsilon_i - \varepsilon_j) = H_{ii}(x) + H_{jj}(x) - 2H_{ij}(x) \ge \underline{S}_{ij}.$$

Taking negative parts, the pairwise maximum, and the supremum over x gives $\beta \le \widehat{\beta}$. Directed outward interval operations preserve every inclusion. Exact rational arithmetic adds zero numerical-roundoff width. □

3.3 Certified rounding

For $q_{d,k} := \min\{k, \lfloor d/2 \rfloor\}$, define the budget coefficient $c_{d,k} := q_{d,k}(d - q_{d,k})/(4d)$.

Theorem 3.3 (Budget-aware certified rounding)**.** *Under Assumptions 2.1 and 2.2, Algorithm 1 returns $y \in \mathcal{C}_k^{\mathrm{bin}}$ for every $x \in \mathcal{C}_k$, with $J(y) \ge J(x) - B_{\mathrm{unif}}$ and $\mathrm{Gap}(\mathcal{C}_k) \le B_{\mathrm{unif}}$, where $B_{\mathrm{unif}} := \widehat{\tau} + c_{d,k}\widehat{\beta}$ for $k > 0$ and $B_{\mathrm{unif}} := 0$ for $k = 0$.*

For comparison with the original experiments, we also report the legacy bound $\mathcal{B}_{\mathrm{Bell}} := \widehat{\tau} + (d-1)\widehat{\beta}/8$. The new bound satisfies $B_{\mathrm{unif}} \le \mathcal{B}_{\mathrm{Bell}}$: for $d \ge 2$, $c_{d,k} \le d/16 \le (d-1)/8$; for $d = 1$, $\widehat{\beta} = 0$. Moreover, $c_{d,k} \le k/4$, so its coefficient stays bounded at fixed budget.

Algorithm 1, stated in Appendix A, first raises coordinates until their sum is integral. It then exchanges mass between fractional coordinates and selects the better endpoint. Each exchange makes at least one coordinate integral. At most $d-1$ exchanges and $2(d-1)$ return evaluations are required. These endpoint evaluations are separate from the $2d+2$ Bellman solves used to construct the reusable certificate. Every endpoint comparison must be exact or certified under Assumption 2.2.

The model-only guarantee depends on the MDP and budget, but not on the candidate. A zero bound therefore proves equality of the continuous and binary optima. The proof uses $\Phi(z) := \sum_i z_i(1 - z_i)$. At the integer-sum point after Phase 1, $\ell = \mathbf{1}^\top z \le k$, and Cauchy–Schwarz gives $\Phi(z) \le \ell(d-\ell)/d \le 4c_{d,k}$. Telescoping the exchange-potential decrease then charges at most $\widehat{\beta}\Phi(z)/4$ in Phase 2. Appendix A gives the proof without additional solves.

The same Bellman intervals also yield a sharper guarantee once a particular continuous candidate is available. Fix once and for all a deterministic Phase-1 fill order (we use increasing coordinate index) and let $z = z(x)$ be the point obtained by raising coordinates in that order until $\mathbf{1}^\top z = \lceil \mathbf{1}^\top x \rceil$. Define

$$\widehat{\tau}_i := [-\inf G_i]_+, \qquad F(z) := \{i : 0 < z_i < 1\}, \qquad \widehat{\beta}_{F(z)} := \max_{\substack{i<j \\ i,j \in F(z)}} \left[-\underline{S}_{ij}\right]_+, \tag{17}$$

with $\widehat{\beta}_{F(z)} = 0$ when $|F(z)| < 2$. Let $\Phi(z) := \sum_{i=1}^d z_i(1 - z_i)$, $\Delta_i := z_i - x_i$ and define the candidate-specific pre-rounding certificate $B_{\mathrm{cand}}(x) := \sum_{i=1}^d \Delta_i \widehat{\tau}_i + \frac{\widehat{\beta}_{F(z)}}{4}\Phi(z)$.

Proposition 3.4 (Candidate-specific pre-rounding certificate)**.** *Let Assumptions 2.1 and 2.2 hold, and let Algorithm 1 use the fixed Phase-1 fill order above. For every $x \in \mathcal{C}_k$, its output $y \in \mathcal{C}_k^{\mathrm{bin}}$ satisfies $J(y) \ge J(x) - B_{\mathrm{cand}}(x)$, $\quad B_{\mathrm{cand}}(x) \le B_{\mathrm{unif}} \le \mathcal{B}_{\mathrm{Bell}}$. Thus $B_{\mathrm{cand}}(x)$ is available once the continuous candidate is known and before any pairwise endpoint comparison.*

The sharper bound retains candidate geometry that Theorem 3.3 discards. For an exchange of coordinates i, j over the maximal interval $[a, b]$, Proposition 3.2 gives the pair-specific curvature bound $\widehat{\beta}_{ij} = \left[-\underline{S}_{ij}\right]_+$, and the proof of Theorem 3.3 yields the one-step charge $\widehat{\beta}_{ij}(-a)b/2$. Moreover, every maximal exchange satisfies the potential identity $\Phi(z^t) - \Phi(z^{t+1}) = 2(-a_t)b_t$. Since the fractional support can only shrink and the final point is binary, $\sum_t (-a_t)b_t = \Phi(z)/2$. This identity yields the second term of B_{cand}. The first term records coordinate-specific charges along the deterministic fill path. Appendix A provides the proof.

The reusable certificate is constructed once per instance. Computing the Bellman envelopes and interval quantities takes $2d+2$ Bellman solves, $O(d^2 S)$ interval operations, and an $O(d^2)$ scan. The total cost is $(2d+2)T_{\mathrm{Bell}}(S) + O(d^2 S)$. Given a candidate x, $B_{\mathrm{cand}}(x)$ uses the stored G_i and $\underline{S}_{ij}$,

so no further Bellman solve is needed. The fill and potential cost $O(d)$ in total, while the support scan costs at most $O(d^2)$. This model-only construction is most attractive when $d \ll S$. When $d = S$, the number of Bellman solves is twice the number of states. Rounding is then performed once for each candidate.

3.4 Computation and interval conservatism

The preceding operation count does not include rational bit growth. With L-bit inputs, exact rational arithmetic has bit complexity polynomial in S, d, and L. Validated floating point instead adds an $O(S^2)$ residual pass to each solve. Its intervals have positive width, while outward endpoints preserve Proposition 3.2.

The interval calculation treats Q_i, A_i, and B_{ij} as independent boxes and can therefore include endpoint combinations that no common x realizes. The excess width vanishes when the Bellman envelopes are tight. Proposition 4.1 gives $\widehat{\beta} = \beta$ on the example family.

4 Dimension dependence and long-horizon conservatism

At fixed k, the budget-aware coefficient $c_{d,k}$ remains bounded. Linear dependence on dimension is nevertheless unavoidable when $k = \Theta(d)$. The family from Section 1.1 also distinguishes exchange-count slack from curvature conservatism.

Proposition 4.1 (Fixed-discount lower-bound family)**.** *For every even $m \geq 2$ and $\gamma \in (0, 1)$, the illustrative family is a transition-only MDP satisfying Assumption 2.1, with $d = m + 1$ and $k = m/2$. Writing $\delta = 1 - \gamma$, it satisfies $\widehat{\tau} = \tau = 0$, $\widehat{\beta} = \beta = 4\gamma^2/(m\delta^3)$ and $\mathrm{Gap}(\mathcal{C}_k) = \gamma^2/[2\delta(2-\gamma)]$. At its continuous optimum, $B_{\mathrm{cand}}(x^\star) = \gamma^2/(4\delta^3)$, $B_{\mathrm{unif}} = \frac{m+2}{m+1} B_{\mathrm{cand}}(x^\star)$, and $\mathcal{B}_{\mathrm{Bell}} = 2B_{\mathrm{cand}}(x^\star)$.*

Corollary 4.2 (A linear dimension coefficient is necessary)**.** *Fix $\gamma \in (0, 1)$. If $c_d \geq 0$ satisfies $\mathrm{Gap}(\mathcal{C}_k) \leq \tau + c_d\beta$ on every member of this family, then $c_d \geq (d-1)(1-\gamma)^2/[8(2-\gamma)]$. Thus $c_d = \Omega(d)$ along this family, whose budget is $k = (d-1)/2$.*

Proof. Proposition 4.1 gives $\tau = 0$ and $\beta > 0$, so $c_d \geq \mathrm{Gap}(\mathcal{C}_k)/\beta = m(1-\gamma)^2/[8(2-\gamma)]$. □

Appendix B derives these formulas. From $x^\star$, Algorithm 1 reaches a binary optimum after $m/2$ exchanges, so its actual loss equals $\mathrm{Gap}(\mathcal{C}_k)$. The legacy bound overcounts the exchange contribution by a factor of two. The uniform bound reduces this slack to $(m+2)/(m+1)$ relative to the candidate bound. Nevertheless, $B_{\mathrm{cand}}(x^\star)/\mathrm{Gap}(\mathcal{C}_k) = (2-\gamma)/(2\delta^2)$ still diverges as $\gamma \uparrow 1$. Here $\widehat{\beta} = \beta$, so interval propagation does not explain this residual slack.

One might suspect that the long-horizon conservatism comes only from evaluating curvature globally. On this family, even exact segment-local extrema retain the same divergent order. An executed exchange has $g(t) = J(z + t(\varepsilon_i - \varepsilon_j))$ with $t \in [-1/2, 1/2]$. Although it starts at two half-integral coordinates, its endpoints set one coordinate to zero, retaining large curvature. The exact segment maximum is $\beta_{\mathrm{seg}} = 2\gamma^2(\delta^{-3} + 1)/m$. Summing its one-step charges gives $B_{\mathrm{seg}} = \gamma^2(\delta^{-3} + 1)/8$, and hence $B_{\mathrm{seg}}/\mathrm{Gap}(\mathcal{C}_k) = (2-\gamma)(1+\delta^3)/(4\delta^2)$. Thus segment-local maximization alone does not remove the long-horizon slack.

On this family, weighted trajectory interpolation removes the residual slack. For any C^2 exchange function on $[a, b]$ with $a < 0 < b$, let $K_{a,b}(t) = b(t-a)/(b-a)$ for $t \leq 0$ and $K_{a,b}(t) = (-a)(b-t)/(b-a)$ for $t \geq 0$. The nonnegative kernel gives the loss bound $[g(0) - \max\{g(a), g(b)\}]_+ \leq \int_a^b K_{a,b}(t)[-g''(t)]_+ \, dt$. A constant curvature bound recovers the old charge because $\int_a^b K_{a,b}(t)\, dt = (-a)b/2$. On this family, $g'' < 0$ and $g(a) = g(b)$, so the integral equals the actual one-step loss. Summing yields $B_{\mathrm{int}} = \mathrm{Gap}(\mathcal{C}_k)$ for every γ. At $\gamma = 0.95$, the candidate and segment-maximum ratios are 210 and 105.013, respectively, whereas the exact integral ratio is 1; reliable 256-subinterval upper sums give 1.044. Appendix B.6 proves the identity and Appendix D.9 reports exact-rational checks. The same interpolation bound can be computed on general exchanges.

Proposition 4.3 (Rational exchange representation)**.** *Under Assumption 2.1, every feasible exchange has $g(t) = p(t)/q(t)$, where $\deg p, \deg q \leq 2$, $q(0) = 1$, and $q(t) > 0$ throughout its interval. Moreover, $g''(t) = h(t)/q(t)^3$ with $\deg h \leq 3$.*

Computable trajectory bound. Appendix B.7 derives the rational coefficients from a rank-two resolvent update. Exact quadratic extrema and Bernstein polynomial bounds enclose $[-g'']_+$ on each subinterval. Multiplication by the exact kernel integrals gives a reliable upper sum U_t for each exchange. With Phase-1 output z, define $B_{\mathrm{loc}} = [J(x) - J(z)]_+ + \sum_t U_t$. Then $[J(x) - J(y)]_+ \le B_{\mathrm{loc}}$. The resulting certificate is available during rounding and does not require global Bellman envelopes. Section 6 evaluates its tightness and cost. Before the exchange trajectory is known, certification still relies on B_{unif} or B_{cand}.

5 Complexity of exact curvature

Theorem 3.3 uses the tractable upper bound $\widehat{\beta}$ in place of β. Could exact curvature provide an efficient, uniformly sharper alternative? The threshold problem is NP-hard even on a strongly restricted MDP class, and its binary restriction is NP-complete.

All N rows in this section are editable. For $v \in \mathbb{R}^N$, define

$$\kappa_{\mathcal{M}}(v) = \sup_{x\in[0,1]^N} \left[-v^\top \nabla^2 J(x) v\right]_+, \qquad \kappa^{\mathrm{bin}}_{\mathcal{M}}(v) = \max_{x\in\{0,1\}^N} \left[-v^\top \nabla^2 J(x) v\right]_+, \tag{18}$$

$$\beta_{\mathcal{M}} = \max_{a<b} \kappa_{\mathcal{M}}(\varepsilon_a - \varepsilon_b), \qquad \beta^{\mathrm{bin}}_{\mathcal{M}} = \max_{a<b} \kappa^{\mathrm{bin}}_{\mathcal{M}}(\varepsilon_a - \varepsilon_b). \tag{19}$$

A Woodbury elimination exposes the relevant dependence on the policy variables. Let $D = [d_1 \cdots d_N]$, $X = \mathrm{diag}(x)$, $R_0 = (I - \gamma P^0)^{-1}$, $u^\top = \mu^\top R_0$, $C = \gamma D^\top R_0$, and $a = \Delta + \gamma D^\top R_0 r^0$. Then $J(x) = J(0) + u^\top (I - XC)^{-1} X a$. Let $\mathcal{G}(C)$ contain the arc $i \to j$ when $C_{ij} \neq 0$.

Definition 5.1 (Restricted MDP class $\mathfrak{B}_3$)**.** The class $\mathfrak{B}_3$ contains rational row-mixture MDPs satisfying Assumption 2.1, with $\gamma = 1/2$, a point-mass initial distribution, two actions per row, rewards in $[0, 1]$, strictly positive transition endpoints, and acyclic $\mathcal{G}(C)$ of depth at most three. Membership is decidable in polynomial time.

Theorem 5.2 (Complexity of exact pairwise curvature)**.** *The language* $\{\langle \mathcal{M}, T\rangle : \mathcal{M} \in \mathfrak{B}_3,\ \beta^{\mathrm{bin}}_{\mathcal{M}} \ge T\}$ *is NP-complete, and thresholding the continuous-cube quantity* $\beta_{\mathcal{M}}$ *is NP-hard.*

The construction. The proof maps unweighted MAX-CUT to a rational instance $\mathcal{M}_G \in \mathfrak{B}_3$. After padding the graph, the construction creates one designated exchange (p, q) whose curvature satisfies $\beta_{\mathcal{M}_G} = \beta^{\mathrm{bin}}_{\mathcal{M}_G} = \kappa_{\mathcal{M}_G}(\varepsilon_p - \varepsilon_q) = \alpha_G\{\Lambda n - m + \mathrm{MaxCut}(G)\}$, where $\Lambda = 2m + 1$ and $\alpha_G > 0$ has polynomial encoding length. The reduced curvature is multi-affine, so a maximizer exists at a vertex. The choice of Λ forces complementary graph controls and recovers the MAX-CUT objective. Coefficient bounds isolate the designated pair from every other exchange. For the binary restriction, that pair and a binary control vector form a polynomial certificate. The construction also preserves local exchange feasibility under the row budget $k = n + 1$.

Appendix C gives the rational transition endpoints, positivity checks, path expansion, pair-separation bounds, budget bridge, and encoding analysis. The reduction concerns exact exchange curvature rather than return maximization. Thus exact thresholding does not provide a polynomial-time alternative to the Bellman-envelope bound in general, unless P = NP.

6 Evaluation

The evaluation addresses four questions: when deployment can be certified before rounding, how much return the rounded policy retains, how tight the bounds remain at long horizons, and what certification costs. Exact rational arithmetic is used for binary optima and theorem checks. The scaling study uses outward-rounded validated floating point. Appendix D contains the generators, fixed seeds, protocols, and complete tables. Structured-suite comparisons retain the legacy $\mathcal{B}_{\mathrm{Bell}}$ baseline. Budget-aware B_{unif} results are reported for the closed-form and coupled suites.

Soundness calibration. On the original exact-rational grid of 144 cases across GRIDWORLD, QUEUE, MAINTENANCE, and GARNET, every witnessed continuous-to-binary advantage is bounded by $\mathcal{B}_{\mathrm{Bell}}$. The smallest nonzero certificate is 2.95×10^{-6}, while naive midpoint FP64 reports a false zero in 97/144 cases. This grid tests soundness rather than active rounding because 141 of the 144 quantized candidates are already binary.

Table 1: Certified deployment improvement on the 139 active structured cases. The legacy B_{Bell} and B_{cand} are pre-rounding certificates, whereas B_{trace} is a post-trace audit. Direct reports the exact check $J(y) > J(0)$, with counts deduced from the stored returns.

Discount	Active	B_{Bell}	B_{cand}	B_{trace}	Direct
$\gamma = 1/2$	68	67	68	68	68
$\gamma = 4/5$	71	0	35	54	71
All	139	67	103	122	139

Active rounding. The closed-form family of Proposition 4.1 uses $m \in \{4, 8, 16, 32, 64\}$ and $\gamma \in \{0.2, 0.5, 0.8, 0.95\}$. Algorithm 1 performs exactly $m/2$ exchanges in all 20 cases, reaches a binary optimum, and incurs realized loss exactly equal to $\mathrm{Gap}(\mathcal{C}_k)$. Moreover, $B_{\mathrm{cand}}(x^\star) = \mathcal{B}_{\mathrm{Bell}}/2$, while $B_{\mathrm{unif}} = \frac{m+2}{m+1} B_{\mathrm{cand}}(x^\star)$. The exact weighted bound equals the gap, as shown in Section 4.

The separately frozen structured suite contains 144 cases, of which 139 have genuinely fractional candidates. Every active candidate has positive witnessed advantage over the exact binary optimum. The median candidate has four fractional coordinates and requires three exchanges. All certified rounding checks pass, and Algorithm 1 reaches the exact binary optimum in 119 of 144 cases. Across the 139 active cases, its mean normalized optimality gap is 0.185%, compared with 0.178% for greedy. The corresponding median return-evaluation counts are 6 and 26. Appendix D reports the complete baseline comparison.

Pre-rounding certification. For a pre-rounding bound B, define the certified margin $M_B(x) := J(x) - B - J(0)$. A positive margin guarantees that Algorithm 1 returns a deployable policy with $J(y) > J(0)$. Table 1 reports coverage on the 139 active structured cases. The legacy model-only certificate covers $67/139 = 48.2\%$ of them. Candidate geometry raises coverage to $103/139 = 74.1\%$ without another Bellman solve. At $\gamma = 4/5$, the corresponding counts are 0/71 and 35/71. All 139 final policies improve over the baseline, so an uncovered case is an abstention rather than a failed deployment. Appendix D gives the full tightness hierarchy. The median bound-to-loss ratio falls from 70.8 for $\mathcal{B}_{\mathrm{Bell}}$ to 12.1 for B_{cand}.

Direct rounding followed by verification. For a fixed candidate and rounded output, every sound bound satisfies $J(x) - B - J(0) \le J(y) - J(0)$. An exact direct check therefore accepts every deployment accepted by the bound. Across the 20 closed-form cases, the legacy, budget-aware, and candidate bounds accept 5, 10, and 10 cases. Direct verification accepts all 20. The return of the final chosen endpoint is already cached, so verification adds no evaluation to the 496 used for rounding. Stored exact returns similarly imply 139/139 direct acceptances on the structured suite. Pre-rounding certification covers 103/139 cases, while the post-trace bound covers 122/139. The distinction is temporal: a loss guarantee is available before deployment, whereas direct verification requires the final policy.

General exchange bounds and complete costs. The 24-instance coupled suite combines two transition structures, $d \in \{6, 8\}$, three discounts, and two seeds. Every candidate has positive advantage over the enumerated binary optimum. At $\gamma = 0.95$, the median bound-to-loss ratio falls from 402.3 for B_{cand} to 2.08 for B_{loc}. The trajectory bound certifies improvement in all 24 cases. It preserves the rounded policy while tightening the loss guarantee.

Including continuous search and exact setup, median total times are 7.60 ms for direct rounding, 9.43 ms for rank-two rounding, and 36.19 ms for local integration. Each instance is timed three times. The tighter guarantee therefore incurs additional computation on this small synthetic suite. Appendix D.14 reports the complete comparison, exact audits, and separate scaling checks.

Scaling and validated arithmetic. We next test certification beyond the exact-enumeration regime. The sparse family uses either $S = d \in \{16, 32, 64, 128\}$, or $d = 16$ with $S \in \{64, 128, 256\}$. Full validated certification takes 4.45 s at $S = d = 128$ and 1.31 s at $(S, d) = (256, 16)$. In the latter case, Bellman solves account for 1.28 s and interval propagation for 0.031 s, consistent with the $(2d + 2)T_{\mathrm{Bell}}(S) + O(d^2 S)$ decomposition.

Validated intervals contain the exact recomputation in all six audited reference cases, whereas naive FP64 passes the Bellman audit in none. This audit covers only the small reference cases. At $S = d = 16$, the validated implementation is about $14.8\times$ faster than the existing exact-rational pipeline.

7 CONCLUSION

Sparse deployment changes the policy-optimization problem because each row edit also changes future visitation. We showed that Bellman envelopes can control this coupling uniformly, while budget and candidate geometry sharpen the guarantee available before rounding. For the realized exchange path, a rank-two representation supports weighted curvature integration and avoids paying for remote worst-case policies.

On the structured suite, candidate geometry raises pre-rounding certification coverage from 48.2% to 74.1%. On coupled instances at $\gamma = 0.95$, trajectory integration lowers the median bound-to-loss ratio from 402.3 to 2.08. These findings distinguish two uses of a certificate: reusable guarantees before deployment and tighter audits once the rounding path is known.

Scope and extensions. The analysis assumes a known finite discounted MDP with two actions per editable row. Larger action sets, estimated models, partial observability, and average reward require new arguments. Open questions include the best universal dimension constant, NP membership of continuous curvature thresholding, and validated floating-point implementations of the trajectory bounds.

AI USE DISCLOSURE

Generative AI tools assisted with language editing and debugging selected parts of the experimental code.

REPRODUCIBILITY STATEMENT

Formal specification. Assumptions 2.1 and 2.2 state the model and numerical requirements. Section 3 defines every certificate quantity. Appendices A, B, and C contain the proofs. The generators, seeds, arithmetic protocols, and complete tables appear in Appendix D.

Arithmetic and continuous search. Enumeration-based experiments use exact rational arithmetic for binary optima, evaluated returns, and theorem checks. The scaling study instead uses directed outward rounding and independently audits Bellman inequalities and interval containment with exact rational arithmetic on reference cases. Continuous multistart search is heuristic. Before certification, candidates used for exact verification are quantized to feasible rational points. Search quality can change the witnessed gap, but it cannot affect certificate soundness. The code artifact reconstructs the rational witnesses and validated checks from the reported instances. No GPU or learned model is required.

Closed-form and structured suites. One script checks the revised budget coefficient, the interpolation identity, and all 20 closed-form cases in exact rational arithmetic. A second script reruns those cases with cached endpoint returns to compare pre-rounding certification and direct verification. Structured-suite coverage retains the identified legacy baseline. Its direct-verification counts are deduced from stored exact returns, without a new timing run.

Coupled suite. The coupled suite includes saved rational instances, candidate vectors, and independent exact checks of exchange lines and derivatives. Each small instance has three timing repetitions. Global certificate construction, continuous search, and local integration are accounted for separately.

REFERENCES

Alessandro Abate, Thom Badings, Giuseppe De Giacomo, and Francesco Fabiano. Best-effort policies for robust markov decision processes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(43):36120–36128, 2026. doi: 10.1609/aaai.v40i43.40929.

Alexander A. Ageev and Maxim Sviridenko. Pipage rounding: A new method of constructing algorithms with proven performance guarantee. *Journal of Combinatorial Optimization*, 8(3): 307–328, 2004. doi: 10.1023/B:JOCO.0000038913.96607.C2.

Thomas W. Archibald, Kenneth I. M. McKinnon, and Lyn C. Thomas. On the generation of Markov decision processes. *Journal of the Operational Research Society*, 46(3):354–361, 1995.

Ali Asadi, Krishnendu Chatterjee, Ehsan Goharshady, Mehrdad Karrabi, Alipasha Montaseri, and Carlo Pagano. Strongly polynomial time complexity of policy iteration for L_∞ robust MDPs. In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 427–457. PMLR, 2026. URL https://proceedings.mlr.press/v336/asadi26a.html.

Muqsit Azeem, Debraj Chakraborty, Sudeep Kanav, Jan Křetínský, Mohammadsadegh Mohagheghi, Stefanie Mohr, and Maximilian Weininger. 1–2–3–go! policy synthesis for parameterized markov decision processes via decision-tree learning and generalization. In *Verification, Model Checking, and Abstract Interpretation*, volume 15530 of *Lecture Notes in Computer Science*, pp. 97–120. Springer, 2025. doi: 10.1007/978-3-031-82703-7_5.

Federico Bianchi, Edoardo Zorzi, Alberto Castellini, Thiago D. Simão, Matthijs T. J. Spaan, and Alessandro Farinelli. Scalable safe policy improvement for factored multi-agent MDPs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3952–3973. PMLR, 2024. URL https://proceedings.mlr.press/v235/bianchi24b.html.

Federico Bianchi, Alberto Castellini, Edoardo Zorzi, Thiago D. Simão, Matthijs T. J. Spaan, and Alessandro Farinelli. Scaling safe policy improvement: Monte Carlo tree search and policy iteration strategies. *Journal of Artificial Intelligence Research*, 84, 2025. doi: 10.1613/jair.1.19649.

Niv Buchbinder and Moran Feldman. Extending the extension: Deterministic algorithm for non-monotone submodular maximization. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*. ACM, 2025. doi: 10.1145/3717823.3718120.

Xi-Ren Cao. The maclaurin series for performance functions of markov chains. *Advances in Applied Probability*, 30(3):676–692, 1998. doi: 10.1239/aap/1035228123.

Alberto Castellini, Federico Bianchi, Edoardo Zorzi, Thiago D. Simão, Alessandro Farinelli, and Matthijs T. J. Spaan. Scalable safe policy improvement via Monte Carlo tree search. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 3732–3756. PMLR, 2023. URL https://proceedings.mlr.press/v202/castellini23a.html.

Ziyi Chen and Heng Huang. Accelerated policy gradient for s-rectangular robust MDPs with large state spaces. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6847–6880. PMLR, 2024. URL https://proceedings.mlr.press/v235/chen24s.html.

Florent Delgrange, Raphaël Avalos, and Willem Röpke. Deep SPI: Safe policy improvement via world models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0212b0dff7b4aa3bd8b8cc4ce24fd384-Abstract-Conference.html.

Thomas Furmston, Guy Lever, and David Barber. Approximate newton methods for policy search in markov decision processes. *Journal of Machine Learning Research*, 17(226):1–51, 2016. URL https://www.jmlr.org/papers/v17/15-414.html.

Mehrdad Ghadiri, Richard Santiago, and Bruce Shepherd. Beyond submodular maximization via one-sided smoothness. *Mathematical Programming*, 2025. doi: 10.1007/s10107-025-02301-5.

Edwin Hamel-De le Court, Francesco Belardinelli, and Alexander W. Goodall. Probabilistic shielding for safe reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39 (15):16091–16099, 2025. doi: 10.1609/aaai.v39i15.33767.

Arnd Hartmanns, Sebastian Junges, Tim Quatmann, and Maximilian Weininger. The revised practitioner's guide to MDP model checking algorithms. *International Journal on Software Tools for Technology Transfer*, 2026. doi: 10.1007/s10009-026-00848-y.

Linus Heck, Tim Quatmann, Jip Spel, Joost-Pieter Katoen, and Sebastian Junges. Generalized parameter lifting: Finer abstractions for parametric markov chains. In *Automated Technology for Verification and Analysis*, volume 16145 of *Lecture Notes in Computer Science*, pp. 207–230. Springer, 2025. doi: 10.1007/978-3-032-08707-2_10.

Linus Heck, Filip Macák, Milan Češka, and Sebastian Junges. Constrained and robust policy synthesis with satisfiability-modulo-probabilistic-model-checking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(43):36253–36261, 2026. doi: 10.1609/aaai.v40i43.40944.

Sebastian Junges, Erika Ábrahám, Christian Hensel, Nils Jansen, Joost-Pieter Katoen, Tim Quatmann, and Matthias Volk. Parameter synthesis for markov models: Covering the parameter space. *Formal Methods in System Design*, 62:181–259, 2024. doi: 10.1007/s10703-023-00442-x.

Mohammadsadegh Khorasani, Saber Salehkaleybar, Negar Kiyavash, Niao He, and Matthias Grossglauser. Efficiently escaping saddle points for policy optimization. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 2143–2162. PMLR, 2025. URL https://proceedings.mlr.press/v286/khorasani25a.html.

Navdeep Kumar, Esther Derman, Matthieu Geist, Kfir Y. Levy, and Shie Mannor. Policy gradient for rectangular robust markov decision processes. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ba8aee784ffe0813890288b334444eda-Abstract-Conference.html.

Mengmeng Li, Daniel Kuhn, and Tobias Sutter. Policy gradient algorithms for robust MDPs with nonrectangular uncertainty sets. *SIAM Journal on Optimization*, 36(1):120–151, 2026. doi: 10.1137/24M1631250.

Zhenwei Lin, Chenyu Xue, Qi Deng, and Yinyu Ye. A single-loop robust policy gradient method for robust markov decision processes. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 30392–30426. PMLR, 2024. URL https://proceedings.mlr.press/v235/lin24u.html.

Mizhaan P. Maniyar, Prashanth L.A., Akash Mondal, and Shalabh Bhatnagar. A cubic-regularized policy Newton algorithm for reinforcement learning. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4708–4716. PMLR, 2024. URL https://proceedings.mlr.press/v238/maniyar24a.html.

Tobias Meggendorfer, Maximilian Weininger, and Patrick Wienhöft. Solving robust markov decision processes: Generic, reliable, efficient. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(25):26631–26641, 2025. doi: 10.1609/aaai.v39i25.34865.

Alessandro Montenegro, Marco Mussi, Alberto Maria Metelli, and Matteo Papini. Learning optimal deterministic policies with stochastic policy gradients. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36160–36211. PMLR, 2024. URL https://proceedings.mlr.press/v235/montenegro24a.html.

Ramon E. Moore, R. Baker Kearfott, and Michael J. Cloud. *Introduction to Interval Analysis*. SIAM, 2009.

Siqiao Mu and Diego Klabjan. On the second-order convergence of biased policy gradient algorithms. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36455–36485. PMLR, 2024. URL https://proceedings.mlr.press/v235/mu24b.html.

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

Luke Rickard, Alessandro Abate, and Kostas Margellos. Learning robust policies for uncertain parametric Markov decision processes. In *Proceedings of the 6th Annual Learning for Dynamics & Control Conference*, volume 242 of *Proceedings of Machine Learning Research*, pp. 876–889. PMLR, 2024. URL https://proceedings.mlr.press/v242/rickard24a.html.

Yannik Schnitzer, Alessandro Abate, and David Parker. Robust parameter learning for uncertain MDPs. In *Proceedings of the 3rd International Joint Conference on Quantitative Evaluation of Systems and Formal Modelling and Analysis of Timed Systems*. Springer, 2026. URL https://arxiv.org/abs/2605.01339. arXiv:2605.01339.

Abhishek Sharma, Leo Benac, Sonali Parbhoo, and Finale Doshi-Velez. Decision-point guided safe policy improvement. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2935–2943. PMLR, 2025. URL https://proceedings.mlr.press/v258/sharma25a.html.

Thiago D. Simão, Marnix Suilen, and Nils Jansen. Safe policy improvement for POMDPs via finite-state controllers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 15109–15117, 2023. doi: 10.1609/aaai.v37i12.26763.

Marnix Suilen, Thom Badings, Eline M. Bovy, David Parker, and Nils Jansen. Robust markov decision processes: A place where AI and formal methods meet. In *Principles of Verification: Cycling the Probabilistic Landscape*, volume 15262 of *Lecture Notes in Computer Science*, pp. 126–154. Springer, 2024. doi: 10.1007/978-3-031-75778-5_7.

Zhongchang Sun, Sihong He, Fei Miao, and Shaofeng Zou. Policy optimization for robust average reward MDPs. In *Advances in Neural Information Processing Systems*, volume 37. Curran Associates, Inc., 2024. doi: 10.52202/079017-0551. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/1f28e9341ab99d8e5a5734f0a76601c7-Abstract-Conference.html.

Qiuhao Wang, Yuqi Zha, Chin Pang Ho, and Marek Petrik. Provable policy gradient for robust average-reward MDPs beyond rectangularity. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 65368–65399. PMLR, 2025. URL https://proceedings.mlr.press/v267/wang25ei.html.

Patrick Wienhöft, Marnix Suilen, Thiago D. Simão, Clemens Dubslaff, Christel Baier, and Nils Jansen. More for less: Safe policy improvement with stronger performance guarantees. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023*, pp. 4406–4415. ijcai.org, 2023. doi: 10.24963/ijcai.2023/490.

Ganzhao Yuan. Smoothing proximal gradient methods for nonsmooth sparsity constrained optimization: Optimality conditions and global convergence. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57842–57870. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/yuan24a.html.

Ganzhao Yuan and Bernard Ghanem. Binary optimization via mathematical programming with equilibrium constraints. *arXiv preprint arXiv:1608.04425*, 2016.

Ganzhao Yuan, Li Shen, and Wei-Shi Zheng. A block decomposition algorithm for sparse optimization. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 275–285. ACM, 2020. doi: 10.1145/3394486.3403070.

Runyu Zhang, Yang Hu, and Na Li. Soft robust MDPs and risk-sensitive MDPs: Equivalence, policy gradient, and sample complexity. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/3f7f788b479a26348605a9f97c465d22-Abstract-Conference.html.

Yixian Zhang, Huaze Tang, Changxu Wei, Chao Wang, and Wenbo Ding. Policy Newton algorithm in reproducing kernel Hilbert space. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10009658.

A Proofs for Section 3

Proof of Theorem 3.3. If $k = 0$, then $x = y = 0$ and the claim is immediate. Assume $k > 0$.

Phase 1. Write $s = \mathbf{1}^\top x$, $\ell = \lceil s \rceil \le k$, and $\eta = \ell - s \in [0, 1)$. The cube has sufficient unused capacity to raise coordinates by total mass η. Since $\partial_i J \ge -\tau$ everywhere, integrating along the monotone fill path gives $J(z) \ge J(x) - \eta\tau \ge J(x) - \tau$. The resulting z is feasible and has integer sum ℓ.

One exchange. An integer-sum nonbinary point has at least two fractional coordinates. For such i, j, set $v = \varepsilon_i - \varepsilon_j$ and $g(t) = J(z + tv)$. The maximal feasible interval has endpoints $a = \max\{-z_i, z_j - 1\} < 0$ and $b = \min\{1 - z_i, z_j\} > 0$, with $b - a \le 1$. If $g'' \ge -\beta$, the function $h(t) = g(t) + \beta t^2/2$ is convex. Apply convexity with endpoint weights $b/(b-a)$ and $(-a)/(b-a)$ to obtain $g(0) \le \max\{g(a), g(b)\} + \beta(-a)b/2$. The chosen endpoint therefore loses at most $\beta(-a)b/2 \le \beta/8$ and makes at least one coordinate integral.

Algorithm 1 Constructive rounding used in Theorem 3.3

Require: $x \in \mathcal{C}_k$, with exact or certified endpoint comparisons of J
Ensure: $y \in \mathcal{C}_k^{\mathrm{bin}}$ with $J(y) \geq J(x) - B_{\mathrm{unif}}$
$z \leftarrow x$
if $\mathbf{1}^\top z \notin \mathbb{Z}$ **then**
  raise coordinates of z in the fixed order $1, 2, \ldots, d$ until $\mathbf{1}^\top z = \lceil \mathbf{1}^\top x \rceil$ ▷ deterministic fill and costs at most τ
end if
while z has a fractional coordinate **do**
  pick two fractional coordinates $i \neq j$ ▷ an integral sum forbids exactly one
  $[a, b] \leftarrow$ maximal interval with $z + t(\varepsilon_i - \varepsilon_j) \in [0,1]^d$
  $z \leftarrow z + t^\star(\varepsilon_i - \varepsilon_j)$, where $t^\star \in \arg\max_{t \in \{a,b\}} J\{z + t(\varepsilon_i - \varepsilon_j)\}$ ▷ costs at most $\beta(-a)b/2$
end while
return $y \leftarrow z$

Potential identity and termination. Define $\Phi(u) = \sum_i u_i(1 - u_i)$. If $z_i + z_j \leq 1$, the endpoints are $(0, z_i + z_j)$ and $(z_i + z_j, 0)$; either choice decreases Φ by $2z_i z_j = 2(-a)b$. If $z_i + z_j \geq 1$, the endpoints are $(1, z_i + z_j - 1)$ and $(z_i + z_j - 1, 1)$; the decrease is $2(1 - z_i)(1 - z_j) = 2(-a)b$. Integral coordinates are never selected again. Thus the algorithm terminates after at most $d - 1$ exchanges at $y \in \mathcal{C}_k^{\mathrm{bin}}$, and telescoping gives $\sum_t (-a_t)b_t = \Phi(z)/2$. The accumulated Phase-2 loss is at most $\beta\Phi(z)/4$.

Budget-aware bound. Cauchy–Schwarz gives $\sum_i z_i^2 \geq \ell^2/d$, hence $\Phi(z) \leq \ell(d - \ell)/d$. The maximum of $\ell(d - \ell)$ over integers $0 \leq \ell \leq k$ is $q_{d,k}(d - q_{d,k})$, attained at $\ell = q_{d,k}$. Therefore $J(y) \geq J(x) - \tau - c_{d,k}\beta \geq J(x) - B_{\mathrm{unif}}$. For $d = 1$, z is already binary and $c_{d,k} = \beta = 0$. Applying the guarantee to a continuous maximizer and using $J(y) \leq \max_{\mathcal{C}_k^{\mathrm{bin}}} J$ proves the gap bound. The older charge of $\beta/8$ for each of at most $d - 1$ exchanges also proves the legacy bound $\mathcal{B}_{\mathrm{Bell}}$ used in the original evaluation. □

Proof of Proposition 3.4. The result is immediate for $k = 0$. Otherwise retain the deterministic Phase-1 point z, increments $\Delta_i = z_i - x_i$, and fractional support $F(z)$. Proposition 3.2 gives $\partial_i J \geq -\widehat{\tau}_i$ on the fill path, so $J(z) \geq J(x) - \sum_i \Delta_i \widehat{\tau}_i$. For an executed pair (i_t, j_t), it also gives $g_t'' \geq \underline{S}_{i_t j_t} \geq -\widehat{\beta}_{i_t j_t}$, where $\widehat{\beta}_{ij} := [-\underline{S}_{ij}]_+$. The one-step argument above thus charges at most $\widehat{\beta}_{i_t j_t}(-a_t)b_t/2$. The support only shrinks, so every executed pair satisfies $\widehat{\beta}_{i_t j_t} \leq \widehat{\beta}_{F(z)}$. Using the already proved potential identity yields $J(y) \geq J(x) - \sum_i \Delta_i \widehat{\tau}_i - \widehat{\beta}_{F(z)}\Phi(z)/4 = J(x) - B_{\mathrm{cand}}(x)$. Finally, $\sum_i \Delta_i < 1$, $\widehat{\tau}_i \leq \widehat{\tau}$, $\widehat{\beta}_{F(z)} \leq \widehat{\beta}$, and $\Phi(z) \leq 4c_{d,k}$ imply $B_{\mathrm{cand}}(x) \leq B_{\mathrm{unif}}$. For $d \geq 2$, $c_{d,k} \leq d/16 \leq (d - 1)/8$ proves $B_{\mathrm{unif}} \leq \mathcal{B}_{\mathrm{Bell}}$; for $d = 1$ the curvature term vanishes. □

B Proof of Proposition 4.1

We prove Proposition 4.1.

B.1 MDP construction

Use one target state with reward one and m bad states with reward zero. The target self-loops under both actions. From bad state i, the baseline action self-loops and the repair action jumps to the target. The initial distribution is uniform over the bad states. Add a dummy editable coordinate at the target whose two actions coincide, giving $d = m + 1$.

B.2 Return in closed form

Direct Bellman solution separates the return across rows. Fix bad state i. Under the mixed policy $x_i \in [0, 1]$, the transition from state i sends mass x_i to the target and mass $1 - x_i$ back to itself. The

discounted occupancy of the target, starting from state i, satisfies the Bellman equation

$$V_i = \gamma\big[(1 - x_i)\, V_i + x_i\, V_{\text{target}}\big],$$

where $V_{\text{target}} = 1/(1-\gamma)$ (the target has reward one and self-loops). Solving for V_i gives

$$\begin{aligned} V_i &= \gamma(1 - x_i)V_i + \frac{\gamma x_i}{1-\gamma} \\ V_i\big[1 - \gamma(1 - x_i)\big] &= \frac{\gamma x_i}{1-\gamma} \\ V_i &= \frac{\gamma x_i}{(1-\gamma)(1-\gamma+\gamma x_i)}, \end{aligned} \tag{20}$$

Define

$$f_\gamma(t) = \frac{\gamma t}{(1-\gamma)(1-\gamma+\gamma t)}. \tag{21}$$

Since the initial distribution is uniform over the m bad states,

$$J(x) = \frac{1}{m}\sum_{i=1}^{m} f_\gamma(x_i). \tag{22}$$

Each summand is increasing and concave. We verify concavity by computing the second derivative:

$$\begin{aligned} f'_\gamma(t) &= \frac{\gamma}{(1-\gamma)} \cdot \frac{(1-\gamma+\gamma t) - \gamma t}{(1-\gamma+\gamma t)^2} = \frac{\gamma}{(1-\gamma+\gamma t)^2}, \\ f''_\gamma(t) &= \frac{-2\gamma^2}{(1-\gamma+\gamma t)^3} < 0, \end{aligned} \tag{23}$$

B.3 Continuous and binary optima

Under budget $k = m/2$, the constraint is $\sum_{i=1}^{m+1} x_i \le m/2$. The dummy coordinate does not affect J, so the optimum sets it to zero and distributes the full budget over bad rows. Each f_γ is strictly concave. By Jensen's inequality applied to the symmetric concave problem, the continuous optimum places equal mass:

$$x_i^\star = \frac{k}{m} = \frac{1}{2} \quad \text{for every bad row } i, \qquad x_{m+1}^\star = 0.$$

A binary optimum selects any $m/2$ bad rows (setting them to 1, the rest to 0), with the dummy row at zero. By symmetry, all such selections achieve the same return. Hence

$$\begin{aligned} \mathrm{Gap}(\mathcal{C}_k) &= f_\gamma(1/2) - \tfrac{1}{2} f_\gamma(1) - \tfrac{1}{2} f_\gamma(0) \\ &= f_\gamma(1/2) - \tfrac{1}{2} f_\gamma(1) \\ &= \frac{\gamma}{(1-\gamma)(2-\gamma)} - \frac{\gamma}{2(1-\gamma)} \\ &= \frac{2\gamma - \gamma(2-\gamma)}{2(1-\gamma)(2-\gamma)} \\ &= \frac{\gamma^2}{2(1-\gamma)(2-\gamma)}, \end{aligned} \tag{24}$$

The gap is independent of m.

B.4 Interval certificate computation

Put $C_\gamma = (1-\gamma)^{-1}$. For a bad row, the canonical construction of Section 3 gives:

$$Q_i = [1/m,\ C_\gamma/m], \tag{25}$$

$$A_i = [\gamma,\ \gamma C_\gamma], \tag{26}$$

$$B_{ii} = [-\gamma C_\gamma,\ -\gamma], \tag{27}$$

and $B_{ij} = \{0\}$ for distinct bad rows $i \neq j$. We derive these as follows.

Derivation of Q_i. By (11), $Q_i = [\mu^\top \underline{h}_i,\ \mu^\top \overline{h}_i]$. The auxiliary reward e_i gives reward one at state i only. Since the initial distribution is uniform over bad states, $\mu^\top h_i = h_i(i)/m$ (the contribution from starting at state i). The minimum over all binary policies redirects state i to the target, so state i is visited only at the initial time step: $\underline{h}_i(i) = 1$. The maximum keeps state i self-looping, so it is revisited at every step: $\overline{h}_i(i) = 1 + \gamma + \gamma^2 + \cdots = C_\gamma$. Therefore $Q_i = [1/m,\ C_\gamma/m]$.

Derivation of A_i. Each bad row has zero reward under both actions and the target has reward one. The transition edit d_i redirects mass from self-loop at state i to the target. By (12), the advantage interval depends on the value envelope at the target: $\underline{V}_{\text{target}} = 1/(1-\gamma) = C_\gamma$ when all bad states self-loop, and $\overline{V}_{\text{target}} = C_\gamma$ identically, since the target self-loops regardless of other rows' actions. However, state i's own value enters through d_i^-: the lower endpoint uses $\overline{V}_i$ (largest value at state i) and the upper uses $\underline{V}_i$ (smallest). Since $\underline{V}_i = 0$ (self-loop at zero reward) and $\overline{V}_i = \gamma C_\gamma$ (redirect to target), the interval is $A_i = [\gamma(C_\gamma - \gamma C_\gamma),\ \gamma C_\gamma] = [\gamma,\ \gamma C_\gamma]$.

Derivation of B_{ii}. Row i's transition edit affects its own visitation in the opposite direction: repairing row i reduces its self-visits. Formally, $B_{ii} = \gamma[(d_i^+)^\top \underline{h}_i - (d_i^-)^\top \overline{h}_i,\ (d_i^+)^\top \overline{h}_i - (d_i^-)^\top \underline{h}_i]$. This gives the negative interval $B_{ii} = [-\gamma C_\gamma,\ -\gamma]$.

Computing $\widehat{\tau}$ and $\widehat{\beta}$. The gradient interval for bad row i is

$$G_i = Q_i \odot A_i = \left[\tfrac{1}{m} \cdot \gamma,\ \tfrac{C_\gamma}{m} \cdot \gamma C_\gamma\right] = \left[\tfrac{\gamma}{m},\ \tfrac{\gamma C_\gamma^2}{m}\right].$$

Since every lower endpoint is positive, $\widehat{\tau} = \max_i [-\inf G_i]_+ = 0$. This also means $\tau = 0$, since $\tau \le \widehat{\tau}$.

For the Hessian diagonal, the triple product $2\, Q_i \odot B_{ii} \odot A_i$ has one negative factor (B_{ii}) and two positive factors (Q_i, A_i). The most negative value is attained at the largest magnitudes: $q = \sup Q_i = C_\gamma/m$, $b = \inf B_{ii} = -\gamma C_\gamma$, $a = \sup A_i = \gamma C_\gamma$:

$$\inf \mathcal{H}_{ii} = 2 \cdot \frac{C_\gamma}{m} \cdot (-\gamma C_\gamma) \cdot \gamma C_\gamma = \frac{-2\gamma^2 C_\gamma^3}{m} = \frac{-2\gamma^2}{m(1-\gamma)^3}, \tag{28}$$

Since $B_{ij} = \{0\}$ for $i \neq j$ (distinct bad rows share no transition coupling), the cross-Hessian intervals are zero: $\mathcal{H}_{ij} = \{0\}$. An exchange between two bad rows i, j gives

$$\begin{aligned} \underline{S}_{ij} &= \inf \mathcal{H}_{ii} + \inf \mathcal{H}_{jj} - 2 \sup \mathcal{H}_{ij} \\ &= \frac{-2\gamma^2}{m(1-\gamma)^3} + \frac{-2\gamma^2}{m(1-\gamma)^3} - 0 = \frac{-4\gamma^2}{m(1-\gamma)^3}, \end{aligned} \tag{29}$$

Therefore

$$\widehat{\beta} = \max_{i<j} \left[-\underline{S}_{ij}\right]_+ = \frac{4\gamma^2}{m(1-\gamma)^3}, \tag{30}$$

which is Proposition 4.1.

Exact curvature. The exact exchange curvature along $\varepsilon_i - \varepsilon_j$ for two bad rows is

$$\begin{aligned} -(\varepsilon_i - \varepsilon_j)^\top \nabla^2 J(x)(\varepsilon_i - \varepsilon_j) &= \frac{1}{m}\left[-f_\gamma''(x_i) - f_\gamma''(x_j)\right] \\ &= \frac{1}{m}\Big[\frac{2\gamma^2}{(1-\gamma+\gamma x_i)^3} + \frac{2\gamma^2}{(1-\gamma+\gamma x_j)^3}\Big]. \end{aligned} \tag{31}$$

The supremum over $[0,1]^2$ is attained at $x_i = x_j = 0$, giving

$$\beta = \frac{4\gamma^2}{m(1-\gamma)^3} = \widehat{\beta}.$$

The envelopes are exact on the worst pair. Exchanges between the dummy row and a bad row have half the magnitude (only one nontrivial diagonal term), so the pairwise maximum is indeed the bad–bad exchange.

B.5 CERTIFICATE AND RATIO

Since $\widehat{\tau} = 0$ and $d - 1 = m$,

$$\mathcal{B}_{\mathrm{Bell}} = \widehat{\tau} + \frac{d-1}{8}\widehat{\beta} = \frac{m}{8} \cdot \frac{4\gamma^2}{m(1-\gamma)^3} = \frac{\gamma^2}{2(1-\gamma)^3}, \tag{32}$$

This is independent of m.

Dividing (24) by (32),

$$\begin{aligned} \frac{\mathrm{Gap}(\mathcal{C}_k)}{\mathcal{B}_{\mathrm{Bell}}} &= \frac{\gamma^2}{2(1-\gamma)(2-\gamma)} \cdot \frac{2(1-\gamma)^3}{\gamma^2} \\ &= \frac{(1-\gamma)^2}{2-\gamma}, \end{aligned} \tag{33}$$

It tends to $1/2$ as $\gamma \downarrow 0$ and satisfies $\mathrm{Gap}(\mathcal{C}_k)/\mathcal{B}_{\mathrm{Bell}} \sim (1-\gamma)^2$ as $\gamma \uparrow 1$. In particular, $\mathrm{Gap}(\mathcal{C}_k) = \Theta(\mathcal{B}_{\mathrm{Bell}})$ as $m \to \infty$ at any fixed γ, so no bound of the form $\tau + c\beta$ with c independent of d can be sound on this family.

At $m = 4$, $\gamma = 1/2$: $\widehat{\beta} = 2$, $\mathcal{B}_{\mathrm{Bell}} = 1$, and $\mathrm{Gap}(\mathcal{C}_k) = 1/6$.

For the candidate bound, $\Phi(x^\star) = m/4$ and $\widehat{\beta}_{F(x^\star)} = \widehat{\beta}$, so $B_{\mathrm{cand}}(x^\star) = \gamma^2/[4(1-\gamma)^3]$. Since $q_{m+1,m/2} = m/2$, the new coefficient is $c_{m+1,m/2} = m(m+2)/[16(m+1)]$ and $B_{\mathrm{unif}} = \frac{m+2}{m+1} B_{\mathrm{cand}}(x^\star)$. This proves the remaining claims of Proposition 4.1.

B.6 WEIGHTED EXCHANGE INTERPOLATION

General identity. Let $g \in C^2([a,b])$ with $a < 0 < b$, and use the kernel $K_{a,b}$ defined in Section 4. Taylor's formula gives $g(a) = g(0) + ag'(0) + \int_a^0 (t-a)g''(t)\,dt$ and $g(b) = g(0) + bg'(0) + \int_0^b (b-t)g''(t)\,dt$. Multiply by $b/(b-a)$ and $(-a)/(b-a)$, respectively, and add. The first-order terms cancel, yielding the interpolation identity

$$g(0) - \frac{b\,g(a) - a\,g(b)}{b-a} = -\int_a^b K_{a,b}(t) g''(t)\,dt. \tag{34}$$

The weighted endpoint average does not exceed the larger endpoint. Since $K_{a,b} \geq 0$, taking positive parts bounds the actual loss by $\int_a^b K_{a,b}(t)[-g''(t)]_+\,dt$. Integration gives $\int_a^b K_{a,b}(t)\,dt = (-a)b/2$, recovering the constant-curvature charge as a special case.

Exact segment maximum on the family. At $x^\star$, every exchange pairs two remaining half-integral bad rows, with interval $[-1/2, 1/2]$. Each endpoint resolves both coordinates; after $m/2$ exchanges the policy selects $m/2$ bad rows and is binary-optimal. Let $\delta = 1 - \gamma$. Up to an additive constant, $g(t) = [f_\gamma(1/2+t) + f_\gamma(1/2-t)]/m$. Thus $-g''(t)$ equals $\frac{2\gamma^2}{m}\{[\delta + \gamma(1/2+t)]^{-3} + [\delta + \gamma(1/2-t)]^{-3}\}$. It is convex and symmetric, so its maximum is attained at either endpoint, giving $\beta_{\mathrm{seg}} = 2\gamma^2(\delta^{-3}+1)/m$. Multiplying by $1/8$ per exchange and by $m/2$ exchanges gives $B_{\mathrm{seg}} = \gamma^2(\delta^{-3}+1)/8$. Dividing by the exact gap yields the ratio stated in Section 4.

Exact weighted loss and reliable upper sums. Here $g'' < 0$ and $g(-1/2) = g(1/2)$, so (34) is an equality for actual loss. Direct substitution gives the one-step integral $[2f_\gamma(1/2) - f_\gamma(0) - f_\gamma(1)]/m = \gamma^2/[m\delta(2-\gamma)]$. Summing proves $B_{\mathrm{int}} = \gamma^2/[2\delta(2-\gamma)] = \mathrm{Gap}(\mathcal{C}_k)$. For a partition into intervals $I = [u,v]$, convexity of $-g''$ on this family implies $r_I := \max\{-g''(u), -g''(v)\} \geq [-g''(t)]_+$ on I. The sum $\sum_I r_I \int_I K_{a,b}(t)\,dt$ is consequently a reliable upper bound. Rational inputs make every term rational.

Scope for general MDPs. If reliable envelopes $[-g_t''(s)]_+ \leq \rho_t(s) \leq \widehat{\beta}_{i_t j_t}$ are available along a trajectory, replacing each charge by $\int_{a_t}^{b_t} K_{a_t,b_t}(s)\rho_t(s)\,ds$ gives a bound no larger than B_{trace} after adding the same Phase-1 charges. For a general MDP, polynomial enclosures bound curvature throughout each subinterval.

B.7 General exchange computation

Proof of Proposition 4.3. Fix the current policy z and an exchange of rows i, j. Write $A = I - \gamma P(z)$, $R = A^{-1}$, $V = Rr(z)$, and $U = [e_i, -e_j]$. Let $W = \gamma[d_i, d_j]$ and $\eta = (\Delta r_i, \Delta r_j)^\top$. Then $A(t) = A - tUW^\top$ and $r(t) = r(z) + tU\eta$. Define $M = W^\top RU$, $w = \eta + W^\top V$, and $c^\top = \mu^\top RU$. The resolvent identity gives $V(t) = V + tRU(I - tM)^{-1}w$, hence $g(t) = J(z) + tc^\top(I - tM)^{-1}w$. Set $q(t) = \det(I - tM) = 1 - t \operatorname{tr} M + t^2 \det M$ and $p(t) = J(z)q(t) + tc^\top \operatorname{adj}(I - tM)w$. Both polynomials have degree at most two. The determinant lemma gives $q(t) = \det A(t)/\det A$. Every feasible policy has an invertible discounted resolvent, so q has no zero on the exchange interval. Since $q(0) = 1$, continuity gives $q > 0$. Finally, let $N = p'q - pq'$. Its cubic coefficients cancel, so $\deg N \le 2$. Differentiating again gives $g'' = h/q^3$, with $h = N'q - 2Nq'$ and $\deg h \le 3$. □

Reliable subinterval integration. For a subinterval $I = [u, v]$, compute exact extrema $0 < q_I^- \le q(t) \le q_I^+$ by checking its endpoints and the quadratic stationary point when it lies inside I. After the substitution $t = u + (v - u)s$, express $-h(t)$ in the Bernstein basis on $s \in [0, 1]$. Nonnegative basis weights sum to one, so the minimum and maximum coefficients give bounds $n_I^- \le -h(t) \le n_I^+$. Consequently, $\rho_I^- := \max\{0, n_I^-\}/(q_I^+)^3$ and $\rho_I^+ := \max\{0, n_I^+\}/(q_I^-)^3$ enclose $[-g''(t)]_+$ on I. Include zero as a partition boundary, so $K_{a,b}$ is linear on each piece. The exact weights $\omega_I = \int_I K_{a,b}(t)\,dt$ are nonnegative. Therefore $L = \sum_I \rho_I^- \omega_I$ and $U = \sum_I \rho_I^+ \omega_I$ enclose the weighted curvature integral. The implementation bisects the piece with largest uncertainty $(\rho_I^+ - \rho_I^-)\omega_I$. Every finite partition is sound, even if the refinement budget is exhausted. Exact rational inputs make coefficients, extrema, weights, and comparisons rational.

Accumulation and relation to global bounds. Let U_t be the upper sum for the tth executed exchange. The interpolation inequality bounds each positive exchange loss by U_t. Adding $[J(x) - J(z)]_+$ for Phase 1 proves $[J(x) - J(y)]_+ \le B_{\mathrm{loc}}$. The value $J(z)$ is obtained by the initial resolvent solve. If global Bellman bounds are already available, one may instead use $B_{\mathrm{hyb}} := \sum_i \Delta_i \widehat{\tau}_i + \sum_t \min\{U_t, \widehat{\beta}_{i_t j_t}(-a_t)b_t/2\}$. Then $[J(x) - J(y)]_+ \le B_{\mathrm{hyb}} \le B_{\mathrm{trace}} \le B_{\mathrm{cand}} \le B_{\mathrm{unif}}$. No ordering between B_{loc} and B_{cand} is required. Both bounds are computed along the chosen rounding trajectory.

Computation. With a stored resolvent, forming the two-by-two matrix and updating the selected endpoint costs $O(S^2)$ arithmetic operations per exchange. One initial dense inverse costs $O(S^3)$. Thus T exchanges and L total subinterval computations cost $O(S^3 + TS^2 + L)$ arithmetic operations. Bit complexity also depends on rational operand lengths. The same representation computes both endpoint returns with constant-size rational operations. The low-rank direct baseline therefore uses precisely the same coefficient construction and resolvent updates, omitting curvature enclosures. It selects the same endpoint as full exact policy evaluation. Integral certification augments these updates with curvature enclosures to sharpen the trajectory loss bound.

C Proof of Theorem 5.2

This appendix proves Theorem 5.2. Begin with an arbitrary simple MAX-CUT instance (G_0, K_0). Add a disjoint edge and four isolated vertices, call the padded graph (G, K), and put $K = K_0 + 1$. Max-cut values add across components, so the preprocessing preserves YES/NO. All notation below refers to the padded graph. In particular,

$$n = |V(G)| \ge 6, \qquad m = |E(G)| \ge 1, \qquad \Lambda = 2m + 1.$$

Orient every edge by increasing endpoint. The proof constructs the restricted MDP, derives its reduced multi-affine objective, identifies the designated exchange with a cut polynomial, separates that pair from all others, and finally checks budget feasibility and polynomial encoding size.

C.1 MDP construction

Use the state and control-row ordering

$$p, q, L_1, \ldots, L_n, R_1, \ldots, R_n, z, \qquad N = 2n + 3,$$

and define

$$\eta = \frac{1}{64N^2}, \qquad \rho = \frac{\eta^4}{64\Lambda N}, \qquad \delta = \eta^3, \tag{35}$$

so that $\rho = 1/(2^{30}\Lambda N^9)$. Let C have the noncompensating entries

$$C_{pq} = \eta, \qquad C_{qL_i} = \eta, \qquad C_{qR_j} = \eta\rho(\Lambda - \deg^-(j)), \tag{36}$$

$$C_{L_iR_i} = -2\Lambda\rho, \qquad C_{L_iR_j} = 2\rho \quad (i \to j \in E), \tag{37}$$

and put the negative row sum in column z for every nonzero row. Thus $C\mathbf{1} = 0$. Set

$$a_{L_i} = \rho(\Lambda - \deg^+(i)), \qquad a_{R_j} = 1, \tag{38}$$

with all other entries of a zero.

The rank-one baseline kernel is $P^0 = \mathbf{1}\pi^\top$, where

$$\pi_q = \pi_{L_i} = \delta, \qquad \pi_{R_i} = 8\Lambda\rho, \qquad \pi_z = \frac{1}{4}, \qquad \pi_p = 1 - \sum_{s \neq p} \pi_s. \tag{39}$$

Take $\gamma = 1/2$, $\mu = \varepsilon_p$, $r^0 = 0$, and define the alternative endpoints by

$$P^1 = P^0 + 2C, \qquad r^1 = a. \tag{40}$$

Verification of the raw promises. We check that both P^0 and P^1 are valid stochastic matrices and that rewards lie in $[0, 1]$.

Baseline positivity. The mass outside state p is

$$\begin{aligned} \sum_{s \neq p} \pi_s &= \frac{1}{4} + (n+1)\delta + n \cdot 8\Lambda\rho = \frac{1}{4} + (n+1)\eta^3 + 8n\Lambda\rho \\ &< \frac{1}{4} + \frac{n+1}{64^3 N^6} + \frac{8n}{2^{30}N^9} < \frac{1}{2}, \end{aligned} \tag{41}$$

The last inequality uses $n < N$ and $N \geq 15$. Hence $\pi_p = 1 - \sum_{s \neq p} \pi_s > 1/2 > 0$, and since every other component of π is manifestly positive, $P^0 = \mathbf{1}\pi^\top$ has all positive entries.

Stochasticity of P^1. Row stochasticity of $P^1 = P^0 + 2C$ follows from $C\mathbf{1} = 0$: $P^1\mathbf{1} = P^0\mathbf{1} + 2C\mathbf{1} = \mathbf{1} + 0 = \mathbf{1}$.

Positivity of P^1. The only negative transition edits occur at $p \to z$, $q \to z$, and $L_i \to R_i$:

1. At $p \to z$: the compensating entry $C_{pz} = -\eta$, so $P^1_{pz} = \pi_z + 2C_{pz} = 1/4 - 2\eta > 0$ because $\eta < 1/8$.
2. At $q \to z$: the compensating entry satisfies

$$\begin{aligned} |C_{qz}| &= \sum_i C_{qL_i} + \sum_j C_{qR_j} = n\eta + \eta\rho \sum_j (\Lambda - \deg^-(j)) \\ &= n\eta + \eta\rho(n\Lambda - m) < 2n\eta, \end{aligned} \tag{42}$$

   Here $\sum_j \deg^-(j) = m$, and the final bound follows from $\rho < 1/\Lambda$. Hence $P^1_{qz} = \pi_z - 2|C_{qz}| > 1/4 - 4n\eta > 0$.
3. At $L_i \to R_i$: $P^1_{L_iR_i} = \pi_{R_i} + 2C_{L_iR_i} = 8\Lambda\rho - 4\Lambda\rho = 4\Lambda\rho > 0$.

All other edits are nonnegative.

Reward bounds. By (38), $0 < a_{L_i} = \rho(\Lambda - \deg^+(i)) \leq \rho\Lambda < 1$, and $a_{R_j} = 1$. All remaining rewards are zero.

Support structure. The nonzero entries of C define a DAG with edges $p \to q$, $q \to L_i$, $q \to R_j$, and $L_i \to R_j$. The longest directed path is $p \to q \to L_i \to R_j$, which has length three. Every compensator edge ends at z, which has no outgoing edges in the DAG. Hence $\mathcal{G}(C)$ has depth at most three, confirming $\mathcal{M}_G \in \mathfrak{B}_3$.

C.2 Resolvent and reduced objects

Since $P^0 = \mathbf{1}\pi^\top$ is rank one and $\gamma = 1/2$, the resolvent has a closed form via the Neumann series:

$$\begin{aligned} R_0 &= (I - \gamma P^0)^{-1} = (I - \tfrac{1}{2}\mathbf{1}\pi^\top)^{-1} \\ &= \sum_{t=0}^{\infty} (\tfrac{1}{2})^t (\mathbf{1}\pi^\top)^t = I + \sum_{t=1}^{\infty} (\tfrac{1}{2})^t \mathbf{1}\pi^\top \\ &= I + \mathbf{1}\pi^\top, \end{aligned} \tag{43}$$

The Neumann series converges because $\rho(\frac{1}{2}\mathbf{1}\pi^\top) = 1/2$. For $t \geq 1$, $(\mathbf{1}\pi^\top)^t = \mathbf{1}\pi^\top$, and the remaining scalar series sums to one.

The initial visitation vector is

$$u^\top = \mu^\top R_0 = \varepsilon_p^\top (I + \mathbf{1}\pi^\top) = \varepsilon_p^\top + \pi^\top. \tag{44}$$

In particular, $U \coloneqq u_p = 1 + \pi_p$ and $u_s = \pi_s$ for $s \neq p$.

Reduced transition matrix. Since $P^1 - P^0 = 2C$, the reduced matrix is

$$\begin{aligned} \gamma(P^1 - P^0)R_0 &= \tfrac{1}{2} \cdot 2C \cdot (I + \mathbf{1}\pi^\top) = C(I + \mathbf{1}\pi^\top) \\ &= C + C\mathbf{1}\pi^\top = C, \end{aligned} \tag{45}$$

The final equality uses $C\mathbf{1} = 0$. Thus the reduced objects are exactly C, a, and u, and the reduced performance identity (5) holds.

C.3 Expansion of the objective

Because $\mathcal{G}(C)$ has depth three, $(XC)^4 = 0$ for any diagonal $X = \mathrm{diag}(x)$ with $x \in [0,1]^N$. The Neumann inverse in (5) therefore truncates:

$$(I - XC)^{-1} = I + XC + (XC)^2 + (XC)^3. \tag{46}$$

Write

$$A_i = \Lambda - \deg^+(i), \quad B_j = \Lambda - \deg^-(j), \quad c_{ij} = 2\rho(\mathbf{1}_{i \to j} - \Lambda \mathbf{1}_{i=j}),$$

and let $U = u_p$, $\ell_i = x_{L_i}$, and $r_j = x_{R_j}$.

Path enumeration. Each nonzero term in $u^\top (I - XC)^{-1} Xa$ corresponds to a directed path in $\mathcal{G}(C)$ ending at a state with nonzero reward. The nonzero rewards are $a_{L_i} = \rho A_i$ and $a_{R_j} = 1$. We enumerate every contributing path by length.

Length zero (the identity term): $u_s \cdot x_s \cdot a_s$ is nonzero only for $s = L_i$ or $s = R_j$. This gives $\sum_i \pi_{L_i} \ell_i \rho A_i + \sum_j \pi_{R_j} r_j = \delta\rho \sum_i A_i \ell_i + 8\Lambda\rho \sum_j r_j$.

Length one (the XC term): paths of the form $s \to t$ with $a_t \neq 0$. The only edges ending at a rewarded state are $L_i \to R_j$ (with coefficient c_{ij}) and $q \to L_i$ (with coefficient η). These contribute $\sum_{i,j} u_{L_i} \ell_i c_{ij} r_j \cdot 1 + \sum_i u_q x_q \eta \ell_i \rho A_i$. The first sum is $\delta \sum_{i,j} c_{ij} \ell_i r_j$. The second is $\delta\eta\rho x_q \sum_i A_i \ell_i$.

Length two and three: paths $q \to L_i \to R_j$ and $p \to q \to L_i \to R_j$ (arriving at R_j with reward 1), plus $p \to q \to L_i$ (arriving at L_i with reward ρA_i). Collecting all terms involving the cut polynomial g_G (defined below) and combining with the shorter paths gives

$$\begin{aligned} J(x) = \delta\rho \sum_i A_i \ell_i + 8\Lambda\rho \sum_j r_j + \delta \sum_{i,j} c_{ij} \ell_i r_j \\ + \delta\eta\rho\, x_q g_G(\ell, r) + U\eta^2 \rho\, x_p x_q g_G(\ell, r), \end{aligned} \tag{47}$$

where

$$g_G(\ell, r) = \Lambda \sum_i (\ell_i + r_i - 2\ell_i r_i) - \sum_{i \to j \in E} (\ell_i + r_j - 2\ell_i r_j). \tag{48}$$

Derivation of the designated Hessian. Every monomial in (47) is multi-affine (degree at most one in each variable). Hence all pure second derivatives vanish: $\partial_{pp}J = \partial_{qq}J = 0$. The mixed derivative $\partial_{pq}J$ extracts only the terms containing both x_p and x_q, which is the last term in (47):

$$\partial_{pq}J(x) = \partial_p\partial_q\big[U\eta^2\rho\, x_p x_q\, g_G(\ell, r)\big] = U\eta^2\rho\, g_G(\ell, r). \tag{49}$$

Therefore the exchange curvature along $\varepsilon_p - \varepsilon_q$ is

$$-H_{\mathcal{M}_G,\varepsilon_p-\varepsilon_q}(x) = -(\varepsilon_p - \varepsilon_q)^\top \nabla^2 J(x)(\varepsilon_p - \varepsilon_q) = 2\partial_{pq}J(x) = \alpha_G g_G(\ell, r), \tag{50}$$

with $\alpha_G = 2U\eta^2\rho > 0$. The factor of 2 arises because the off-diagonal entry of the Hessian appears twice (the (p, q) and (q, p) entries) and with a sign flip from the exchange direction $\varepsilon_p - \varepsilon_q$.

C.4 Maximizing the cut polynomial

The maximum of a multi-affine polynomial on a cube occurs at a vertex. We show that every vertex maximizer of g_G over $\{0, 1\}^{2n}$ is complementary ($r_i = 1 - \ell_i$ for all i).

Non-complementary vertices are suboptimal. At a binary vertex $(\ell, r) \in \{0, 1\}^{2n}$, let D count the number of indices with $r_i \neq 1 - \ell_i$. For such an index, the contribution to the Λ-term is:

- If $(\ell_i, r_i) = (0, 0)$: $\Lambda(0 + 0 - 0) = 0$.
- If $(\ell_i, r_i) = (1, 1)$: $\Lambda(1 + 1 - 2) = 0$.

In both cases, vertex i contributes zero to the Λ-sum, compared to Λ for a complementary pair. The edge terms $-(\ell_i + r_j - 2\ell_i r_j)$ have absolute value at most 1 per edge. Therefore

$$\begin{aligned} g_G(\ell, r) &\leq \Lambda(n - D) + m \leq \Lambda(n - 1) + m \\ &= \Lambda n - \Lambda + m = \Lambda n - (2m + 1) + m = \Lambda n - m - 1, \end{aligned} \tag{51}$$

The first inequality bounds the edge terms by m. The second uses $D \geq 1$, and the last equality substitutes $\Lambda = 2m + 1$.

Complementary vertices achieve the cut value. A complementary assignment $r_i = 1 - \ell_i$ gives $\ell_i + r_i - 2\ell_i r_i = 1$ for every i, so the Λ-sum equals Λn. For each edge $i \to j \in E$, let $S_i = \ell_i$ encode the partition. Then

$$\begin{aligned} \ell_i + r_j - 2\ell_i r_j &= S_i + (1 - S_j) - 2S_i(1 - S_j) \\ &= S_i + 1 - S_j - 2S_i + 2S_iS_j = 1 - S_i - S_j + 2S_iS_j \\ &= 1 - \mathbf{1}_{S_i \neq S_j}, \end{aligned} \tag{52}$$

Hence $\sum_{i\to j}(\ell_i + r_j - 2\ell_i r_j) = m - \mathrm{cut}_G(S)$, and

$$g_G(\ell, r) = \Lambda n - (m - \mathrm{cut}_G(S)) = \Lambda n - m + \mathrm{cut}_G(S). \tag{53}$$

Since $\mathrm{cut}_G(S) \geq 0$, a complementary vertex achieves at least $\Lambda n - m > \Lambda n - m - 1$, which exceeds the non-complementary bound (51). Therefore every maximizer is complementary, and

$$\max_{[0,1]^{2n}} g_G = \Lambda n - m + \mathrm{MaxCut}(G). \tag{54}$$

The right side is positive, so the positive part in κ does not alter the supremum.

C.5 Bounds for all row pairs

Equation (47) lists every possible nonzero mixed derivative. We show that the designated pair (p, q) strictly dominates all others.

Zero pairs. Pairs containing z are zero because x_z appears in no term of (47). Pairs within the L block (L_i, L_j with $i \neq j$) are zero because no monomial contains $\ell_i\ell_j$ (multi-affinity and the DAG structure preclude such products). Similarly, pairs within the R block are zero.

Coefficient bounds for nonzero pairs. For a multi-affine polynomial $h(x_1, \ldots, x_N)$, the supremum of $|h|$ over $[0,1]^N$ is at most the ℓ_1-norm of its coefficients (the sum of absolute values of all coefficients). Define

$$M = U\eta^2\rho\Lambda = \frac{\alpha_G\Lambda}{2}. \tag{55}$$

The absolute column sums of c_{ij} are

$$\sum_j |c_{ij}| = 2\rho(\Lambda + \deg^+(i)), \qquad \sum_i |c_{ij}| = 2\rho(\Lambda + \deg^-(j)), \tag{56}$$

since each row i of c has one entry $-2\Lambda\rho$ (the diagonal) and $\deg^+(i)$ entries of $+2\rho$.

We bound each pair type. Let d denote the relevant in- or out-degree.

Pair (p, L_i) or (p, R_j): The only term in (47) involving both x_p and a graph variable is the last term $U\eta^2\rho\, x_p x_q\, g_G(\ell, r)$. The coefficient ℓ_1-norm of g_G restricted to terms involving ℓ_i (or r_j) is at most $3\Lambda + \deg^+(i)$:

$$\|g_G|_{\ell_i}\|_1 = (\Lambda - \deg^+(i)) + 2\Lambda + 2\deg^+(i) = 3\Lambda + \deg^+(i). \tag{57}$$

The three terms come from the linear coefficient, the bilinear $\ell_i r_i$ coefficient, and the outgoing-edge bilinears, respectively. Hence $|\partial_{p,L_i} J| \le U\eta^2\rho(3\Lambda + \deg^+(i))$. Since $\deg^+(i) \le m < 2m+1 = \Lambda$,

$$U\eta^2\rho(3\Lambda + d) < U\eta^2\rho \cdot 4\Lambda = 4M. \tag{58}$$

Pair (q, L_i) or (q, R_j): Both the $\delta\eta\rho\, x_q\, g_G$ and $U\eta^2\rho\, x_p x_q g_G$ terms contribute. The combined coefficient is $(\delta\eta + U\eta^2)\rho(3\Lambda + \deg^+(i))$. Since $\delta/(\eta U) = \eta^3/(\eta U) = \eta^2/U$, and $U = 1 + \pi_p > 1$ while $\eta^2 < 1/16$, we have $\delta\eta < U\eta^2/16$. Therefore

$$(\delta\eta + U\eta^2)\rho(3\Lambda + \deg^+(i)) < \frac{17}{16} U\eta^2\rho \cdot 4\Lambda = \frac{17}{4} M. \tag{59}$$

Pair (L_i, R_j): The terms involving both ℓ_i and r_j come from $\delta c_{ij}\ell_i r_j$, $\delta\eta\rho x_q$(terms in g_G), and $U\eta^2\rho x_p x_q$(terms in g_G). The coefficient bound is $|c_{ij}|(\delta + \delta\eta + U\eta^2)$. Since $\delta(1+\eta)/(U\eta^2) = \eta(1+\eta)/U < 1/16$ (using $\eta < 1/8$ and $U > 1$),

$$|c_{ij}|\{\delta(1+\eta) + U\eta^2\} < \frac{17}{16}|c_{ij}|U\eta^2. \tag{60}$$

The largest $|c_{ij}|$ is $2\Lambda\rho$ (the diagonal entry). Hence

$$|\partial_{L_i,R_j} J| < \frac{17}{16} \cdot 2\Lambda\rho \cdot U\eta^2 = \frac{17}{8} \cdot M < \frac{9}{4} M. \tag{61}$$

These bounds are summarized in the following table:

pair	coefficient-ℓ_1 bound for $\partial_{ab} J$
(p, L_i), (p, R_j)	$U\eta^2\rho(3\Lambda + \deg^+(i)) < 4M$
(q, L_i), (q, R_j)	$(\delta\eta + U\eta^2)\rho(3\Lambda + \deg^+(i)) < 17M/4$
(L_i, R_j)	$\lvert c_{ij}\rvert\{\delta(1+\eta) + U\eta^2\} < 9M/4$

Separation of the designated pair. Exchange curvature is twice the corresponding mixed derivative. Thus every undesignated pair satisfies

$$\kappa(\varepsilon_a - \varepsilon_b) < 2 \cdot \frac{17}{4} M = \frac{17}{2} M. \tag{62}$$

In contrast, the designated pair (p, q) achieves

$$\begin{aligned} \kappa(\varepsilon_p - \varepsilon_q) &= \alpha_G \max_{[0,1]^{2n}} g_G = \alpha_G(\Lambda n - m + \mathrm{MaxCut}(G)) \\ &= 2M(n - m/\Lambda) \ge 2M\left(n - \frac{m}{2m+1}\right) \\ &> 2M\left(6 - \frac{1}{2}\right) = 11M, \end{aligned} \tag{63}$$

Here $\alpha_G\Lambda = 2M$, $m/\Lambda < 1/2$, and $n \ge 6$. Since $11M > 17M/2$, the designated pair strictly realizes the all-pairs maximum β, proving (5).

C.6 Feasibility under a row budget

For completeness, define the exchange-feasible and binary curvature restricted to the budget set by

$$\beta^{\leftrightarrow}_{\mathcal{M},k} := \max_{a<b} \sup_{\substack{x\in\mathcal{C}_k:\\ x\pm\delta(\varepsilon_a-\varepsilon_b)\in\mathcal{C}_k \text{ for some } \delta>0}} [-H_{\mathcal{M},\varepsilon_a-\varepsilon_b}(x)]_+ , \tag{64}$$

$$\beta^{\mathrm{bin}}_{\mathcal{M},k} := \max_{a<b} \max_{x\in\mathcal{C}^{\mathrm{bin}}_k} [-H_{\mathcal{M},\varepsilon_a-\varepsilon_b}(x)]_+ . \tag{65}$$

These quantities are used only to state the budget bridge for the reduction image. Theorem 5.2 itself concerns the unrestricted-cube quantities.

Equation (50) shows that the designated exchange curvature depends only on (ℓ, r), not on x_p, x_q, x_z. We show that the curvature maximum is attainable within $\mathcal{C}_{n+1}$.

Mass accounting. The proof of (54) shows that every maximizing graph assignment is complementary:

$$\ell_i + r_i = 1 \quad (i = 1, \dots, n).$$

The total mass of such an assignment is

$$\sum_{i=1}^{n} (\ell_i + r_i) = \sum_{i=1}^{n} 1 = n. \tag{66}$$

Setting $x_p = x_q = 1/2$ and $x_z = 0$ gives total mass

$$\mathbf{1}^\top x = n + \frac{1}{2} + \frac{1}{2} + 0 = n + 1 \le k, \tag{67}$$

so $x \in \mathcal{C}_{n+1}$.

Exchange feasibility. The exchange direction $\varepsilon_p - \varepsilon_q$ satisfies $\mathbf{1}^\top(\varepsilon_p - \varepsilon_q) = 0$, so moving along it preserves $\mathbf{1}^\top x = n + 1$. Starting from $x_p = x_q = 1/2$, both $x_p + \varepsilon \le 1$ and $x_q - \varepsilon \ge 0$ hold for $\varepsilon \in (-1/2, 1/2)$. Therefore the exchange can move a nonzero distance in both directions while remaining in $\mathcal{C}_{n+1}$, and the maximum of g_G is attained. It follows that

$$\beta^{\leftrightarrow}_{\mathcal{M}_G,k} = \beta_{\mathcal{M}_G}. \tag{68}$$

Binary witness. For the binary quantity, set $x_p = x_q = x_z = 0$ and choose any complementary binary assignment for the graph variables. The total mass is $n \le n + 1 = k$, so $x \in \mathcal{C}^{\mathrm{bin}}_k$. The designated exchange evaluates to $\alpha_G g_G(\ell, r)$, which equals $\beta_{\mathcal{M}_G}$ at any cut-maximizing complementary assignment. Hence $\beta^{\mathrm{bin}}_{\mathcal{M}_G,k} = \beta_{\mathcal{M}_G}$. Combined with (68),

$$\beta^{\leftrightarrow}_{\mathcal{M}_G,k} = \beta_{\mathcal{M}_G} = \beta^{\mathrm{bin}}_{\mathcal{M}_G,k}. \tag{69}$$

The bridge is specific to the reduction image.

C.7 Reduction and encoding size

Map the padded threshold $K = K_0 + 1$ to

$$T_G = \alpha_G(\Lambda n - m + K).$$

Correctness of the reduction. Equations (54) and (5) yield a chain of equivalences:

$$\begin{aligned} \mathrm{MaxCut}(G_0) \ge K_0 &\iff \mathrm{MaxCut}(G) \ge K \\ &\iff \Lambda n - m + \mathrm{MaxCut}(G) \ge \Lambda n - m + K \\ &\iff \alpha_G(\Lambda n - m + \mathrm{MaxCut}(G)) \ge T_G \\ &\iff \beta_{\mathcal{M}_G} \ge T_G, \end{aligned} \tag{70}$$

The first equivalence uses the disjoint padding edge. The remaining steps add the constant $\Lambda n - m$, multiply by $\alpha_G > 0$, and apply (5).

Encoding size. All parameters are recomputed from the padded graph. From (35) and $m = O(n^2)$:

$$\begin{aligned}\log_2 \eta &= O(\log N),\\ \log_2 \rho &= 4\log_2 \eta + O(\log(\Lambda N)) = O(\log N),\\ \log_2 \delta &= 3\log_2 \eta = O(\log N).\end{aligned} \tag{71}$$

Each rational entry of P^0, P^1, r^0, r^1 has numerator and denominator with $O(\log N)$ bits. The MDP has $N = 2n + 3$ states, N control rows, two actions per row, and N^2 transition entries per action, giving $O(N^2)$ rational entries total. The dense raw MDP therefore has encoding size $O(N^2 \log N) = \mathrm{poly}(n)$.

NP-completeness of the binary language. A certificate for the binary threshold problem $\beta_{\mathcal{M}}^{\mathrm{bin}} \geq T$ consists of:

1. A pair of rows (a, b).
2. A binary control vector $x \in \{0, 1\}^N$.

Given this certificate, verification proceeds in three steps: (i) check $\mathcal{M} \in \mathfrak{B}_3$ (polynomial from the rational input, by Definition 5.1). (ii) compute $R_0 = (I - \gamma P^0)^{-1}$ via exact Gaussian elimination (polynomial intermediate bit length). (iii) evaluate $-H_{\mathcal{M},\varepsilon_a - \varepsilon_b}(x)$ from the reduced form and compare with T. All steps run in polynomial time with exact rational arithmetic. Hence the binary language is in NP, and the reduction from MAX-CUT shows it is NP-hard. Therefore the binary language is NP-complete.

On the reduction image, the continuous maximum is attained at binary graph controls (a maximizer of g_G can be chosen at a vertex), proving continuous NP-hardness.

D Evaluation protocol

This appendix describes the protocol behind Section 6. Unless explicitly marked B_{unif}, the recorded model-only values use the legacy $\mathcal{B}_{\mathrm{Bell}} = \widehat{\tau} + (d-1)\widehat{\beta}/8$. The revised uniform bound is evaluated on the closed-form and coupled suites. The structured-suite coverage and recorded scaling times retain their original baseline.

D.1 Arithmetic discipline

The evaluation uses exact and validated quantities for different purposes. Whenever binary optima, witnessed advantages, or exact theorem checks are reported, all model entries, returns, and certificates are represented with Python's fractions. Fraction. Continuous multistart search is heuristic and supplies a feasible candidate quantized to a rational point before exact evaluation. Certificate checks use this fixed candidate. The large-scale study uses directed outward-rounded binary64 arithmetic: Bellman solutions are corrected until their inequalities hold outward, and every downstream interval operation is outward rounded. Naive midpoint floating point supplies the uncertified comparison.

D.2 Grid specification

The evaluation grid crosses five factors:

1. **Domain** (4 levels): GRIDWORLD, QUEUE, MAINTENANCE, and GARNET.
2. **Dimension** (3 levels): $d \in \{6, 8, 10\}$ with $S = d$ states.
3. **Budget** (2 levels): a smaller budget $k_{\mathrm{lo}} = \lceil d/4 \rceil$ and a larger budget $k_{\mathrm{hi}} = \lfloor d/2 \rfloor$.
4. **Discount factor** (2 levels): $\gamma \in \{1/2,\ 4/5\}$.
5. **Edit type** (3 levels): transition-only, reward-only, and mixed (both transition and reward).

The Cartesian product contains $4 \times 3 \times 2 \times 2 \times 3 = 144$ cases. The domain generator also assigns a repair scale of $\frac{3}{5}$ or 1 and a weak or strong baseline. Each domain appears in 36 cases, each budget, discount, and edit-type level appears in 72 cases. Case identifiers and random seeds are deterministic functions of the grid indices. The base seed is 2026081300, fixed before execution.

D.3 PER-CASE EXECUTION SEQUENCE

Each case follows the same sequence of seven stages.

Stage 1: model construction. The code constructs the rational MDP from the domain generator. Every transition probability and reward is stored as an exact rational. The domain generator uses the case seed to produce the state graph, transition kernels, and reward vectors.

Stage 2: Bellman bounds. The code computes the $2d + 2$ extremal Bellman value and visitation vectors: $\overline{V}, \underline{V}$ (the upper and lower value bounds) and $\overline{h}_i, \underline{h}_i$ for each editable row i (the upper and lower auxiliary visitation bounds). Each solve is a rational linear system $(I - \gamma P)\, V = r$ under a specific binary policy. Every Bellman inequality is verified exactly after computation:

$$\underline{V} \leq V^{\pi} \leq \overline{V} \quad \text{for every binary policy } \pi,$$

using the fact that the extremal policies attain the bounds.

Stage 3: certificate computation. The code computes the interval quantities (14) through (16):

- The gradient intervals $G_i = Q_i \odot A_i$ for each row i, yielding $\widehat{\tau} = \max_i \left[-\inf G_i\right]_+$.
- The Hessian diagonal intervals $\mathcal{H}_{ii}$ and cross intervals $\mathcal{H}_{ij}$, yielding the exchange bound $\underline{S}_{ij}$ for each pair (i, j).
- The curvature bound $\widehat{\beta} = \max_{i<j} \left[-\underline{S}_{ij}\right]_+$.
- The combined certificate $\mathcal{B}_{\mathrm{Bell}} = \widehat{\tau} + \frac{d-1}{8}\widehat{\beta}$.

As a cross-check, the calculation is repeated from saved rational inputs and verified to yield identical results.

Stage 4: binary enumeration. The exact binary optimum is computed by complete enumeration. The smaller budget contains $|\mathcal{C}^{\mathrm{bin}}_{k_{\mathrm{lo}}}| = 22$, 37, and 176 feasible binary vectors for $d = 6, 8, 10$, respectively. The largest budget contains at most 638 feasible vectors. Every return is evaluated in exact rational arithmetic.

Stage 5: fractional search. A multistart SLSQP solver proposes a continuous point $\tilde{x} \in \mathcal{C}_k$. The solver runs 50 random starts per case with tolerance 10^{-12} and a maximum of 500 iterations, using SciPy 1.14. The best feasible point is projected onto $\mathcal{C}_k$ and then quantized to a rational with denominator at most 10^6 using nearest-fraction rounding. Budget feasibility ($\mathbf{1}^{\top} x \leq k$) is verified exactly after quantization. If violated, the smallest nonzero coordinate is reduced until feasibility holds.

Stage 6: rounding methods. Four rounding and selection methods are applied, each starting from the quantized fractional point or from scratch:

- *Algorithm 1* (pipage rounding): the constructive rounding procedure from Theorem 3.3, which produces a deployable $y \in \mathcal{C}^{\mathrm{bin}}_k$ from the fractional point.
- *Top-k selection*: rank the d coordinates by their marginal return $J(\varepsilon_i) - J(\mathbf{0})$ and select the k highest.
- *Greedy forward selection*: start with $y = \mathbf{0}$ and iteratively add the coordinate giving the largest return increase, breaking ties by index. This is a sequential greedy that accounts for interaction effects.
- *Uniform random selection*: draw a uniformly random k-subset of coordinates and repeat 100 times, reporting the best return found.

All methods use exact rational return evaluations. Every method has an explicit return-query limit, and all query counts are recorded.

Stage 7: verification. The witnessed advantage $\Delta_{\mathrm{wit}} = [J(x_{\mathrm{cand}}) - J^{\star}_{\mathrm{bin}}]_+$ is computed using the main-text definition, where x_{cand} is the quantized fractional candidate and $J^{\star}_{\mathrm{bin}}$ is the enumerated binary optimum. The certificate inequality $\Delta_{\mathrm{wit}} \le \mathcal{B}_{\mathrm{Bell}}$ is verified exactly. Algorithm 1 achieves the exact binary optimum in every case.

D.4 Calibration records

Table 2 reports one case per domain at common factor levels. The values of $\widehat{\tau}$, $\widehat{\beta}$, $\mathcal{B}_{\mathrm{Bell}}$, and Δ_{wit} are exact rationals reported to four significant digits.

Table 2: One case per domain at $d = S = 8$, $k = 2$, $\gamma = 1/2$, mixed edits, repair scale $3/5$, and a strong baseline. Each case has $|\mathcal{C}^{\mathrm{bin}}_k| = 37$.

Domain	$\widehat{\tau}$	$\widehat{\beta}$	$\mathcal{B}_{\mathrm{Bell}}$	Δ_{wit}	wall
GRIDWORLD	2.159×10^{-1}	2.819×10^{-2}	2.405×10^{-1}	0	1.64 s
QUEUE	4.547×10^{-4}	2.409×10^{-2}	2.153×10^{-2}	3.261×10^{-6}	2.02 s
MAINTENANCE	1.361×10^{-1}	1.567×10^{-2}	1.498×10^{-1}	0	2.45 s
GARNET	1.305×10^{-1}	6.263×10^{-3}	1.360×10^{-1}	0	3.11 s

D.5 Environment

All experiments ran on a single workstation with an Intel Core i9 processor and 64 GiB RAM, using Python 3.12 and fractions.Fraction for exact rational arithmetic. SciPy 1.14 provided the SLSQP solver. Memory was measured by the peak resident set size of the Python process. No GPU was used.

D.6 Domain generators

Each domain generator takes a seed, state count $S = d$, and an edit type (transition-only, reward-only, or mixed) and outputs a rational MDP satisfying Assumption 2.1:

- GRIDWORLD: cyclic nearest-neighbour transitions on a one-dimensional ring. Repair rows redirect to a distant target.
- QUEUE: birth–death chain with load-dependent service. Repair rows increase the service rate.
- MAINTENANCE: deterioration chain with repair actions that reset to a better state.
- GARNET: random sparse transition matrices drawn from a Dirichlet distribution (Archibald et al., 1995). Repair rows replace one action with an independently drawn row.

The companion code artifact contains the full generator source.

D.7 Runtime breakdown

Across all 144 cases, the Bellman solves take 23.5 s, the interval calculation takes 15.8 s, and binary enumeration takes 39.1 s. Model construction takes less than 0.2 s in total. The complete sweep, including the fractional search and four rounding methods, takes 519 s. Certification costs more than enumeration at $d = 6$ and $d = 8$. Enumeration is the larger term at $d = 10$, reflecting its combinatorial growth $\binom{d}{k}$.

D.8 Exact-zero affine construction

The exact-zero examples use a separate affine construction designed so that $J(x)$ has no curvature and the certificate is provably zero.

Construction. Row i sends mass $\varepsilon = 1/10$ to every state and sends the remaining mass $1 - S\varepsilon$ to itself. The two actions have identical transition rows: $P^0_{i\cdot} = P^1_{i\cdot}$ for every editable row i. The

proposed action increases only the reward: $r_i^1 > r_i^0$. Since $P^0 = P^1$, the policy-induced transition matrix $P(x) = (1-x_i)P_{i\cdot}^0 + x_i P_{i\cdot}^1 = P_{i\cdot}^0$ is independent of x. Therefore the resolvent $(I - \gamma P(x))^{-1}$ is constant and $J(x)$ is affine in x.

Consequence for the certificate. An affine function has zero second derivatives everywhere:

$$\nabla^2 J(x) = 0 \quad \text{for all } x \in [0,1]^d.$$

Therefore $\beta = 0$. Moreover, each directional derivative along a coordinate-increasing direction is constant and nonnegative (the proposed action increases reward), so $\tau = 0$. The interval propagation inherits this: $\widehat{\tau} = \widehat{\beta} = 0$, hence $\mathcal{B}_{\mathrm{Bell}} = 0$.

Verification. All twelve combinations of $d \in \{4, 6, 8, 10\}$ and $\gamma \in \{\frac{1}{2}, \frac{4}{5}, \frac{9}{10}\}$ give $\widehat{\tau} = \widehat{\beta} = \mathcal{B}_{\mathrm{Bell}} = 0$. Enumeration confirms that the continuous and binary optima have the same exact return: the affine maximum on $\mathcal{C}_k$ is attained at a vertex, so $\mathrm{Gap}(\mathcal{C}_k) = 0$. This validates that the interval propagation correctly identifies zero-curvature instances.

D.9 Mechanistic active-family study

We use the family in Proposition 4.1. The dimensions are $m \in \{4, 8, 16, 32, 64\}$; the discount factors are $\gamma \in \{0.2, 0.5, 0.8, 0.95\}$. These give 20 deterministic cases. We retain the dummy coordinate, so $d = m + 1$ and $k = m/2$. The initial candidate has $x_i = 1/2$ on the m bad rows and zero at the dummy. No optimizer or random seed is used. Every trajectory has exactly $m/2$ exchanges and reaches a binary optimum.

The new verification script evaluates every chosen endpoint in exact rational arithmetic, checks the resulting loss against the closed-form gap, and independently integrates the rational second derivative against the interpolation kernel. It also computes reliable upper sums on 16, 64, and 256 uniform subintervals; all sums dominate the exact integral and decrease under refinement. All 20 cases and 248 executed exchanges pass. The script additionally checks 1,325 dimension-budget combinations and 27 polynomial instances of the kernel identity.

Table 3: Recomputed mechanistic ratios to $\mathrm{Gap}(\mathcal{C}_k)$, shown at $m = 4$. Only the B_{unif} ratio depends on m; B_{256} uses reliable 256-subinterval upper sums.

γ	$\mathcal{B}_{\mathrm{Bell}}$	B_{unif}	B_{cand}	B_{seg}	B_{int}	B_{256}
0.20	2.8125	1.6875	1.4063	1.0631	1	1.0002
0.50	6	3.6	3	1.6875	1	1.0019
0.80	30	18	15	7.56	1	1.0099
0.95	420	252	210	105.0131	1	1.0439

For every tested m, $B_{\mathrm{unif}}/B_{\mathrm{cand}} = (m+2)/(m+1)$, whereas $\mathcal{B}_{\mathrm{Bell}}/B_{\mathrm{cand}} = 2$. Thus the new uniform ratio has a mild dimension dependence, while the other ratios in Table 3 are unchanged across the tested dimensions. The weighted calculation is exact on this family even at long horizons.

D.10 Structured active suite

The original 144-case calibration grid is intentionally broad but produces mostly vertex candidates. To test nontrivial rounding under structured dynamics, we freeze a separate active-geometry grid before evaluation. It uses four domains (GRIDWORLD, QUEUE, MAINTENANCE, and GARNET), $d \in \{6, 8, 10\}$, $\gamma \in \{1/2, 4/5\}$, two budgets, and three fixed holdout seeds, giving $4 \times 3 \times 2 \times 2 \times 3 = 144$ cases.

Starting from each domain's baseline dynamics and rewards, each editable baseline row receives a seeded persistence coefficient $\lambda_i \in \{0.55, 0.60, 0.65\}$. The repair endpoint redirects a seeded fraction $s_i \in \{0.45, 0.50, 0.55\}$ toward a fixed favorable state chosen by the domain semantics. The initial distribution is uniform. Persistence creates repeated visits and diminishing returns, while the common favorable region induces competition among repairs. No case is selected or discarded based on its outcome.

Table 4: Deployment baselines on the 139 active structured cases. Gap is the mean normalized binary-optimality gap and Queries is the median number of return evaluations.

Method	Binary optimum	Gap	Queries
Algorithm 1	114/139 (82.0%)	0.185%	6
Top-k	61/139 (43.9%)	3.066%	9
Greedy	125/139 (89.9%)	0.178%	26
Random-100	98/139 (70.5%)	0.727%	100

The continuous search is completed before the exact binary optimum is computed, unlike the original runner, the binary optimum is never used as a continuous-optimizer start. A raw coordinate is classified as fractional when it lies in $(10^{-6}, 1-10^{-6})$. The candidate is then quantized to a feasible rational point. The maximum coordinate change from quantization is 1.27×10^{-9}.

Of the 144 cases, 139 are active after continuous search. All 139 have strictly positive witnessed advantage over the exact binary optimum. The median active candidate has four fractional coordinates, fractional mass $\sum_i \min\{x_i, 1-x_i\} = 1.00$, and three pairwise exchanges. The maximum exchange count is nine. Domain-level active counts are 35/36 for GRIDWORLD, 36/36 for QUEUE, 36/36 for MAINTENANCE, and 32/36 for GARNET. By editable dimension, the active counts for $d = 6, 8, 10$ are 44/48, 47/48, and 48/48, with median fractional-coordinate counts 3, 4, 5 and median exchange counts 2, 3, 4, respectively.

All exact certificate and rounding checks pass. Algorithm 1 reaches the exact binary optimum in 119/144 cases, or 114/139 active cases. The 25 active non-optimal rounded policies have median binary-optimality gap approximately 0.017 and maximum approximately 0.064.

Deployment baselines. We compare Algorithm 1 with the three simple deployment rules defined in Section D.3. For each method we report the fraction of active cases that attain the exact binary optimum, the mean normalized binary-optimality gap

$$\frac{J^\star_{\mathrm{bin}} - J(y)}{J^\star_{\mathrm{bin}} - J(0)},$$

and the median number of exact return evaluations. Random selection reports the best of 100 uniformly sampled feasible k-subsets. Greedy attains the binary optimum most often, but at substantially higher query cost. Algorithm 1 has nearly the same mean normalized optimality gap while using fewer than one quarter as many median return evaluations. Relative to the cheaper top-k heuristic, Algorithm 1 more than halves the number of non-optimal deployments and reduces the mean normalized gap by over an order of magnitude. The table measures deployment quality only. Unlike the three heuristics, Algorithm 1 is coupled to the pre-rounding certificate of Theorem 3.3 and Proposition 3.4.

D.11 TIGHTNESS AND DEPLOYMENT DIAGNOSTICS

For each active structured case, let

$$L_{\mathrm{rnd}} := [J(x_{\mathrm{cand}}) - J(y)]_+$$

be the realized rounding loss. In addition to the model-only $\mathcal{B}_{\mathrm{Bell}}$ and the candidate-specific pre-rounding certificate B_{cand}, we record the post-trace pair-specific audit bound

$$B_{\mathrm{trace}} := \sum_i \Delta_i \widehat{\tau}_i + \frac{1}{2} \sum_t \widehat{\beta}_{i_t j_t}(-a_t) b_t.$$

The proof of Proposition 3.4 gives

$$L_{\mathrm{rnd}} \le B_{\mathrm{trace}} \le B_{\mathrm{cand}} \le \mathcal{B}_{\mathrm{Bell}}.$$

All 139 active cases satisfy this recorded hierarchy exactly. The revised theorem additionally places B_{unif} between B_{cand} and $\mathcal{B}_{\mathrm{Bell}}$, but its per-instance values and coverage are not reported here.

The median ratios to realized loss are

$$\operatorname{med} \frac{\mathcal{B}_{\mathrm{Bell}}}{L_{\mathrm{rnd}}} = 70.8, \qquad \operatorname{med} \frac{B_{\mathrm{cand}}}{L_{\mathrm{rnd}}} = 12.1, \qquad \operatorname{med} \frac{B_{\mathrm{trace}}}{L_{\mathrm{rnd}}} = 9.20.$$

The median shrinkage $\mathcal{B}_{\mathrm{Bell}}/B_{\mathrm{cand}}$ is 6.33. Candidate geometry reduces the exchange-count and support slack, while the remaining gap is consistent with the global-versus-local curvature mismatch of Section 4.

For deployment, define $M_B(x) = J(x) - B - J(0)$. A positive margin certifies that the rounded deployable policy improves over the baseline. Table 5 includes the post-trace audit in addition to the two pre-rounding certificates.

Table 5: Deployment-improvement coverage on the 139 active structured cases. Direct is inferred from the existing exact-return records.

Discount	Active	$\mathcal{B}_{\mathrm{Bell}}$	B_{cand}	B_{trace}	Direct
$\gamma = 1/2$	68	67	68	68	68
$\gamma = 4/5$	71	0	35	54	71
All	139	67	103	122	139

Coverage is 48.2% for the legacy bound, 74.1% for the candidate bound, and 87.8% for the post-trace audit. At $\gamma = 4/5$, the corresponding counts are 0/71, 35/71, and 54/71. All 139 rounded policies actually improve over the baseline. A nonpositive certified margin is therefore an abstention event in this suite.

D.12 Direct verification baseline

Protocol. The new experiment reruns the 20 closed-form cases in Appendix D.9. Both methods receive the same candidate, baseline return, and candidate return, and use the same deterministic pair and endpoint selection. The pre-rounding tests evaluate $J(x) - B - J(0) > 0$ for $B \in \{\mathcal{B}_{\mathrm{Bell}}, B_{\mathrm{unif}}, B_{\mathrm{cand}}\}$ before any exchange. All candidates are then rounded, including those not certified, so both tests refer to the same output. The direct test uses its cached exact return to check $J(y) > J(0)$. All arithmetic and comparisons use rational numbers.

Coverage and accounting. At $\gamma = 0.2$, all three pre-rounding bounds certify all five dimensions. At $\gamma = 0.5$, only the budget-aware and candidate bounds certify them. At $\gamma = 0.8$ and 0.95, none of these three bounds certifies an instance. Direct verification certifies all 20 outputs. There are 40 shared setup evaluations for candidate and baseline returns, followed by 496 endpoint evaluations across 248 exchanges. Direct verification uses the retained selected-endpoint values and adds no return evaluations. Certificate values are computed from the family's closed forms. In a general run with no exchange and no cached output return, a final policy evaluation may be needed.

Dominance under exact evaluation. The rounding guarantee gives $J(y) \geq J(x) - B$. Subtracting $J(0)$ shows that the direct margin dominates the certified margin. Therefore, a positive certified margin implies a positive direct margin. This observation applies to any sound candidate-level loss bound for the same output, including a post-trace bound. The pre-rounding guarantee is available before y is constructed. Under validated interval arithmetic, confirming a positive direct margin may require refining return enclosures. Zero additional evaluation cost is established here for the exact-rational runs with cached endpoint returns.

Structured-suite inference. The original records report $J(y) > J(0)$ under exact arithmetic in all 139 active cases. These records imply direct-verification coverage of 139/139, including 68/68 at $\gamma = 1/2$ and 71/71 at $\gamma = 4/5$. The table's Direct column reinterprets those recorded returns as a verification baseline using the original experiment records. Structured-suite coverage is reported for the legacy, candidate, and trace bounds.

D.13 Scaling and validated arithmetic

The scaling experiment uses a sparse ring MDP with S states and only the first d rows editable. Transition probabilities are dyadic, so the binary64 model input is represented exactly. Axis A uses $S = d \in \{16, 32, 64, 128\}$. Axis B fixes $d = 16$ and uses $S \in \{64, 128, 256\}$, directly testing the $d \ll S$ regime. Each configuration uses three fixed seeds. Large cases omit binary enumeration.

Validated floating point first solves the Bellman systems in binary64, then applies outward residual corrections until every Bellman inequality is satisfied. All subsequent interval operations are also rounded outward. Memory is measured in an independent process for each run to avoid inherited high-water marks. Table 6 reports median runtime and peak resident-set size.

Table 6: Validated-certification scaling. Times are medians over three fixed seeds. RSS is measured in independent processes.

Axis	S	d	Total (s)	Interval (s)	RSS (MiB)
A	16	16	0.059	0.007	118.3
A	32	32	0.223	0.030	118.7
A	64	64	0.937	0.122	120.0
A	128	128	4.448	0.553	125.7
B	64	16	0.215	0.012	118.5
B	128	16	0.487	0.019	119.0
B	256	16	1.311	0.031	120.3

For numerical auditing we use six small cases with $(S, d) \in \{(8, 8), (16, 16)\}$. Each binary64 validated envelope is reinterpreted as an exact dyadic rational, and every Bellman inequality is then rechecked with exact arithmetic. All 6/6 validated cases pass this audit. Recomputing the interval primitives from the same endpoints using exact Fraction arithmetic also confirms outward containment of the validated certificate in all 6/6 cases. In contrast, the corresponding naive midpoint-FP64 solutions pass the exact Bellman audit in 0/6 cases even though their printed certificate values are numerically close to the validated values. At $S = d = 16$, the validated implementation is approximately $14.8\times$ faster than the existing exact-rational envelope-and-certificate pipeline.

D.14 General exchange study and complete costs

New instances. This suite is separate from the original structured suite. We evaluate the full Cartesian product of two transition structures, $d \in \{6, 8\}$, $\gamma \in \{1/2, 4/5, 19/20\}$, and seeds 11 and 29, giving 24 instances. Each has $S = d + 1$ states and budget $k = d/2$. The last state is an absorbing target with reward one. A baseline bad-state row puts a seeded fraction between $72/100$ and $90/100$ on itself and the remainder on one other bad state. That successor follows a ring or is randomly selected. The proposed row redirects a seeded fraction between $40/100$ and $65/100$ of its baseline transition mass to the target. Baseline bad-state rewards belong to $\{0, 1/100, 2/100, 3/100\}$, and proposed reward changes belong to $\{0, -1/100, -2/100, -3/100\}$. The initial distribution is uniform over bad states. The off-diagonal transitions couple the editable rows. All configurations are retained. This suite is synthetic and designed to exercise fractional rounding.

Candidates and exact references. Continuous search uses SLSQP with analytic gradients, three deterministic starts, tolerance 10^{-10}, and at most 150 iterations per start. No binary optimum is supplied to the search. Coordinates are rounded to denominator 10^6 and any excess budget is removed in index order. Every resulting candidate has strictly positive return advantage over the exactly enumerated binary optimum. All subsequent returns, Bellman envelopes, endpoint choices, low-rank updates, and integral enclosures use rational arithmetic. Every one of the 135 exchanges is checked against direct policy evaluation at both endpoints and an interior point. All rounded outputs and returns agree between the full-solve and rank-two implementations.

Integral refinement and independent audit. The refinement target is $U - L \le 0.05U + 10^{-12}$ with a cap of 64 cells per exchange. The target holds on every executed exchange, with 2,003 cells in total. This tolerance concerns the curvature integral, not its ratio to actual rounding loss. The standalone method uses B_{loc} and no global envelopes. The hybrid method also computes global Bellman bounds and clips each integral charge as defined in Appendix B.7. The two bounds coincide on this suite. All instances satisfy $[J(x) - J(y)]_+ \le B_{\mathrm{hyb}} \le B_{\mathrm{trace}} \le B_{\mathrm{cand}} \le B_{\mathrm{unif}}$. An independent audit uses 16 exchanges in fully dense five-state MDPs, with strictly positive rational transition endpoints and signed rewards. At five points per exchange, direct resolvent calculations agree exactly with the rational representation for both return and second derivative, giving 80 exact checks. No absorbing-target structure is assumed in this audit.

Table 7: New coupled suite. Ratios are medians of bound divided by actual rounding loss. Coverage counts certified baseline improvements. Each discount has eight instances. Local and hybrid bounds coincide here and use the trajectory.

γ	$B_{\mathrm{cand}}/L_{\mathrm{rnd}}$	$B_{\mathrm{trace}}/L_{\mathrm{rnd}}$	$B_{\mathrm{loc}}/L_{\mathrm{rnd}}$	Candidate	Local
$1/2$	5.05	3.82	1.81	8/8	8/8
$4/5$	24.68	13.56	1.96	1/8	8/8
$19/20$	402.31	170.93	2.08	0/8	8/8

Cost accounting and baselines. All methods use the same exact return arithmetic. Direct rounding solves a policy-evaluation system for each endpoint. Its rank-two baseline uses the same rational exchange coefficients and resolvent updates as local integration, but omits derivative enclosures. Candidate top-k selects the largest candidate coordinates. Greedy starts from the baseline and adds the best positive-return edit until the budget is reached or no improvement remains. The global-certificate pipeline adds all $2d + 2$ Bellman envelope problems before rank-two rounding. Local integration needs no global preprocessing. Hybrid integration includes both global preprocessing and local enclosures. Candidate-based totals include continuous optimization, exact candidate and baseline evaluations, and all subsequent computation. Greedy is timed from the model and does not pay for continuous optimization. Final verification reuses evaluated returns. Exact enumeration and independent audit evaluations are reference checks and are excluded from method runtimes. Each small instance is run three times with numerical linear-algebra libraries restricted to one thread. We first take the median over repetitions for each instance, then the median across instances.

Table 8: Complete costs on the new 24-instance suite. The normalized gap is $(J^{\star}_{\mathrm{bin}} - J(y))/(J^{\star}_{\mathrm{bin}} - J(0))$, averaged across instances.

Method	Total (ms)	Binary optimum	Mean gap (%)
Direct rounding	7.60	20/24	0.088
Rank-two direct	9.43	20/24	0.088
Global certificate + rank-two	19.14	20/24	0.088
Local integration	36.19	20/24	0.088
Hybrid integration	48.81	20/24	0.088
Candidate top-k	4.33	21/24	0.169
Greedy	4.89	21/24	0.130

The table quantifies the computation cost of tightening trajectory bounds. All four rounding variants attain the same policy quality on this suite.

Exact scaling checks. We also test the ring construction at $(S, d) = (16, 8), (32, 8), (64, 16)$, with seed 11 and $\gamma = 19/20$. Additional noneditable states retain their baseline rows. All candidates are fractional and all exact rounding and bound checks pass. Binary optima are not enumerated. These timing checks use one run per configuration and are reported separately from the repeated small-suite comparison. Table 9 includes initial inversion and all per-exchange work in each rounding time, but excludes continuous search. Saved results also report full end-to-end times.

Table 9: Exact scaling checks. Ratios use realized rounding loss. Rounding times are in seconds and are single-run measurements.

S	d	$B_{\mathrm{cand}}/L_{\mathrm{rnd}}$	$B_{\mathrm{loc}}/L_{\mathrm{rnd}}$	Direct	Rank-two	Local
16	8	123.30	1.60	0.022	0.034	0.681
32	8	121.92	2.35	0.099	0.133	0.473
64	16	291.73	1.55	1.562	1.898	4.109

The rational implementation provides exact numerical certification. Its runtime includes arithmetic on the accumulated numerators and denominators.