AUDITING RECTANGULARIZATION IN ROBUST MDPS: SELECTIVE RECOUPLING OF SHARED TRANSITION UNCERTAINTY

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust MDPs are often made tractable by assuming transition uncertainty is independent across time, states, and actions. Under this *rectangular* assumption, an adversary chooses the worst transition at each time–state–action triple, enabling Bellman recursion. This can be overly pessimistic when transitions share an episode-persistent factor, because locally adverse choices may not be jointly realizable. We study the intervening design problem: which declared transition dependencies yield the largest certified reduction when at most k modeling blocks are recoupled? For a fixed policy and transitions affine in shared scalar factors, we derive an exact decomposition of the rectangularization gap into nonnegative defects at reachable rows. These defects yield computable certificates and a Global-Balance rule for jointly selecting dependencies, including complementarities between groups. We also provide a mixed-integer formulation, approximation and selection guarantees, and finite-data certificates valid after data-dependent selection. In generated routing and queue systems with binary persistent regimes, Global Balance captures cross-block certificate gains on 47.3% of policy–subset pairs. Restoring eight of ten blocks reduces held-out candidate-menu regret relative to full release, while an incorrect sharing map reverses the benefit of full recoupling. With six independent factors, certificate-selected blocks recover 89.5% of the full recoupling gain using 4 of the 64 conditional evaluations required by full recoupling.

1 INTRODUCTION

Robust control protects a policy against an *ambiguity set*, a collection of plausible transition models. We focus on uncertainty in transition probabilities, while rewards and the evaluated policy are fixed. For a time step h, state s, and action a, we call $P_h(\cdot \mid s, a)$ a *transition row*. A complete model P contains one row for every time–state–action triple, and $\mathcal{C}$ denotes the set of plausible models.

The central modeling choice is how the adversary—conventionally called *nature*—may choose from $\mathcal{C}$. Under the *committed* model, nature selects one $P \in \mathcal{C}$ before the episode and uses all rows from that model. Under the state–action rectangular assumption standard in robust MDPs (Iyengar, 2005; Nilim & El Ghaoui, 2005), each row may instead be chosen independently from its possible values across $\mathcal{C}$. This makes the worst-case problem separable and enables Bellman dynamic programming, but removes dependencies requiring rows governed by the same physical parameter to remain compatible.

This matters when uncertainty is driven by an episode-persistent factor, such as weather or demand affecting transitions at many states. Rectangular nature can use different regimes across rows, even when no physical environment can switch this way, thereby *pasting* locally plausible rows into an infeasible model. The resulting robust value is tractable but may be overly pessimistic. Preserving dependencies avoids this artifact but generally destroys the row-wise decomposition underlying robust Bellman recursion (Iyengar, 2005; Nilim & El Ghaoui, 2005; Wiesemann et al., 2013).

Prior work develops dynamic programming for rectangular robust MDPs, studies coupled uncertainty, and releases repeated parameters to obtain verification bounds (Schnitzer et al., 2026; Quatmann et al., 2016; Heck et al., 2025). We instead ask: *Which dependency releases most increase a fixed policy's worst-case loss, and which should be reversed under a block budget?*

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.

This is a fixed-policy *evaluation and ambiguity-design* problem. Given probe policies, affine endpoint transitions, a domain-supplied factor map, and a block budget, we return a partially recoupled ambiguity set with a certified reduction in pessimism. The budget counts predeclared modeling blocks whose occurrences are tied to a shared parameter. A downstream solver may optimize over the selected set, or the audit may compare a finite policy menu; Appendix D reviews both uses.

Three features make this challenging. Coupling prevents row-wise Bellman evaluation, row importance depends on model-dependent reachability, and groups can be complementary when occurrences prefer opposite factor endpoints. Independent rankings can miss groups that matter only when restored together. With estimated transitions, certification must remain valid after data-dependent selection. We address these issues for a layered process, fixed Markov probes, additive loss, and a predeclared map of affine two-endpoint factors. Operationally, we start from the released model, identify decision-relevant pessimism, and jointly restore the most consequential dependencies.

The paper contributes three main ideas:

- An **exact audit identity** that decomposes the gap between the rectangular worst-case value and the coherent shared-factor worst-case value into nonnegative defects at reachable transition rows.
- A **selective recoupling method** that determines which shared dependencies to restore under a limited budget, including complementarities that row-wise rankings can miss.
- **Optimization and statistical guarantees**: a mixed-integer selection formulation, an approximation guarantee under endpoint overlap, and finite-data certificates valid after data-dependent selection.

Evaluating the transition-kernel performance-difference identity (Kearns & Singh, 2002; Modhe et al., 2021) at the released Bellman solution yields nonnegative residuals that Global Balance pools for selection; the overlap theorem relates the certificate to the best same-budget gain. Numerical audits show correct recoupling can reduce held-out regret, whereas incorrect sharing can increase it.

2 Related Work

Rectangular and coupled uncertainty. Rectangular multiple-prior models support dynamic consistency (Epstein & Schneider, 2003), and state–action rectangularity enables Bellman recursions in robust MDPs (Iyengar, 2005; Nilim & El Ghaoui, 2005; Shapiro, 2016; Li & Shapiro, 2025). Nonrectangular ambiguity preserves cross-row dependence but generally loses this decomposition; structured factor sets can recover tractability, while coupled budgets show that discarding dependence can be conservative (Goyal & Grand-Clément, 2023; Mannor et al., 2016). Related work studies the price of lost correlations (Agrawal et al., 2012), shared disturbances and latent uncertainty (Zouitine et al., 2024; Li et al., 2026; Wang et al., 2025; Schnitzer et al., 2026), and known parameter sharing for data-efficient policy improvement (Engelen et al., 2025). These methods learn or exploit coupled models; we instead quantify the value distortion created when a declared coupling is released. In particular, the value-aware support gap of Li et al. (2026) is anchor-dependent, whereas our released reference yields an exact reachable-defect identity for a fixed policy.

Schnitzer et al. (2026) learn a shared-parameter confidence set and compare sound rectangular relaxations; Heck et al. (2025) tighten parameter-lifting bounds by grouping and shortcutting transitions. We keep the local row sets fixed and select which declared parameter equalities to restore under a block budget. Gadot et al. (2023) instead use sensitivity to uncertainty radii to guide additional sampling. Our selection targets the exact release gap in Equation (8) through the selection rule in Equation (17).

Parameter release and dependency selection. Our selector uses two-scenario max–min optimization (Yu, 1996; Kawase & Sumita, 2019), with weights derived from reachable transition defects. The contribution is the budgeted choice of shared-factor equalities, its overlap guarantee, and certification after data-dependent selection.

3 Preliminaries

Finite-horizon MDP and evaluation target. We consider a finite layered controlled process with decision stages $h = 0, \dots, H-1$ and terminal stage H. Let $\mathcal{S}_h$ and $\mathcal{A}_h$ denote the finite state and

action sets at stage h, respectively. Define $S_{\max} = \max_h |\mathcal{S}_h|$, $A_{\max} = \max_{h<H} |\mathcal{A}_h|$, and the set of time–state–action triples $\mathcal{X} = \{(h, s, a) : 0 \le h < H,\ s \in \mathcal{S}_h,\ a \in \mathcal{A}_h\}$. For $x = (h, s, a) \in \mathcal{X}$, the corresponding *transition row* is the probability vector $P_x = P_h(\cdot \mid s, a) \in \Delta(\mathcal{S}_{h+1})$, where $\Delta(E)$ denotes the probability simplex over a finite set E. A complete transition model is the collection of all such rows, written as $P = (P_x)_{x \in \mathcal{X}}$.

A fixed *probe* is a policy-evaluation experiment $e = (d_0, \pi)$, where $d_0 \in \Delta(\mathcal{S}_0)$ is the initial-state distribution and $\pi = (\pi_h)_{h=0}^{H-1}$ is a randomized Markov policy with $\pi_h(\cdot \mid s) \in \Delta(\mathcal{A}_h)$. We use the term *probe* to emphasize that the policy is fixed while the transition model is audited. In particular, the theoretical comparison in this paper does not optimize over policies.

The complete trajectory space is $\mathcal{T} = \{(s_0, a_0, \ldots, a_{H-1}, s_H) : s_h \in \mathcal{S}_h, a_h \in \mathcal{A}_h\}$. Trajectories that are impossible under a particular model simply receive probability zero. For $\tau \in \mathcal{T}$, the trajectory distribution induced by probe e and transition model P is

$$\mu_e(P)(\tau) = d_0(s_0) \prod_{h=0}^{H-1} \pi_h(a_h \mid s_h) P_h(s_{h+1} \mid s_h, a_h). \tag{1}$$

Given bounded loss $g : \mathcal{T} \to \mathbb{R}$, its expectation under P is $\langle g, \mu_e(P) \rangle := \sum_{\tau \in \mathcal{T}} g(\tau) \mu_e(P)(\tau)$, which is the fixed policy's robust evaluation target. We also define $[v]_+ = \max\{v, 0\}$, $\mathbf{1}\{E\}$ as the indicator of E, and $\mathrm{supp}(p) = \{s : p(s) > 0\}$.

Robust transition models and rectangularization. Let $\mathcal{C}$ be a nonempty compact family of complete transition models, with row dependencies encoding physical restrictions. Under the committed interpretation, nature chooses one $P \in \mathcal{C}$ before the rollout. Its worst-case value is

$$W_{\mathcal{C}}(e, g) := \sup_{P \in \mathcal{C}} \langle g, \mu_e(P) \rangle. \tag{2}$$

For each $x \in \mathcal{X}$, let $\mathcal{U}_x := \mathrm{proj}_x \mathcal{C} = \{P_x : P \in \mathcal{C}\}$ be the set of transition rows that occur at x in some plausible complete model. The standard rectangular relaxation is the Cartesian product

$$\mathcal{R} = \mathrm{rect}(\mathcal{C}) := \prod_{x \in \mathcal{X}} \mathrm{proj}_x \mathcal{C}. \tag{3}$$

A member of $\mathcal{R}$ may combine one row from one model in $\mathcal{C}$ with another row from a different model. Thus $\mathcal{C} \subseteq \mathcal{R}$. The worst-case value is (when e and g are clear, we omit them from the notation):

$$W_{\mathcal{R}}(e, g) := \sup_{Q \in \mathcal{R}} \langle g, \mu_e(Q) \rangle, \tag{4}$$

and $W_{\mathcal{R}} - W_{\mathcal{C}} \ge 0$ is the *rectangularization* or *release* gap audited in this paper.

We compare two information structures. Committed nature selects one $P \in \mathcal{C}$ before the episode, whereas row-adaptive nature selects from $\mathrm{proj}_x \mathcal{C}$ after observing row x; richer histories must be encoded in the state. Under this timing, randomization does not improve on a pure contingent array in $\mathcal{R}$. For additive losses from Section 4 onward, Equation (4) follows from rectangular Bellman recursion. Appendix E proves this normal form and explains why complete counterfactual rows, rather than committed path laws, are needed to identify the release gap.

4 Certifying the Effect of Dependency Release

This section develops the certificate used throughout the paper. We first derive an exact decomposition of the value increase caused by releasing cross-row dependencies. We then specialize it to shared affine transition factors and replace model-dependent occupancies by parameter-uniform lower bounds. Finally, we aggregate these row-level quantities into Global Balance, a computable lower certificate for any proposed set of restored dependencies.

4.1 Exact identity for dependency release

The inclusion $\mathcal{C} \subseteq \mathcal{R}$ gives row-wise nature at least as much freedom as committed nature, but does not identify the resulting value gap or the rows responsible for it. We therefore solve the released Bellman problem and measure, at each reachable row, the gain from replacing a coherent transition by its row-wise worst choice. Assume additive trajectory loss, $g(\tau) = r_H(s_H) + \sum_{h=0}^{H-1} r_h(s_h, a_h, s_{h+1})$,

where r_H is terminal loss and r_h the one-step loss. For $x = (h, s, a)$, define $r_x(s') := r_h(s, a, s')$. Under $\mathcal{R}$, let $\overline{V}_h(s)$ denote the worst-case loss-to-go and define

$$\overline{V}_H(s) = r_H(s), \qquad u_x(s') = r_x(s') + \overline{V}_{h+1}(s'), \tag{5}$$

$$\overline{V}_h(s) = \sum\nolimits_a \pi_h(a \mid s) \max_{q \in \mathcal{U}_{hsa}} q^\top u_{hsa}. \tag{6}$$

Thus $q^\top u_x$ is the expected one-step loss plus released continuation loss when row q is used at x. Let $\overline{W} := \sum_s d_0(s)\overline{V}_0(s) = W_{\mathcal{R}}(e, g)$. Now fix a coherent model $P \in \mathcal{C}$. Write $J(P) := \langle g, \mu_e(P)\rangle$ and define the row occupancy $d_x^P := \Pr_{d_0,\pi,P}(s_h = s, a_h = a)$, for $x = (h, s, a)$. For any row $p \in \mathcal{U}_x$, define the local defect $\delta_x(p) = \max_{q \in \mathcal{U}_x} q^\top u_x - p^\top u_x \geq 0$. It is the local advantage to row-wise nature at x, measured using the released continuation value.

Theorem 1 (Exact reachable-defect identity)**.** *For every nonempty compact coherent family $\mathcal{C}$,*

$$\overline{W} - J(P) = \sum\nolimits_x d_x^P \delta_x(P_x), \qquad \forall P \in \mathcal{C}. \tag{7}$$

$$\textit{Consequently, } W_{\mathcal{R}}(e, g) - W_{\mathcal{C}}(e, g) = \min_{P \in \mathcal{C}} \sum\nolimits_x d_x^P \delta_x(P_x). \tag{8}$$

The two models have the same worst-case value if and only if there exists a worst-case coherent model $P^\star \in \mathcal{C}$ that is locally row-optimal at every row it makes reachable:

$$d_x^{P^\star} > 0 \Longrightarrow \delta_x(P_x^\star) = 0. \tag{9}$$

The identity weights local defects by coherent reachability. The shared affine structure below converts this model-dependent expression into a certificate from the released solution and endpoint rows.

4.2 From the exact identity to a computable certificate

We specialize the coherent family to shared affine transition factors. The difficulty is that both the row choice and its occupancy depend on the same coupled parameter. The affine representation separates the local endpoint preference from a parameter-uniform lower bound on reachability.

Associate each row x with endpoint rows P_x^0 and P_x^1. If the factor affects x, its transition lies between these endpoints; otherwise set $P_x^1 = P_x^0$. Define $\Delta_x := P_x^1 - P_x^0$, $\mathcal{X}_+ := \{x \in \mathcal{X} : \Delta_x \neq 0\}$. Thus $\mathcal{X}_+$ contains rows with factor-varying transitions; rows outside $\mathcal{X}_+$ remain fixed.

A partition Π of $\mathcal{X}_+$ specifies which rows share a scalar parameter. Each block $D \in \Pi$ uses one common $\theta_D \in [0, 1]$. Writing $D_\Pi(x)$ for the block containing x, define

$$\mathcal{C}_\Pi = \left\{ P^{\theta,\Pi} : \theta \in [0,1]^\Pi, \quad P_x^{\theta,\Pi} = \begin{cases} P_x^0 + \theta_{D_\Pi(x)}\Delta_x, & x \in \mathcal{X}_+, \\ P_x^0, & x \notin \mathcal{X}_+. \end{cases} \right\}. \tag{10}$$

Let $W_\Pi := W_{\mathcal{C}_\Pi}(e, g)$ and let d_x^θ be the occupancy under $P^{\theta,\Pi}$. The finest partition $\Pi^\top$ places every uncertain row in a singleton block and coincides with full row release. If Π' refines Π, then

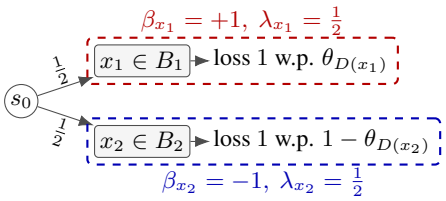
$$\mathcal{C}_\Pi \subseteq \mathcal{C}_{\Pi'}, \qquad W_\Pi \leq W_{\Pi'} \leq W_{\Pi^\top}, \qquad \mathrm{rect}(\mathcal{C}_\Pi) = \mathcal{C}_{\Pi^\top}. \tag{11}$$

For comparison, define the binary family $\mathcal{C}_\Pi^{\mathrm{bin}}$ by restricting each θ_D to $\{0, 1\}$, with worst-case value W_Π^{bin}. Under full row release, the continuous and binary families have the same value because each row-wise objective is affine in its interpolation parameter. The released continuation u_x assigns each uncertain row the signed slope $\beta_x := \Delta_x^\top u_x = (P_x^1 - P_x^0)^\top u_x$. If $\beta_x > 0$, row-wise nature prefers endpoint 1; if $\beta_x < 0$, it prefers endpoint 0; and if $\beta_x = 0$, the two endpoints are equivalent.

Corollary 2 (Affine partition certificate)**.** *Every dependency partition satisfies*

$$W_{\Pi^\top} - W_\Pi = \min_{\theta \in [0,1]^\Pi} \sum\nolimits_{x \in \mathcal{X}_+} d_x^\theta \left[\max\{0, \beta_x\} - \theta_{D_\Pi(x)}\beta_x\right]. \tag{12}$$

The bracketed term is the local defect. If two reachable rows in a block have opposite signs of β_x, row-wise nature prefers different endpoints, while the coherent model must use one common θ_D. The remaining difficulty is that d_x^θ varies with the same coherent parameter θ. A local conflict does not imply a global effect if some parameter choice can avoid that row. We therefore construct



row-wise release: $\theta_{x_1} = 1,\ \theta_{x_2} = 0 \Rightarrow W_{\Pi^\top} = 1$
one shared θ**:** $\max_\theta \left[\frac{1}{2}\theta + \frac{1}{2}(1-\theta)\right] = \frac{1}{2}$
certificates: $\mathcal{L}(\{1\}) = \mathcal{L}(\{2\}) = 0$
$\mathcal{L}(\{1,2\}) = \min\{\frac{1}{2}, \frac{1}{2}\} = \frac{1}{2} = W_{\Pi^\top} - W_{\Pi(\{1,2\})}$

Figure 1: The complementary pair: row-wise nature pastes the adverse endpoint on both rows (loss 1), one shared factor caps the loss at $1/2$, and only the pooled certificate sees the gain.

an occupancy lower bound that holds uniformly over all θ. For $x = (h, s, a)$, let Γ_x denote the state-action prefixes $\gamma = (s_0, a_0, s_1, a_1, \ldots, s_h = s, a_h = a)$ ending at x, and define

$$\lambda_x = \sum\nolimits_{\gamma \in \Gamma_x} d_0(s_0)\pi_h(a \mid s) \prod\nolimits_{t=0}^{h-1} \left[\pi_t(a_t \mid s_t) \min\nolimits_{e \in \{0,1\}} P^e_{t,s_t,a_t}(s_{t+1})\right]. \tag{13}$$

The minimum in Equation (13) is between the two endpoint probabilities of the specific next state.

Lemma 3 (Uniform common-prefix occupancy)**.** *For every row x, $d^\theta_x \geq \lambda_x$ for every coherent parameter vector θ. Consequently, if $m \geq 1$ pairwise row-disjoint pairs (x_i^+, x_i^-) lie in common partition blocks, with $\beta_{x_i^+} > 0 > \beta_{x_i^-}$ and positive prefix masses, then*

$$W_{\Pi^\top} - W_\Pi \geq \sum\nolimits_{i=1}^{m} \min\left\{\lambda_{x_i^+}\beta_{x_i^+}, \lambda_{x_i^-}(-\beta_{x_i^-})\right\} > 0. \tag{14}$$

The lemma replaces model-dependent occupancies by quantities uniform over the coherent factor. It shows opposite endpoint preferences matter only when both sides are guaranteed to be reachable.

4.3 Global Balance: A Lower Certificate for Recoupling Atomic Blocks

The pairwise condition does not aggregate conflicts across a candidate dependency structure. Global Balance does so by pooling signed evidence: two blocks can each have zero certificate alone yet become valuable when one prefers endpoint 1 and the other endpoint 0 under a shared factor.

Predeclared atomic recoupling blocks. Suppose domain knowledge indicates row occurrences share a scalar transition factor, but this dependence can be retained or released only in predeclared units. We call each unit an *atomic recoupling block* $B \subseteq \mathcal{X}_+$ and write $\mathcal{B} = \{B_1, \ldots, B_m\}$ for pairwise-disjoint blocks in one latent-factor pool. Blocks are fixed before inspecting λ_x or β_x and represent the smallest independently modeled units; singletons recover row-level selection.

For a selected subset $K \subseteq \mathcal{B}$, define $R_K := \bigcup_{B \in K} B$. All rows in R_K share an episode-persistent scalar, while unselected rows remain singleton blocks. Although R_K behaves as one larger recoupled group, the budget counts selected atomic blocks, so choosing q blocks consumes q units.

For an atomic block B, define its positive and negative signed masses $A_B = \sum_{x \in B} \lambda_x[\beta_x]_+$, $C_B = \sum_{x \in B} \lambda_x[-\beta_x]_+$. A_B collects reachable rows that prefer endpoint 1, while C_B collects those that prefer 0. Let $\Pi(K)$ be the partition that ties all rows in R_K to one common episode-persistent scalar and leaves all unselected uncertain rows as singletons. Define the Global-Balance certificate

$$\mathcal{L}(K) = \min\left\{\sum\nolimits_{B \in K} A_B, \sum\nolimits_{B \in K} C_B\right\}. \tag{15}$$

Equivalently, $\mathcal{L}(K) = \min\left\{\sum_{x \in R_K} \lambda_x[\beta_x]_+, \sum_{x \in R_K} \lambda_x[-\beta_x]_+\right\}$. The two totals are certified pressure toward endpoints 1 and 0. Once the selected rows share one scalar, nature cannot realize both simultaneously, so the smaller total is a guaranteed loss of adversarial flexibility. For example, in Figure 1, two branches reached with probability $1/2$, with terminal loss-one probabilities θ and $1 - \theta$. Their block masses are $(A_1, C_1) = (1/2, 0)$ and $(A_2, C_2) = (0, 1/2)$, so each singleton certificate is zero but $\mathcal{L}(\{1, 2\}) = 1/2$, exactly the reduction from released loss one to coherent loss $1/2$.

Theorem 4 (Factor-consistent Global-Balance certificate)**.** *For every $K \subseteq \mathcal{B}$,*

$$W_{\Pi^\top} - W_{\Pi(K)} \geq \mathcal{L}(K), \qquad W^{\text{bin}}_{\Pi^\top} - W^{\text{bin}}_{\Pi(K)} \geq \mathcal{L}(K). \tag{16}$$

Corollary 5 (Several declared factor pools)**.** *Let the candidate row blocks be pairwise disjoint and partitioned into pools $\mathcal{B}_f$, one per predeclared scalar. If $K_f \subseteq \mathcal{B}_f$ is recoupled through its pool-specific scalar while different pools remain independent, then the continuous and binary commitment gaps are both at least $\sum_f \mathcal{L}_f(K_f)$, with $\mathcal{L}_f$ given by Equation* (15) *on pool f.*

5 Dependency Selection and Guarantees

We now turn from *certification* to *design*: which blocks should be restored under a budget, and how closely does the certificate track the true recoupling gain? Section 5.1 formulates budgeted selection for one or several policies, while Section 5.2 quantifies the certificate's approximation quality.

5.1 Selecting atomic recoupling blocks for one or multiple evaluation policies

The atomic blocks in $\mathcal{B}$ are the selectable units declared before optimization. Selecting $K \subseteq \mathcal{B}$ recouples the row union R_K, and the budget $|K| \le k$ limits how many blocks can share the factor. For one evaluation policy, the natural Global-Balance selector is $\max_{K\subseteq\mathcal{B}:\,|K|\le k} \mathcal{L}(K)$.

Because $\mathcal{L}(K)$ jointly balances positive and negative mass, the best multi-block set cannot generally be found by ranking singleton certificates. Both λ_x and β_x depend on the evaluation policy, so selecting blocks for a single policy may over-specialize the restored structure. We thus consider one common subset for several policies. Let $j \in \mathcal{J}_+$ index the policies, and (A_{jB}, C_{jB}) be their signed masses. Because release gaps may differ across policies, normalize by a fixed $G_j > 0$. A natural choice is $G_j := W_{\Pi^\top,j} - W_{\Pi_{\mathrm{all}},j}$, where Π_{all} restores all declared atomic blocks. When $G_j > 0$, $\mathcal{L}_j(K)/G_j$ is the guaranteed fraction of policy j's full recoupling gain addressed by K. Zero-gap policies may be omitted from $\mathcal{J}_+$; other positive normalizers can encode different policy weights. The measured certificate-selection studies use $G_j = 1$.

Let $y_B \in \{0,1\}$ indicate if block B is selected, and z_j be the certified fraction for policy j.

Proposition 6 (Certificate-maximizing multi-policy selection)**.** *For a nonempty policy index set $\mathcal{J}_+$, policy-specific masses A_{jB}, C_{jB}, and fixed normalizers $G_j > 0$, the subset of at most k atomic blocks maximizing the mean normalized Global-Balance certificate is exactly the solution of*

$$\begin{aligned} \max_{y,z} \quad & \frac{1}{|\mathcal{J}_+|}\sum\nolimits_{j\in\mathcal{J}_+} z_j \\ \textit{s.t.} \quad & z_j \le \frac{\sum_B A_{jB} y_B}{G_j}, \qquad z_j \le \frac{\sum_B C_{jB} y_B}{G_j}, \qquad j \in \mathcal{J}_+, \\ & \sum\nolimits_B y_B \le k, \qquad y_B \in \{0,1\}, \quad z_j \ge 0. \end{aligned} \tag{17}$$

For fixed y, maximizing z_j gives $z_j = \mathcal{L}_j(K)/G_j$. Thus the mixed-integer program optimizes the certificate itself while allowing the same block to contribute differently across policies.

5.2 How informative is the Global-Balance certificate?

Global Balance replaces the model-dependent occupancy d_x^θ by the conservative common-prefix mass λ_x, ensuring validity for every coherent parameter value. The resulting certificate may nevertheless be smaller than the true recoupling gain because endpoint differences compound along a trajectory. Without sufficient endpoint overlap, no finite multiplicative comparison is possible.

We therefore impose a uniform overlap condition. Assume the two endpoint rows have *common support* and, for some $\chi \ge 1$, their nonzero probabilities satisfy

$$\max_{e\in\{0,1\}} P_x^e(s') \le \chi \min_{e\in\{0,1\}} P_x^e(s') \quad \text{for every } s' \text{ in the common support.} \tag{18}$$

Thus $\chi = 1$ means endpoints agree, while larger χ permits greater mismatch. Define continuous recoupling gain $G(K) := W_{\Pi^\top} - W_{\Pi(K)}$ and the endpoint-only gain $G^{\mathrm{bin}}(K) := W^{\mathrm{bin}}_{\Pi^\top} - W^{\mathrm{bin}}_{\Pi(K)}$.

Theorem 7 (Overlap sandwich and horizon dependence)**.** *Under Eq.* (18)*, every subset K satisfies*

$$\mathcal{L}(K) \le G(K) \le G^{\mathrm{bin}}(K) \le \chi^{H-1}\mathcal{L}(K). \tag{19}$$

Without common support, the assumptions imply no finite multiplicative factor. Moreover, for every $\chi > 1$ and $H \ge 2$, a one-action two-branch family satisfying Equation (18) *realizes*

$$\frac{G(K)}{\mathcal{L}(K)} = \Big(\frac{\chi+1}{2}\Big)^{H-1}, \qquad \frac{G^{\mathrm{bin}}(K)}{\mathcal{L}(K)} = \chi^{H-1}. \tag{20}$$

Hence the binary factor is sharp, and the continuous family admits no horizon-free guarantee under the row-wise overlap condition alone.

The theorem shows when Global Balance is quantitatively informative: under uniform overlap, its gap from the true recoupling benefit is controlled by the explicit horizon factor. It also implies a corresponding $\chi^{-(H-1)}$ approximation guarantee for certificate-based budgeted selection. Appendix B gives the sharpness constructions, the full selector guarantee, and their proofs.

6 Certification from Finite Transition Data

Sections 4–5 assume known endpoint rows. We now ask whether the same data can select K and certify that recoupling lowers worst-case loss. For each uncertain row x and endpoint $e \in \{0, 1\}$, assume two known successor states with unknown probability $p_{x,e}$. Samples from each row–endpoint pair form an identifiable *endpoint data stream*; at each predictable visit, the successor indicator is conditionally Bernoulli with mean $p_{x,e}$. One-step losses are known and bounded. Because stream counts and K depend on data, we require simultaneous coverage over streams and observed counts.

Anytime row confidence intervals. Index the N_{str} identifiable endpoint streams by r. After $n \geq 1$ observations from stream r, let $\widehat{p}_{r,n}$ be its empirical Bernoulli mean and define

$$\eta_{r,n} = \frac{6\delta}{\pi^2 N_{\mathrm{str}} n^2}, \qquad \rho_{r,n} = \sqrt{\frac{\log(2/\eta_{r,n})}{2n}}, \qquad I_{r,n} = \left[\widehat{p}_{r,n} - \rho_{r,n}, \widehat{p}_{r,n} + \rho_{r,n}\right] \cap [0, 1].$$

An unvisited stream receives $I_{r,0} = [0, 1]$. Since $\sum_{r=1}^{N_{\mathrm{str}}} \sum_{n \geq 1} \eta_{r,n} = \delta$, Hoeffding's inequality and a union bound give $\Pr(p_r \in I_{r,n} \text{ for every stream } r \text{ and count } n) \geq 1 - \delta$. Thus sample counts need not be fixed in advance. The intervals define a confidence set $\widehat{\mathcal{P}}$ for endpoint rows $\{P_x^0, P_x^1\}_{x \in \mathcal{X}_+}$.

Two recoupling certificates. Let $P^\star$ denote the true endpoint rows. For any candidate collection P, let $W_R(P)$ be the robust loss under full row release and $W_K(P)$ the robust loss after recoupling blocks K. Write $W_R := W_R(P^\star),$ and $W_K := W_K(P^\star)$. For the direct certificate, define only the extrema needed to lower-bound the gain, $\underline{W}_R := \inf_{P \in \widehat{\mathcal{P}}} W_R(P), \;\; \overline{W}_K := \sup_{P \in \widehat{\mathcal{P}}} W_K(P)$. Full release uses Bellman recursion; if exact recoupled evaluation is hard, any sound lower bound on W_R and upper bound on W_K may replace these extrema.

The same confidence set yields a simpler Global-Balance certificate. For $P \in \widehat{\mathcal{P}}$, let $\lambda_x(P)$ and $\beta_x(P)$ be the common-prefix mass and signed slope from Sections 4.2–4.3, and define

$$\underline{\lambda}_x = \inf_{P \in \widehat{\mathcal{P}}} \lambda_x(P), \qquad \underline{\beta}_x = \inf_{P \in \widehat{\mathcal{P}}} \beta_x(P), \qquad \overline{\beta}_x = \sup_{P \in \widehat{\mathcal{P}}} \beta_x(P).$$

For each atomic block B, set $\underline{A}_B = \sum_{x \in B} \underline{\lambda}_x [\underline{\beta}_x]_+$, $\underline{C}_B = \sum_{x \in B} \underline{\lambda}_x [-\overline{\beta}_x]_+$, and define $\underline{L}(K) = \min\left\{\sum_{B \in K} \underline{A}_B, \sum_{B \in K} \underline{C}_B\right\}$. A row contributes positive or negative pressure only when its entire slope interval has the corresponding sign. If $[\underline{\beta}_x, \overline{\beta}_x]$ crosses zero, its certified contribution is zero.

Proposition 8 (Finite-sample certification after data-dependent selection)**.** *Under the data model above, construct the anytime row intervals with total failure probability δ. Then, with probability at least $1 - \delta$, simultaneously for every evaluation policy and candidate recoupling subset K,*

$$W_R - W_K \geq \underline{W}_R - \overline{W}_K \qquad \textit{and} \qquad W_R - W_K \geq \underline{L}(K).$$

The direct value inequality applies to whichever family defines $W_R(P)$ and $W_K(P)$, while the Global-Balance inequality holds for both the continuous and binary affine families. Both guarantees remain valid when $K = \widehat{K}$ is selected using the same transition data.

The two bounds are complementary: one propagates the row-confidence set through robust values, while the other certifies only the signed mass needed by Global Balance.

7 Experiments

We test five aspects of the method: downstream decisions under correct versus incorrect sharing maps; certificate validity and complementarity; finite-data certification; transfer/scaling beyond the main generators; and the gain–cost trade-off with several factors.

Full protocols and numerical checks are in Appendix F. The main certificate study uses 24 routing/queue systems, six groups, nine fitting probes, six held-out policies, and all 64 subsets; endpoint

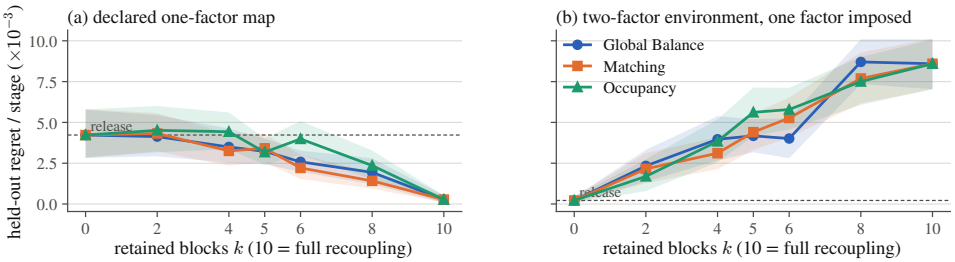


Figure 2: Held-out candidate-menu regret per stage. (a) Under the declared one-factor map, restoring blocks lowers regret and full recoupling ($k = 10$) removes most of it. (b) With two factors in the environment and one imposed, the same operation raises regret. Dashed line: release; bands: pointwise 95% stratified percentile-bootstrap intervals over 24 systems.

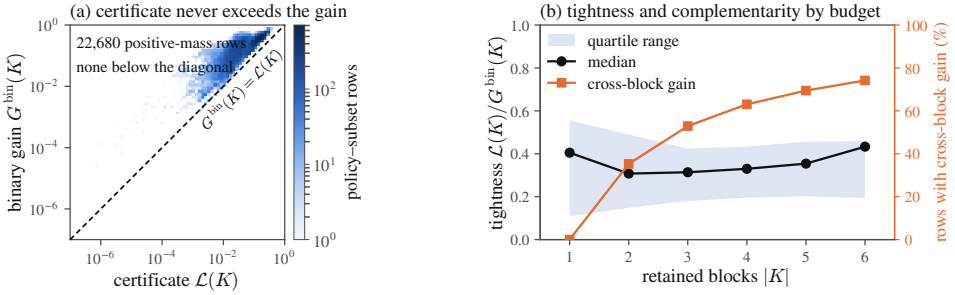


Figure 3: Certificate validity and tightness on the 23,040 policy–subset rows of the certificate study. (a) Joint density of $\mathcal{L}(K)$ and $G^{\text{bin}}(K)$ over the 22,680 positive-mass rows; no row lies below the diagonal. (b) Tightness ratio $\mathcal{L}(K)/G^{\text{bin}}(K)$ by retained-block count (median and shaded quartile range) and the share of rows on which $\mathcal{L}(K)$ strictly exceeds the sum of within-block certificates (orange, right axis): zero at $|K| = 1$, 74% at $|K| = 6$, 47.3% overall.

transitions are estimated at 8,192 and 131,072 episodes per endpoint, with a missing-coverage intervention. A separate decision study uses 24 ten-group systems, 64 probes, a frozen 256-policy menu, and sixteen held-out perturbations. The stochastic studies use $\mathcal{C}_{\Pi}^{\text{bin}}$; with F retained binary factor pools the evaluator enumerates 2^F conditional assignments. A continuous-factor check is within 3.36% of the binary gain on the certificate-study rows (Appendix F.1). Downstream quality is candidate-menu regret per stage; ranking quality is the AUC of the normalized residual release gap (lower is better). Intervals use 10,000 system- or dataset-level percentile-bootstrap draws.

Q1: Full recoupling lowers held-out candidate-menu regret under the declared one-factor map, and retaining eight of ten groups recovers part of the benefit. Holding local rows, menus, and budgets fixed but using two factors reverses the effect: imposing the wrong sharing map increases regret (Figure 2). Appendix F.3 reports matched comparisons for all selectors.

Q2: All 23,040 policy–subset rows satisfy $\mathcal{L}(K) \le G^{\text{bin}}(K) \le \chi^{H-1}\mathcal{L}(K)$ (maximum residual 5.998×10^{-16}), and enumeration agrees with the two-scenario formulation on all 1,344 selections. Global Balance exceeds the sum of within-block certificates on 47.3% of rows, with median $\mathcal{L}(K)/G^{\text{bin}}(K) = 0.325$. Across the horizon/overlap sweep the median ratio remains 0.867–0.994 with 24 uncertain rows, but falls as uncertain rows become dense (Appendix F.6). Figure 3 summarizes validity, tightness, and complementarity. A paired mixed-sign control isolates this complementarity: at $k = 2$, Global Balance improves held-out gain over Matching by 0.0128 $[0.0085, 0.0170]$ per stage and over Occupancy by 0.0088 $[0.0042, 0.0134]$, while every method has zero gain in the same-sign control. Appendix F.4 gives the protocol and mechanism plot.

Q3: At the 90% time-uniform joint confidence level, the direct test passes $0/144$ covered cases at 8,192 episodes per endpoint and $110/144$ at 131,072; after removing observations for one action it passes only $3/144$. The balance test passes all 144 cases in each condition because other observed rows retain enough opposing sign mass. At the larger covered budget the direct test certifies a median 53.3% of the true gain on passed cases. Simulator truth lies in all 69,120 propagated intervals with no false positive recorded; full finite-data results are in Appendix F.

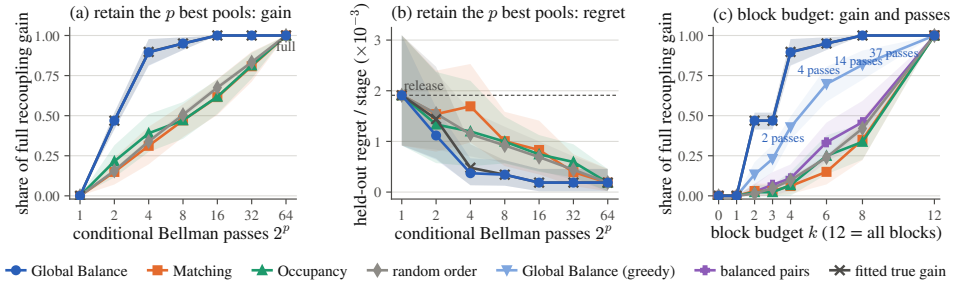


Figure 4: Selective recoupling with six independent factors on 24 systems. (a) Held-out share of the full recoupling gain versus conditional Bellman passes when retaining the p best-ranked pools; (b) candidate-menu regret per stage on the same axis; (c) gain share versus block budget k, with mean passes for the Global-Balance choice. Bands are 95% bootstrap intervals over systems.

Table 1: Central numerical findings for binary-regime systems and the selector-only scaling sweep. Intervals are 95% bootstrap intervals; method contrasts are paired within system. Lower regret and AUC are better; GB denotes global balance.

Finding / estimand	Estimate [interval]
Correct map: full recoupling − release regret/stage	-0.00396 $[-0.00541, -0.00266]$
Correct map: release − GB regret/stage at $k = 8$	$+0.00229$ $[0.00063, 0.00397]$
Wrong map: full recoupling − release regret/stage	$+0.00838$ $[0.00678, 0.00996]$
GB exceeds summed within-group certificates	$10{,}887/23{,}040$ (47.3%)
Complementarity control: GB − Matching gain/stage	$+0.0128$ $[0.0085, 0.0170]$
Defect − random AUC: inventory / cliff	-0.2954 / -0.03006
Six factors: GB share of full gain with 4 of 64 passes	0.895 $[0.810, 0.978]$
Six factors: GB − Matching share at $k = 4$ blocks	$+0.834$ $[0.721, 0.939]$
Eight factors, state-region map: GB − Matching share at $k = 8$	$+0.031$ $[0.018, 0.045]$
Greedy/MIP ratio at 256 groups; time at 128 probes	0.99995; $0.068/0.880$ s

Q4: Defect-based rankings outperform random ordering in seasonal inventory and trajectory-estimated CliffWalking; Appendix F.7 also includes crosswind navigation. In the selector-only sweep, all 120 MIPs reach the prescribed 10^{-4} relative-gap tolerance, and 32 small cases match exhaustive enumeration. At 256 groups the greedy/MIP ratio is 0.99995 $[0.99987, 0.99999]$; at 128 probes median times are $0.068/0.880$ seconds. Appendices F.7–F.9 give the detailed results.

Q5: With six binary factor pools, retaining the two largest pooled certificates recovers 89.5% of the full recoupling gain on held-out policies using four Bellman passes, versus 31.2% for Matching, 39.0% for Occupancy, and 33.9% for random orders; full recoupling uses 64 passes. Candidate-menu regret per stage drops from 0.00191 under release to 0.00037 with those pools and 0.00019 under full recoupling. At four blocks, exact Global Balance again recovers 89.5% of the gain, exceeding Matching by 0.834 $[0.721, 0.939]$. Figure 4 shows the gain–cost trade-off.

A harder eight-factor benchmark uses a structure-derived state-region/stage-band map with varied within-block sign patterns. Global Balance has higher mean gain share than Matching at $k \in \{4, 6, 8, 12, 16\}$ (paired differences $+0.015$ to $+0.037$); the mean shares of greedy and exact Global Balance differ by at most 0.0071. The four best pools recover 67.1% of the gain with 16 of 256 passes, and six recover 90.0% with 64. Appendix F.5 reports both benchmarks, all comparators, and policy-pass accounting.

8 Conclusion

Reachable defects give an exact, policy-specific audit of rectangularization pessimism. For declared affine shared factors, Global Balance provides a budgeted recoupling certificate with overlap and selection guarantees, while finite-data validity is preserved after selection. Experiments show that correct sharing can reduce held-out regret, incorrect sharing can reverse the benefit, and multi-factor selection can recover most of the full gain with far fewer Bellman passes. Thus transition dependence becomes an auditable, optimizable part of robust MDP design.

Statement of Use of AI

The research questions, central ideas, theoretical claims, methodology, and experimental protocols of this paper were determined by the authors, who divided the work among themselves, supervised every stage of the project, and made all decisions regarding the claims and conclusions. Under the authors' direction, we used generative AI tools as assistants in several disclosed roles. In particular, for the theoretical development, generative AI was used as a form of "theory attacker": it assisted in implementing computational searches and large-scale simulations over very large collections of generated examples with the aim of identifying potential counterexamples, boundary cases, or weaknesses in candidate theorem statements. These computational stress tests provided additional empirical evidence that helped the authors examine and refine the statements of the theorems. However, the failure of such searches to find a counterexample does not constitute a proof, and no theorem in this paper is regarded as established on the basis of AI-generated testing or simulation. All formal proofs were inspected, evaluated, and accepted by the authors on their mathematical merits.

Generative AI tools were also used to discuss experimental design, assist with implementing the proposed methods and procedural generators for the synthetic benchmark systems, and support the interpretation of experimental results. The authors inspected the primary sources and executed and independently verified all analyses. No experimental data or reported results were produced by generative AI tools themselves: every reported number comes from experiments that were actually executed, and translation and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools for literature-search triage, drafting and editing the text and LaTeX for readability, checking notation and the consistency of reported numbers, and figure layout. The authors take full responsibility for every claim, proof, citation, and experimental result in this paper; no AI system is an author.

Reproducibility Statement

Appendix F specifies the generators, policy study, estimands, sampling units, solver settings, and numerical cross-checks for the reported experiments. All estimates and intervals used by the paper are reported in this PDF. The supplementary material provides the source code and configurations for all experiments, together with summaries of the recorded outputs.

References

Shipra Agrawal, Yichuan Ding, Amin Saberi, and Yinyu Ye. Price of correlations in stochastic optimization. *Operations Research*, 60(1):150–162, 2012. doi: 10.1287/opre.1110.1011.

Kasper Engelen, Guillermo A. Pérez, and Marnix Suilen. Data-efficient safe policy improvement using parametric structure. *arXiv preprint arXiv:2507.15532*, 2025. URL https://arxiv.org/abs/2507.15532. Accepted at ECAI 2025.

Larry G. Epstein and Martin Schneider. Recursive multiple-priors. *Journal of Economic Theory*, 113 (1):1–31, 2003. doi: 10.1016/S0022-0531(03)00097-8.

Uri Gadot, Kaixin Wang, Esther Derman, Navdeep Kumar, Kfir Levy, and Shie Mannor. Targeted uncertainty reduction in robust MDPs. In *NeurIPS 2023 Workshop on Generalization in Planning*, 2023. URL https://openreview.net/pdf?id=5kO572pUvr.

Vineet Goyal and Julien Grand-Clément. Robust Markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 48(1):203–226, 2023. doi: 10.1287/moor.2022.1259.

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.

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

Yasushi Kawase and Hanna Sumita. Randomized strategies for robust combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7876–7883, 2019. doi: 10.1609/aaai.v33i01.33017876.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49:209–232, 2002. doi: 10.1023/A:1017984413808.

Jean B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001. doi: 10.1137/S1052623400366802.

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.

Yan Li and Alexander Shapiro. Rectangularity and duality of distributionally robust Markov decision processes. *Mathematical Programming*, 2025. doi: 10.1007/s10107-025-02297-y.

Shie Mannor, Ofir Mebel, and Huan Xu. Robust MDPs with k-rectangular uncertainty. *Mathematics of Operations Research*, 41(4):1484–1509, 2016. doi: 10.1287/moor.2016.0786.

Nirbhay Modhe, Harish Kamath, Dhruv Batra, and Ashwin Kalyan. Model-advantage and value-aware models for model-based reinforcement learning: Bridging the gap in theory and practice. *arXiv preprint arXiv:2106.14080*, 2021. URL https://arxiv.org/abs/2106.14080.

Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005. doi: 10.1287/opre.1050.0216.

Tim Quatmann, Christian Dehnert, Nils Jansen, Sebastian Junges, and Joost-Pieter Katoen. Parameter synthesis for Markov models: Faster than ever. In *Automated Technology for Verification and Analysis*, volume 9938 of *Lecture Notes in Computer Science*, pp. 50–67. Springer, 2016. doi: 10.1007/978-3-319-46520-3_4.

Anirudh Satheesh, Ziyi Chen, Furong Huang, and Heng Huang. Provably efficient algorithms for s- and non-rectangular robust MDPs with general parameterization. *arXiv preprint arXiv:2602.11387*, 2026. URL https://arxiv.org/abs/2602.11387.

Yannik Schnitzer, Alessandro Abate, and David Parker. Robust parameter learning for uncertain MDPs. *arXiv preprint arXiv:2605.01339*, 2026. URL https://arxiv.org/abs/2605.01339.

Alexander Shapiro. Rectangular sets of probability measures. *Operations Research*, 64(2):528–541, 2016. doi: 10.1287/opre.2015.1466.

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.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013. doi: 10.1287/moor.1120.0566.

Gang Yu. On the max-min 0-1 knapsack problem with robust optimization applications. *Operations Research*, 44(2):407–415, 1996. doi: 10.1287/opre.44.2.407.

Adil Zouitine, David Bertoin, Pierre Clavier, Matthieu Geist, and Emmanuel Rachelson. Time-constrained robust MDPs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 35574–35611. Curran Associates, Inc., 2024. doi: 10.52202/079017-1122.

Appendix outline. The first part proves the dependency-release and operations-research results. The second fixes the timing and counterfactual-representation semantics. The final part records the experimental protocol, binary evaluation recursions, numerical validation criteria, and quantitative audits for the reported studies.

A Supplementary Pairwise Setup

Retain the fixed Markov probe and additive loss, and let $z = (z_i)_{i\in\Pi} \in [0,1]^{\Pi}$ be the block-parameter vector. A partition is *path-simple* when no parameter block occurs twice on any trajectory supported by the union of its endpoint kernels, and has interaction order at most two when every such trajectory visits at most two uncertain blocks. For this class, expanding expected loss over rewarded prefixes gives the pairwise multi-affine objective

$$p_{\Pi}(z) = c_0 + \sum_i c_i z_i + \sum_{i<j} c_{ij} z_i z_j. \tag{21}$$

The signed interaction graph has vertex set Π, an edge ij whenever $c_{ij} \neq 0$, and edge sign $\operatorname{sign}(c_{ij})$. The proof appendix states the associated coefficient recursion, sign-switchable min-cut result, and shallow-class hardness theorem.

B Proofs for Dependency Release

This section proves the certificate in four layers. We first derive an exact one-step residual identity and telescope it under coherent occupancies. We then specialize the identity to affine dependency partitions and construct a parameter-uniform prefix measure. The prefix measure leads to the two-scenario global-balance certificate. Finally, we study the selector's pairwise computational results, the overlap-dependent approximation factor, and the finite-data event on which every post-selection certificate remains valid.

Proof of Theorem 1. Step 1 (one-step defect recursion). Fix $P \in \mathcal{C}$ and let V_h^P be its Bellman value. Put $D_h(s) = \bar{V}_h(s) - V_h^P(s)$. Inserting and subtracting $P_x^\top(r_x + \bar{V}_{h+1})$ in the two recursions gives, exactly,

$$D_h(s) = \sum_a \pi_h(a \mid s)\left[\delta_{hsa}(P_{hsa}) + P_{hsa}^\top D_{h+1}\right]. \tag{22}$$

Step 2 (occupancy telescoping). Let $\rho_h^P(s) = \Pr_P(s_h = s)$, so $d_{hsa}^P = \rho_h^P(s)\pi_h(a \mid s)$. Multiply (22) by $\rho_h^P(s)$ and sum over s. The continuation term is

$$\sum_{s,a,s'} \rho_h^P(s)\pi_h(a \mid s)P_{hsa}(s')D_{h+1}(s') = \sum_{s'} \rho_{h+1}^P(s')D_{h+1}(s').$$

Summing over stages telescopes; $D_H = 0$, proving (7).

Step 3 (minimization and the equality case). Lemma 15 and the Bellman recursion (6) give $\bar{W} = W_{\mathcal{R}}(e,g)$, while $\max_{P\in\mathcal{C}} J(P) = W_{\mathcal{C}}(e,g)$. Since $\bar{W}$ is independent of P and $\mathcal{C}$ is compact,

$$\bar{W} - \max_{P\in\mathcal{C}} J(P) = \min_{P\in\mathcal{C}}\big[\bar{W} - J(P)\big],$$

which proves (8). Every summand is nonnegative and the minimum is attained, so it vanishes exactly under (9). □

Proof of Equation (11)*, Corollary 2, and Lemma 3. Step 1 (partition nesting).* If Π' refines Π, assign the same value to every Π' block contained in a common Π block. This embeds $\mathcal{C}_\Pi$ into $\mathcal{C}_{\Pi'}$. Every partition has the same endpoint segment at an individual row; the Cartesian product of those segments is exactly $\mathcal{C}_{\Pi^\top}$. This proves (11).

Step 2 (affine row defects). For a row in block B,

$$\max_{\eta\in[0,1]} (P_x^0 + \eta\Delta_x)^\top u_x - (P_x^0 + \theta_B\Delta_x)^\top u_x$$
$$= \max\{0, \beta_x\} - \theta_B\beta_x.$$

Rows outside $\mathcal{X}_+$ have zero defect, so substitution into Theorem 1 and restriction of the sum to $\mathcal{X}_+$ prove (12) and its equality condition.

Step 3 (uniform prefix masses). For every $\gamma \in \Gamma_x$ underlying (13), each affine transition coordinate is at least the smaller endpoint coordinate. The events defined by distinct state–action prefixes ending at x are disjoint, so summing their lower probabilities gives $d_x^\theta \geq \lambda_x$ uniformly over every coherent parameter vector. For selected pair i in block $B(i)$, the two rows contribute at least

$$\lambda_{x_i^+}\beta_{x_i^+}(1 - \theta_{B(i)}) + \lambda_{x_i^-}(-\beta_{x_i^-})\theta_{B(i)},$$

which is at least the corresponding term in (14) for every $\theta_{B(i)} \in [0, 1]$. Pairwise row-disjointness prevents double counting. Summing and minimizing over θ completes the proof, whether or not partition blocks repeat.

The same calculation gives a direct forward recursion for all prefix masses. Put $\mu_0 = d_0$ and

$$\mu_{h+1}(s') = \sum_{s,a} \mu_h(s)\pi_h(a \mid s) \min_{e\in\{0,1\}} P_{hsa}^e(s'), \qquad \lambda_{hsa} = \mu_h(s)\pi_h(a \mid s).$$

Expanding this recursion recovers exactly the disjoint prefixes in (13); dense contractions take $O(HS^2A)$ arithmetic operations. □

Proof of Theorem 4. Step 1 (two-scenario balance). For an arbitrary feasible parameter vector, discard the nonnegative defect terms from rows outside $\bigcup_{B\in K} B$. Write θ for the scalar shared by the selected rows. Uniformly over all remaining parameters, every selected-row occupancy satisfies $d_x^\theta \geq \lambda_x$. Their positive-slope rows therefore contribute at least $(1 - \theta)\sum_{B\in K} A_B$, and their negative-slope rows contribute at least $\theta\sum_{B\in K} C_B$. Minimization over $\theta \in [0, 1]$ gives

$$\min_{\theta\in[0,1]} \left\{ (1 - \theta) \sum_{B\in K} A_B + \theta \sum_{B\in K} C_B \right\} = \mathcal{L}(K).$$

This proves the continuous bound by Corollary 2. The binary family is a subset of the continuous family, whereas full row release has the same endpoint value in both; hence its commitment gap is no smaller and obeys the same bound. If each group has a separate scalar, the same calculation minimizes group by group and gives $\sum_{B\in K} \min\{A_B, C_B\}$.

Step 2 (sharpness under the retained information). Under the retained rowwise information, the available constraints are $d_x^\theta \geq \lambda_x$ and nonnegativity of every omitted contribution. Every defect coefficient is nonnegative, so the relaxed minimum sets each selected occupancy to its lower endpoint λ_x. What remains is exactly the scalar affine function minimized in the preceding display, whose value is $\mathcal{L}(K)$. Thus $\mathcal{L}(K)$ is sharp for this algebraic relaxation of the occupancy variables. □

Proof of Proposition 6. For each menu probe, z_j in (17) is bounded above by both sign totals and is maximized, so at any fixed binary vector it equals their minimum divided by G_j. The cardinality constraint gives exactly the declared subset family. Maximizing over y therefore proves the program. □

B.1 Pure and convexified selectors for one probe

The balance objective has two different optimization geometries. A deployed recoupling menu is a single subset and must balance both endpoint scenarios inside that subset. Convexifying over menus instead balances the two scenarios in expectation. The next result identifies the exact gap between these views. We use the convexified problem as a computable upper-bounding oracle. Its

optimizer has a two-set support. More strongly, the marginal inclusion vector can be rounded through a two-cover linear program to a deterministic menu using only one group beyond the original budget.

For $K \subseteq \mathcal{B}$, abbreviate $A(K) = \sum_{B \in K} A_B$, $C(K) = \sum_{B \in K} C_B$, and let $\mathfrak{K}_k = \{K \subseteq \mathcal{B} : |K| \leq k\}$. Define

$$V_k^{\text{pure}} = \max_{K \in \mathfrak{K}_k} \min\{A(K), C(K)\}, \tag{23}$$

$$V_k^{\text{mix}} = \max_{\nu \in \Delta(\mathfrak{K}_k)} \min\{\mathbb{E}_\nu A(K), \mathbb{E}_\nu C(K)\}. \tag{24}$$

The next lemma is the geometric step that makes the convexified problem different from a generic linear program over exponentially many menus. For $\lambda \in [0, 1]$, define

$$\mathcal{M}_k(\lambda) = \arg\max_{K \in \mathfrak{K}_k} \{\lambda A(K) + (1 - \lambda) C(K)\} \tag{25}$$

and write

$$d(K) = A(K) - C(K) \tag{26}$$

for the signed imbalance of a menu.

Lemma 9 (Active-menu subgradient certificate)**.** *Let*

$$T_k(\lambda) = \max_{K \in \mathfrak{K}_k} \{\lambda A(K) + (1 - \lambda) C(K)\}. \tag{27}$$

Then T_k is convex and piecewise affine, with subdifferential

$$\partial T_k(\lambda) = \operatorname{conv}\{d(K) : K \in \mathcal{M}_k(\lambda)\}. \tag{28}$$

If λ^* minimizes T_k on $[0, 1]$, one of the following mutually exclusive certificates exists.

(a) If $0 < \lambda^* < 1$, there are active menus K^-, K^+ and $\rho \in [0, 1]$ such that

$$d(K^-) \leq 0 \leq d(K^+) \tag{29}$$

and

$$\rho d(K^-) + (1 - \rho) d(K^+) = 0. \tag{30}$$

The corresponding mixture has both expected sign masses equal to $T_k(\lambda^*)$.

(b) If $\lambda^* = 0$, an active menu K^+ satisfies

$$C(K^+) = T_k(0) \qquad \text{and} \qquad A(K^+) \geq C(K^+). \tag{31}$$

(c) If $\lambda^* = 1$, an active menu K^- satisfies

$$A(K^-) = T_k(1) \qquad \text{and} \qquad C(K^-) \geq A(K^-). \tag{32}$$

Thus every minimizer of the scalar envelope produces either one pure menu or a mixture of two menus whose smaller expected coordinate equals the envelope minimum.

Proof. For every feasible menu, expand its scalarized score as

$$\lambda A(K) + (1 - \lambda) C(K) = C(K) + \lambda d(K). \tag{33}$$

Hence T_k is the pointwise maximum of finitely many affine functions. It is therefore convex and piecewise affine. The subdifferential of such a maximum is the convex hull of the slopes of the active affine functions, giving Equation (28).

Suppose first that λ^* is interior. Fermat's rule gives

$$0 \in \partial T_k(\lambda^*). \tag{34}$$

Because the subdifferential is an interval in one dimension, its smallest and largest elements are achieved by two active menus. Call them K^- and K^+. They satisfy Equation (29), and a convex weight ρ can be chosen to satisfy Equation (30). Every active menu obeys

$$\lambda^* A(K) + (1 - \lambda^*) C(K) = T_k(\lambda^*). \tag{35}$$

After mixing this equality over K^- and K^+, the zero-imbalance identity forces the two expected coordinates to coincide. Their common value must be $T_k(\lambda^*)$.

At the left endpoint, constrained optimality is equivalent to the existence of a subgradient $g \in \partial T_k(0)$ with $g \geq 0$. Since the largest active slope is attained, there is an active K^+ with $d(K^+) \geq 0$. Activity at $\lambda = 0$ gives $C(K^+) = T_k(0)$, which proves Equation (31). The right-endpoint argument is symmetric: an active nonpositive slope exists, activity gives $A(K^-) = T_k(1)$, and Equation (32) follows. □

Theorem 10 (Two-scenario specialization and one-extra rounding). *Let $A_B, C_B \geq 0$ and $1 \leq k \leq |\mathcal{B}|$.*

(i) For arbitrary nonnegative integer score pairs, the threshold problem for V_k^{pure} is weakly NP-complete. This is a downstream score-vector statement, not a claim that every pair is induced by the MDP construction. An exact frontier recursion nevertheless runs in $O(|\mathcal{B}|k(1 + \min\{A_\Sigma, C_\Sigma\}))$ time and $O(k(1 + \min\{A_\Sigma, C_\Sigma\}))$ memory.

(ii) The convexified value has the exact one-dimensional dual

$$V_k^{\mathrm{mix}} = \min_{0 \leq \lambda \leq 1} T_k(\lambda), \qquad T_k(\lambda) = \sum_{B \in \mathrm{Top}_k(s(\lambda))} s_B(\lambda), \qquad s_B(\lambda) = \lambda A_B + (1-\lambda) C_B. \tag{36}$$

Here ties may be resolved arbitrarily in the value of T_k. Thus an exponentially supported linear program reduces to a parametric top-k sweep.

(iii) There is an optimal distribution in Equation (24) supported on at most two subsets $K^-, K^+ \in \mathfrak{K}_k$. There is also a deterministic menu $K^\uparrow$ that realizes the mixed upper bound with only one additional group:

$$V_k^{\mathrm{pure}} \leq V_k^{\mathrm{mix}} \leq V_{\min\{k+1, |\mathcal{B}|\}}^{\mathrm{pure}} \tag{37}$$

and

$$|K^\uparrow| \leq \min\{k+1, |\mathcal{B}|\}. \tag{38}$$

Moreover,

$$\min\{A(K^\uparrow), C(K^\uparrow)\} \geq V_k^{\mathrm{mix}}. \tag{39}$$

For every $1 \leq k < |\mathcal{B}|$, the extra group is necessary on some instance.

(iv) The value in Equation (36), the two supporting menus, and their mixing weight can be recovered in $O(|\mathcal{B}|^2 \log |\mathcal{B}|)$ arithmetic operations by sweeping the pairwise crossings of the affine scores s_B. The menu $K^\uparrow$ is recovered by one linear program with two covering constraints followed by rounding at most two fractional coordinates upward.

Parts (ii) and the two-menu statement in (iii) are the two-scenario modular specialization of randomized robust combinatorial optimization (Kawase & Sumita, 2019); the proof is included for completeness. The two-cover rounding argument gives a deterministic menu using at most one group beyond the mixed budget.

Proof. The proof has four layers. We first establish the discrete obstruction and its pseudo-polynomial recursion. We then exchange a distribution over menus with the two-scenario adversary. The geometry of the resulting planar convex hull gives the two-menu support. Finally, a parametric-order-statistic sweep turns the dual identity into an explicit algorithm.

Step 1 (the pure selector is weakly NP-complete). We reduce from exact-k subset sum. Let nonnegative integers $w_1, \ldots, w_n$, a positive target T, and a cardinality k be given. Put $w_{\max} = \max_i w_i$ and $M = kw_{\max} + T$. For group i, define

$$A_i = kM + kw_i. \tag{40}$$

Define its second-scenario mass by

$$C_i = kM + 2T - kw_i. \tag{41}$$

Finally, set the decision threshold to

$$\tau = k^2 M + kT. \tag{42}$$

All masses are nonnegative. If $|K| = k$ and $w(K) = \sum_{i \in K} w_i$, then

$$\begin{aligned} \min\{A(K), C(K)\} &= \min\{k^2 M + k w(K), k^2 M + 2kT - k w(K)\} \\ &= \tau - k|w(K) - T|. \end{aligned} \tag{43}$$

Hence a size-k set reaches τ exactly when its weights sum to T.

The budget in Equation (23) is "at most k," so we must also rule out smaller witnesses. If $m = |K| < k$, then

$$\begin{aligned} \min\{A(K), C(K)\} &\le A(K) \le mk(M + w_{\max}) \\ &\le (k-1)k(M + w_{\max}) < k^2 M + kT = \tau. \end{aligned} \tag{44}$$

The strict inequality is equivalent to $M + T > (k-1) w_{\max}$ and follows from the definition of M. The reduction has polynomial encoding length. Membership in NP is immediate from a selected subset, which proves weak NP-completeness.

For completeness, assume first that $A_\Sigma \le C_\Sigma$. After processing the first i groups, store

$$D_i(m, a) = \max\{C(K) : K \subseteq \{1, \dots, i\},\ |K| = m,\ A(K) = a\}, \tag{45}$$

with value $-\infty$ for an unreachable state. Starting from $D_0(0, 0) = 0$, inclusion or exclusion of group i gives

$$D_i(m, a) = \max\{D_{i-1}(m, a), D_{i-1}(m-1, a - A_i) + C_i\}. \tag{46}$$

The exact readout is

$$\max_{0 \le m \le k} \max_{0 \le a \le A_\Sigma} \min\{a, D_n(m, a)\}. \tag{47}$$

Two layers use $O(k(1 + A_\Sigma))$ memory. Interchanging A and C when $C_\Sigma < A_\Sigma$ proves the stated bounds.

Step 2 (minimax collapses the convexification to one coordinate). For any two nonnegative scalars a, c,

$$\min\{a, c\} = \min_{0 \le \lambda \le 1} \{\lambda a + (1 - \lambda) c\}. \tag{48}$$

Apply this identity to the two expectations in Equation (24). The distribution simplex and the interval $[0, 1]$ are compact and convex, and the displayed payoff is bilinear. Finite-dimensional minimax therefore permits the exchange

$$\begin{aligned} V_k^{\mathrm{mix}} &= \max_{\nu \in \Delta(\mathfrak{K}_k)} \min_{0 \le \lambda \le 1} \mathbb{E}_\nu[\lambda A(K) + (1 - \lambda) C(K)] \\ &= \min_{0 \le \lambda \le 1} \max_{K \in \mathfrak{K}_k} \sum_{B \in K} \{\lambda A_B + (1 - \lambda) C_B\}. \end{aligned} \tag{49}$$

At fixed λ, every score is nonnegative. The inner maximum therefore selects the k largest scores, which proves Equation (36).

Step 3 (two-menu support and one-extra deterministic rounding). Associate each feasible menu with the point $q_K = (A(K), C(K)) \in \mathbb{R}_+^2$ and let $\mathcal{Q}_k = \mathrm{conv}\{q_K : K \in \mathfrak{K}_k\}$. A distribution ν produces exactly one point of $\mathcal{Q}_k$, and the mixed objective is the smaller coordinate of that point. Let q^* be an optimizer and let λ^* minimize Equation (36). The minimax saddle conditions place q^* on the exposed face

$$\mathcal{F}^* = \arg\max_{q \in \mathcal{Q}_k} \{\lambda^* q_1 + (1 - \lambda^*) q_2\}. \tag{50}$$

An exposed face of a planar polytope is a point or a line segment. Hence q^* is a convex combination of at most two vertices of $\mathcal{F}^*$, say q_{K^-} and q_{K^+}. This proves the support claim without invoking a distribution over exponentially many menus.

We next obtain a deterministic realization that is sharper than taking the union. For any optimal distribution ν^*, let

$$x_B = \Pr_{K \sim \nu^*}\{B \in K\}. \tag{51}$$

The marginal vector satisfies

$$\sum_B x_B \le k, \tag{52}$$

$$\sum_B A_B x_B \geq V_k^{\mathrm{mix}}, \tag{53}$$

and

$$\sum_B C_B x_B \geq V_k^{\mathrm{mix}}. \tag{54}$$

Consider the two-cover linear program

$$\begin{aligned} \min_{0 \leq z \leq 1} \quad & \sum_B z_B \\ \text{subject to} \quad & \sum_B A_B z_B \geq V_k^{\mathrm{mix}}, \\ & \sum_B C_B z_B \geq V_k^{\mathrm{mix}}. \end{aligned} \tag{55}$$

The marginal vector is feasible, so this program has an optimal extreme point z^* with

$$\sum_B z_B^* \leq k. \tag{56}$$

At most two coordinates of z^* are strictly between zero and one. Indeed, if three or more coordinates were fractional, the coefficient vectors of the active covering constraints would impose at most two homogeneous linear conditions on perturbations supported on those coordinates. A nonzero perturbation satisfying both conditions would exist. A sufficiently small positive or negative multiple would preserve all box constraints and every active cover, while inactive covers remain slack. This would express z^* as the midpoint of two distinct feasible points, contradicting extremality.

Round every fractional coordinate of z^* upward and call the selected set $K^{\uparrow}$. Nonnegativity of the masses preserves both covers. If there are no fractional coordinates, Equation (56) gives $|K^{\uparrow}| \leq k$. With one fractional coordinate, rounding gives an integer strictly smaller than $k+1$, hence again at most k. With two fractional coordinates,

$$|K^{\uparrow}| < \sum_B z_B^* + 2 \leq k + 2, \tag{57}$$

so integrality gives $|K^{\uparrow}| \leq k+1$. This proves Equations (37) and (38)– (39); the first inequality in the sandwich holds because every pure menu is a degenerate distribution.

It remains to prove sharpness. For $k = 1$, take the two groups with masses $(1, 0)$ and $(0, 1)$. Their uniform mixture has value $1/2$, while either one-group menu has value zero. For $k \geq 2$, take $k-1$ common groups with

$$(A_B, C_B) = (2, 2) \tag{58}$$

and two complementary groups with masses $(1, 0)$ and $(0, 1)$. At $\lambda = 1/2$, every common group has scalarized score 2 and each complementary group has score $1/2$. Symmetry and convexity of T_k make $1/2$ a minimizer, whence

$$V_k^{\mathrm{mix}} = 2(k-1) + \frac{1}{2}. \tag{59}$$

Any menu of size at most k that retains all common groups can take only one complementary group and has smaller mass at most $2(k-1)$. If it omits a common group, even taking both complementary groups gives smaller mass at most

$$2(k-2) + 1 = 2k - 3. \tag{60}$$

Thus no k-menu reaches the mixed value, whereas the full $(k+1)$-group menu does. The one-extra-group guarantee is sharp for every admissible k.

Step 4 (recovering the saddle by a top-k sweep). Each $s_B(\lambda)$ is affine. The identity

$$T_k(\lambda) = \max_{K \in \mathfrak{K}_k} \sum_{B \in K} s_B(\lambda) \tag{61}$$

shows that T_k is convex and piecewise affine. Its active top-k set can change only at an endpoint or at a pairwise score crossing. There are $O(|\mathcal{B}|^2)$ such locations. Sort them once, maintain the score

order and the sum of the largest k entries in a balanced order-statistic structure, and inspect the affine envelope at each event. This costs $O(|\mathcal{B}|^2 \log |\mathcal{B}|)$ arithmetic operations.

At a minimizing crossing, the left and right slopes of the convex envelope bracket zero. Choose active top-k menus K^-, K^+ attaining those extreme slopes and write $d^- = A(K^-) - C(K^-)$ and $d^+ = A(K^+) - C(K^+)$. When $d^- < 0 < d^+$, the explicit weight

$$\rho = \frac{d^+}{d^+ - d^-} \tag{62}$$

on K^- makes $\rho d^- + (1-\rho)d^+ = 0$. The common exposed value then makes both expected coordinates equal to V_k^{mix}; if either slope is zero, its menu alone has this property. At an endpoint minimizer, an active menu whose one-sided slope has the required sign already attains the smaller endpoint coordinate; it can be used alone. Ties at the kth score are handled by choosing, within the tied block, the smallest and largest total imbalance. This recovers the claimed menus and weight and completes the top-k part of the algorithm. Solving Equation (55) at the recovered value and selecting an optimal basic feasible solution exposes at most two fractional coordinates. Rounding them upward recovers $K^{\uparrow}$ in polynomial time and completes the proof. □

Proof of Corollary 5. Fix arbitrary values for every unselected singleton row. For each factor pool f, give all rows in $\bigcup_{B \in K_f} B$ one scalar θ_f and discard the nonnegative defect terms of the unselected rows. Lemma 3 holds uniformly over the complete vector of factor values. The selected rows in pool f therefore contribute at least

$$(1-\theta_f) \sum_{B \in K_f} A_B + \theta_f \sum_{B \in K_f} C_B.$$

Different pools use different coordinates, so minimizing the sum separates over f and gives $\sum_f \mathcal{L}_f(K_f)$. The binary conclusion follows by restricting every θ_f to its two endpoints. □

Proof of Theorem 7. The proof couples four ingredients. First we compare every coherent prefix occupancy with the identical-support prefix measure. We then use a feasible shared parameter to upper-bound both the continuous and binary gaps and place these bounds beside the global-balance lower certificate. Finally, a two-branch family attains the binary horizon factor and reduces the continuous ratio to a scalar convex-symmetrization problem.

Step 1 (occupancy sandwich). Every affine mixture lies coordinatewise between its two endpoints. Hence (18) bounds the probability of any length-h prefix under any coherent parameter vector by χ^h times the product of the endpoint-coordinate minima along that prefix. Summing the disjoint prefixes ending at row $x = (h, s, a)$ gives

$$d_x^\theta \le \chi^h \lambda_x \le \chi^{H-1} \lambda_x. \tag{63}$$

Step 2 (upper bounds for continuous and binary coupling). Fix K and assign every unselected singleton row its locally adverse endpoint, so all of its defects vanish. If the selected rows share $\theta \in [0,1]$, Theorem 1 and (63) give

$$\bar{W} - J(P^\theta) \le \chi^{H-1} \left[(1-\theta) \sum_{B \in K} A_B + \theta \sum_{B \in K} C_B \right].$$

The continuous gap is the minimum of the left side over its coherent family, so minimizing this displayed upper bound over $\theta \in [0,1]$ gives $G(K) \le \chi^{H-1} \mathcal{L}(K)$. For the binary family, choose the better of $\theta = 0$ and $\theta = 1$; the same minimum is $\mathcal{L}(K)$, proving $G^{\mathrm{bin}}(K) \le \chi^{H-1} \mathcal{L}(K)$.

Step 3 (ordering and lower bounds). Finally, $G(K) \le G^{\mathrm{bin}}(K)$ because the binary coherent family is contained in the continuous one, and both lower bounds follow from Theorem 4.

Step 4 (sharpness construction). It remains to prove (20). Put $h = H - 1$ and $p = 1/(\chi + 1)$. Start with probability $1/2$ at each of two one-action chains, labelled $+$ and $-$. Before the last row, a

transition either continues along its chain or enters a compensating state. At every prefix row, the continuation probabilities are

$$q_+(\theta) = p\{\chi - (\chi - 1)\theta\}, \qquad q_-(\theta) = p\{1 + (\chi - 1)\theta\}.$$

Equivalently, the uncertain rows form the following finite gadget:

Row	endpoint-0/endpoint-1 probability	remaining transition
$+$ prefix, $t < h$	$\chi p \,/\, p$ to next $+$ state	complement to compensation
$-$ prefix, $t < h$	$p \,/\, \chi p$ to next $-$ state	complement to compensation
$+$ terminal row	$p \,/\, \chi p$ to loss-one terminal	complement to loss-zero terminal
$-$ terminal row	$\chi p \,/\, p$ to loss-one terminal	complement to loss-zero terminal

The complementary probabilities go to the compensating state. Give that transition a loss equal to the fully released continuation of the next chain state. Its two successor continuations are then equal, so every prefix row has $\beta_x = 0$, while a path that enters compensation incurs no release defect. At the last $+$ row, let the probability of terminal loss one rise from p to χp; at the last $-$ row let it fall from χp to p. All other terminal loss is zero. Every uncertain row belongs to the one selected shared block.

Every endpoint row has the two positive coordinates p and χp, in opposite order, so (18) holds with equality. Only the two last rows have nonzero released slopes, of common magnitude $b = (\chi - 1)p$. Each has common-prefix mass $\lambda = p^h/2$; hence $\mathcal{L}(K) = bp^h/2$. The exact continuous defect is

$$\frac{b}{2} \min_{0 \le \theta \le 1} \{(1-\theta)q_+(\theta)^h + \theta q_-(\theta)^h\}.$$

Writing $c = \chi - 1$ and $\phi(t) = t(1 + ct)^h$, the expression inside braces, after division by p^h, is $\phi(\theta) + \phi(1-\theta)$. Since ϕ is convex on $[0, 1]$, Jensen's inequality makes its minimum occur at $\theta = 1/2$ and gives $((\chi + 1)/2)^h$. Restricting to binary θ instead gives χ^h. Dividing by $\mathcal{L}(K)$ proves both identities.

The construction is finite, and its compensation losses are defined backwards after the next-stage released values are fixed. A common stagewise shift followed by one positive global rescaling places all losses in a bounded nonnegative range. This transformation acts equally on every model value and leaves both ratios unchanged. The same gadget also gives an explicit failure without common support. Set the prefix continuation endpoints to $(1, 0)$ on the $+$ branch and $(0, 1)$ on the $-$ branch, and reverse them at the terminal rows. Then $\mathcal{L}(K) = 0$, whereas $G(K) = \frac{1}{2} \min_{\theta \in [0,1]} \{(1-\theta)^H + \theta^H\} = 2^{-H} > 0$. Thus no finite multiplicative bound follows in that case. □

Corollary 11 (Budgeted selector guarantee)**.** *For j, let $\mathcal{L}_j(K)$ and $D_j^t(K)$ be its Global-Balance certificate and true recoupling gain, respectively, for $t \in \{\mathrm{con}, \mathrm{bin}\}$, with fixed $G_j > 0$. If every policy obeys the overlap bound χ, the certificate-maximizing subset $\widehat{K}$ from Proposition 6 satisfies*

$$\frac{1}{|\mathcal{J}_+|} \sum\nolimits_j \frac{D_j^t(\widehat{K})}{G_j} \ge \chi^{-(H-1)} \max_{|K| \le k} \frac{1}{|\mathcal{J}_+|} \sum\nolimits_j \frac{D_j^t(K)}{G_j}, \qquad t \in \{\mathrm{con}, \mathrm{bin}\}. \tag{64}$$

Proof of Corollary 11. For $t \in \{\mathrm{con}, \mathrm{bin}\}$, define

$$C(K) = \frac{1}{|\mathcal{J}_+|} \sum_j \frac{\mathcal{L}_j(K)}{G_j}, \qquad F_t(K) = \frac{1}{|\mathcal{J}_+|} \sum_j \frac{D_j^t(K)}{G_j}.$$

Theorem 7 holds probe by probe, so $C(K) \le F_t(K) \le \chi^{H-1} C(K)$. If K_t^* maximizes F_t under the budget, Proposition 6 gives

$$F_t(\widehat{K}) \ge C(\widehat{K}) \ge C(K_t^*) \ge \chi^{-(H-1)} F_t(K_t^*),$$

which is (64). □

For the local comparison, put $B^+ = \{x \in B : \lambda_x > 0, \beta_x > 0\}$, $B^- = \{x \in B : \lambda_x > 0, \beta_x < 0\}$, and

$$\ell_B = \max_{M \in \mathsf{Match}(B^+, B^-)} \sum_{(x^+, x^-) \in M} \min\{\lambda_{x^+} \beta_{x^+}, \lambda_{x^-}(-\beta_{x^-})\}. \tag{65}$$

Corollary 12 (Matching is dominated by global balance)**.** *For every $K \subseteq \mathcal{B}$,*

$$\mathcal{L}(K) \geq \sum_{B \in K} \min\{A_B, C_B\} \geq \sum_{B \in K} \ell_B. \tag{66}$$

Here ℓ_B is the largest independently paired, row-disjoint contribution inside B. It can discard unmatched within-group mass and complementary mass across groups. The first inequality can be strict even when every $\ell_B = 0$; on other instances the second is strict when within-group sign mass cannot be fully paired.

Proof of Corollary 12. Fix a group B. Every row-disjoint collection of positive/negative conflicts inside B is a bipartite matching between B^+ and B^-, and every such matching is an admissible collection in (14). That equation assigns an edge (x^+, x^-) exactly the weight displayed in (65). Maximizing its independently paired contributions is therefore precisely the maximum-weight matching value ℓ_B.

Every matched edge weight is bounded by the weight of each incident vertex. Because a matching uses a vertex at most once, $\ell_B \leq A_B$ and $\ell_B \leq C_B$. Thus $\ell_B \leq \min\{A_B, C_B\}$. Summing over $B \in K$ gives the second inequality in (66); the first is $\sum_B \min\{A_B, C_B\} \leq \min\{\sum_B A_B, \sum_B C_B\}$. □

Index the N_{str} separately identifiable two-successor endpoint streams by r, let $\widehat{p}_{r,n}$ be the empirical Bernoulli mean at its nth visit, and set

$$\eta_{r,n} = \frac{6\delta}{\pi^2 N_{\mathrm{str}} n^2}, \qquad \rho_{r,n} = \sqrt{\frac{\log(2/\eta_{r,n})}{2n}}. \tag{67}$$

An unvisited row receives $[0,1]$; a visited row receives the clipping of $[\widehat{p}_{r,n} - \rho_{r,n}, \widehat{p}_{r,n} + \rho_{r,n}]$ to $[0,1]$. The Cartesian product of these intervals with the known supports defines a compact joint endpoint-row confidence set $\widehat{\mathcal{P}}$. For every complete endpoint array $P \in \widehat{\mathcal{P}}$, construct its released continuation, slopes $\beta_x(P)$, and prefix masses $\lambda_x(P)$ as in the population analysis. Define

$$\underline{\beta}_x = \inf_{P \in \widehat{\mathcal{P}}} \beta_x(P), \qquad \overline{\beta}_x = \sup_{P \in \widehat{\mathcal{P}}} \beta_x(P), \qquad \underline{\lambda}_x = \inf_{P \in \widehat{\mathcal{P}}} \lambda_x(P). \tag{68}$$

Compactness makes these extrema well defined. We compute an outward enclosure by propagating row and continuation intervals and taking coordinatewise lower endpoint probabilities for $\underline{\lambda}_x$; any outward interval calculation containing the same extrema yields the same guarantee. For a declared binary protocol Π, let $\mathcal{Z}_\Pi$ be its finite set of globally admissible complete regime assignments: one common label is used where Π retains a dependency, while released rows may receive distinct labels. Let $\widehat{\mathcal{U}}_x(z)$ contain the confidence-compatible rows allowed at x after fixing $z \in \mathcal{Z}_\Pi$. Starting from $\underline{V}^z_H = \overline{V}^z_H = r_H$, and writing r_{hsa} for the vector of successor-dependent one-step losses, the interval Bellman operator is

$$\begin{aligned} \underline{V}^z_h(s) &= \sum_a \pi_h(a \mid s) \min_{q \in \widehat{\mathcal{U}}_{hsa}(z)} q^\top \big(r_{hsa} + \underline{V}^z_{h+1}\big), \\ \overline{V}^z_h(s) &= \sum_a \pi_h(a \mid s) \max_{q \in \widehat{\mathcal{U}}_{hsa}(z)} q^\top \big(r_{hsa} + \overline{V}^z_{h+1}\big). \end{aligned} \tag{69}$$

The protocol enclosure is $L_\Pi = \max_{z \in \mathcal{Z}_\Pi} d_0^\top \underline{V}^z_0$ and $U_\Pi = \max_{z \in \mathcal{Z}_\Pi} d_0^\top \overline{V}^z_0$; retained assignments keep their prescribed common labels throughout the interval calculation. If $[L_K, U_K]$, $[L_\mathcal{R}, U_\mathcal{R}]$, and $[L_C, U_C]$ are the interval Bellman enclosures for a retained subset, release, and commitment, respectively, write W_C for the true committed value. Then $L_\mathcal{R} - U_K > 0$ certifies a true gain. The stronger condition

$$U_K - L_C \leq \tau(L_\mathcal{R} - U_C), \qquad L_\mathcal{R} > U_C, \tag{70}$$

certifies normalized residual at most τ.

Proof of Proposition 8. Step 1 (one anytime event). Let $T_{r,n}$ be the predictable time of the nth visit to stream r. By the observation contract, conditional on the pre-transition filtration at $T_{r,n}$, the successor indicator has the same fixed Bernoulli mean p_r. Conditional Hoeffding's lemma applied to the first n observations of each predictable stream gives the usual Bernoulli tail bound; equivalently,

their centered indicators form a bounded martingale-difference sequence and Hoeffding–Azuma bounds interval failure by $\eta_{r,n}$. The union bound over streams and counts is

$$\sum_{r=1}^{N_{\mathrm{str}}} \sum_{n\geq 1} \eta_{r,n} = \frac{6\delta}{\pi^2} \sum_{n\geq 1} n^{-2} = \delta.$$

Predictability of the visit times makes the same bound valid at the random visit counts.

Step 2 (simultaneous Bellman enclosures). Because the known vector r_{hsa} remains inside each row contraction, Bellman minimization and maximization over the resulting row intervals preserve outer bounds for transition-dependent losses by backward induction. Policies, subsets, scores, and selection are deterministic post-processing of the same event and require no further union bound. Equation (68) supplies simultaneous lower prefix masses and slope intervals. The sign-separated lower weights are $\underline{\lambda}_x[\underline{\beta}_x]_+$ and $\underline{\lambda}_x[-\overline{\beta}_x]_+$.

Step 3 (certificate preservation). The proof of Theorem 4 then applies pointwise on the confidence event.

Step 4 (decision guarantees). Finally, $W_{\mathcal{R}} - W_K \geq L_{\mathcal{R}} - U_K$. Also $W_K - W_C \leq U_K - L_C$ and $W_{\mathcal{R}} - W_C \geq L_{\mathcal{R}} - U_C$; these inequalities prove (70). □

C Proof of Pairwise Tractability and Shallow-Class Hardness

Call a transition endpoint-supported when it has positive probability under at least one endpoint row. A union-support trajectory has positive initial and probe probability and uses only endpoint-supported transitions. The path-simple and interaction-order promises in Section A refer to all such trajectories.

Theorem 13 (Sign-switchable pairwise tractability and shallow-class hardness)**.** *For every path-simple, interaction-order-two partition,*

$$\max_{z\in[0,1]^{\Pi}} p_{\Pi}(z) = \max_{z\in\{0,1\}^{\Pi}} p_{\Pi}(z), \tag{71}$$

and, after pruning states outside all union-support trajectories, the coefficients in (21) *are obtainable by an exact $O(HS^2A|\Pi|^2)$-arithmetic-operation recursion.*

If variable complements make every pair coefficient nonnegative, one s–t min-cut gives the optimum. Such complements exist exactly when every cycle in the signed interaction graph has positive edge-sign product.

Without that sign promise, deciding whether the coherent value is at least a rational threshold is NP-complete even for $H = 2$, one action, zero running loss, binary terminal loss, pure-box parameters, two successors per uncertain row, and two distinct blocks per path. The reduction embeds normalized Max-Cut and has row-released value one, so deciding whether $W_{\Pi^\top} - W_{\Pi} \leq \eta$ is NP-complete as well.

Multi-affine reduction and coefficient recursion. Expand expected additive loss over rewarded prefixes. Every prefix probability is a product of affine row coordinates. Path simplicity makes each monomial square-free, and interaction order two bounds its degree by two, so the result has form (21). Fixing all but one coordinate makes the polynomial affine in the remaining coordinate; moving that coordinate to an endpoint cannot decrease it. Repeating proves (71).

For constructive coefficient propagation, first prune every stage–state pair that lies on no union-support trajectory; this is a forward/backward graph reachability calculation on the layered support graph. At each remaining pair, store the coefficients indexed by the empty set, all singleton blocks, and all unordered block pairs. Contract a successor table against P_x^0. For a row in block b, also contract against Δ_x and shift each suffix monomial I to $I \cup \{b\}$. The promises ensure $b \notin I$ on structurally supported suffixes and $|I \cup \{b\}| \leq 2$. Averaging under π and proceeding backwards gives $1 + |\Pi| + \binom{|\Pi|}{2}$ coefficients per state. Dense contractions therefore use $O(HS^2A|\Pi|^2)$ arithmetic operations and $O(S|\Pi|^2)$ rolling storage.

One-cut formula. Assume first $c_{ij} \geq 0$. For binary labels,

$$z_i z_j = \tfrac{1}{2}(z_i + z_j - |z_i - z_j|),$$

and put

$$w_{ij} = \tfrac{1}{2}c_{ij}, \qquad d_i = c_i + \tfrac{1}{2}\sum_{j \neq i} c_{ij}, \qquad \kappa_i = \min\{0, -d_i\}. \tag{72}$$

so, with (72),

$$p_\Pi(z) = c_0 + \sum_i d_i z_i - \sum_{i<j} w_{ij}|z_i - z_j|. \tag{73}$$

Let $U_i(0) = 0$, $U_i(1) = -d_i$ and $\widetilde{U}_i(\ell) = U_i(\ell) - \kappa_i \geq 0$. Use label convention $z_i = 0$ on the source side and $z_i = 1$ on the sink side. Add an undirected edge of capacity w_{ij}, an arc $s \to i$ of capacity $\widetilde{U}_i(1)$, and an arc $i \to t$ of capacity $\widetilde{U}_i(0)$. The cut for z has capacity

$$\sum_{i<j} w_{ij}|z_i - z_j| + \sum_i \widetilde{U}_i(z_i) = -p_\Pi(z) + c_0 - \sum_i \kappa_i,$$

so if C^* is the minimum cut,

$$\max p_\Pi = c_0 - C^* - \sum_i \kappa_i. \tag{74}$$

For sign switching, choose $a_i \in \{-1, 1\}$ and write $z_i = b_i + a_i y_i$ with $b_i = (1 - a_i)/2$. The transformed coefficients are

$$\begin{aligned} c'_{ij} &= a_i a_j c_{ij}, \\ c'_i &= a_i \left(c_i + \sum_{j \neq i} c_{ij} b_j \right), \qquad c'_0 = c_0 + \sum_i c_i b_i + \sum_{i<j} c_{ij} b_i b_j. \end{aligned} \tag{75}$$

Thus all pair terms are positive when $a_i a_j = \operatorname{sign}(c_{ij})$. Multiplying these equalities around a cycle proves necessity of positive edge-sign product. Conversely, propagate signs from one root per connected component; the cycle condition makes every non-tree edge consistent.

Depth-two Max-Cut construction. Take a simple undirected graph $G = (V, E)$ with $m = |E| \geq 1$ and both orientations $\vec{E}$. There is one parameter block B_v with $z_v \in [0, 1]$ for each nonisolated vertex. For every $(u, v) \in \vec{E}$, create an initial state s_{uv} of mass $1/(2m)$ and private stage-one states y^0_{uv}, y^1_{uv}. The unique stage-zero row, in block B_u, is

$$P(y^1_{uv} \mid s_{uv}) = z_u, \qquad P(y^0_{uv} \mid s_{uv}) = 1 - z_u. \tag{76}$$

At stage one, both rows are in block B_v and transition to common terminals Good and Bad according to

$$\begin{aligned} P(\text{Bad} \mid y^0_{uv}) &= z_v, & P(\text{Good} \mid y^0_{uv}) &= 1 - z_v, \\ P(\text{Bad} \mid y^1_{uv}) &= 1 - z_v, & P(\text{Good} \mid y^1_{uv}) &= z_v. \end{aligned} \tag{77}$$

Only Bad has terminal loss one. Conditional on s_{uv}, its probability is $(1 - z_u)z_v + z_u(1 - z_v)$. Averaging both orientations gives

$$p_G(z) = \frac{1}{|E|} \sum_{\{u,v\} \in E} \big[z_u(1 - z_v) + (1 - z_u)z_v \big] = \frac{\mathrm{Cut}_G(z)}{|E|} \quad (z \in \{0, 1\}^V). \tag{78}$$

Each path visits the distinct blocks B_u, B_v once, and every uncertain row has two successors.

At binary z, Equation (78) is the normalized cut size; multi-affinity gives the same box optimum. The construction has $6m + 2$ states and rational encoding size $O(|V| + |E|)$, so Max-Cut threshold reduces in polynomial time. Membership in NP because an optimal binary vector is a polynomial witness whose two-step value is evaluated exactly. Under full row release, each initial row selects y^0_{uv} and that row independently selects Bad, so the value is one. The exact gap is therefore $1 - \mathrm{MaxCut}(G)/m$. Thus $\mathrm{MaxCut}(G) \geq K$ if and only if the gap is at most $1 - K/m$; the same binary vector is a polynomial witness for this upper-gap decision problem.

D How to Compute a Robust Policy

This section is a self-contained guide to how the fixed-policy audit can be used inside a robust-control workflow. It summarizes established solution routes and is not a new algorithmic contribution of this paper.

D.1 The policy-optimization problem

Let $\mathfrak{P}_{\mathrm{adm}}$ be a specified class of admissible policies. With our loss convention, a robust policy for the committed ambiguity set solves

$$\pi^\star_{\mathcal{C}} \in \operatorname*{arg\,min}_{\pi\in\mathfrak{P}_{\mathrm{adm}}} \max_{P\in\mathcal{C}} J(\pi, P), \qquad J(\pi, P) := \mathbb{E}_{d_0,\pi,P}[g(\tau)]. \tag{79}$$

This problem contains two distinct tasks: evaluating the worst-case loss of a given policy and improving the policy. The main paper addresses the first task for a fixed π and audits how its value changes when dependencies in $\mathcal{C}$ are released or restored. It does not optimize over $\mathfrak{P}_{\mathrm{adm}}$.

The answer to Equation (79) depends critically on the policy class and on when nature chooses the model. The discussion below uses the paper's finite-horizon loss, assumes a fully observed MDP, and first restricts attention to Markov policies. If a controller may use the history to learn an unobserved persistent model, a sufficient statistic for that history must be added to the state; the resulting partially observed or adaptive robust control problem is different from the fixed-Markov-policy setting studied here.

D.2 Rectangular ambiguity: robust Bellman recursion

For the rectangular set $\mathcal{R} = \prod_{h,s,a} \mathcal{U}_{hsa}$, nature's choice at one row does not restrict its choices at other rows. Writing $r_{hsa}(s') = r_h(s, a, s')$, backward induction gives

$$\begin{aligned} V^\star_H(s) &= r_H(s), \\ V^\star_h(s) &= \min_{a\in\mathcal{A}_h} \max_{q\in\mathcal{U}_{hsa}} q^\top \big(r_{hsa} + V^\star_{h+1}\big), \qquad h = H-1, \dots, 0. \end{aligned} \tag{80}$$

Any minimizing action in Equation (80) defines a robust Markov policy. Because the adversary observes the realized action in our timing convention, randomizing over actions cannot improve on the best action at a state. The rowwise maximizations are independent, which is the source of tractability. Robust value iteration, policy iteration, and duality-based variants for rectangular uncertainty are developed by Iyengar (2005) and Nilim & El Ghaoui (2005); state-rectangular formulations and their policy structure are treated by Wiesemann et al. (2013).

Equation (80) also gives a natural baseline for this paper: compute a policy under the fully released set $\mathcal{R}$, audit that policy under $\mathcal{C}$, restore the dependencies selected by global balance, and then re-optimize under the resulting ambiguity model if that model admits a suitable solver.

D.3 Committed or coupled ambiguity

For a general coupled set $\mathcal{C}$, the inner maximization cannot be separated across rows, so Equation (80) is not valid. Several established routes remain, each requiring additional structure.

Finite policy menu. If $\mathfrak{P}_{\mathrm{adm}} = \{\pi^1, \dots, \pi^L\}$ is a finite menu, evaluate $W_{\mathcal{C}}((d_0, \pi^\ell), g)$ for every candidate and choose

$$\pi^\star_{\mathrm{menu}} \in \operatorname*{arg\,min}_{1\le\ell\le L} W_{\mathcal{C}}((d_0, \pi^\ell), g). \tag{81}$$

This is the policy-selection operation used in our experiments. It is exact for the declared menu but makes no claim of global optimality outside it.

Finite model or endpoint scenarios. If $\mathcal{C} = \{P^1, \dots, P^M\}$ is finite, the inner maximum is obtained by evaluating all scenarios. Joint optimization can be written in epigraph form as

$$\min_{\pi\in\mathfrak{P}_{\mathrm{adm}},\, t} \; t \quad \text{subject to} \quad J(\pi, P^m) \le t, \qquad m = 1, \dots, M. \tag{82}$$

For deterministic Markov policies this is a finite combinatorial problem; for randomized Markov policies, $J(\pi, P^m)$ is generally multilinear across stages. Enumeration, mixed-integer formulations, or global branch-and-bound are therefore practical only for modest state, action, horizon, or scenario counts. A binary model with F independent persistent factors can already have 2^F complete scenarios.

Structured coupled uncertainty. Some non-row-rectangular models retain enough structure for specialized algorithms. Examples include k-rectangular uncertainty (Mannor et al., 2016) and factor-matrix models with rectangular uncertainty over the latent factors (Goyal & Grand-Clément, 2023). These methods can optimize robust policies within their respective assumptions, but they do not solve arbitrary cross-row coupling. Policy-gradient approaches provide another route for certain parameterized nonrectangular problems (Li et al., 2026; Wang et al., 2025); their guarantees and objectives depend on the specific uncertainty representation and policy parameterization.

Continuous affine factors. For a compact affine family $P(\theta)$ with $\theta \in \Theta \subseteq [0,1]^F$, the committed objective becomes

$$\min_{\pi \in \Pi_{\mathrm{adm}}} \max_{\theta \in \Theta} J(\pi, P(\theta)). \tag{83}$$

For a fixed finite-horizon Markov policy, the inner objective is a polynomial in θ; when policy probabilities are variables, it is also polynomial in those probabilities. Small instances can be handled by deterministic global optimization, and moment/sum-of-squares hierarchies provide convergent bounds for compact polynomial formulations (Lasserre, 2001). These tools are general but do not yield the Bellman-scale complexity of rectangular robust dynamic programming.

D.4 Where the present audit fits

A practical workflow is therefore:

1. declare the coherent transition family, dependence map, policy class, and adversary timing;
2. compute a rectangular robust policy by Equation (80), or construct a policy menu by any domain-appropriate method;
3. run the fixed-policy audit to measure which released dependencies affect each candidate's worst-case value;
4. restore the selected dependencies and solve the resulting structured, finite-scenario, or menu problem using one of the routes above; and
5. validate the selected policy under the committed model family and under plausible misspecifications of the dependence map.

Our results provide Step 3 and the dependency choice in Step 4. They can reduce the size of the coupled model that a downstream solver must handle, but they do not establish convergence or global optimality for that downstream policy-optimization problem.

Remark 14 (Residual pessimism bounds menu regret)**.** *Fix one endpoint array and a finite policy menu $\mathcal{P}_{\mathrm{opt}}$. Let $W_{\Pi(K)}(\pi)$ and $W_{\mathcal{C}}(\pi)$ be the worst-case losses under the selected and the fully coherent ambiguity sets, and let $\widehat{\pi}_K$ minimize $W_{\Pi(K)}$ over the menu. Because $\mathcal{C} \subseteq \mathcal{C}_{\Pi(K)}$, every menu policy π satisfies $W_{\mathcal{C}}(\widehat{\pi}_K) \le W_{\Pi(K)}(\widehat{\pi}_K) \le W_{\Pi(K)}(\pi)$, so*

$$W_{\mathcal{C}}(\widehat{\pi}_K) - \min_{\pi \in \mathcal{P}_{\mathrm{opt}}} W_{\mathcal{C}}(\pi) \le \max_{\pi \in \mathcal{P}_{\mathrm{opt}}} \big[W_{\Pi(K)}(\pi) - W_{\mathcal{C}}(\pi)\big] \le \max_{\pi \in \mathcal{P}_{\mathrm{opt}}} \big[G_\pi(\mathcal{B}) - \mathcal{L}_\pi(K)\big], \tag{84}$$

where $G_\pi(\mathcal{B}) = W_{\Pi^\top}(\pi) - W_{\mathcal{C}}(\pi)$ is the full release gap of π and the last step is Theorem 4. The regret of the deployed policy under the coherent model is therefore at most the largest pessimism that the selected blocks leave on the menu, which the certificate bounds policy by policy.

E Declared-Cell Row-Adversary Semantics

Committed nature selects one $P \in \mathcal{C}$ before the rollout and applies all of its rows coherently. At stage h, the controller observes its history and samples a_h; row-adaptive nature then observes exactly $x = (h, s_h, a_h)$ and selects from $\mathcal{U}_x := \mathrm{proj}_x \mathcal{C}$. Nature may correlate its choices through a private

seed but has no additional history or private state. Any richer information must be represented by augmenting the state before $\mathcal{C}$ and $\mathcal{R}$ are formed. Controller and nature seeds are independent, and the fixed Markov probe uses conditionally independent action draws $\pi_h(\cdot \mid s_h)$.

Lemma 15 (x-measurable row play has a pure contingent normal form)**.** *For every fixed probe e and bounded path loss $g : \mathcal{T} \to \mathbb{R}$,*

$$\sup_{x\text{-measurable randomized } \sigma} \mathbb{E}_{e,\sigma}[g(\tau)] = \max_{Q \in \mathcal{R}} \langle g, \mu_e(Q) \rangle, \tag{85}$$

where σ ranges over admissible randomized row-adaptive strategies. Shared or independent randomization therefore cannot improve on one pure contingent plan $Q = (q_x)_x \in \mathcal{R}$. The same statement holds for committed nature randomizing over $\mathcal{C}$.

The proof removes the timing ambiguity by conditioning on nature's private randomization, identifying the resulting pure contingent array, and then averaging back over that randomization.

Proof of Lemma 15. Let U collect all private randomization available to an admissible nature. Because its row choice can depend only on U and the declared information state x, fixing $U = u$ determines a row $q_x(u) \in \mathcal{U}_x$ at every cell. Hence it determines a pure contingent array

$$Q(u) = (q_x(u))_{x \in \mathcal{X}} \in \prod_{x \in \mathcal{X}} \mathcal{U}_x = \mathcal{R}.$$

Fresh behavioral randomization can be pre-sampled at all finitely many cells; shared randomization is already part of U. Conditional on $U = u$, the controller acts according to the fixed probe and nature uses $Q(u)$, so the conditional trajectory law is exactly $\mu_e(Q(u))$. Therefore

$$\mathbb{E}_{e,\sigma}[g(\tau)] = \mathbb{E}_U[\langle g, \mu_e(Q(U)) \rangle] \le \max_{Q \in \mathcal{R}} \langle g, \mu_e(Q) \rangle.$$

The maximum exists because $\mathcal{R}$ is compact and $Q \mapsto \mu_e(Q)$ is continuous. Conversely, any pure $Q \in \mathcal{R}$ is implemented by a degenerate seed, proving equality. Replacing $\mathcal{R}$ by $\mathcal{C}$ gives the committed statement. The argument would fail for a nature whose choice can depend on a history not encoded in x; that history must instead create refined cells. □

Counterfactual-row non-identifiability. Take $H = 2$, one action, stage-one states A, B, terminals G, D, and terminal loss one only at D. Let P^A send the root to A and both stage-one rows to G, and let P^B send the root to B and both rows to G. Let P^{A*} agree with P^A except that its unreachable B row goes to D. For $\mathcal{C}_0 = \{P^A, P^B\}$ and $\mathcal{C}_1 = \mathcal{C}_0 \cup \{P^{A*}\}$,

$$\mu_e(\mathcal{C}_0) = \mu_e(\mathcal{C}_1), \qquad W_{\mathcal{C}_0}(e, g) = W_{\mathcal{C}_1}(e, g) = W_{\mathrm{rect}(\mathcal{C}_0)}(e, g) = 0,$$

but $W_{\mathrm{rect}(\mathcal{C}_1)}(e, g) = 1$: paste the root-to-$B$ row from P^B with the B-to-D row from P^{A*}. Thus $\mathcal{C} \mapsto \mathrm{rect}(\mathcal{C})$ does not descend to the committed path-law quotient. The complete counterfactual rows and factor map are declared structural input; passive committed rollouts alone need not identify the release gap. The identifiable endpoint streams and resets in Proposition 8 supply stronger information.

Table 2: Closest classical ingredients and the incremental claims used here.

Component	Classical ingredient	Role or increment here
Row release	Rectangularity and parameter lifting (Iyengar, 2005; Schnitzer et al., 2026; Quatmann et al., 2016; Heck et al., 2025)	Exact loss audit and selective restoration after a declared release.
Reachable defect	Model/transition performance difference (Kearns & Singh, 2002; Modhe et al., 2021; Satheesh et al., 2026)	Foundational Bellman-optimal specialization with nonnegative terms and a reachable-row iff.
Global balance	Two-scenario max–min selection (Yu, 1996; Kawase & Sumita, 2019)	Defect-derived masses, joint dependency selection, and the true-gap overlap sandwich.

F Experimental Protocol and Reproducibility

This section specifies the estimands, generators, decision rules and numerical checks. The independent unit is a generated system, or a paired trajectory dataset for the fixed CliffWalking task. Every interval in the main study resamples systems within generator family using 10,000 percentile-bootstrap draws unless stated otherwise; method contrasts are computed within system before resampling. All configurations, policy menus, thresholds, and random streams were fixed before the corresponding scientific studies were opened.

F.1 Binary evaluation and computational accounting

The routing, queue, inventory, crosswind, and CliffWalking performance studies evaluate binary persistent regimes. Fix a probe and retained set K. For $b \in \{0,1\}$, start from $V_H^{b,K} = r_H$ and define $v_{hsa}^{b,K} = r_{hsa} + V_{h+1}^{b,K}$. The exact recursion is

$$V_h^{b,K}(s) = \sum_a \pi_h(a \mid s) \begin{cases} (P_{hsa}^b)^\top v_{hsa}^{b,K}, & (h,s,a) \in R_K, \\ \max_{e\in\{0,1\}} (P_{hsa}^e)^\top v_{hsa}^{b,K}, & (h,s,a) \notin R_K, \end{cases} \tag{86}$$

$$W_{\Pi(K)}^{\mathrm{bin}} = \max_{b\in\{0,1\}} d_0^\top V_0^{b,K}. \tag{87}$$

For fixed b, the remaining row choices are independent, so backward induction maximizes over them exactly. Maximizing over b covers the two admissible retained assignments. Fixed rows have identical endpoints. With F independently retained binary factors the same evaluator enumerates 2^F assignments, at dense arithmetic cost $O(2^F H S_{\max}^2 A_{\max})$ per policy and subset. For the one-factor studies this is two passes at every nonzero budget, including full recoupling.

The six-group study enumerates 64 subsets and the ten-group study selects over the 1,024 possible masks. Both studies maximize the mean unnormalized certificate ($G_j = 1$), on nine and 64 selector probes, respectively. Each certificate computation uses one released backward pass and one prefix forward pass per probe; Residual-AUC evaluation additionally uses the full recoupling value. Candidate deployment and holdout evaluation are separate calls on the 256-policy menu. The reported scaling timings measure selection from already constructed score arrays.

The stochastic-system overlap audit checks $\mathcal{L}(K) \le G^{\mathrm{bin}}(K) \le \chi^{H-1}\mathcal{L}(K)$. The continuous sharpness calculation is a separate five-fixture exact-rational check of Equation (20). Value recursions use float64; the finite-data validation threshold is 2×10^{-9} and certificate positivity requires 10^{-8}. The maximum observed binary-bound excess is 5.998×10^{-16}; the five sharpness fixtures satisfy their rational identities exactly. The MIP tolerance and independent numerical checks are specified below.

Continuous factor check. On the 24 certificate-study systems, every held-out policy and retained subset was re-evaluated with one shared factor value $\theta \in [0,1]$ on the retained rows and the worse endpoint at every released row. For an interval $I \subseteq [0,1]$, letting each retained row choose its own value in I gives a released Bellman pass in which retained rows choose between the two interval ends; its value upper-bounds the coherent value on I. Starting from 64 equal intervals, branch and bound uses evaluated endpoint and midpoint values as lower bounds. It retains the upper bounds of pruned intervals when forming the final enclosure $[W^-, W^+]$. Each row stops when $W^+ - W^- \le \max\{10^{-9}, 10^{-4} G^{\mathrm{bin}}(K)\}$. The float64 calculation adds 2×10^{-12} to interval upper values for numerical roundoff and uses at most eight further bisections. All 9,216 rows meet the stopping tolerance; the largest enclosure width is below 6.54×10^{-5}. The induced gain interval is $[W_{\Pi^\top} - W^+, W_{\Pi^\top} - W^-]$, intersected with the nonnegative axis. These intervals are consistent with $\mathcal{L}(K) \le G(K) \le G^{\mathrm{bin}}(K)$. An endpoint value is within the stopping tolerance of W^+ in 9,076 rows (98.5%). Among rows with positive binary gain, the enclosure bounds $(G^{\mathrm{bin}}(K) - G(K))/G^{\mathrm{bin}}(K)$ by 3.36%. Over positive-certificate rows, both ends of the ratio intervals give median 0.312 and quartiles 0.162 and 0.447 at the reported precision.

F.2 Finite-data generator and decision rule

The main study contains twelve layered-routing and twelve queue-control systems. The factor map and the two-successor support of every uncertain row are declared inputs; the experiment

estimates only the endpoint probabilities. This separates uncertainty about transition probabilities from uncertainty about the sharing semantics. The latter is varied explicitly in the misspecification audit rather than fitted from the same trajectories.

The routing generator has six layers, three transient lanes per layer, three actions, two absorbing outcomes, and horizon six. Write a transient state as (h, ℓ), let $a \in \{0, 1, 2\}$, and let $b_{\mathrm{sys}} \in \{0, 1\}$ be the system's fixed orientation bit. Before the last layer, its two successors are

$$(h + 1, a \bmod 3), \qquad \big(h + 1, (a + 1 + \mathbf{1}\{h + \ell + a \text{ odd}\}) \bmod 3\big); \tag{88}$$

at the last layer they are success and failure. Low and high right-successor probabilities are drawn independently from $[0.18, 0.34]$ and $[0.66, 0.82]$ and their endpoint orientation is the parity of $h + 2\ell + a + b_{\mathrm{sys}}$. The declared routing-group map is $f_{\mathrm{route}}(h, \ell) = (2\lfloor h/2 \rfloor + \ell) \bmod 6$. Step and lane coefficients are uniform on $[0.01, 0.04]$ and $[0.03, 0.13]$; terminal failure and success coefficients are uniform on $[0.65, 1]$ and $[0, 0.08]$. The stage loss is the step coefficient plus half the action-to-target lane distance, the successor-to-next-target lane distance, and the applicable terminal cost. The layer targets are independent uniform draws from the three lane labels; the initial law is uniform over the three first-layer lanes, and all 18 transient lane states form the routing behavior-state registry.

The queue generator has thirteen levels, three actions, and horizon eight. Action a at level s moves to either $\max(0, s - a - 1)$ or $\min(12, s + 1)$. Its low and high upward-move probabilities are uniform on $[0.20, 0.38]$ and $[0.62, 0.80]$, with orientation given by the parity of $s + a + b_{\mathrm{sys}}$. The queue-group map is $f_{\mathrm{queue}}(s) = \lfloor 6s/13 \rfloor$. Holding, service, and overflow coefficients are uniform on $[0.025, 0.065]$, $[0.005, 0.035]$, and $[0.55, 0.95]$. The stage loss is holding times the normalized successor level, plus squared service effort and the overflow indicator; the terminal loss adds a quartic normalized-level penalty. The queue initial law is uniform on levels 4–8, and all thirteen levels form its behavior-state registry.

Each system supplies nine fitting policies, six disjoint evaluation policies, and all 64 retained subsets. For endpoint mixture w, form $P_w = (1 - w)P^0 + wP^1$ by applying the same $w \in [0, 1]$ to every uncertain row. Let $Q_h^w(s, a)$ be the resulting stage-h state–action loss-to-go. For temperature $\tau > 0$, set

$$\pi_h^{w,\tau}(a \mid s) \propto \exp\left\{ -\frac{Q_h^w(s, a) - \min_{a' \in \mathcal{A}_h} Q_h^w(s, a')}{\tau} \right\}. \tag{89}$$

Fitting policies use weights $0, 1/2, 1$ and temperatures $0.08, 0.2, 0.5$; evaluation policies use weights $0.2, 0.8$ and temperatures $0.12, 0.3, 0.7$. The two acquisition budgets are 8,192 and 131,072 episodes per endpoint. The missing-coverage condition retains the larger budget but never samples action 2 in groups 1–4. A 90% time-uniform Hoeffding union sequence is constructed over every registered binary row stream, with $[0, 1]$ assigned to zero-count rows. The numerical, certificate-positivity, and normalized-residual thresholds are respectively 2×10^{-9}, 10^{-8}, and 0.1. For each endpoint, trajectories start uniformly over behavior states. Before action selection at each of eight steps, the current state is redrawn uniformly if an independent $\mathrm{Bernoulli}(0.35)$ reset fires or if the two action-0 support indices coincide; otherwise it retains its initialized state on the first step and the previous successor thereafter. Actions are uniform with probability 0.75 and otherwise equal $(s + e) \bmod 3$. In the missing-coverage condition, action 2 is replaced by action 0 whenever the current state's group is in $\{1, 2, 3, 4\}$.

System generation, acquisition and bootstrap resampling use separate random streams. Matched comparisons reuse the same generated system and acquired observations.

Let $\mathcal{P}_+$ denote the evaluation policies whose full release gap exceeds 10^{-10}, and let $|\mathcal{B}|$ be the number of candidate groups. For selector m, let $K_{m,k}$ be its retained set of size k, let $W_\varnothing(\pi)$ and $W_{\mathrm{all}}(\pi)$ denote the released and fully recoupled worst-case values of policy π. The reported normalized-residual curve and its trapezoidal area under the curve (AUC) are

$$\begin{aligned} R_m(k) &= \frac{1}{|\mathcal{P}_+|} \sum_{\pi \in \mathcal{P}_+} \frac{W_{\Pi(K_{m,k})}(\pi) - W_{\mathrm{all}}(\pi)}{W_\varnothing(\pi) - W_{\mathrm{all}}(\pi)}, \\ \mathrm{AUC}_m &= \frac{1}{|\mathcal{B}|} \sum_{k=0}^{|\mathcal{B}|-1} \frac{R_m(k) + R_m(k+1)}{2}. \end{aligned} \tag{90}$$

Thus $R_m(0) = 1$, $R_m(|\mathcal{B}|) = 0$, and lower AUC means that the selector removes the release gap earlier.

Table 3: Held-out binary-family normalized-residual AUC at 131,072 episodes per endpoint. Selectors use only the nine-policy fitting menu; intervals bootstrap the 24 systems. Lower is better.

Method	Mean AUC	95% interval
Truth exhaustive best-k	0.3005	$[0.2811, 0.3201]$
Sequential marginal	0.3013	$[0.2808, 0.3217]$
Singleton marginal	0.3023	$[0.2817, 0.3230]$
Exact defect	0.3117	$[0.2920, 0.3308]$
Confidence global balance	0.3319	$[0.3115, 0.3527]$
Global balance, greedy	0.3644	$[0.3281, 0.4019]$
Global balance, exact	0.3655	$[0.3295, 0.4045]$
Matching prefix	0.3685	$[0.3312, 0.4072]$
One-pair prefix	0.3880	$[0.3452, 0.4312]$
Occupancy	0.4383	$[0.4155, 0.4620]$
Random	0.4806	$[0.4645, 0.4970]$

Selectors and deterministic ties. For each of the nine fitting probes $j \in \mathcal{J}_+$, write π_j for its policy and set $G_j = 1$. For a selected subset $K \subseteq \mathcal{B}$, let

$$F(K) = \frac{1}{|\mathcal{J}_+|} \sum_{j \in \mathcal{J}_+} \frac{\min\{\sum_{B \in K} A_{jB}, \sum_{B \in K} C_{jB}\}}{G_j}. \tag{91}$$

In this six-group study, exact global balance maximizes $F(K)$ at each cardinality by enumeration. The ten-group decision study uses $G_j = 1$; the scaling study uses the MIP in Equation (17). Let K_t be the set after t greedy additions, starting with $K_0 = \varnothing$. The next group is chosen from $B \in \mathcal{B} \setminus K_t$ to maximize $F(K_t \cup \{B\}) - F(K_t)$, breaking ties by the smaller group index. With maintained sign totals this costs $O(k|\mathcal{B}||\mathcal{J}_+|)$ arithmetic; F is not generally submodular, and no approximation ratio is claimed. Matching and one-pair rank the mean of their per-probe group certificates. Exact defect first chooses a worst fully coherent binary regime $\widehat{P}_j^\star$ for each fitting probe, then ranks the group scores $\frac{1}{|\mathcal{J}_+|} \sum_j \sum_{x \in B} d_x^{\widehat{P}_j^\star} \delta_x(\widehat{P}_{j,x}^\star)$. Occupancy ranks the same coherent occupancies without the defect factor. Singleton marginal instead ranks $\frac{1}{|\mathcal{J}_+|} \sum_j [\widehat{W}_{\varnothing,j} - \widehat{W}_{\{B\},j}]$; sequential marginal recomputes the fitted value reduction after each addition. Confidence Global Balance maximizes $\frac{1}{|\mathcal{J}_+|} \sum_j \underline{L}_j(K)$, using the same $G_j = 1$. All these fitted scores use the estimated endpoint rows. Score rankings break ties by group index, exhaustive optimizers by the smallest integer mask, and deployed-policy ties by the first frozen candidate index.

The primary paired exact-global-balance-minus-matching AUC contrast is -0.0030 with interval $[-0.0145, 0.0091]$: it favors global balance on nine systems, ties numerically on twelve, and favors matching on three. The routing and queue contrasts are -0.0067 $[-0.0280, 0.0167]$ and 0.0008 $[-0.0034, 0.0070]$. Exhaustive, marginal, and exact-defect rows use additional coherent or subset-value evaluations beyond the prefix-based scores.

All selectors receive the same estimated endpoint rows and nine fitting probes. Prefix scores use one released backward pass and one prefix forward pass per probe. Exact defect and Occupancy additionally choose a worst coherent binary regime and propagate its occupancy. Singleton and sequential marginal use seven and 22 subset-value evaluations per probe, respectively. These are algorithmic call counts, separate from acquisition, subset selection and held-out evaluation. Across all $24 \times 15 \times 64 = 23{,}040$ policy–subset rows, the binary overlap bounds hold to 5.998×10^{-16}. All 1,344 exact-cardinality selections match the two-scenario optimization authority. Global balance strictly exceeds the sum of within-group certificates on 10,887 rows (47.3%). Among 22,680 positive-mass rows, $\mathcal{L}(K)/G^{\mathrm{bin}}(K)$ has median 0.3246 and IQR $[0.1710, 0.4573]$ over the positive-mass policy–subset instances.

For Table 4 and Figure 5, point-estimate Global Balance selects one size-three subset using the nine fitting probes in each system and acquisition condition. Both tests evaluate that same subset on the six held-out policies. Simulator truth lies in all 69,120 propagated policy–subset intervals. There is no false positive among 15,595 direct and 57,593 balance decisions. Removing action coverage raises direct abstention by 0.7431 with paired interval $[0.6181, 0.8542]$ under the unchanged threshold. The balance test passes 144/144 policies in every row of Table 4 because it pools lower sign mass. The observed rows retain sufficient certified positive and negative mass even under the coverage

Table 4: Finite-data decisions for the six held-out policies in each system at $k = 3$. A pass requires a strictly positive certified gain.

Condition	Direct pass	Direct abstain	Balance pass	Balance abstain
Covered, 8,192	0/144	144/144	144/144	0/144
Covered, 131,072	110/144	34/144	144/144	0/144
Missing coverage, 131,072	3/144	141/144	144/144	0/144

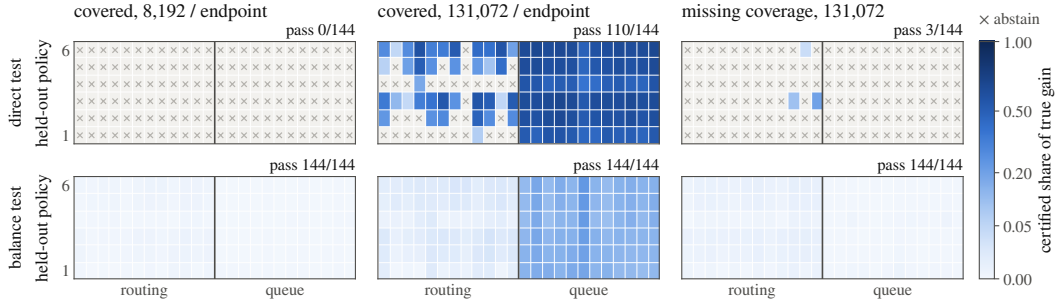


Figure 5: Finite-data decisions at $k = 3$ for the six held-out policies (rows) of each of the 24 systems (columns; routing left, queue right) under the three acquisition conditions. Each cell shows the certified lower bound as a share of the true recoupling gain on a square-root colour scale; crosses mark abstentions. The direct test propagates row intervals through the value recursion and passes $0/144$, $110/144$, and $3/144$ cases; the balance test lower-bounds pooled sign mass and passes all 144 cases in every condition. No certificate exceeds the true gain.

intervention. A balance decision abstains when its pooled lower certificate fails the positivity threshold; a direct decision abstains when its value enclosure fails that threshold. These counts describe the acquired datasets at the common 90% time-uniform confidence level. Figure 5 displays every decision behind Table 4: the direct test becomes more decisive with narrower row intervals and broader coverage. At 131,072 episodes it certifies a median 53.3% of the true gain on the cases it passes (quartiles 40.4% and 60.8%), at least a quarter of it in 98 of 144 cases, and at least half in 67. The balance test certifies a strictly positive gain in every case and at least a tenth of the true gain in 72 of 144 cases at that budget.

F.3 Downstream candidate-menu consequence

This independently generated study uses twelve routing and twelve queue systems with $H = 8, 16$ states, three actions, ten groups, and four uncertain rows per group. A uniform random permutation of the $H \cdot 16 \cdot 3$ rows assigns the first 40 to group labels repeated four times. For row jitter $j \in \{0, 1, 2, 3\}$, the two distinguished successors are

$$\ell = (s + a + h + j) \bmod 16 \quad \text{(routing)}, \qquad \ell = \mathrm{clip}_{[0,15]}(s + a - 1 + j) \quad \text{(queue)},$$

and $r = (\ell + 9 + (j \bmod 4)) \bmod 16$. Each endpoint puts base mass $0.10/16$ on every successor and adds $(0.78, 0.12)$ or $(0.12, 0.78)$ to (ℓ, r); fixed rows add $(0.45, 0.45)$ at those successors. Odd groups swap the endpoint labels, producing the alternating orientation. The cost at successor q is $\mathrm{clip}_{[0.08,0.82]}(0.22 + 0.08h/7 + 0.04a + 0.18\sin(0.71(q + s + h + 1)))$, except that the two distinguished costs are $0.04 + 0.01a$ and $0.94 - 0.01a$. Initial logits decrease linearly from one to minus one and receive independent $0.08\,\mathrm{Unif}[0, 1]$ jitter before softmax. Here $\mathrm{clip}_I(v)$ truncates v to the displayed interval I, and $\mathrm{Unif}[0, 1]$ is the uniform law on $[0, 1]$.

The 64 selector probes and disjoint 256-policy optimization menu use logits

$$0.9\zeta_{hsa} + 0.18\cos((h + 1)(a + 1)) - 0.12\sin((s + 1)(a + 1)), \qquad \zeta_{hsa} \overset{\text{iid}}{\sim} N(0, 1),$$

followed by softmax; $N(0, 1)$ denotes the standard normal law. The optimization menu's first three policies replace these logits by $+5$ on one action and -5 on the others; the remaining 253 are independent soft policies. Nested samples contain 4,096 and 32,768 transitions per endpoint row and use a 0.5 pseudocount. For each holdout index $\omega \in \{1, \ldots, 16\}$, endpoint label $b \in \{0, 1\}$, and row $x \in \mathcal{X}$, set

$$P^b_{\omega,x} = 0.92P^b_x + 0.08Z^b_{\omega,x}, \tag{92}$$

Table 5: Held-out candidate-menu regret per stage under the declared one-factor binary semantics. Fitted gain uses complete subset-value comparisons. Global Balance and Matching use prefix certificates; Occupancy uses released row occupancy.

k	Global balance	Matching	Occupancy	Fitted gain
2	0.00413	0.00432	0.00451	0.00317
4	0.00349	0.00325	0.00443	0.00316
5	0.00323	0.00341	0.00318	0.00355
6	0.00258	0.00221	0.00399	0.00241
8	0.00194	0.00142	0.00236	0.00145

where $Z^b_{\omega,x}$ is obtained by normalizing a vector of i.i.d. $\mathrm{Unif}[0,1]$ entries to sum to one; the two copies of every fixed row remain equal. Let $\mathcal{P}_{\mathrm{opt}}$ be the frozen 256-policy optimization menu, and let $\widehat{W}_{\Pi(K)}(\pi)$ denote its estimated worst-case value when the groups in K are recoupled according to partition $\Pi(K)$. For selector m and retained budget k, deployment chooses

$$\widehat{\pi}_{m,k} \in \operatorname*{arg\,min}_{\pi\in\mathcal{P}_{\mathrm{opt}}} \widehat{W}_{\Pi(K_{m,k})}(\pi). \tag{93}$$

System generation, acquisition, holdout perturbations and bootstrap resampling use separate random streams. For holdout environment ω and declared factor map f, let $W_{\omega,f}(\pi)$ be the true worst-case value of policy π under the holdout endpoint rows and sharing map f. The reported candidate-menu regret per stage is

$$\mathrm{Reg}_f(m,k) = \frac{1}{16H}\sum_{\omega=1}^{16}\left[W_{\omega,f}(\widehat{\pi}_{m,k}) - \min_{\pi\in\mathcal{P}_{\mathrm{opt}}} W_{\omega,f}(\pi)\right]. \tag{94}$$

The correct condition uses one factor for all groups. The semantic intervention uses separate even- and odd-orientation factors while leaving every local row set unchanged. Full recoupling changes regret by -0.00396 $[-0.00541, -0.00266]$ in the declared condition and by 0.00838 $[0.00678, 0.00996]$ after the intervention. This matched reversal is the controlled sensitivity audit for the supplied factor semantics. At $k = 8$, the release-minus-global-balance regret contrast is 0.00229 $[0.00063, 0.00397]$. At the secondary budget of 4,096 transitions per endpoint row, full recoupling minus release is -0.00395 $[-0.00540, -0.00264]$ under the correct map and 0.00864 $[0.00692, 0.01034]$ under the wrong map. For GB at $k = 8$, the corresponding contrasts are -0.00246 $[-0.00417, -0.00086]$ and 0.00846 $[0.00699, 0.00998]$.

In this ten-group study, Occupancy ranks mean row occupancy under the released worst-case model over the 64 fitting probes. The six-group study instead uses the worst coherent regime. The Fitted gain comparator in Table 5 selects the size-k subset maximizing $\frac{1}{64}\sum_j[\widehat{W}_{\varnothing,j} - \widehat{W}_{K,j}]$ by complete subset enumeration, then uses the same 256-policy deployment rule. It is separate from the block-wise exact-defect ranking in the other studies.

At $k = 8$, global balance minus matching is 0.000523 with paired interval $[-0.000193, 0.001372]$, and global balance minus occupancy is -0.000416 $[-0.000883, 0.000041]$. Fitted gain uses those additional subset-value evaluations; its regret minus Global Balance is -0.000490 $[-0.001205, 0.000038]$ at $k = 8$ and -0.000964 $[-0.001905, -0.000170]$ at $k = 2$. Torch and CUDA select identical masks, with maximum discrepancy 1.78×10^{-15}.

F.4 Controlled complementarity and factor-count workload

This additional known-kernel experiment uses the downstream generator in Appendix F.3 with $H = 4$, eight states, three actions, six groups, and four uncertain rows per group. Twelve fresh systems per family give 24 paired systems. The generator's initial endpoint ordering gives the same direction of adverse preference across blocks. The paired mixed-sign condition reverses both endpoint rows at every occurrence of three groups selected uniformly without replacement. Coordinatewise endpoint minima and maxima, row ambiguity sets, losses, initial distributions, and policy menus are identical within each pair.

Nine fitting and six disjoint evaluation policies use independent softmax-logit draws with the same formula as the downstream study; the deterministic-action anchors are omitted. Both menus are

Table 6: Controlled complementarity at $k = 2$ on 24 paired systems. Entries average six held-out policies per system. Higher is better for certificate and true gain. The same-sign control gives 0.0000 in both columns for every method.

Mixed-sign condition	Certificate/stage	True gain/stage
Global Balance	0.0181	0.0223
Matching	0.0070	0.0095
Within-group balance	0.0070	0.0095
Occupancy	0.0109	0.0135
Fitted true-gain selection	0.0167	0.0216

frozen before the orientation intervention. At $k = 2$, GB maximizes mean certificate with $G_j = 1$; Matching, within-group balance, and Occupancy rank their group scores. Ties use the smallest group index. A fitted true-gain comparator maximizes the actual binary gain on the nine fitting policies over all size-two subsets. All methods are evaluated on the six held-out policies. The measured gain is $(W_{\Pi^\top} - W^{\mathrm{bin}}_{\Pi(K)})/H$; the certificate is $\mathcal{L}(K)/H$.

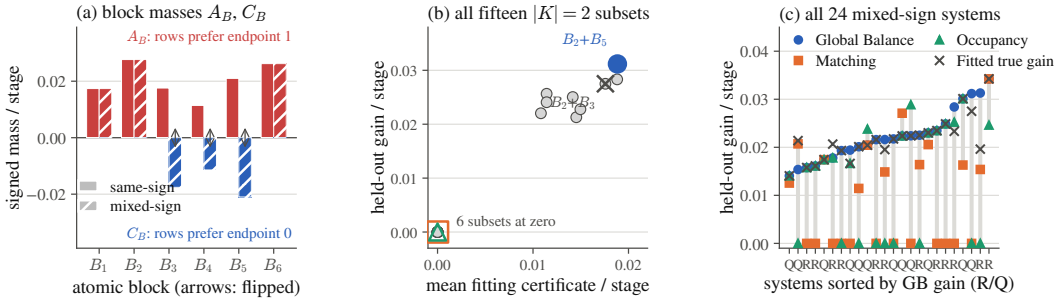


Figure 6: Why Global Balance pools signed mass. (a) Block masses A_B (red, up) and C_B (blue, down) for one queue system, averaged over the nine fitting probes: in the same-sign condition every block prefers endpoint 1 and $\mathcal{L}(K) = 0$ for all K; the mixed-sign condition reverses B_3–B_5. (b) All fifteen size-two subsets, with the mean fitting certificate $\frac{1}{9H}\sum_j \mathcal{L}_j(K)$ on the horizontal axis and mean held-out gain per stage on the vertical axis. The minimum is taken separately for each fitting probe before averaging. Matching and Occupancy choose zero-gain subsets; Global Balance chooses B_2+B_5. (c) Held-out true gain at $k = 2$ on all 24 mixed-sign systems; Global Balance strictly exceeds Matching on 19.

The primary paired GB-minus-Matching true-gain contrast is 0.0128 $[0.0085, 0.0170]$ per stage; GB has strictly larger gain on 19 of 24 systems. The GB-minus-Occupancy contrast is 0.0088 $[0.0042, 0.0134]$. Intervals use 10,000 pooled system-bootstrap draws, with contrasts formed within system. All 18,432 held-out policy–subset certificates hold within 2.36×10^{-16}; a separate matrix-vector evaluator replays all 18,432 binary values with maximum discrepancy 8.88×10^{-16}. Generation, fitting probes, evaluation policies, orientation interventions and resampling use separate random streams.

For a separate evaluator workload check, the final queue fixture assigns one independent binary factor to each of its six groups and reverses two of the four row occurrences within each group. Retaining the first F groups and releasing the remainder requires 2^F conditional assignments. On one fixed evaluation policy, CPU evaluation with one thread takes 0.5299 ms at $F = 2$ and 15.0444 ms at $F = 6$ (one warmup, median of five repetitions). This measures the conditional evaluator on the specified fixture; the block budget and the number of active factors are recorded separately.

F.5 Selective recoupling with several independent factors

This known-kernel study uses the downstream generator of Appendix F.3 with $H = 8$, sixteen states, three actions, twelve atomic blocks of four uncertain rows, and odd blocks with reversed endpoint labels. Twelve fresh systems per family give 24 systems. A uniformly random pairing assigns the twelve blocks to six factor pools of two blocks; each pool has its own episode-persistent binary factor. For a retained block set K, every retained block uses its pool's endpoint, unretained rows are released, and the evaluator of Appendix F.1 enumerates one assignment per touched pool, so evaluating K uses

Table 7: Share of the full recoupling gain on held-out policies with six independent factors, averaged over 24 systems. Pool-level columns retain all blocks of the p best-ranked pools at 2^p passes; block-level columns retain k blocks and show the mean number of conditional passes in parentheses.

	pools p (passes)			blocks k (mean passes)			
Selector	1 (2)	2 (4)	3 (8)	2	4	6	8
Global Balance (exact)	0.469	0.895	0.950	0.469 (2)	0.895 (4)	0.950 (14)	1.000 (37)
Global Balance (greedy)	–	–	–	0.133 (3)	0.428 (8)	0.698 (16)	0.817 (40)
Balanced pairs	–	–	–	0.025 (4)	0.109 (12)	0.334 (28)	0.460 (46)
Matching	0.149	0.312	0.471	0.030 (4)	0.061 (14)	0.150 (32)	0.349 (50)
Occupancy	0.215	0.390	0.471	0.023 (4)	0.067 (12)	0.249 (27)	0.337 (51)
Random order	0.156	0.339	0.506	0.014 (4)	0.098 (12)	0.239 (27)	0.420 (47)
Fitted true gain	0.469	0.898	0.950	0.469 (2)	0.898 (4)	0.950 (14)	1.000 (37)

$2^{\text{pools}(K)}$ conditional Bellman passes and full recoupling uses 64. Sixteen fitting probes, six disjoint held-out evaluation policies, a disjoint frozen menu of 256 candidate policies, and sixteen holdout perturbations of radius 0.08 follow the decision-study construction; holdout worst cases use the true six-factor semantics.

Pool-level selection retains all blocks of p pools. Global Balance ranks pools by the mean over probes of $\min\{\sum_{B\in f} A_{jB}, \sum_{B\in f} C_{jB}\}$; Matching and Occupancy rank pools by their summed block scores; random uses 32 pool orders per system; fitted true gain enumerates all p-pool sets on the fitting probes. Block-level selection retains k blocks: exact Global Balance maximizes the pooled certificate $\frac{1}{16}\sum_j \sum_f \min\{\sum_{B\in K\cap f} A_{jB}, \sum_{B\in K\cap f} C_{jB}\}$ by complete enumeration, greedy Global Balance adds one block at a time, balanced pairs alternate the block with the largest mean positive mass and the block with the largest mean negative mass, Matching and Occupancy rank block scores, random uses 32 block orders, and fitted true gain enumerates all k-subsets on the fitting probes. Every selection is scored on the six held-out policies by its share of the full recoupling gain and, through the 256-policy deployment rule, by held-out candidate-menu regret per stage. Gain shares are computed per held-out policy, averaged over the six policies within a system, then averaged over the 24 systems. The original evaluator check reproduces the one-factor recursion on four subsets per system with all blocks assigned to one pool, and the complete six-factor enumeration on each full subset, within 1.8×10^{-15}.

Table 7 reports every comparator. The paired GB-minus-Matching share of the full gain is 0.320 $[0.209, 0.422]$ at one pool, 0.583 $[0.430, 0.729]$ at two pools, and $+0.834$ $[0.721, 0.939]$ at four blocks; GB-minus-random is 0.556 $[0.465, 0.636]$ at two pools. Held-out regret per stage is 0.00191 under release, 0.00112 for Global Balance at one pool, 0.00037 at two pools, and 0.00019 under full recoupling; the GB-minus-Matching regret contrast at four blocks is -0.00132 $[-0.00244, -0.00042]$ and GB-minus-random is -0.00133 $[-0.00228, -0.00053]$. Because every block's rows prefer the same endpoint, all singleton certificates are zero; exact selection therefore matters, and greedy addition, which must start from a zero singleton, recovers 13.3% of the gain at two blocks against 46.9% for the exact pooled-certificate selector. For the batch of six evaluation policies, the GPU evaluator takes 0.90 ms under release, 2.06 ms with two retained pools, and 5.91 ms under full recoupling. These are means across 24 systems of the median of five synchronized repetitions after one warmup. Timing retains pools in index order and includes assignment construction in the evaluator call. Intervals use 10,000 system-bootstrap draws with contrasts formed within system; generation, pool assignment, probes, menus, holdouts, and resampling use separate random streams.

Eight factors with a structure-derived map. The second benchmark uses the same row targets and losses with $H = 16$, sixteen states, and three actions. Pool $f \in \{0, \dots, 7\}$ is the state region $\{2f, 2f+1\}$ and carries one episode-persistent binary factor; its three atomic blocks are the stage bands $[0, 5]$, $[6, 10]$, and $[11, 15]$. Each block has eight uncertain rows drawn uniformly from its region, band, and action cell, so 192 of the 768 rows are uncertain. Each uncertain row keeps the generator's endpoint orientation with probability 0.5, 0.7, or 0.9 according to $f \bmod 3$ and is reversed otherwise; on average 2.6 of the eight rows of a block are reversed. Twelve fresh systems per family give 24 systems; sixteen fitting probes, six held-out policies, the 256-policy menu, and sixteen holdout perturbations follow the six-factor protocol, and holdout worst cases use the true eight-factor semantics, so full recoupling uses 256 passes. Exact Global Balance maximizes the summed certificate of Corollary 5, extending Proposition 6 with one certificate variable per probe and

Table 8: Share of the full recoupling gain on held-out policies in the eight-factor benchmark, averaged over 24 systems. Pool-level columns retain all blocks of the p best-ranked pools at 2^p passes; block-level columns retain k blocks with the mean number of conditional passes in parentheses.

	pools p (passes)			blocks k (mean passes)		
Selector	2 (4)	4 (16)	6 (64)	4	8	16
Global Balance (exact)	0.367	0.671	0.900	0.218 (14)	0.440 (60)	0.839 (165)
Global Balance (greedy)	–	–	–	0.219 (14)	0.433 (57)	0.839 (165)
Balanced pairs	–	–	–	0.147 (16)	0.314 (189)	0.670 (256)
Matching	0.367	0.666	0.894	0.204 (16)	0.409 (87)	0.815 (203)
Occupancy	0.265	0.506	0.751	0.183 (12)	0.361 (58)	0.700 (216)
Random order	0.251	0.504	0.756	0.151 (12)	0.310 (65)	0.646 (233)
Sequential fitted gain	0.396	0.692	0.914	0.282 (6)	0.520 (22)	0.893 (117)

pool and a common block budget. The MIP uses a relative-gap tolerance of 10^{-9}. Sequential fitted gain adds, at each step, the block with the largest fitted gain on the probes and is charged the passes of every candidate it evaluates. The per-stage row evaluation is one fused GPU kernel that computes both endpoint expectations, the released maximum, and the retained endpoint pick; an independent plain evaluator replays selected values within 3.6×10^{-15}.

Table 8 and Figure 7 report every comparator. The paired GB-minus-Matching share of the full gain is $+0.015$ $[0.006, 0.024]$ at four blocks, $+0.031$ $[0.018, 0.045]$ at eight, and $+0.024$ $[0.009, 0.040]$ at sixteen; GB-minus-Occupancy and GB-minus-random are $+0.079$ $[0.052, 0.106]$ and $+0.130$ $[0.111, 0.149]$ at eight blocks. At pool level Global Balance and Matching rank the pools almost identically and both exceed Occupancy and random orders by 0.165 and 0.167 at four pools. Held-out regret per stage is 0.00596 under release, 0.00486 for Global Balance at two pools, 0.00207 at four, 0.00072 at six, and 0.00007 under full recoupling; at sixteen blocks the GB-minus-random regret contrast is -0.00188 $[-0.00267, -0.00112]$. Sequential fitted gain evaluates candidate subsets with a mean of 419, 1,725, and 8,753 conditional passes per fitting policy to select four, eight, and sixteen blocks. It also computes one released reference pass per policy. Certificate weights use two passes per policy, or 32 across the sixteen fitting probes. The mean held-out gain shares of greedy and exact Global Balance differ by at most 0.0071 at the tested budgets.

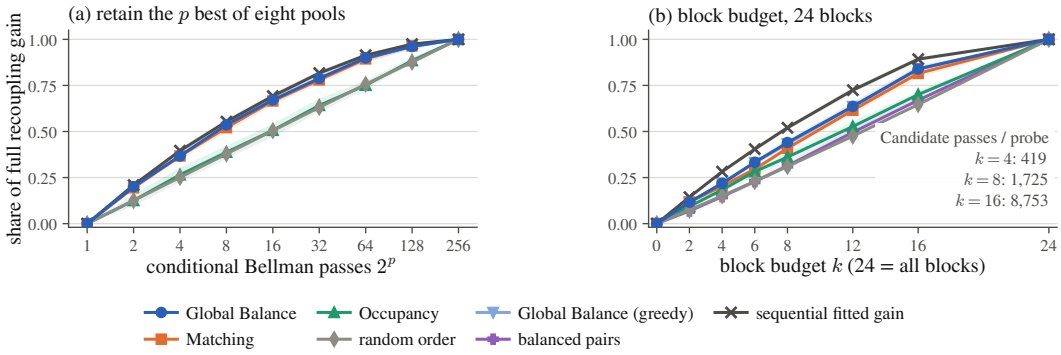


Figure 7: Eight-factor benchmark with a structure-derived sharing map and varied row orientations, 24 systems. (a) Share of the full recoupling gain against conditional passes when the p best-ranked pools are retained; (b) share against the block budget k for the block-level selectors, with a text box giving the mean candidate-evaluation passes per fitting policy used by sequential fitted gain. Its released reference adds one pass per policy. Bands are 95% bootstrap intervals over systems.

F.6 Tightness across horizon, overlap, and uncertain-row density

This known-kernel sweep uses the downstream generator with sixteen states, three actions, six blocks, odd blocks reversed, nine fitting and six evaluation policies, and all 63 nonempty retained subsets. The two distinguished successors of every uncertain row receive probabilities c and $0.9 - c$ at the two endpoints on top of the common mass, so the overlap ratio is $\chi = (c + 1/160)/(0.9 - c + 1/160)$; $c \in \{0.78, 0.66, 0.54, 0.47\}$ gives $\chi \approx 6.2, 2.7, 1.5, 1.1$. Horizons are $H \in \{4, 8, 16, 32\}$ with four

uncertain rows per block, and a density axis raises the rows per block to 16 and 32 at $H \in \{8, 32\}$ and $\chi \in \{6.2, 1.5\}$. Each cell has twelve routing and twelve queue systems and 22,680 policy–subset rows, of which 17,640 have positive certificate mass; $\mathcal{L}(K) \le G^{\mathrm{bin}}(K)$ holds on all 544,320 rows within 1.9×10^{-15}. Figure 8 reports the median and quartiles of $\mathcal{L}(K)/G^{\mathrm{bin}}(K)$ over positive-mass rows. With 24 uncertain rows the median moves from 0.895 to 0.867 between $H = 4$ and $H = 32$ at $\chi = 6.2$ and stays above 0.991 at $\chi = 1.1$, whereas the worst-case factor χ^{H-1} grows from 242 to 4×10^{24}. Raising the uncertain rows from 24 to 192 of the 384 rows at $H = 8$ lowers the median to 0.356 at $\chi \approx 6.2$ and 0.738 at $\chi \approx 1.5$. The certificate-study systems have uncertain transitions at every transient row and a median ratio of 0.325. The share of rows with a strict cross-block gain is 77.8% in every cell, as it depends only on the sign pattern of the six blocks.

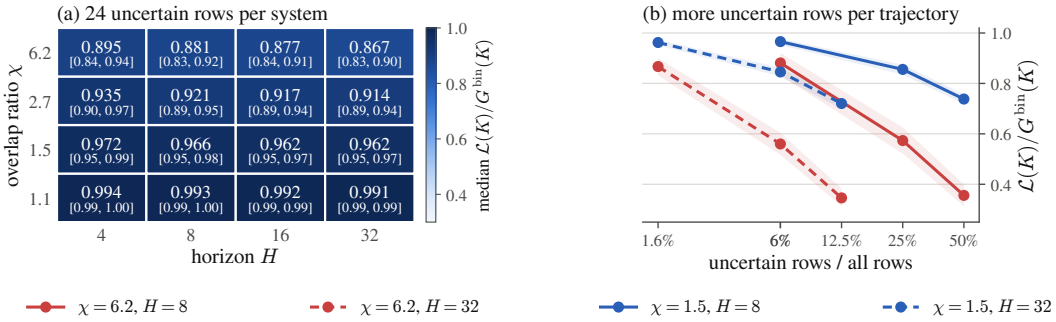


Figure 8: Certificate tightness across horizon, endpoint overlap, and uncertain-row density. (a) Median of $\mathcal{L}(K)/G^{\mathrm{bin}}(K)$ over positive-mass rows, with quartiles in brackets, for 24 uncertain rows per system. (b) Median and quartile band against the share of uncertain rows at two overlap ratios and two horizons.

F.7 Application-shaped breadth audits

The same released-gap metric is evaluated without changing its definition on two additional families. The inventory family has horizon 14, stock levels -10 through 8, order limit seven, six stock-level recoupling blocks shared across time, and nine base-stock policies. Its loss combines ordering, holding, backlog, and terminal inventory. The navigation family has a 6×5 grid, horizon 16, ten spatial dependency groups, persistent left/right crosswind, three hazards, four actions, and a two-cell goal. Its loss combines time, distance, hazards, and terminal failure. All family-level intervals use the generated system as the resampling unit.

Inventory transitions and probes. For stock s and target $b \in \{0, \dots, 8\}$, order $a = \min\{7, \max\{b - s, 0\}\}$ and set $s' = \mathrm{clip}_{[-10,8]}(s + a - D)$. The six recoupling blocks are stock ranges $[-10, -7], [-6, -4], [-3, -1], [0, 2], [3, 5], [6, 8]$ at every stage. For demand $D = 0, \dots, 4$, the two endpoint distributions in each seasonal class are the following weights divided by 20:

Class	Endpoint 0	Endpoint 1
0	$(10, 7, 2, 1, 0)$	$(0, 1, 3, 7, 9)$
1	$(8, 8, 3, 1, 0)$	$(0, 2, 5, 8, 5)$
2	$(11, 5, 3, 1, 0)$	$(0, 1, 4, 9, 6)$
3	$(7, 9, 3, 1, 0)$	$(0, 2, 4, 6, 8)$
4	$(9, 6, 4, 1, 0)$	$(0, 1, 5, 8, 6)$
5	$(12, 5, 2, 1, 0)$	$(0, 3, 4, 7, 6)$

The 14-stage class sequence is $(0, 1, 0, 2, 3, 0, 1, 4, 2, 0, 3, 1, 5, 2)$. Each system starts at one uniformly drawn stock in $\{-2, -1, 0, 1, 2\}$. Order, holding, backlog and terminal coefficients (c_o, c_h, c_b, c_T) are independently uniform on grids of spacing 0.1 in $[0.1, 0.4]$, $[0.4, 1.2]$, $[2.0, 5.0]$ and $[1.0, 2.5]$. Stage loss is $c_o a + c_h u_h [s']_+ + c_b v_h [-s']_+$; terminal loss is $c_T(c_h [s]_+ + c_b [-s]_+)$. The holding and backlog multipliers are

$$u = (4, 3, 2, 1, 0.5, 4, 3, 2, 1, 0.5, 4, 3, 1, 5),$$
$$v = (0.5, 0.75, 1, 2, 4, 0.5, 0.75, 1, 2, 4, 0.5, 1, 4, 0.5).$$

Table 9: Binary-regime breadth audits under the normalized residual in Equation (90). Target k is the first retained budget with residual at most 0.1. Lower is better.

Family	Systems	Exact-defect AUC	Random AUC	Target k: defect / random
Seasonal inventory	12	0.208	0.503	3.00/6.00
Persistent crosswind	16	0.338	0.363	5.75/7.50

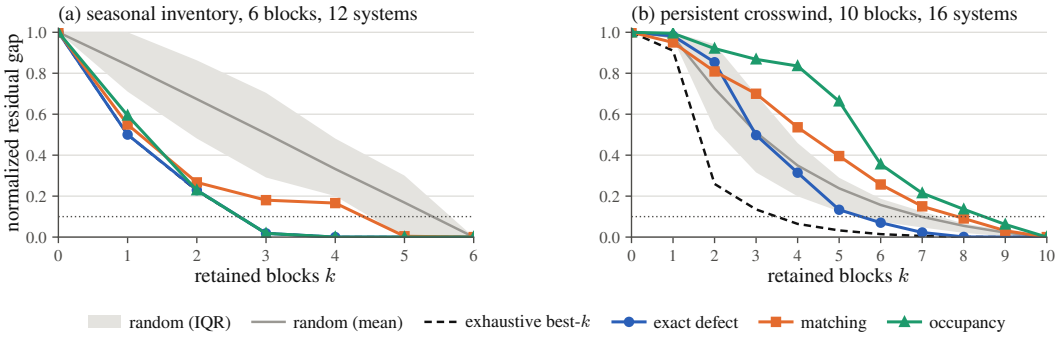


Figure 9: Application-shaped breadth audits: mean normalized residual gap against retained blocks for the exact-defect, Matching, and Occupancy rankings, exhaustive best-k, and 256 random orders per system (mean and interquartile band) on (a) seasonal inventory and (b) persistent crosswind. AUCs are family means; the dotted line is the 0.1 target.

Crosswind transitions and probes. Coordinates satisfy $0 \le x \le 5$, $0 \le y \le 4$; goals are $(2,0),(3,0)$; each system starts at one uniformly drawn point in $\{(2,4),(3,4)\}$. Block $2y + \mathbf{1}\{x \ge 3\}$ is shared across time. Each system independently draws action-success probability from $\{0.72, 0.82, 0.90\}$ and wind probability from $\{0.45, 0.55, 0.65\}$. The intended cardinal move succeeds with the first probability; the two perpendicular moves share the remainder. After clipping that move to the grid, a wind event adds $(-1,0)$ or $(1,0)$ at the two endpoints and clips again. Goals are absorbing. Three distinct hazards are uniform draws from $x \in \{1,2,3,4\}$, $y \in \{1,2,3\}$.

Let $d(s)$ be Manhattan distance to the nearer goal. Outside a goal, stage loss is $c_s + c_d d(s')/9 + c_z \mathbf{1}\{s' \text{ is a hazard}\}$ and terminal loss is $c_f + d(s)/9$; both are zero at a goal. Coefficients are independently uniform on $c_s \in \{0.02, 0.04, 0.06\}$, $c_d \in \{0.08, 0.12, 0.16\}$, $c_z \in \{0.3, 0.5, 0.8\}$ and $c_f \in \{2.0, 2.5, 3.0\}$. The 18 deterministic probes pair $t \in \{0.75, 1.25\}$ with $o \in \{-2.0, -1.5, \ldots, 2.0\}$ and choose the clipped successor (x', y') minimizing $t\, d(x', y') + |x' - \text{clip}_{[0,5]}(2.5 + oy/4)|$. Ties follow north, west, east, south.

In both breadth families, exact defect averages each probe's coherent-occupancy defect contribution divided by its full release gap; Matching uses its matching score divided by that gap. Occupancy averages coherent block occupancy divided by its total over blocks. All rankings use the same frozen probes. Random bands average the within-system 25th and 75th percentiles of 256 random orders.

In inventory, exact-defect ordering nearly coincides with exhaustive best-k search (AUC 0.207), and exact defect minus random AUC is -0.2954 with paired interval $[-0.3104, -0.2795]$. Crosswind is more interaction-heavy: exhaustive search attains AUC 0.192, while exact defect minus random is -0.0258 $[-0.0615, 0.0080]$. These comparisons evaluate exact-defect ranking through both whole-curve AUC and the number of blocks needed to reach the fixed residual target. Figure 9 shows the full residual curves of both families.

F.8 Trajectory-estimated CliffWalking

The standard 4×12 geometry is lifted to 39 states by representing a cliff fall as an observable transient event that incurs cost 100 and returns to the standard start. The two labeled episode-persistent regimes share intended-action probability 0.72 and exchange the lateral-slip probabilities 0.23 and 0.05. The horizon is 28, eight spatial groups and fifteen soft-DP controllers are frozen before sampling, and endpoint kernels are estimated with a 0.5 self-loop pseudocount. Sixteen paired datasets are generated at 128, 512, and 2,048 episodes per endpoint; a disjoint 512-episode sample evaluates likelihood.

Table 10: Trajectory-estimated CliffWalking with binary persistent regimes. Intervals bootstrap paired datasets; lower AUC, total variation (TV), and Kendall distance are better. Δ is learned exact-defect minus random AUC.

Episodes/endpoint	TV	Kendall	Learned AUC	Δ AUC	Target k
128	0.05975	0.19643	0.57665	-0.00044 $[-0.01763, 0.01768]$	7.375 [7.125, 7.625]
512	0.02936	0.10045	0.55990	-0.01719 $[-0.02972, -0.00547]$	7.063 [7, 7.188]
2048	0.01434	0.05357	0.54703	-0.03006 $[-0.03907, -0.01976]$	7.000 [7, 7]

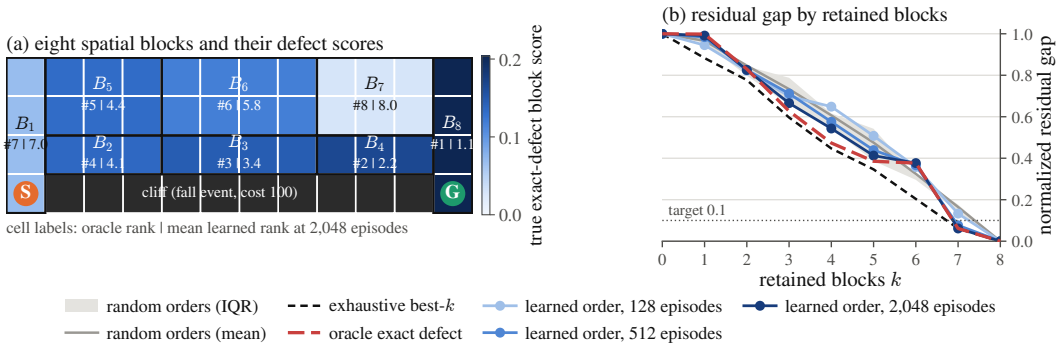


Figure 10: Trajectory-estimated CliffWalking. (a) The eight spatial blocks, shaded by the simulator-true exact-defect score and labelled with the oracle rank and the mean learned rank at 2,048 episodes over 16 datasets; dark cells are the cliff. (b) Mean normalized residual gap for the learned order at each budget, the oracle order, exhaustive best-k, and 256 random orders (mean and interquartile band). With more data the learned order approaches the oracle and meets the 0.1 target with seven blocks.

Simulator kernels define the fixed probe menu and evaluate the learned subset choices; the learned rankings themselves use estimated kernels.

The 15 probes combine endpoint-mixture weights $\{0, 0.25, 0.5, 0.75, 1\}$ with planning fall costs $\{30, 100, 300\}$. Backward soft-DP uses temperature 0.35 and policy $\pi_h(a \mid s) \propto \exp[-(Q_h(s,a) - \min_b Q_h(s,b))/0.35]$, with continuation averaged under this policy. Evaluation uses step cost one, fall-entry cost 100, and terminal cost $35+$ Manhattan distance to the goal; the goal has zero cost and the fall state's terminal cost is 36. The initial state is the standard start. The left and right boundary columns are separate blocks; the remaining columns form bands 1–3, 4–7 and 8–10, each split into upper and lower grid halves. The fall state belongs to the right-boundary block.

Each acquisition trajectory contains 48 transitions at one fixed endpoint. It chooses a route row uniformly from $\{0, 1, 2\}$, moves vertically to that row, then right, then down at the right boundary. With probability 0.35 it instead takes a uniform action. Reaching the goal resets to the start and draws a new route row; a fall event returns to the start on the next transition. Row counts receive a 0.5 self-loop pseudocount and are normalized; the goal row remains absorbing. TV is row total-variation error averaged with training visit counts as weights, then averaged across datasets. Kendall distance is the number of discordant pairs divided by $\binom{8}{2} = 28$. Exact-defect rankings use the gap-normalized scores specified for the breadth studies.

At 2,048 episodes, the simulator-oracle exact order has AUC 0.53152 and the learned-order excess is 0.01551 $[0.00669, 0.02528]$. The learned selector matches the exhaustive seven-group target in every dataset. Occupancy has AUC 0.53195 but needs all eight groups to reach the same 0.1 residual target.

F.9 Scaling and Numerical Validation

The scaling sweep uses group counts $|\mathcal{B}| \in \{16, 32, 64, 128, 256\}$, probe counts $|\mathcal{J}_+| \in \{8, 32, 128\}$, eight independently generated instances per cell, and budget $k = \lceil 0.25|\mathcal{B}| \rceil$. To preserve the empirical score geometry, each large instance samples its $|\mathcal{B}|$ group columns with replacement from the 24×6 fitted-system pool and samples each of its probe rows with replacement from the nine-policy fitting pool. Positive and negative scores are divided by their joint mean. This completely specifies the synthetic large-menu distribution; it does not draw independent arbitrary score matrices. Greedy

implements Equation (91) by maintaining both sign totals and adding the group with largest resulting objective, with lower-index ties. It is empirical heuristic evidence, not a submodular-greedy guarantee.

Runs used an NVIDIA RTX 4080 SUPER, Torch 2.13.0+cu130 and CUDA 13.0; the exact MIP used HiGHS 1.15.1 with 16 threads, a five-second limit, and relative-gap tolerance 10^{-4}. GPU timings use two warmups and the median of seven synchronized repetitions. All 120 MIPs receive solver-reported optimal status under this tolerance. At 256 groups, pooling the three probe counts, the ratio of the greedy objective to the returned MIP incumbent is 0.99995 $[0.99987, 0.99999]$. At 128 probes, median selector time is 0.88010 seconds for the exact formulation and 0.06804 seconds for greedy; matching and occupancy take 0.000358 and 0.000365 seconds. These times exclude score construction, data acquisition, confidence propagation, Bellman calls, and end-to-end memory. Thirty-two small fixtures agree with exhaustive enumeration. Figure 11 summarizes the sweep.

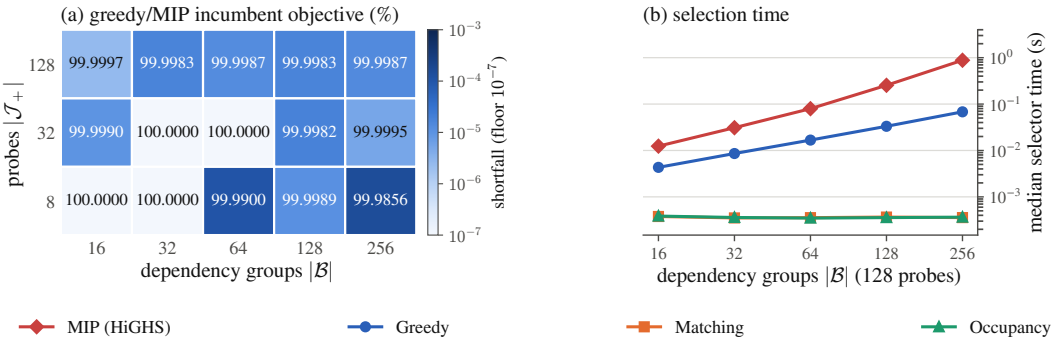


Figure 11: Selector-only scaling sweep. (a) Greedy/MIP certified-objective ratio to the returned MIP incumbent in percent over group and probe counts, eight instances per cell. Colour shows $\max\{1 - \text{ratio}, 10^{-7}\}$ on a logarithmic scale; the MIP relative-gap tolerance is 10^{-4}. (b) Median wall time at 128 probes for the MIP (HiGHS) and the three score-based selectors. Timings exclude score construction, data acquisition, confidence propagation, and Bellman calls.

The independent finite-data check reconstructs every row interval and all 69,120 propagated bounds; for the policy study it recomputes every selected policy and holdout regret; for CliffWalking it regenerates all paired trajectory datasets and simulator-truth subset values; and for scaling it reconstructs all 120 score tensors and mixed-integer objectives. Torch and CUDA evaluations share inputs but use independent implementations. Their maximum discrepancies are 1.14×10^{-13} for value recursion and 1.78×10^{-15} for the downstream mask audit.