WHAT CAN BAD DATA IDENTIFY? NEGATIVE DEMONSTRATIONS IN OFFLINE DECISION MAKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Offline logs often contain crashes, violations, rejected outputs, and abandoned episodes even when expert demonstrations and dense rewards are unavailable. Can these "bad" trajectories reveal what an agent should do, rather than only what it should avoid? Within a finite-MDP, linear-reward framework, we separate three requirements for using failure frequencies to guide control: an observation model that connects latent return to the selection of failures, variation that reveals reward contrasts between competing policies, and transition coverage for the policy ultimately selected. Labels without such an observation model generally identify only actions or trajectories to avoid. Under a negative-rational exponential-family selection model, we characterize reward ambiguity exactly by the null space of centered trajectory features and give an exact test for whether one policy is optimal throughout an affine reward-ambiguity set. The same geometry determines when failures add local Fisher information to an existing log, yields a finite-sample policy-recovery bound under exact-model assumptions, and supports a locally E-optimal design for collecting complementary failure sources. Under matched exponential channels, neither positive nor negative demonstrations are intrinsically more informative: the channel with greater feature variance along the contrasts needed for control has greater local Fisher information, with an exponential advantage possible in either direction. Extensions to negative–unlabeled mixtures and occupancy repulsion expose the additional assumptions behind common uses of bad data. Controlled synthetic tabular experiments isolate the predicted rank, alignment, and coverage effects and demonstrate offline hazard control using failures as the only reward supervision in that setting.

1 INTRODUCTION

Learning from demonstrations normally begins with examples of successful behavior, while reward learning from human feedback typically asks for comparisons between alternatives (Ng & Russell, 2000; Ziebart et al., 2008; Christiano et al., 2017). Both forms of supervision can be expensive when good behavior is difficult to produce or alternatives must be labeled retrospectively. By contrast, many deployed systems already record collisions, human interventions, rejected outputs, alarms, shutdowns, and abandoned runs (Kelly et al., 2019). These events are often easy to recognize after the fact. If their informational content were understood, existing failure logs could guide policy design, data collection, and decisions about when an offline learner should abstain—without first requiring an expert demonstrator or dense reward.

Yet a failure is not an inverted expert demonstration. Observing that one trajectory is bad rules out treating it as demonstrated optimal behavior, but it need not distinguish among the many ways of behaving differently. Even the frequency of a failure is ambiguous: a trajectory may appear often because it is undesirable, because the behavior policy proposes it often, or because it is easy to reach. Finally, identifying a preferred alternative does not imply that an offline learner has transition data with which to evaluate it. These are different obstructions, and increasing the number of negative examples resolves only some of them.

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.

Our primary question is: *when do bad demonstrations provide enough information to identify, and with finite data recover, a good policy that an offline transition log can support?* Reward identification is an intermediate diagnostic, not an end in itself. Several rewards may explain the same failures yet agree on the best covered policy; conversely, identifying a reward does not make a policy learnable if its transitions are absent from the log. We therefore study both reward and policy recovery, with the policy decision as the ultimate target. Our answer separates three requirements. *Semantics* specify how latent return affects the probability that a trajectory is observed as negative. *Excitation* requires failures to vary along reward directions that distinguish competing policies. *Coverage* requires the transition log to support the policy selected using that information. The distinction is operational: a semantic failure calls for a better observation model, missing excitation calls for complementary failure sources, and missing coverage calls for more transition data or a more conservative target policy.

Existing approaches show that non-preferred demonstrations can enforce safety or avoidance (Jang et al., 2023; Hoang et al., 2025), that demonstrations can reveal latent constraints (Malik et al., 2021), and that pairwise preferences can guide offline control (Kang et al., 2023). None of these facts implies that a collection of failures, observed without preferred comparisons, orders the unobserved alternatives. We therefore begin with a deliberately weak observation: a trajectory is labeled bad, with no model for how it was generated. Even with infinitely many labels and known dynamics, this supervision generally identifies only avoidance information. We then introduce a transparent negative-selection model in which a known proposal distribution is exponentially tilted toward low-return trajectories. This model represents a logging or curation mechanism that preferentially retains poor outcomes. Within it, the population distribution of failures reveals reward contrasts, and its centered feature geometry determines which contrasts are identifiable. We then ask whether any reward ambiguity left by those failures can actually change the best covered policy.

Our contributions are:

- We establish the semantic boundary of failure supervision: labels alone identify avoidance, not an optimal alternative, and an unrestricted, unobserved proposal distribution completely confounds reward with availability.
- For a negative-rational trajectory model, we characterize reward ambiguity exactly through centered feature null spaces. We further give an exact common-policy test over an affine reward-ambiguity set, showing when invisible reward directions cannot change a covered control decision.
- We derive statistical and data-collection consequences of this geometry: a finite-sample policy-recovery bound under the exact selection model, a directional Fisher-information comparison with positive demonstrations, and a locally E-optimal allocation across complementary failure sources. We also characterize the extra assumptions required by negative–unlabeled mixtures and occupancy-repulsion objectives.
- Controlled tabular studies test each mechanism separately. They verify the rank transition, show that policy-relevant alignment rather than label sign determines sample efficiency, separate semantic misspecification from transition coverage, and recover a hazard-control policy using failures as the only reward supervision.

The paper uses three levels of claim. *Population identification* asks what the exact negative-trajectory distribution reveals about reward and, more importantly, about the best covered policy. *Finite-sample recovery* asks whether that policy can be selected reliably from n_- sampled negative trajectories under the assumed model. *Offline planning* additionally asks whether the transition log supports evaluation of the selected policy. Identifying every reward weight is not required for a policy decision, and a result at one level does not automatically imply a result at the next.

The remainder of the paper formalizes these three requirements: Section 2 positions the question, Section 3 defines the target and proves the label-only limit, Section 4 develops the identification theory, recovery procedure, collection design, and extensions, and Section 5 tests the predicted mechanisms.

2 RELATED WORK

Reward and policy identification. IRL can leave reward ambiguous even when behavior is observed exactly (Ng & Russell, 2000; Ziebart et al., 2008; Kim et al., 2021; Cao et al., 2021; Shehab et al., 2024). Observations from multiple environments or experts can remove some ambiguity (Rolland et al., 2022), while constrained and finite-sample analyses distinguish reward identification from transfer and control (Schlaginhaufen & Kamgarpour, 2023; Zhao et al., 2024). Our channel is different: it preferentially retains low-return trajectories. This is not merely maximum-entropy IRL with the sign reversed, because the proposal law determines which frequency ratios have reward meaning and the covered policy class determines whether residual reward ambiguity matters for control.

Observation models and imperfect demonstrations. Reward inference depends on how behavior becomes data: demonstrator bias, perceptual corruption, or uncertain dynamics can otherwise be absorbed into the reward (Shah et al., 2019; Suresh & Doshi, 2022; Herman et al., 2016). Rankings and preferences provide explicit comparisons (Brown et al., 2019; 2020; Christiano et al., 2017; Kang et al., 2023), whereas an isolated negative label does not. Negative-demonstration and inverse-constraint methods instead support avoidance or constraint inference (Jang et al., 2023; Hoang et al., 2025; Malik et al., 2021; Lindner et al., 2024). We characterize what the negative observation law identifies before optimization, thereby separating evidence for reward or policy choice from avoidance alone.

Collection and offline control. Active IRL designs expert queries or exploration (Brown et al., 2018; Lindner et al., 2022; Guo et al., 2022); our design instead allocates a fixed budget among available failure sources. Offline RL poses a separate transition-coverage problem (Levine et al., 2020; Foster et al., 2022): reward information cannot create missing transitions, and extra transition data do not explain how negative selection relates to reward. Appendix H develops these distinctions in detail.

3 PROBLEM FORMULATION AND THE LIMIT OF LABELS

Decision model. Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \gamma, d_0)$ be a finite discounted Markov decision process (MDP), where $\mathcal{S}$ and $\mathcal{A}$ are the state and action sets, P is the transition kernel, $\gamma \in (0, 1)$ is the discount factor, and d_0 is the initial-state distribution. A policy π has normalized state–action occupancy

$$\rho_\pi(s, a) = (1 - \gamma)\mathbb{E}_\pi \left[\sum_{t \geq 0} \gamma^t \mathbf{1}\{s_t = s, a_t = a\} \right].$$

Here $\mathbf{1}\{\cdot\}$ is the indicator function and the expectation follows d_0, P, π. We assume a linear reward, $r_\theta(s, a) = \theta^\top \phi(s, a)$, where $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is a known feature map and $\theta \in \Theta \subseteq \mathbb{R}^d$ contains unknown feature weights. This assumption makes the reward information in negative trajectories interpretable through feature differences and permits an exact finite-dimensional null-space characterization. It is not a claim that real rewards must be linear: a fixed nonlinear feature map can still be used, but learning the representation or allowing an unrestricted nonlinear reward would change the identifiability question and is outside our results. The normalized feature occupancy is $\mu_\pi = \mathbb{E}_{(s,a)\sim\rho_\pi}[\phi(s, a)] \in \mathbb{R}^d$, so $J_\theta(\pi) = (1 - \gamma)^{-1}\theta^\top \mu_\pi$. Thus policy choice depends only on inner products between the unknown reward weights and differences in policy feature occupancy.

Two data sources, two roles. Let $\mathcal{D}$ denote an offline transition dataset, and let $\mathcal{B}_\mathcal{D} \subseteq \mathcal{S} \times \mathcal{A}$ be the population support of its logging behavior. Define

$$\Pi_\mathcal{D} := \{\pi : \rho_\pi(s, a) > 0 \Longrightarrow (s, a) \in \mathcal{B}_\mathcal{D}\}$$

and let $\Pi_\mathcal{D}^{\text{det}}$ be its finite set of deterministic policies. The optimization target remains $\Pi_\mathcal{D}$, including randomized policies. In a finite discounted MDP with linear expected return, every covered occupancy is a convex combination of deterministic covered occupancies. Hence checking this finite subclass suffices for optimality, and it permits a positive gap to suboptimal deterministic competitors. This reduction need not hold for objectives that favor randomization, such as entropy-regularized control.

A covered policy uses only state–action pairs represented by the population logging process. This support condition is necessary but does not by itself guarantee accurate finite-sample evaluation; the amount of transition data and the strength of coverage still control dynamics error. Separately, negatively selected trajectories provide information about reward. Our default target is therefore a *covered* optimum $\pi^*_\theta \in \operatorname{argmax}_{\pi \in \Pi_\mathcal{D}} J_\theta(\pi)$, not an unconstrained optimum. The data-generating reward is θ_*, and we abbreviate one corresponding covered optimum by $\pi_* := \pi^*_{\theta_*}$.

Depending on the setting, we may also observe an unlabeled trajectory distribution. The *negative channel* is the process that selects or records trajectories as negative demonstrations—for example, a logger retaining episodes associated with failures or interventions. Calling this channel "negative" describes how its records are designated; it does not by itself say how likely any trajectory is to be selected, or imply that every selected trajectory has low reward. A *negative-observation model* q^-_θ specifies the resulting distribution over recorded trajectories. Specifically, $q^-_\theta(\tau)$ is the probability mass (or density) assigned to trajectory τ when the reward parameter is θ. The superscript "$-$" labels the channel, not a negative numerical value; the subscript θ records its dependence on reward. The later negative-rational model adds the particular assumption that selection favors low-return trajectories. Thus q^-_θ is not the probability that an individual trajectory is objectively bad. Before discussing reward identification, fix an accepted invariance subspace $\mathcal{G}$ (for example, the representable additive-constant or potential-shaping directions) and write $\theta \sim_r \theta'$ when $\theta - \theta' \in \mathcal{G}$. Write $\theta \sim_\pi \theta'$ when the two parameters induce the same set of covered optimal policies.

We assume $\Pi_\mathcal{D} \neq \varnothing$, so the finite supported MDP has a deterministic covered policy.

Definition 1 (Reward, policy, and avoidance identification)**.** *Let $\theta, \theta' \in \Theta$ be any two candidate reward parameters. Negative data* identify reward up to $\mathcal{G}$ *if $q^-_\theta = q^-_{\theta'}$ implies $\theta \sim_r \theta'$: two rewards that predict exactly the same negative-trajectory distribution may differ only by an accepted invariance. They* identify the covered optimal policy *if $q^-_\theta = q^-_{\theta'}$ implies $\theta \sim_\pi \theta'$: observationally indistinguishable rewards must choose the same covered optima. Under label-only supervision, we use* avoidance information *more modestly to mean that a label rules out a specified choice as optimal without ranking the remaining choices. For example, in a one-state problem, labeling action a_0 suboptimal rules out the policy that always chooses a_0, but says nothing about whether a_1 or a_2 is better.*

Policy identification is weaker than reward identification and often sufficient for control. Both are population notions: they compare exact observation distributions, not estimates from finite samples. The summary in Table 2 of Appendix A contrasts these notions across observation models.

3.1 What labels alone identify

First consider the deliberately weak semantics that a label tells us only that the demonstrated choice is *strictly suboptimal*; it may still be nearly optimal, because the label supplies no numerical gap. We make this assumption in order to test what the label alone can establish, without using how frequently choices are labeled, a ranking, or a comparison with a preferred trajectory. The next theorem gives a finite-action counterexample: even knowing that one action is not best need not reveal which alternative is best. The same ambiguity can occur with continuous actions if the reward class permits three feasible actions whose rewards can be reordered, but the stated theorem concerns a finite action set.

Theorem 1 (Bad labels do not identify what is best)**.** *Consider a known one-state MDP with at least three actions and a reward class containing every vector in $[-1, 1]^{|\mathcal{A}|}$. Suppose every observed negative demonstration takes action a_0, and the observation reveals only that a_0 is suboptimal. Then, even with infinitely many demonstrations, the optimal policy is not identifiable: two rewards can agree that a_0 is suboptimal while making different actions uniquely optimal, so no rule using only these labels can always choose the correct one. More generally, identification fails if the label-compatible reward set contains two observationally equivalent rewards with different sets of covered optimal policies; reversing two* suboptimal *alternatives alone is not enough.*

The proof constructs two rewards that agree on every negative observation but swap the values of the two alternatives. The result specifies the extra information needed for a positive conclusion: negative labels shrink a version space and identify avoidance, while a probability model, ranking,

score, or comparison must orient the remaining alternatives. This boundary is semantic and persists independently of sample size, dynamics knowledge, or function approximation.

4 IDENTIFICATION FROM STRUCTURED BAD BEHAVIOR

We now replace label-only supervision with a model of *how often* each trajectory appears in the negative channel. The extra information comes from this selection model, not from the word "bad."

Let $\mathcal{T}_-$ be the support of trajectories that can enter the negative-selection process; it may be a strict subset of trajectories composable from the offline transition support. Let $q_0(\tau) > 0$ be a known reference probability mass function on $\mathcal{T}_-$, and define a trajectory's discounted feature sum by $\Phi(\tau) = \sum_{t\geq 0} \gamma^t \phi(s_t, a_t)$, with the sum truncated at termination for finite logged episodes. For notational clarity, the formal statements below use a finite or countable support, which includes an empirical pool of logged trajectories, and assume that the normalizing constant is finite. On a general trajectory space, the same model is written as an exponential Radon–Nikodym density with respect to q_0, and equalities hold q_0-almost surely. When τ is random, we abbreviate $\Phi(\tau)$ by Φ inside expectations and variances. Generatively, q_0 describes candidate trajectories produced by a logging process, including its action choices and stochastic transitions. We then posit a separate curation step that retains a realized candidate τ with weight proportional to $\exp\{-\beta\theta^\top \Phi(\tau)\}$. Conditional on retention, a trajectory follows the *negative-rational* law

$$q_\theta^-(\tau) = \frac{q_0(\tau)\exp\{-\beta\theta^\top \Phi(\tau)\}}{Z_-(\theta)}, \qquad \beta > 0. \tag{1}$$

Here $Z_-(\theta) = \sum_{\tau\in\mathcal{T}_-} q_0(\tau)\exp\{-\beta\theta^\top\Phi(\tau)\}$ is the normalizing constant, and β is the selection strength. Low-return trajectories are preferentially retained. The exponential tilt is an explicit assumption about selection, not a consequence of deterministic dynamics or a maximum-entropy principle; it also covers stochastic dynamics when q_0 is the reference law over the resulting finite or terminated trajectories. Nor does it claim that the behavior policy intentionally minimizes reward. Its central implication is transparent: for any $\tau, \tau' \in \mathcal{T}_-$,

$$\log \frac{q_\theta^-(\tau)/q_0(\tau)}{q_\theta^-(\tau')/q_0(\tau')} = -\beta\theta^\top\big(\Phi(\tau) - \Phi(\tau')\big).$$

This log-ratio identity follows algebraically from the assumed trajectory-selection law: the normalizer $Z_-(\theta)$ cancels, regardless of whether transitions are deterministic or stochastic. After correcting for availability under q_0, relative failure frequency therefore reveals a reward contrast at population level. The sign and temperature are part of these semantics; without a known orientation, θ and $-\theta$ cannot be distinguished as desirable versus undesirable.

A one-step example: which way does the inferred reward point? Suppose a decision has three actions a, b, c, the reference chooses each with probability $1/3$, and $\beta = 1$. Imagine that their population frequencies among negatively selected trajectories are $q^-(a) = 0.6$, $q^-(b) = 0.3$, and $q^-(c) = 0.1$. Writing $r(x)$ for the reward of action x, the log-ratio identity gives

$$r(b) - r(a) = -\log\frac{q^-(b)/q_0(b)}{q^-(a)/q_0(a)} = \log 2, \qquad r(c) - r(a) = \log 6.$$

Thus, after choosing the arbitrary baseline $r(a) = 0$, the reward consistent with this model is $(r(a), r(b), r(c)) = (0, \log 2, \log 6)$, and the reward-maximizing policy chooses c. The minus sign in (1) belongs to the *selection mechanism*: it makes low-reward actions appear more often in the negative sample. It does not mean that the recovered reward should be negated before planning. This inference relies on the equal reference probabilities; with a nonuniform q_0, raw negative frequencies alone need not have this ordering.

The reference law is essential. If q_0 is unrestricted and unobserved, any candidate reward can be paired with a reference law that reproduces the observed failures; if q_0 is known but β is not, the data identify $\beta\theta$ only up to the invisible directions $\mathcal{N}$. Proposition 1 in Appendix A states this confounding result formally. It motivates recording the proposal policy, estimating a restricted reference from a separate unlabeled log while propagating its uncertainty, or using matched comparisons in which the

proposal factor cancels. Raw failure frequency without this correction confounds availability with undesirability.

Fix $\tau_0 \in \mathcal{T}_-$, and let

$$M = \operatorname{span}\{\Phi(\tau) - \Phi(\tau_0) : \tau \in \mathcal{T}_-\} \subseteq \mathbb{R}^d, \qquad \mathcal{N} = M^\perp.$$

Here $M^\perp$ denotes the orthogonal complement of M:

$$\mathcal{N} = M^\perp = \left\{u \in \mathbb{R}^d : u^\top \big(\Phi(\tau) - \Phi(\tau_0)\big) = 0 \text{ for every } \tau \in \mathcal{T}_-\right\}.$$

Thus M contains the feature directions along which supported negative trajectories differ, whereas $\mathcal{N}$ contains reward-weight directions that are perpendicular to every observed feature difference. Adding any $u \in \mathcal{N}$ to the reward parameter changes every supported negative trajectory's return by the same amount, so it does not change their relative probabilities and cannot be detected from the negative data. For two covered policies, we call $g_{\pi,\pi'} := \mu_\pi - \mu_{\pi'}$ their *policy feature contrast*, because $\theta^\top g_{\pi,\pi'}$ determines their return ordering. References below to a policy-relevant direction mean one of these contrasts, or a direction in their span, for policies under consideration.

Theorem 2 (Exact identification by bad trajectories)**.** *Under* (1)*, $q_\theta^- = q_{\theta'}^-$ if and only if $\theta - \theta' \in \mathcal{N}$. Consequently:*

1. *reward is identified modulo exactly the null directions $\mathcal{N}$;*
2. *on any parameter set Θ, a reward θ is point identified exactly when $(\theta + \mathcal{N}) \cap \Theta = \{\theta\}$. In particular, on a normalized affine parameter space $\bar{\theta} + V$, global point identification is equivalent to $V \cap \mathcal{N} = \{0\}$;*
3. *the covered optimal policy is identified if and only if every pair $\theta, \theta' \in \Theta$ with $\theta - \theta' \in \mathcal{N}$ is policy equivalent.*

The third statement can hold without reward identification: unobserved reward directions may be irrelevant to choosing a policy in the covered MDP. The next theorem asks a control-focused question: is one candidate policy optimal under *every* reward still compatible with the negative data? If so, the decision is justified even though the reward weights are not unique. It turns this question into a finite test against deterministic covered competitors. Let $\operatorname{Proj}_{\mathcal{N}} x$ be the orthogonal projection of a vector x onto $\mathcal{N}$: it extracts the component of x along reward directions invisible to the negative data. For this theorem we deliberately relax the parameter-restricted equivalence class $(\bar{\theta} + \mathcal{N}) \cap \Theta$ to the full, possibly unbounded affine set $\mathcal{V} = \bar{\theta} + \mathcal{N}$. The resulting test checks whether an invisible reward change can reverse the candidate's ordering against any competitor and, if not, whether the visible reward component favors the candidate. This certificate is conservative: bounded prior information can shrink the version set and certify additional policies.

Theorem 3 (Exact common-policy test on an affine version set)**.** *A deterministic covered policy $\pi^\dagger$ is optimal for every $\theta \in \mathcal{V}$ if and only if, for every $\pi \in \Pi_{\mathcal{D}}^{\det}$,*

$$\operatorname{Proj}_{\mathcal{N}}(\mu_{\pi^\dagger} - \mu_\pi) = 0 \quad \textit{and} \quad \bar{\theta}^\top(\mu_{\pi^\dagger} - \mu_\pi) \geq 0. \tag{2}$$

Under these conditions, it is the unique commonly optimal deterministic policy if the second inequality is strict for every $\pi \neq \pi^\dagger$. Hence a common control decision over $\mathcal{V}$ can be certified using feature occupancies and a null-space projection, without identifying reward.

The result distinguishes reward recovery from the actual control decision. The likelihood equivalence class inside the parameter space is $(\bar{\theta} + \mathcal{N}) \cap \Theta$, while Theorem 3 uses the larger affine set $\bar{\theta} + \mathcal{N}$ for a conservative certificate. The test is necessary and sufficient for common optimality over that affine set; it is not, in general, necessary for policy identification over a smaller or bounded parameter set. An unidentified direction is harmless exactly when it gives the candidate and every competitor the same feature-return difference. Policy identification therefore depends not on the dimension of the version set, but on whether moving within it can cross a policy decision boundary.

4.1 Information and collection implications

The same subspace geometry answers three practical questions. First, under the correctly specified negative-rational model and finite-moment regularity, one bad trajectory has local Fisher information

$F(\theta) = \beta^2 \mathrm{Cov}_{q_\theta^-}[\Phi]$, whose null space is exactly $\mathcal{N}$. Thus failures add local information to a baseline log only in previously unresolved directions along which their features vary (Corollary 1 and Proposition 2 in Appendix A).

Second, label sign does not by itself determine sample efficiency. Under matched, correctly specified positive and negative exponential tilts, local Fisher information in a decision direction v is proportional to the variance of $v^\top \Phi$ in that channel. Negative data have greater local Fisher information in v exactly when this variance is larger; the reverse is equally possible. This comparison does not by itself guarantee smaller finite-sample error or faster policy recovery, which also depend on how the informative directions align with all relevant policy contrasts. Proposition 3 and the three-point construction in Appendix A show that the directional Fisher-information ratio can be exponential in β.

Third, diversity is valuable when it closes missing directions. Suppose K known failure mechanisms have per-sample information matrices $F_1, \ldots, F_K$. Repeating one source reduces uncertainty within its existing range but cannot change its null space; combining sources can increase rank when their null spaces are complementary. For a fixed parameter value, target subspace, and E-optimality criterion, Theorem 5 in Appendix A maximizes the smallest information eigenvalue on that subspace. Corollary 2 gives a separate greedy rank-gain guarantee when only a fixed number of mechanisms can be used. Because the matrices depend on the unknown reward, a practical implementation uses pilot estimates and is only approximately locally optimal.

4.2 From bad trajectories to a covered policy

The recovery procedure under (1) has two stages. First, given negative trajectories $\tau_1, \ldots, \tau_n$, a known reference law q_0, selection strength β, and feature map ϕ, form the identifiable subspace $U = \mathcal{N}^\perp = M$ from the feature differences on $\mathcal{T}_-$. Project the feasible reward set to $\Theta_U := \{\mathrm{Proj}_U \theta : \theta \in \Theta\} \subseteq U$, which retains every feasible reward's visible component, and fit the negative-channel likelihood:

$$\widehat{\theta}_n \in \operatorname*{argmin}_{\theta \in \Theta_U} \left\{ \frac{\beta}{n} \sum_{i=1}^{n} \theta^\top \Phi(\tau_i) + \log Z_-(\theta) \right\}. \tag{3}$$

This uses the frequencies of the negative trajectories relative to q_0, not observed reward values. The projection is valid for likelihood fitting because $\theta - \mathrm{Proj}_U \theta \in \mathcal{N}$ adds the same constant to every supported trajectory's return and leaves (1) unchanged; we evaluate that law on projected representatives even if they lie outside Θ. When q_0 is an empirical distribution on a finite trajectory pool, $Z_-(\theta)$ is a finite sum over that pool. If β is unknown, the likelihood identifies $\beta\theta$ only up to $\mathcal{N}$, so a reward scale cannot be recovered without an additional normalization.

Second, use the offline transition log to estimate dynamics $\widehat{P}$, compute each candidate policy's feature occupancy $\widehat{\mu}_\pi$ under $\widehat{P}$, and restrict planning to the empirically supported policy class $\widehat{\Pi}_\mathcal{D}$. The plug-in decision is

$$\widehat{\pi} \in \operatorname*{argmax}_{\pi \in \widehat{\Pi}_\mathcal{D}} \widehat{\theta}_n^\top \widehat{\mu}_\pi. \tag{4}$$

With known dynamics and population support, replace $\widehat{\mu}_\pi, \widehat{\Pi}_\mathcal{D}$ by $\mu_\pi, \Pi_\mathcal{D}$. If $\mathcal{N} \neq \{0\}$, the fitted representative alone need not justify a policy: the control test in Theorem 3 can certify agreement across all rewards in its affine version set. If that certificate is inconclusive, one may plan against a separately calibrated, bounded reward set as in (10), or abstain. The theorem below analyzes reward sampling error for (3) with known q_0, β, U and fixed policy occupancies; it does not bound error from estimating q_0, U, or P, nor does this paper supply a finite-sample calibration rule for the robust reward set.

4.3 Finite-sample policy recovery

The population results above say which reward directions are visible with unlimited data. We now ask how many negative trajectories are needed to recover a good policy in finite samples using (3). For a matrix or linear operator A, the notation $A|_U$ means that we evaluate it only on vectors in U. Assume $\dim(U) \geq 1$, $\Pi_\mathcal{D} \neq \varnothing$, and $\tau_1, \ldots, \tau_n$ are independent draws from $q_{\theta_*}^-$. This theorem assumes the

true parameter itself lies in $\Theta \cap U$; it therefore does not cover a true reward with an unidentified component in $\mathcal{N}$. We also assume Θ_U is compact and convex and each trajectory feature vector has norm at most B, so an individual observation cannot have arbitrarily large influence. Finally, we require the failure features to retain enough variation throughout the candidate parameter set:

$$\lambda_{\min}\Big(\mathrm{Cov}_{q_\theta^-}[\Phi]|_U\Big) \geq \lambda > 0 \quad \text{for every } \theta \in \Theta_U.$$

Here $\lambda_{\min}$ is the smallest eigenvalue, so $\lambda > 0$ lower-bounds the feature covariance in every direction the model claims to identify. Equivalently, the negative log-likelihood cannot become arbitrarily flat on U. This is stronger than merely knowing that a parameter is identifiable at the true reward. For this theorem choose π_* to be a deterministic covered optimum, which exists in the finite supported MDP. To translate reward-estimation error into a policy guarantee, define $L = \sup_{\pi,\pi' \in \Pi_{\mathcal{D}}^{\mathrm{det}}} \|\mu_\pi - \mu_{\pi'}\|_2$, the largest feature-occupancy difference between deterministic covered policies, and the smallest positive deterministic policy gap

$$\Delta_* = \min_{\substack{\pi \in \Pi_{\mathcal{D}}^{\mathrm{det}}: \\ J_{\theta_*}(\pi) < J_{\theta_*}(\pi_*)}} \theta_*^\top(\mu_{\pi_*} - \mu_\pi).$$

The gap Δ_* is the true reward advantage of an optimal policy over its closest suboptimal deterministic competitor, expressed before the common $(1-\gamma)^{-1}$ factor. If no deterministic covered policy is suboptimal, every covered policy is optimal and recovery is automatic. Otherwise $\Delta_* > 0$ and $L > 0$ because $\Pi_{\mathcal{D}}^{\mathrm{det}}$ is finite. There is generally no positive minimum gap over all randomized policies: mixtures with a vanishing weight on a suboptimal policy approach the optimal return.

Theorem 4 (From identification to policy recovery)**.** *For any confidence parameter $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\|\widehat{\theta}_n - \theta_*\|_2 \leq \frac{4\sqrt{2}B}{\beta\lambda}\sqrt{\frac{\dim(U)\log(2\dim(U)/\delta)}{n}}.$$

In the nontrivial case with a suboptimal deterministic covered policy, if the right-hand side is smaller than $\Delta_*/(2L)$, every policy optimal for $\widehat{\theta}_n$ is optimal for θ_*. A sufficient sample size for exact policy recovery therefore scales as

$$n = O\left(\frac{B^2L^2}{\beta^2\lambda^2\Delta_*^2}\dim(U)\log\frac{\dim(U)}{\delta}\right).$$

The dependence is interpretable: weak negative selection (small β), collinear failures (small λ), or a small decision margin makes identification statistically difficult.

Exact negative rationality is not required to preserve every conclusion. If the negative log-odds deviate uniformly by at most ε from the model, any trajectory comparison whose scaled return gap exceeds 2ε retains the correct orientation. Appendix A states this result formally (Proposition 4) and also develops an alternative statewise anti-Boltzmann model in which reference-corrected bad-action frequencies identify the best action (Theorem 6). These extensions reinforce the same point without changing the main theory: useful failures require known selection semantics and coverage of the relevant comparisons.

4.4 Discussion: extensions and implications

Three extensions delimit the core model. Negative–unlabeled mixtures require a known mixing proportion or an identifying restriction such as an anchor (Proposition 5). Occupancy repulsion selects behavior far from failures, not necessarily behavior with high task return (Theorem 7). Robust planning over a bounded reward version set separates residual reward ambiguity from transition support (Proposition 6). These alternatives identify different objects, and none removes the transition-coverage requirement. Appendix A gives the formal limitations, including the absence of a finite-sample calibration rule for the reward set.

5 Experiments

Questions and design. We test four predictions in controlled finite MDPs: (E1) centered feature rank governs recovery; (E2) policy-relevant alignment, not label sign, determines the useful channel; (E3) complementary failure sources improve identification; and (E4) semantic corruption and

transition-support loss are distinct. We compare negative-rational likelihood fitting with label-only, positive-data, source-allocation, occupancy-repulsion, behavior-cloning, and oracle baselines. Metrics include identifiable-subspace and full reward error, policy recovery or regret, and support loss; the selected-ridge penalty uses held-out negative likelihood, never the hidden reward. Appendices D–E give the complete protocol, intervals, and ablations. These synthetic tests isolate mechanisms; they are not evidence about real failure logs.

Representative sequential-control setting. E4b uses a 15×15 hazard gridworld with twelve absorbing hazards, two goals, $d = 20$ features, horizon 60, and slip probability 0.1. A single log of $3{,}000$ noisy-controller trajectories supplies both the estimated dynamics used by every method and the pool from which negatives are selected. We draw $n_- \in \{100, 500, 2000\}$ negatives under an exact tilt with $\beta = 8$, evaluate every learned policy on the true dynamics, and repeat the comparison over 50 seeds. Table 1 reports both reward-guided control and hazard violations; Appendix E adds the deliberately misspecified bottom-quantile channel.

Table 1: E4b, exact tilt channel, 50 seeds: normalized within-support regret and hazard violation rate. The selected-ridge penalty is chosen using held-out negative likelihood; oracle reward on $\widehat{P}$ is the model-error floor.

	within-support regret			violation rate		
n_-	100	500	2000	100	500	2000
Negative-rational MLE	0.289	0.200	0.047	0.096	0.070	0.041
+ selected ridge	0.103	0.041	**0.026**	0.043	0.037	**0.037**
Occupancy repulsion	0.075	0.075	0.075	0.038	0.038	0.038
Behavior cloning (complement)	0.135	0.115	0.100	0.096	0.070	0.050
Behavior cloning (log)	0.147	0.147	0.147	0.111	0.111	0.111
Bad-count (label-only)	0.472	0.480	0.488	0.210	0.188	0.203
Uniform covered policy	0.613	0.613	0.613	0.689	0.689	0.689
Oracle reward on $\widehat{P}$	0.017	0.017	0.017	0.035	0.035	0.035

Findings. E1's identifiable-subspace error decreases in all 18 cells (slopes -0.43 to -0.57), while full error retains the predicted null-space floor; at 2,500 failures, recovery is 0.94–1.00 when invisible directions are policy-irrelevant and zero in deficient-rank cases where they matter. E2 falsifies the pre-specified prediction that negatives lead in the collapse family in 7 of 8 decided cells, showing that information must align with the contrast needed for control. E3 improves projected error at budgets up to 500, with recovery differences vanishing at saturation, and E4 separates reward-model corruption from support loss. In hazard control, occupancy repulsion leads at 100 failures (regret 0.075), whereas selected-ridge likelihood fitting leads the non-oracle methods at 500 and 2,000 failures (0.041 and 0.026). The affine certificate has no false positives in the tested problems but certifies only 33.5% while empirical recovery reaches 95%, exposing its conservatism. Table 1 and Appendix Table 3 report the representative control comparison and the full set of outcomes and deviations, including the post-hoc avoidance family; these results are mechanism tests under specified channels, not a general benchmark claim.

6 Conclusion

Bad demonstrations are neither useless labels nor inverted expert data. In our framework, failure frequencies become useful for control when three conditions align: a known or sufficiently restricted selection mechanism turns relative frequency into reward comparisons, failures vary along policy feature contrasts, and the transition log supports the resulting policy. Under these conditions, centered feature geometry characterizes what the population negative channel reveals and guides complementary failure collection; without them, the justified output is an avoidance set, a reward version set, or abstention rather than a point reward or uniquely recommended policy.

Our theory makes these boundaries explicit, and the controlled experiments isolate their consequences. Appendix F discusses modeling and empirical scope, out-of-scope questions, and the resulting directions for future work.

Reproducibility statement

All assumptions and population objectives are stated in the main paper, with proofs in Appendix B. Appendices D–E report the experimental protocol, deviations, sample counts, uncertainty intervals, failures, and scope. Full reproduction additionally requires the experiment code, configurations, and seed-level outputs to accompany the submission artifact. Because all environments are synthetic, the manuscript should not be interpreted as a real-data benchmark.

Ethics statement.

Negative demonstrations may encode reporting bias, inequity, or confounding rather than genuine undesirability. Our assumptions expose but do not remove this risk; high-stakes use requires domain validation and must not treat adverse outcomes as causal action labels.

Artificial intelligence use statement

The authors conceived the research, developed the methodology, allocated and supervised research tasks, conducted the experiments, and independently verified the results. All scientific contributions, analyses, interpretations, and conclusions were developed and validated by the authors. Claude AI was used for supporting tasks, including literature searches and related-work suggestions, coding assistance, and manuscript editing to improve readability and check spelling, mathematical notation, LaTeX formatting, and consistency in the presentation of results. The authors reviewed and verified all AI-assisted material. Reported empirical findings are based on experiments actually conducted and, where applicable, cited, publicly available data—not on experimental data or results generated by AI. The authors take full responsibility for the paper's content. No AI tool is a co-author.

References

Daniel S. Brown, Yuchen Cui, and Scott Niekum. Risk-aware active inverse reinforcement learning. In *Proceedings of CoRL*, 2018.

Daniel S. Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. *Proceedings of ICML*, 2019.

Daniel S. Brown, Wonjoon Goo, and Scott Niekum. Better-than-demonstrator imitation learning via automatically-ranked demonstrations. In *Proceedings of CoRL*, 2020.

Haoyang Cao, Samuel N. Cohen, and Lukasz Szpruch. Identifiability in inverse reinforcement learning. *arXiv preprint arXiv:2106.03498*, 2021.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Dylan J. Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. In *Proceedings of COLT*, 2022.

Wenshuo Guo, Kumar Krishna Agrawal, Aditya Grover, Vidya K. Muthukumar, and Ashwin Pananjady. Learning from an exploring demonstrator: Optimal reward estimation for bandits. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pp. 6357–6386, 2022.

Michael Herman, Tobias Gindele, Jörg Wagner, Felix Schmitt, and Wolfram Burgard. Inverse reinforcement learning with simultaneous estimation of rewards and dynamics. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pp. 102–110, 2016.

Huy Hoang, Tien Mai, and Pradeep Varakantham. No experts, no problem: Avoidance learning from bad demonstrations. In *Advances in Neural Information Processing Systems*, 2025.

Youngsoo Jang, Geon-Hyeong Kim, Jongmin Lee, Sungryull Sohn, Byoungjip Kim, Honglak Lee, and Moontae Lee. SafeDICE: Offline safe imitation learning with non-preferred demonstrations. In *Advances in Neural Information Processing Systems*, 2023.

Yachen Kang, Diyuan Shi, Jinxin Liu, Li He, and Donglin Wang. Beyond reward: Offline preference-guided policy optimization. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 15753–15768, 2023.

Michael Kelly, Chelsea Sidrane, Katherine Driggs-Campbell, and Mykel J. Kochenderfer. HG-DAgger: Interactive imitation learning with human experts. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 8077–8083, 2019.

Kuno Kim, Shivam Garg, Kirankumar Shiragur, and Stefano Ermon. Reward identification in inverse reinforcement learning. In *Proceedings of ICML*, 2021.

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

Anqi Li, Byron Boots, and Ching-An Cheng. MAHALO: Unifying offline reinforcement learning and imitation learning from observations. In *Proceedings of ICML*, 2023.

David Lindner, Andreas Krause, and Giorgia Ramponi. Active exploration for inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022.

David Lindner, Xin Chen, Sebastian Tschiatschek, Katja Hofmann, and Andreas Krause. Learning safety constraints from demonstrations with unknown rewards. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, pp. 2386–2394, 2024.

Shehryar Malik, Usman Anwar, Alireza Aghasi, and Ali Ahmed. Inverse constrained reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 7390–7399, 2021.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14:265–294, 1978.

Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of ICML*, 2000.

Paul Rolland, Luca Viano, Norman Schuerhoff, Boris Nikolov, and Volkan Cevher. Identifiability and generalizability from multiple experts in inverse reinforcement learning. *arXiv preprint arXiv:2209.10974*, 2022.

Andreas Schlaginhaufen and Maryam Kamgarpour. Identifiability and generalizability in constrained inverse reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 30224–30251, 2023.

Rohin Shah, Noah Gundotra, Pieter Abbeel, and Anca Dragan. On the feasibility of learning, rather than assuming, human biases for reward inference. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 5670–5679, 2019.

Mohamad Louai Shehab, Antoine Aspeel, Nikos Aréchiga, Andrew Best, and Necmiye Ozay. Learning true objectives: Linear algebraic characterizations of identifiability in inverse reinforcement learning. In *Proceedings of L4DC*, 2024.

Prasanth Sengadu Suresh and Prashant Doshi. Marginal MAP estimation for inverse RL under occlusion with observer noise. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, pp. 1907–1916, 2022.

Yueh-Hua Wu, Nontawat Charoenphakdee, Han Bao, Voot Tangkaratt, and Masashi Sugiyama. Imitation learning from imperfect demonstration. In *Proceedings of ICML*, 2019.

Haoran Xu, Xianyuan Zhan, Honglei Yin, and Huiling Qin. Discriminator-weighted offline imitation learning from suboptimal demonstrations. In *Proceedings of ICML*, 2022.

Tianhe Yu, Aviral Kumar, Yevgen Chebotar, Karol Hausman, Chelsea Finn, and Sergey Levine. How to leverage unlabeled data in offline reinforcement learning. In *Proceedings of ICML*, 2022.

Lei Zhao, Mengdi Wang, and Yu Bai. Is inverse reinforcement learning harder than standard reinforcement learning? a theoretical perspective. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 60957–61020, 2024.

Brian D. Ziebart, Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of AAAI*, 2008.

A Additional theoretical results

This section collects extensions and technical consequences of the main identification theory. They support the claims summarized in Section 4, especially the discussion in Section 4.4, but are not needed for the central progression from observation semantics to reward and policy identification. Read the results in three groups: reference confounding, Fisher information, and source design explain *when failure frequencies reveal reward contrasts*; statewise and mixture models show what changes under *different observation semantics*; and repulsion and robust planning distinguish *avoiding failures* from recovering a high-return covered policy. None removes the separate requirement of transition coverage.

Table 2: Population-level interpretation of different negative-data models. The rows make different assumptions and do not form a progression in which more data automatically imply the next row.

Observation	What it can determine	Assumption needed for that conclusion
Suboptimality label only	The labeled choice is not best; unlabeled alternatives remain unordered.	The label reliably means that the particular choice is suboptimal; no selection-frequency model is assumed.
Negative trajectory frequencies	Reward contrasts in the span of centered trajectory features.	The exponential selection model, a known reference q_0, and known β for reward scale; directions in $\mathcal{N}$ remain invisible.
Bad-action frequencies by state	An optimal action ordering at observed states.	The statewise anti-Boltzmann model, known action reference, and coverage of the relevant states and actions.
Negative and unlabeled mixture	The desirable trajectory component, except in degenerate cases.	A known negative mixing proportion, or a valid negative anchor/irreducibility restriction.
Occupancy repulsion	The occupancy maximizing a separation score plus entropy, not necessarily task return.	The desired policy is recovered only if its occupancy uniquely maximizes that particular objective.

Table 2 distinguishes what an observation model identifies from what an optimization objective selects. The rows require different assumptions, and none removes the separate need for offline transition coverage.

A.1 Reference confounding and Fisher geometry

This subsection separates the role of the reference law from the role of feature variation. It first shows how an unrestricted reference can absorb any candidate reward, then characterizes the reward directions in which negative trajectories add information beyond a baseline data source.

Proposition 1 (Reference confounding)**.** *If q_0 is unrestricted and unobserved, a population bad-trajectory law q^- imposes no restriction on reward beyond the candidate set Θ: for every $\theta' \in \Theta$ there exists a full-support reference q_0' for which* (1) *produces exactly q^-. If q_0 is known but $\beta > 0$ is unknown, the law identifies the product $\vartheta = \beta\theta$ only modulo $\mathcal{N}$; it still identifies return orderings among supported trajectories because β is positive.*

Corollary 1 (Information is contrast)**.** *Assume the log partition function is twice differentiable near θ and the feature second moments are finite there (as on finite support). The Fisher information matrix of one bad trajectory is $F(\theta) := \beta^2 \mathrm{Cov}_{q_\theta^-}[\Phi(\tau)] \in \mathbb{R}^{d\times d}$, and $\ker F(\theta) = \mathcal{N}$. Bad data add identifying information precisely in feature directions along which supported bad trajectories vary.*

Let $I_0(\theta) \in \mathbb{R}^{d\times d}$ be the Fisher information in baseline observations, such as sparse rewards or preferences, and assume that these observations and the bad trajectories are conditionally independent given θ.

Proposition 2 (When bad data strictly help)**.** *After n bad trajectories, joint Fisher information is $I_n(\theta) = I_0(\theta) + nF(\theta)$, and*

$$\ker I_n(\theta) = \ker I_0(\theta) \cap \mathcal{N}.$$

Thus bad data strictly reduce the locally unidentified subspace if and only if there exists $v \in \ker I_0(\theta)$ such that $v^\top F(\theta)v > 0$. Let $\mathcal{G} \subseteq \mathbb{R}^d$ denote the linear subspace of accepted reward invariances. Under the identification convention of Definition 1, the combined data yield local reward identification up to $\mathcal{G}$ if and only if $\ker I_0(\theta) \cap \mathcal{N} \subseteq \mathcal{G}$. If every accepted invariance is itself observationally invisible, so $\mathcal{G} \subseteq \ker I_0(\theta) \cap \mathcal{N}$, this condition is equivalently equality.

A.2 Positive versus negative information

For comparison, let $q_\theta^+(\tau) \propto q_0(\tau)\exp\{+\beta\theta^\top\Phi(\tau)\}$ model positive demonstrations under the same reference and temperature. Assume both tilts have differentiable log partition functions and finite feature second moments near θ. Write $F_+(\theta)$ and $F_-(\theta) = F(\theta)$ for their per-sample Fisher information matrices.

Proposition 3 (Bad data can be more informative)**.** *For either sign, $F_\pm(\theta) = \beta^2\mathrm{Cov}_{q_\theta^\pm}[\Phi(\tau)]$. Therefore negative demonstrations are more informative than positive demonstrations in direction v exactly when*

$$\mathrm{Var}_{q_\theta^-}(v^\top\Phi(\tau)) > \mathrm{Var}_{q_\theta^+}(v^\top\Phi(\tau)).$$

This advantage can be exponential in β.

A.3 Optimal failure-source allocation

Suppose K known proposal mechanisms have finite information matrices $F_j(\theta)$. With $\mathbb{R}_+ := [0,\infty)$ and $\mathbf{1} \in \mathbb{R}^K$ the all-ones vector, let $\Delta_K = \{w \in \mathbb{R}_+^K : \mathbf{1}^\top w = 1\}$ be the allocation simplex, and let $\mathrm{Proj}_\mathcal{U}$ be the orthogonal projector onto a nonzero target subspace $\mathcal{U} \subseteq \mathbb{R}^d$.

Theorem 5 (Optimal failure-diversity design)**.** *For $w \in \Delta_K$, the per-sample information is*

$$F(w;\theta) = \sum_{j=1}^K w_j F_j(\theta), \qquad \ker F(w;\theta) = \bigcap_{j:w_j>0} \ker F_j(\theta).$$

Consequently, the design identifies every direction in $\mathcal{U}$ if and only if $\mathcal{U} \cap \bigcap_{j:w_j>0}\ker F_j(\theta) = \{0\}$. With $A \succeq B$ denoting that $A - B$ is positive semidefinite (the Löwner order), the locally E-optimal allocation at θ solves

$$\max_{w,t}\ t \quad \textit{s.t.} \quad \mathrm{Proj}_\mathcal{U}\left(\sum_{j=1}^K w_j F_j(\theta)\right)\mathrm{Proj}_\mathcal{U} \succeq t\,\mathrm{Proj}_\mathcal{U}, \quad w \geq 0, \quad \mathbf{1}^\top w = 1. \tag{5}$$

If $d_\mathcal{U} = \dim(\mathcal{U})$, an optimal allocation exists using at most $d_\mathcal{U}(d_\mathcal{U}+1)/2 + 1$ mechanisms.

For a hard budget on the number of mechanisms, fix θ, suppress it in $I_0(\theta)$ and $F_j(\theta)$, and define $G(S) = \mathrm{rank}(I_0 + \sum_{j\in S} F_j) - \mathrm{rank}(I_0)$.

Corollary 2 (Greedy failure-source selection)**.** *The rank gain G is normalized, monotone, and submodular. Under $|S| \leq k$, the classical greedy guarantee (Nemhauser et al., 1978) gives*

$$G(S_{\mathrm{gr}}) \geq (1 - 1/e)\max_{|S|\leq k} G(S).$$

The theorem allocates a fractional budget to maximize the weakest information direction on $\mathcal{U}$; the corollary instead closes as many directions as possible under a mechanism-count budget. Both are local geometry results: pilot error, incorrect selection semantics, finite-sample precision, and transition coverage remain separate.

Alternative statewise semantics. As a compact alternative to trajectory selection, let $\mathcal{A}_{\mathcal{D}}(s) := \{a : (s,a) \in \mathcal{B}_{\mathcal{D}}\}$, let $Q^{*,\mathcal{D}}_{\theta_*}$ denote the optimal action value with future choices restricted to supported actions, and suppose

$$\pi^-(a \mid s) = \frac{\mu_{\mathrm{ref}}(a \mid s)\exp\{-\beta Q^{*,\mathcal{D}}_{\theta_*}(s,a)\}}{\sum_{b\in\mathcal{A}_{\mathcal{D}}(s)} \mu_{\mathrm{ref}}(b \mid s)\exp\{-\beta Q^{*,\mathcal{D}}_{\theta_*}(s,b)\}}, \tag{6}$$

where the supported-action reference $\mu_{\mathrm{ref}}(a \mid s) > 0$ is known and $\beta > 0$ may be unknown.

Theorem 6 (Optimal actions from bad-action frequencies)**.** *At every state for which $\pi^-(\cdot \mid s)$ is observed on all of $\mathcal{A}_{\mathcal{D}}(s)$,*

$$\operatorname*{argmax}_{a\in\mathcal{A}_{\mathcal{D}}(s)} Q^{*,\mathcal{D}}_{\theta_*}(s,a) = \operatorname*{argmin}_{a\in\mathcal{A}_{\mathcal{D}}(s)} \frac{\pi^-(a \mid s)}{\mu_{\mathrm{ref}}(a \mid s)}.$$

Thus the optimal action is point identified at that state whenever the corrected bad-action likelihood has a unique minimizer. Covered-optimal behavior from d_0 is identified if this condition holds at every state reachable under the resulting policy; actions at unreachable states remain irrelevant and need not be identified. The conclusion does not require knowing β, but it requires the reference correction when μ_{ref} is nonuniform.

Corrected frequencies therefore give a second route to a policy without recovering trajectory-level reward. This is an alternative model, not a consequence of (1), and it still requires state and transition coverage.

Approximate negative selection. Suppose for some scalar offset b_0,

$$\left|\log \frac{q^-(\tau)}{q_0(\tau)} + \beta\theta_*^\top \Phi(\tau) - b_0\right| \le \varepsilon \quad \text{for every } \tau \in \mathcal{T}_-. \tag{7}$$

Proposition 4 (Orientation is robust up to the gap)**.** *For any $\tau_i, \tau_j \in \mathcal{T}_-$, the ordering inferred from negative log-odds agrees with the true return ordering whenever $\beta|\theta_*^\top(\Phi(\tau_i) - \Phi(\tau_j))| > 2\varepsilon$. If the link between selection and return is unknown even in orientation, no such guarantee is possible.*

This is only a pairwise orientation guarantee: it does not identify the full reward or excuse an unknown selection sign.

A.4 Boundary extensions

This subsection considers three boundary cases in which the clean negative-trajectory model is insufficient: mixed unlabeled data, objectives that repel a policy from failures, and residual reward ambiguity at planning time. The results clarify which object each extension recovers and which assumptions remain necessary.

Negative plus unlabeled mixtures. Suppose an unlabeled trajectory law satisfies

$$m = \alpha q^- + (1-\alpha)q^{\mathrm{des}}, \qquad 0 < \alpha < 1. \tag{8}$$

Proposition 5 (Decontamination and anchors)**.** *If α is known in* (8)*, then $q^{\mathrm{des}} = (m - \alpha q^-)/(1-\alpha)$ is identified wherever the mixture equation holds. If α is unknown, define*

$$\alpha_{\max} = \inf_{\tau : q^-(\tau) > 0} \frac{m(\tau)}{q^-(\tau)}.$$

Then every $\alpha \in (0, \alpha_{\max}] \cap (0,1)$ produces the valid candidate $q^{\mathrm{des}}_\alpha := (m - \alpha q^-)/(1-\alpha)$. The mixing proportion is not identified; the desirable component also varies with α unless $m = q^-$, in which degenerate case it equals q^- for every feasible α. In the discrete case, the negative-anchor condition $q^{\mathrm{des}}(\tau_a) = 0 < q^-(\tau_a)$ for some τ_a identifies $\alpha = m(\tau_a)/q^-(\tau_a) = \alpha_{\max} < 1$.

Occupancy repulsion. Several negative-demonstration methods choose an occupancy far from failures (Jang et al., 2023; Hoang et al., 2025). Let $\mathcal{P}_{\mathcal{D}} = \{\rho_\pi : \pi \in \Pi_{\mathcal{D}}\}$ and consider

$$\inf_{\rho \in \mathcal{P}_{\mathcal{D}}} \inf_{r} \left\{ \langle \rho^- - \rho, r \rangle - \mathcal{H}(\rho) + \psi(r) \right\}. \tag{9}$$

Here ρ^- is normalized bad occupancy, $\mathcal{H}$ is discounted causal entropy, ψ is closed, proper, and convex, and $\psi^*(x) = \sup_r \{\langle x, r\rangle - \psi(r)\}$.

Theorem 7 (Repulsion equals farthest-feasible occupancy)**.** *Assume $\psi^*(\rho - \rho^-)$ is finite on $\mathcal{P}_{\mathcal{D}}$ and the following maximum is attained. The occupancies of policy solutions of* (9) *are exactly*

$$\operatorname*{argmax}_{\rho \in \mathcal{P}_{\mathcal{D}}} \left\{ \psi^*(\rho - \rho^-) + \mathcal{H}(\rho) \right\}.$$

Therefore a desired occupancy ρ^{des} is uniquely selected *by* (9) *if and only if it is the unique maximizer above; this is a condition on the objective, not observational identification of task reward. For the quadratic regularizer $\psi(r) = \frac{\eta}{2}\|r\|_2^2$, with $\eta > 0$, the criterion is $\frac{1}{2\eta}\|\rho - \rho^-\|_2^2 + \mathcal{H}(\rho)$.*

Planning over a reward version set. For a bounded set $\widehat{\Theta}$ of rewards compatible with the negative data, choose

$$\widehat{\pi} \in \operatorname*{argmax}_{\pi \in \Pi_{\mathcal{D}}} \min_{\theta \in \widehat{\Theta}} \theta^\top \mu_\pi. \tag{10}$$

Proposition 6 (Planning loss from identification radius)**.** *Assume $\theta_* \in \widehat{\Theta}$ and $\sup_{\theta \in \widehat{\Theta}} \|\theta - \theta_*\|_2 \le \varepsilon_\theta$. If $\|\mu_\pi\|_2 \le R_\mu$ for all $\pi \in \Pi_{\mathcal{D}}$, then the robust solution in* (10) *satisfies*

$$J_{\theta_*}(\pi_*) - J_{\theta_*}(\widehat{\pi}) \le \frac{2R_\mu \varepsilon_\theta}{1-\gamma}.$$

With an estimated transition model, the usual simulation error adds separately; negative labels cannot remove that term without transition coverage.

Together these auxiliary results delimit the main theory. Reference correction and feature variation determine whether failures add reward information; label sign alone does not determine information; complementary sources can close null directions; and bounded misspecification preserves only sufficiently separated comparisons. Mixture subtraction, occupancy repulsion, and robust planning answer different questions and do not remove the need for transition coverage. The robust bound assumes a calibrated set is already available; this paper does not construct one from finite bad data.

B PROOFS

This appendix supplies the proofs of the main identification, control, and finite-sample results, followed by proofs of the auxiliary propositions in Appendix A. We begin with an occupancy fact used repeatedly in the control arguments.

We use a standard occupancy fact in the control arguments. For the finite discounted MDP, covered normalized occupancies are exactly the nonnegative solutions of the discounted flow equations

$$\sum_a \rho(s,a) = (1-\gamma) d_0(s) + \gamma \sum_{s',a'} P(s \mid s', a') \rho(s', a'), \qquad \rho(s,a) = 0 \text{ if } (s,a) \notin \mathcal{B}_{\mathcal{D}}.$$

This is a compact convex polytope. Every feasible occupancy induces a stationary covered policy by normalizing its action masses at visited states. For any linear reward, discounted dynamic programming with supported actions has a deterministic covered optimum. If the polytope contained a point outside the convex hull of deterministic covered occupancies, a separating linear reward would have a strictly larger maximum over the polytope than over that hull, contradicting deterministic optimality. Thus the covered occupancy polytope is their convex hull; comparisons with all deterministic covered policies suffice for linear-return claims.

B.1 Proof of Theorem 1

The impossibility result is established by constructing two rewards that generate the same label-only observation but have different optimal actions, and then extending this indistinguishability argument to the general case.

Let the three actions be a_0, a_1, a_2. Consider the two reward functions

	a_0	a_1	a_2
r_1	-1	1	0
r_2	-1	0	1

which both belong to the stated reward class. A label-only observation carries the constraint that the demonstrated action a_0 is not optimal; it carries neither the probability with which a_0 was selected nor a comparison between a_1 and a_2. Both rewards therefore produce exactly the same observation, even if the labeled action is repeated infinitely often. Under r_1, however, a_1 is uniquely optimal, whereas under r_2, a_2 is uniquely optimal. Hence no estimator measurable with respect to the labels can identify the optimal action in both models.

For the general statement, take the two stipulated observationally equivalent rewards whose sets of covered optimal policies differ. Any estimator based only on the common observation must return the same answer in both instances and therefore cannot identify both optimal-policy sets. Merely reversing the ranking of two alternatives would not prove this if a third policy remained optimal under both rewards. □

B.2 Proof of Theorem 2

The proof converts equality of negative-trajectory laws into equality of centered feature contrasts. This yields the null-space characterization in both directions.

Suppose first that $q_\theta^- = q_{\theta'}^-$. Because $q_0(\tau) > 0$ on $\mathcal{T}_-$, taking the logarithm of the ratio of the two densities is valid. For every supported τ, equality gives

$$-\beta(\theta - \theta')^\top \Phi(\tau) = \log Z_-(\theta) - \log Z_-(\theta').$$

The right-hand side is independent of τ. Subtracting the corresponding equation for τ_0 and using $\beta > 0$ yields

$$(\theta - \theta')^\top \big(\Phi(\tau) - \Phi(\tau_0)\big) = 0 \quad \text{for every } \tau \in \mathcal{T}_-.$$

Thus $\theta - \theta'$ is orthogonal to the span M, so $\theta - \theta' \in \mathcal{N}$.

Conversely, let $u = \theta - \theta' \in \mathcal{N}$. Orthogonality to every centered feature implies $u^\top \Phi(\tau) = u^\top \Phi(\tau_0) =: c$ for all supported τ. Hence

$$q_0(\tau) e^{-\beta\theta^\top \Phi(\tau)} = e^{-\beta c} q_0(\tau) e^{-\beta\theta'^\top \Phi(\tau)} \quad \text{and} \quad Z_-(\theta) = e^{-\beta c} Z_-(\theta').$$

The common factor cancels after normalization, proving $q_\theta^- = q_{\theta'}^-$.

The observational equivalence class of θ is therefore exactly $(\theta + \mathcal{N}) \cap \Theta$, proving the point-identification criterion for an arbitrary parameter set. If the parameter set is the full normalized affine space $\bar{\theta} + V$, its equivalence class at any point is a singleton exactly when $V \cap \mathcal{N} = \{0\}$, or equivalently when centered features have full rank on V. This equivalence need not hold for an arbitrarily restricted or curved Θ. Finally, policy identification requires every parameter in each observational equivalence class to induce the same set of covered optimal policies, giving the third claim. □

Proof of Proposition 1. Because q^- arises under a full-support reference, it is positive on $\mathcal{T}_-$. Fix any candidate θ' and positive β. In the finite discounted MDP, Φ is uniformly bounded by $\max_{s,a} \|\phi(s,a)\|_2/(1-\gamma)$; hence $C := \sum_{\tau'} q^-(\tau') \exp\{\beta\theta'^\top \Phi(\tau')\}$ is finite and positive. Define

$$q_0'(\tau) = \frac{q^-(\tau) \exp\{\beta\theta'^\top \Phi(\tau)\}}{\sum_{\tau'} q^-(\tau') \exp\{\beta\theta'^\top \Phi(\tau')\}}.$$

The numerator is positive and the denominator is $C \in (0, \infty)$, so q_0' is a full-support probability law. Substitution into the unnormalized negative law gives

$$q_0'(\tau) e^{-\beta \theta'^\top \Phi(\tau)} = q^-(\tau)/C.$$

Summing over τ shows that the corresponding partition function is $1/C$, and normalization therefore recovers exactly q^-. Since this construction works for every $\theta' \in \Theta$, the observation rules out no candidate reward when the reference is unrestricted.

When q_0 is known but β is not, both the exponent and the partition function depend on the pair (β, θ) only through $\vartheta = \beta\theta$. Applying Theorem 2 with unit temperature shows that the law determines ϑ only up to $\mathcal{N}$. In particular, (β, θ) and $(\beta/c, c\theta)$ are observationally equivalent for every $c > 0$ allowed by the parameter space. The supported trajectory return differences are $\vartheta^\top(\Phi(\tau) - \Phi(\tau'))/\beta$; division by an unknown positive β changes their scale but not their signs. □

Proof of Corollary 1. Differentiating the log density gives

$$\nabla_\theta \log q_\theta^-(\tau) = -\beta\Phi(\tau) - \nabla_\theta \log Z_-(\theta) = -\beta\big(\Phi(\tau) - \mathbb{E}_{q_\theta^-}[\Phi]\big),$$

where $\nabla_\theta \log Z_-(\theta) = -\beta \mathbb{E}_{q_\theta^-}[\Phi]$. Taking the expected outer product of the score yields $F(\theta) = \beta^2 \mathrm{Cov}_{q_\theta^-}[\Phi]$.

For any $v \in \mathbb{R}^d$,

$$v^\top F(\theta) v = \beta^2 \mathrm{Var}_{q_\theta^-}(v^\top \Phi).$$

This quantity is zero if and only if $v^\top \Phi(\tau)$ is constant almost surely. The negative law has full support on $\mathcal{T}_-$, so "almost surely" is equivalent to "for every supported trajectory." Subtracting the value at τ_0 shows that this is equivalent to $v \perp M$, or $v \in \mathcal{N}$. □

B.3 Proof of Theorem 3

The proof reduces uniform optimality over the affine reward version set to whether every policy-feature contrast is orthogonal to the invisible reward subspace. A separating invisible direction proves necessity, while orthogonality proves sufficiency.

Fix a deterministic competitor π and write $g_\pi = \mu_{\pi^\dagger} - \mu_\pi$. Since the factor $(1-\gamma)^{-1}$ in return is positive, $\pi^\dagger$ weakly dominates π for every reward in $\mathcal{V}$ exactly when

$$(\bar{\theta} + v)^\top g_\pi \geq 0 \qquad \text{for every } v \in \mathcal{N}.$$

Suppose $\mathrm{Proj}_{\mathcal{N}} g_\pi \neq 0$. Taking $v = -\xi \,\mathrm{Proj}_{\mathcal{N}} g_\pi$ gives

$$(\bar{\theta} + v)^\top g_\pi = \bar{\theta}^\top g_\pi - \xi \| \mathrm{Proj}_{\mathcal{N}} g_\pi \|_2^2,$$

which is negative for sufficiently large $\xi > 0$. Common optimality is therefore impossible. Conversely, if $\mathrm{Proj}_{\mathcal{N}} g_\pi = 0$, then $v^\top g_\pi = 0$ for all $v \in \mathcal{N}$, and the inequality reduces to $\bar{\theta}^\top g_\pi \geq 0$.

Applying this equivalence to every $\pi \in \Pi_{\mathcal{D}}^{\mathrm{det}}$ proves the stated test against deterministic competitors. The covered discounted-occupancy set is the convex hull of deterministic covered-policy occupancies, and return is linear in occupancy, so these comparisons also establish optimality against every randomized covered policy. If the second inequality is strict for each distinct deterministic competitor, $\pi^\dagger$ strictly dominates every other deterministic covered policy for every $\theta \in \mathcal{V}$, proving the stated deterministic uniqueness. If any competitor attains equality, it is also optimal whenever $\pi^\dagger$ is, so common deterministic uniqueness fails. □

Proof of Proposition 2. Conditional independence makes log likelihoods, and hence Fisher information matrices, additive. The baseline contributes $I_0(\theta)$, while n independent bad trajectories contribute $nF(\theta)$, giving $I_n(\theta) = I_0(\theta) + nF(\theta)$.

For any positive-semidefinite matrix A, $v^\top A v = 0$ if and only if $Av = 0$, as follows from a square-root factorization $A = A^{1/2}A^{1/2}$. Therefore

$$v \in \ker(I_0 + nF) \iff v^\top I_0 v + n v^\top F v = 0 \iff v \in \ker I_0 \cap \ker F.$$

Corollary 1 gives $\ker F = \mathcal{N}$. The intersection is strictly smaller than $\ker I_0$ exactly when some $v \in \ker I_0$ has $v^\top F v > 0$. Under Definition 1, local identification up to $\mathcal{G}$ requires the remaining kernel to be contained in $\mathcal{G}$. If $\mathcal{G}$ is known to consist of observational invariances, the reverse inclusion also holds and the containment becomes equality. □

Proof of Proposition 3. Introduce a sign $s \in \{-1, +1\}$ and write $q_\theta^s(\tau) \propto q_0(\tau) e^{s\beta\theta^\top \Phi(\tau)}$. The score is $s\beta(\Phi - \mathbb{E}_{q_\theta^s}\Phi)$. Squaring removes the sign, so

$$F_s(\theta) = \beta^2 \mathrm{Cov}_{q_\theta^s}[\Phi], \qquad v^\top F_s(\theta) v = \beta^2 \mathrm{Var}_{q_\theta^s}(v^\top \Phi).$$

The directional comparison follows immediately.

For completeness, consider the uniform-reference construction $\Phi \in \{0, 1, 10\}$ and $\theta = 1$. Under the negative tilt, the unnormalized weights are $(1, e^{-\beta}, e^{-10\beta})$. The mass at feature 1 is of order $e^{-\beta}$, so the leading contribution to the variance is $\Theta(e^{-\beta})$. Under the positive tilt, divide the weights $(1, e^{\beta}, e^{10\beta})$ by $e^{10\beta}$. The mass at feature 1 is then of order $e^{-9\beta}$, while the mass at feature 0 is smaller, of order $e^{-10\beta}$; hence the leading variance is $\Theta(e^{-9\beta})$. Multiplication by the common β^2 gives $F_-/F_+ = \Theta(e^{8\beta})$. □

B.4 Proof of Theorem 4

The finite-sample argument combines concentration of the likelihood gradient with uniform curvature on the identifiable subspace, then translates reward error into exact policy recovery through the deterministic policy margin.

Write $d_U = \dim(U)$, let I_U be the identity operator on U, and choose an orthonormal basis for U. Up to an additive constant independent of θ, the empirical negative log-likelihood is

$$\widehat{\ell}_n(\theta) = \frac{\beta}{n} \sum_{i=1}^{n} \theta^\top \Phi(\tau_i) + \log Z_-(\theta).$$

Its gradient and Hessian, restricted to U, are

$$\nabla \widehat{\ell}_n(\theta) = \beta \left(\frac{1}{n} \sum_{i=1}^{n} \Phi(\tau_i) - \mathbb{E}_{q_\theta^-}[\Phi] \right), \qquad \nabla^2 \widehat{\ell}_n(\theta) = \beta^2 \mathrm{Cov}_{q_\theta^-}[\Phi]|_U.$$

The curvature assumption therefore implies $\nabla^2 \widehat{\ell}_n(\theta) \succeq \beta^2 \lambda I_U$ throughout Θ_U.

At θ_*, the gradient is an average of independent centered vectors. Since $\|\Phi\|_2 \le B$, each coordinate of $\Phi - \mathbb{E}\Phi$ has range at most $4B$. Applying Hoeffding's inequality to each of the d_U coordinates and a union bound gives, with probability at least $1 - \delta$,

$$\|\nabla \widehat{\ell}_n(\theta_*)\|_2 \le \beta B \sqrt{8 d_U \log(2 d_U/\delta)/n}.$$

Let $h = \widehat{\theta}_n - \theta_*$. Strong convexity along the segment joining the two feasible points gives

$$\widehat{\ell}_n(\widehat{\theta}_n) \ge \widehat{\ell}_n(\theta_*) + \nabla \widehat{\ell}_n(\theta_*)^\top h + \frac{\beta^2 \lambda}{2} \|h\|_2^2.$$

Because $\widehat{\theta}_n$ minimizes the constrained objective, the left-hand side is no larger than $\widehat{\ell}_n(\theta_*)$. Cauchy–Schwarz then yields $(\beta^2\lambda/2)\|\widehat{\theta}_n - \theta_*\|_2^2 \le \|\nabla \widehat{\ell}_n(\theta_*)\|_2 \|\widehat{\theta}_n - \theta_*\|_2$, and division by $\|h\|_2$ (with the zero case immediate) proves the stated parameter bound.

It remains to connect parameter error to control. If every deterministic covered policy is truly optimal, every covered occupancy is a convex combination of their occupancies and the conclusion is immediate. Otherwise, for any suboptimal deterministic π,

$$\widehat{\theta}_n^\top (\mu_{\pi_*} - \mu_\pi) \ge \theta_*^\top (\mu_{\pi_*} - \mu_\pi) - \|\widehat{\theta}_n - \theta_*\|_2 \|\mu_{\pi_*} - \mu_\pi\|_2 \ge \Delta_* - L\|h\|_2.$$

Under the stated condition this gap remains strictly positive. Thus no truly suboptimal deterministic policy can maximize estimated return. The supported discounted-occupancy polytope is the convex hull of deterministic covered-policy occupancies, and return is linear in occupancy; every estimated optimal occupancy is therefore a convex combination of estimated-optimal deterministic occupancies. All of these are truly optimal, so every estimated optimal covered policy is truly optimal. Solving the error inequality for n gives the displayed sufficient sample-size scaling. □

Proof of Theorem 6. Fix a state s at which the bad-action distribution is observed. Positivity of $\mu_{\mathrm{ref}}(a \mid s)$ makes the reference-corrected likelihood well defined for every supported action. For any supported a, b, equation (6) gives

$$\log \frac{\pi^-(a \mid s)/\mu_{\mathrm{ref}}(a \mid s)}{\pi^-(b \mid s)/\mu_{\mathrm{ref}}(b \mid s)} = -\beta\big(Q^{*,\mathcal{D}}_{\theta_*}(s,a) - Q^{*,\mathcal{D}}_{\theta_*}(s,b)\big).$$

Because $\beta > 0$ and the logarithm is increasing, the corrected likelihood and covered optimal Q-value have exactly opposite orderings on $\mathcal{A}_{\mathcal{D}}(s)$. Hence every maximizer of $Q^{*,\mathcal{D}}_{\theta_*}(s,\cdot)$ is a minimizer of $\pi^-(\cdot \mid s)/\mu_{\mathrm{ref}}(\cdot \mid s)$, and conversely. A unique corrected-likelihood minimizer therefore identifies the unique covered-optimal action at s.

If the corrected law is observed and has a unique minimizer at every state reachable under the resulting policy, choosing that action at each such state satisfies the *restricted* Bellman optimality condition on the entire reachable set and yields covered-optimal behavior from d_0. The policy may remain unspecified at unreachable states, which does not affect its occupancy or return from d_0. The value of β affects the magnitude of each log ratio but not its sign, so it need not be known. Without division by μ_{ref}, proposal frequency remains in the ratio and can reverse the ordering. □

Proof of Theorem 5. If a total of n independent trajectories is allocated with $n_j/n = w_j$ to source j, additivity gives total information $\sum_j n_j F_j$; division by n gives $F(w;\theta) = \sum_j w_j F_j(\theta)$.

Each F_j is positive semidefinite. Consequently, for any v,

$$v^\top F(w;\theta) v = \sum_{j:w_j>0} w_j v^\top F_j(\theta) v = 0$$

if and only if every term is zero. As in the proof of Proposition 2, a zero quadratic form is equivalent to membership in the kernel. This proves $\ker F(w;\theta) = \bigcap_{j:w_j>0} \ker F_j(\theta)$. The design identifies $\mathcal{U}$ exactly when no nonzero vector in $\mathcal{U}$ belongs to this intersection.

For a unit vector $v \in \mathcal{U}$, the directional information is $v^\top F(w;\theta)v$. Requiring the projected matrix to dominate $t\,\mathrm{Proj}_{\mathcal{U}}$ is equivalent to requiring this quantity to be at least t for every such v. Maximizing t therefore maximizes the smallest eigenvalue on $\mathcal{U}$. The constraint is affine in (w,t), so (5) is a semidefinite program.

Finally, after choosing an orthonormal basis of $\mathcal{U}$, each projected F_j is a point in the $d_{\mathcal{U}}(d_{\mathcal{U}}+1)/2$-dimensional vector space of symmetric matrices. Any feasible information matrix is in the convex hull of these points. Carathéodory's theorem represents the information matrix of an optimal allocation using at most $d_{\mathcal{U}}(d_{\mathcal{U}}+1)/2+1$ sources. Because the objective depends on w only through that matrix, the sparse representation is also optimal. □

Proof of Corollary 2. For positive-semidefinite matrices, $\ker(A+B) = \ker A \cap \ker B$; taking orthogonal complements yields $\mathrm{range}(A+B) = \mathrm{range}(A) + \mathrm{range}(B)$. Define

$$\mathcal{R}_S := \mathrm{range}(I_0) + \sum_{j\in S} \mathrm{range}(F_j).$$

Then $G(S) = \dim(\mathcal{R}_S) - \dim(\mathcal{R}_\varnothing)$. Normalization follows from $G(\varnothing) = 0$, and monotonicity follows because adding a source can only enlarge $\mathcal{R}_S$.

For submodularity, take $S \subseteq T$ and a source $j \notin T$. The marginal gain is the dimension of the part of $\mathrm{range}(F_j)$ not already contained in $\mathcal{R}_S$. Since $\mathcal{R}_S \subseteq \mathcal{R}_T$, this missing dimension cannot be larger after conditioning on T:

$$G(S\cup\{j\}) - G(S) \geq G(T\cup\{j\}) - G(T).$$

Thus G has diminishing returns. The classical result for greedy maximization of a normalized monotone submodular function under a cardinality constraint gives the $(1-1/e)$ guarantee (Nemhauser et al., 1978). □

Proof of Proposition 5. If α is known, rearranging (8) gives the only possible desirable component, $q^{\mathrm{des}} = (m - \alpha q^-)/(1-\alpha)$. Since the observed distributions satisfy the mixture equation, this candidate is nonnegative and normalized.

Now let α be unknown. For any proposed value $a \in (0,1)$, define $q_a^{\mathrm{des}} = (m - aq^-)/(1-a)$. At a point where $q^-(\tau) > 0$, nonnegativity is equivalent to $a \le m(\tau)/q^-(\tau)$; where $q^-(\tau) = 0$, it holds automatically. Hence $q_a^{\mathrm{des}} \ge 0$ exactly when $a \le \alpha_{\max}$. Moreover,

$$\sum_\tau q_a^{\mathrm{des}}(\tau) = \frac{\sum_\tau m(\tau) - a\sum_\tau q^-(\tau)}{1-a} = 1,$$

so every $a \in (0, \alpha_{\max}] \cap (0,1)$ yields a valid and observationally equivalent decomposition. For two distinct feasible values a, b, equality $q_a^{\mathrm{des}} = q_b^{\mathrm{des}}$ implies $(a-b)(m-q^-) = 0$, so the component changes with the mixing weight unless $m = q^-$. The weight itself remains unidentified in either case.

Under the anchor condition, $q^{\mathrm{des}}(\tau_a) = 0$, so the mixture equation gives $m(\tau_a) = \alpha q^-(\tau_a)$. Thus $\alpha = m(\tau_a)/q^-(\tau_a)$. For all other τ, nonnegativity of the true desirable component implies $m(\tau)/q^-(\tau) \ge \alpha$, so the anchor attains the infimum and $\alpha = \alpha_{\max}$. □

Proof of Proposition 4. Define the residual

$$e(\tau) := \log \frac{q^-(\tau)}{q_0(\tau)} + \beta\theta_*^\top \Phi(\tau) - b_0, \qquad |e(\tau)| \le \varepsilon.$$

Subtracting the identities for τ_i and τ_j cancels b_0 and yields

$$\log \frac{q^-(\tau_i)/q_0(\tau_i)}{q^-(\tau_j)/q_0(\tau_j)} = -\beta\theta_*^\top\big(\Phi(\tau_i) - \Phi(\tau_j)\big) + e(\tau_i) - e(\tau_j).$$

The residual difference has magnitude at most 2ε. If the scaled return gap is larger, this perturbation cannot change its sign, proving orientation recovery. If orientation is unspecified, a decreasing link applied to θ_* and an increasing link applied to $-\theta_*$ produce the same ordering of observed probabilities while assigning opposite meanings to desirable return; no data-only rule can distinguish them. □

Proof of Theorem 7. For fixed ρ, the inner objective involving r is

$$\langle \rho^- - \rho, r\rangle + \psi(r) = -\big(\langle \rho - \rho^-, r\rangle - \psi(r)\big).$$

By the definition of the convex conjugate, its infimum over r is $-\psi^*(\rho - \rho^-)$. Substitution leaves

$$\inf_{\rho \in \mathcal{P}_\mathcal{D}} \big[-\psi^*(\rho - \rho^-) - \mathcal{H}(\rho)\big],$$

whose minimizers are exactly the maximizers of $\psi^*(\rho - \rho^-) + \mathcal{H}(\rho)$ under the stated finiteness and attainment assumptions. Therefore the original objective uniquely selects ρ^{des} precisely when that occupancy is the unique maximizer; duality alone supplies no connection between geometric distance from ρ^- and task optimality.

For $\psi(r) = \eta\|r\|_2^2/2$, complete the square:

$$\langle x, r\rangle - \frac{\eta}{2}\|r\|_2^2 = -\frac{\eta}{2}\left\|r - \frac{x}{\eta}\right\|_2^2 + \frac{1}{2\eta}\|x\|_2^2.$$

The supremum is attained at $r = x/\eta$, so $\psi^*(x) = \|x\|_2^2/(2\eta)$, giving the quadratic specialization. □

Proof of Proposition 6. Write $V_{\min}(\pi) = \inf_{\theta \in \widehat{\Theta}} \theta^\top \mu_\pi$. Robust optimality gives $V_{\min}(\widehat{\pi}) \ge V_{\min}(\pi_*)$. For any $\pi \in \Pi_\mathcal{D}$ and any $\theta \in \widehat{\Theta}$, Cauchy–Schwarz gives

$$|\theta^\top \mu_\pi - \theta_*^\top \mu_\pi| \le \|\theta - \theta_*\|_2 \|\mu_\pi\|_2 \le \varepsilon_\theta R_\mu.$$

Consequently, $|V_{\min}(\pi) - \theta_*^\top \mu_\pi| \le \varepsilon_\theta R_\mu$. Insert $V_{\min}(\pi_*)$ and $V_{\min}(\widehat{\pi})$ between the two true feature returns:

$$\begin{aligned}\theta_*^\top \mu_{\pi_*} - \theta_*^\top \mu_{\widehat{\pi}} &\le \big(\theta_*^\top \mu_{\pi_*} - V_{\min}(\pi_*)\big) + \big(V_{\min}(\widehat{\pi}) - \theta_*^\top \mu_{\widehat{\pi}}\big) \\ &\le 2R_\mu \varepsilon_\theta,\end{aligned}$$

where the robust-optimality inequality makes the omitted middle term nonpositive. Dividing by $1-\gamma$ converts normalized feature return to J_{θ_*}. Dynamics error is separate because this argument holds for fixed occupancies; replacing them by occupancies under an estimated model requires an additional simulation bound. □

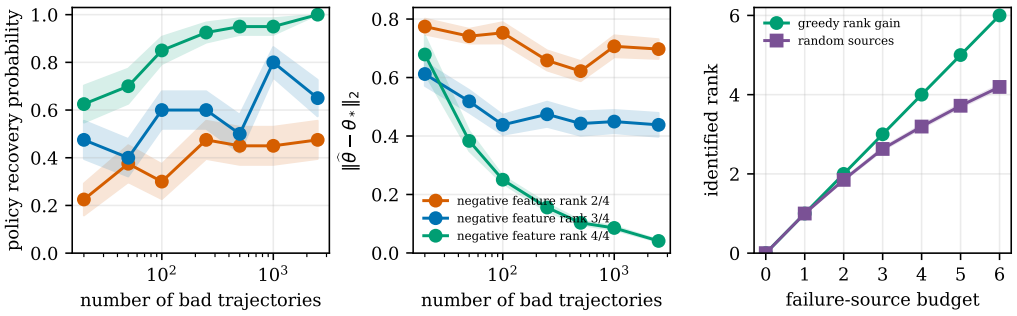

Figure 1: Bad-data diversity controls identification. Left and center: downstream policy recovery and parameter error over 40 seeded problems; bands show one standard error. Curves lacking feature directions have an error floor. Right: greedy failure-source selection closes independent directions faster than random selection (200 trials).

C Illustrative diagnostics

This appendix uses minimal examples and a synthetic rank experiment to make the main identification mechanisms concrete. The diagnostics separately expose reference correction, the rank phase transition, and the benefit of complementary failure sources.

A one-state diagnostic. For a bandit with reference $q_0(a) > 0$, write $r(a)$ for the reward of action a. Equation (1) implies

$$r(a) - r(b) = -\beta^{-1} \log \frac{q^-(a)/q_0(a)}{q^-(b)/q_0(b)}.$$

Hence the best action is the action least likely under the bad demonstrator after correcting for the reference distribution. Labels without frequencies do not yield this conclusion.

Rank phase transition. If bad trajectories vary in only $k < d$ independent feature directions, Theorem 2 leaves a $(d-k)$-dimensional ambiguity. Adding more samples of the same trajectories reduces variance but not that dimension. A new failure mode helps only when its centered feature vector increases rank or reduces ambiguity among optimal policies.

Synthetic verification. Figure 1 samples bad trajectories from (1) and fits the maximum-likelihood reward. The downstream planner chooses among candidates whose features need not occur in the negative-selection support. Full-rank negative variation yields consistent parameter and policy recovery; missing feature directions create an error floor that more samples cannot remove. The experiment is diagnostic rather than a benchmark and is reproduced by the supplied script.

D Detailed experimental protocol

This appendix records the evaluation protocol fixed before the experiments in Appendix E were run. It distinguishes confirmatory tests from the explicitly post-hoc E2 avoidance family. Deviations from this protocol are reported at the start of Appendix E.

D.1 Common protocol and reporting

For every environment, draw a transition log from a behavior policy and retain both its trajectories and one-step transitions. The empirical distribution over the retained trajectory pool is the reference law q_0 supplied to the learner. Define the hidden trajectory return as $R_{\theta_*}(\tau) := \theta_*^\top \Phi(\tau)$. When testing the correctly specified model, draw negative trajectories with replacement from this fixed pool using probabilities proportional to $\exp\{-\beta R_{\theta_*}(\tau)\}$; construct matched positive trajectories analogously with $\exp\{+\beta R_{\theta_*}(\tau)\}$. Conditional on the pool, these are independent draws from the corresponding empirical exponential tilt. The learner receives features, the negative or positive

channel where applicable, the empirical reference probabilities, and the transition log, but not R_{θ_*}. Planning is restricted to empirically supported state–action pairs; unlike the population class $\Pi_{\mathcal{D}}$, this empirical restriction depends on the realized log. Hyperparameters are selected using negative-channel likelihood or a held-out validation channel, never the hidden evaluation reward.

The comparison rules have distinct information inputs. Label-only elimination excludes a labeled bad choice without ranking the remaining choices; bad-count ranks actions by how often they appear in the negative log. Behavior cloning from the transition log imitates its behavior, while complement cloning imitates trajectories in the reference pool not selected as negative. Occupancy repulsion maximizes distance from negative occupancy under the stated objective rather than fitting a reward. Selected-ridge fits the negative-rational likelihood with an ℓ_2 penalty chosen by held-out negative likelihood. Oracle-reward planning uses the hidden reward only as an evaluation benchmark, not as learner input.

Report at least 50 seeds for tabular problems and enough seeds in larger environments to give bootstrap 95% confidence intervals. The pre-specified primary metrics are: (i) identifiable-subspace error $\| \mathrm{Proj}_U(\widehat{\theta} - \theta_*)\|_2$ and full parameter error $\|\widehat{\theta} - \theta_*\|_2$; (ii) principal-angle or null-space recovery error; (iii) the smallest eigenvalue of Fisher information on the span of the policy feature contrasts being tested; (iv) exact covered-policy recovery; and (v) normalized regret under the hidden true reward. To define the regret decomposition, let π^*_{all} be the unconstrained true optimum, π_* the true covered optimum, and $V^{\mathrm{range}}_{\mathcal{D}} := \max_{\pi \in \Pi_{\mathcal{D}}} J_{\theta_*}(\pi) - \min_{\pi \in \Pi_{\mathcal{D}}} J_{\theta_*}(\pi)$, which is positive in the reported instances. Within-support regret is $[J_{\theta_*}(\pi_*) - J_{\theta_*}(\widehat{\pi})]/V^{\mathrm{range}}_{\mathcal{D}}$, support loss is $[J_{\theta_*}(\pi^*_{\mathrm{all}}) - J_{\theta_*}(\pi_*)]/V^{\mathrm{range}}_{\mathcal{D}}$, and total normalized regret is their sum. Secondary metrics are held-out negative log likelihood, source diversity, computation time, and empirical transition coverage. In addition to aggregate curves, report all deviations from predicted patterns and seed-level distributions.

D.2 Main experiments

The four main experiments map the theory to distinct empirical questions: rank and coverage, relative information in positive and negative channels, source allocation, and sequential control with estimated dynamics. Each experiment changes the factor needed for its target claim while holding the other roles as fixed as possible.

E1: rank, samples, margins, and coverage. Generate random finite MDPs with $10 - -30$ states, $3 - -6$ actions, horizon $20 - -50$, and linear rewards of dimension $d \in \{4, 8, 16\}$, together with a gridworld whose feature semantics are inspectable. Control the centered negative-feature rank $r_M := \mathrm{rank}(M) \in \{1, \ldots, d\}$ by restricting which feature-bearing regions the negative selector can visit. Cross r_M/d, n_-, β, and policy gap Δ_*; separately thin the transition log to vary supported-action coverage. Compare the negative-rational maximum-likelihood estimator with no-negative-data estimation, a label-only elimination rule, and an oracle-reward planner. The confirmatory tests are: identifiable-subspace error decreases proportionally to $n_-^{-1/2}$; deficient rank leaves a nonvanishing component of full parameter error in $\mathcal{N}$; the null-space certificate predicts policy recovery more accurately than full reward recovery; and deleting transition coverage increases regret without changing negative-channel Fisher rank. Exact transition, feature, and reward generators, including all sampling distributions, are part of the required submission artifact described in the reproducibility statement.

E2: negative versus positive sample efficiency. Use paired random seeds so q_0, θ_*, transition data, and sample count are identical across channels. Create a success-mode-collapse family with one narrow high-reward trajectory cluster and several distinct failure clusters, and a mirror family with diverse successes and concentrated failures. For each tested policy feature contrast $v = \mu_\pi - \mu_{\pi'}$, compare the local directional information $v^\top F_- v$ and $v^\top F_+ v$, then separately compare identifiable-subspace reward error, policy recovery, regret, and the interpolated sample count needed to reach 90% recovery. Sweep the reward separation and β to test the exponential information-ratio construction. Equal trajectory count, equal environment interactions, and equal feature representation are required; otherwise the comparison is not causal.

E3: budgeted failure-source design. Instantiate K sources using initial-state strata, distinct exploratory controllers, action-noise types, and environment perturbations, with source index $j \in$

$\{1, \ldots, K\}$ as in Theorem 5. Take a small equal pilot sample from each source, estimate F_j, and then freeze the selection rule before acquiring the remaining budget. Compare E-optimal allocation, greedy marginal rank gain, uniform allocation, random source selection, largest trace, and most-common-source allocation. Evaluate both oracle F_j and pilot-estimated F_j to expose design-estimation error. Plot rank, restricted minimum eigenvalue, policy recovery, and regret against total budget and pilot fraction. Include redundant sources so that repeatedly sampling a frequent failure reduces variance but cannot close missing directions.

E4: offline sequential-control validation. On one higher-dimensional logged-control domain, construct the negative channel from complete low-return trajectories while withholding dense rewards from the learner. Define failure sources before inspecting results, for example by initial-state cluster or intervention type. Compare the proposed estimator and robust planner with label-only elimination, direct avoidance/occupancy repulsion, behavior cloning on the complement when available, a no-negative-data offline learner, and an oracle-reward upper bound. Report negative-channel fit, return under the hidden reward, violation rate, and coverage statistics. This experiment is supportive rather than an exact test if the exponential-tilt model is only approximate.

D.3 Ablations and falsification tests

These tests vary one modeling or decision assumption at a time to distinguish failures of observation semantics, feature geometry, robustness, and offline planning. Their purpose is to identify where the theory should fail rather than merely to compare average performance.

Observation semantics. Compare frequency-aware negative likelihoods with binary labels only; estimate with the correct q_0, a perturbed q_0, and an unrestricted learned reference; vary label flips and mixtures of failure mechanisms. For unlabeled mixtures, cross known versus unknown mixture proportion and the presence versus absence of an anchor set. The predicted boundary cases are avoidance-only information under labels alone, reference confounding when the reference is unrestricted, and mixture ambiguity without proportions or anchors.

Geometry and design. Ablate feature dimension, centered rank, source redundancy, pilot size, covariance regularization, SDP versus greedy selection, and E-optimal versus trace-based design. Report both reward-space rank and the smaller policy-relevant rank. This tests whether the control certificate correctly treats some unidentifiable reward directions as harmless.

Robustness and planning. Replace exact exponential selection with mixtures of temperatures, bounded adversarial contamination, temporally correlated trajectories, and nonlinear selection logits. Independently vary support thinning and the conservatism of the offline planner. Plot likelihood misspecification, reward error, and planning regret separately so semantic error is not confused with dynamics or support error. Finally, repeat the main comparisons across at least two feature representations to reveal representation-dependent occupancy conclusions.

Decision rules. Use confidence intervals and effect sizes rather than isolated best-seed means. Treat the theory as supported only if the pre-specified qualitative transitions occur: adding samples cannot repair a fixed rank deficiency; adding complementary sources can; bad data outperform good data only in directions where their covariance is larger; and transition coverage changes downstream regret without changing reward identifiability. Negative results and violations of these patterns should be retained.

E Experimental results

This appendix reports the evaluation pre-specified in Appendix D. Each experiment states its question, its setup, its result, and the scope of its conclusion. Every design was frozen before its own data existed. Two qualifications: the avoidance family of E2 was specified only after the two pre-specified families had been run, and is reported as post hoc; and the crossing of β and the policy gap that Appendix D specifies for E1 is reported as separate sweeps rather than inside the main grid.

Shared setup. All environments are finite MDPs with a known linear feature map. In this appendix, information matrices without an explicit argument are evaluated at θ_*. One *record* is one random-seed–condition pair, and one *cell* aggregates all seeds at a fixed condition. Features are scaled so the largest trajectory feature has unit norm and θ_* is drawn on the unit sphere, so β is comparable across environments. The learner sees the feature map, the transition log and negative trajectories drawn under (1), never the hidden trajectory return $R_{\theta_*}(\tau)$ defined in Appendix D. Reward is estimated by a maximum-likelihood estimator (MLE) under (1), restricted to $U = \mathcal{N}^{\perp}$; planning is restricted to empirically supported state–action pairs; regret splits into support loss and within-support regret, each normalized by the covered value range. Exact policy recovery is informative only where the policy gaps are large. In the trees of E2 they are of order 10^{-1} and the exact criterion is reached on two thirds of records, but in the room and hazard families the median gap is 10^{-8} or smaller and exact recovery is reached on 10% of E1 records and none of E4b's. We therefore report a threshold criterion throughout— normalized within-support regret at most $\varepsilon_{\mathrm{pol}} = 0.02$—in place of primary metric (iv) of Appendix D. This is the one endpoint change we make and the threshold was fixed before the runs. Varying it over $[0.005, 0.05]$ leaves the ordering of the methods unchanged in E1 and among the E4b comparators of Table 1; in E3 the differences among E-optimal, greedy and uniform at budgets other than $n = 500$ are within noise and can change sign with the tolerance. Cells carry 50 seeds, except the E3 replicate and the certificate study (200) and the bound calibration (2,000). Intervals are seed-level bootstrap intervals over 2,000 resamples, paired where the design is paired. No family-wise correction is applied, so intervals are per-comparison and counts of the form k/n are descriptive.

Table 3: Claims and outcomes. "Exact" rows are algebraic identities checked on every record and must hold up to numerical precision when implemented correctly; the remaining rows are statistical statements declared before the runs.

Claim	Test	Outcome
Identification up to $\mathcal{N}$ (Theorem 2)	E1, exact	holds on every record
Control certificate (Theorem 3)	E1, oracle version	agrees exactly with the version-set criterion on all 8,400 records
Identifiable-subspace rate $n_-^{-1/2}$, with a separate full-error null-space floor (Theorem 4)	E1, 18 cells	identifiable-subspace error decreases in 18/18 (slopes -0.43 to -0.57); slope intervals contain $-1/2$ in 17/18, while deficient-rank problems retain full error in $\mathcal{N}$
Recovery under strong selection	E1, bar 0.9	met in 8/9 cells at $\beta = 8$; 1/9 at $\beta = 2$
Per-sample information ordering (Proposition 3)	E2, exact	met in 3,598/3,600 directions; two construction-level exceptions exceed the declared 1% resolution
Positives reach the bar first, mirror family	E2, 12 cells	holds, 7/7 decided cells
Negatives reach the bar first, collapse family	E2, 12 cells	pre-specified downstream prediction not met; 1/8 decided cells
Negatives reach the bar first, avoidance family (post hoc)	E2, 12 cells	holds, 7/7 decided cells
Design theorem and greedy guarantee (Theorem 5, Corollary 2)	E3, exact	holds on every record
E-optimal maximizes $\lambda_{\min}$ over the rules	E3, SDP check	holds; certified duality gap $\leq 10^{-8}$
E-optimal against baselines downstream	E3, 200 seeds	lower identifiable-subspace error at $n \leq 500$; recovery differences within noise at saturation
Semantics separate from coverage	E4a, 600 records	holds; estimate unchanged at stored precision under thinning
Control from negatives alone	E4b, 300 records	ahead of every baseline at $n_- \geq 500$
Labels alone are insufficient (Theorem 1)	200 instances	holds; r_1-optimum recovered in 0.525 [0.45, 0.60]
Statewise ordering (Theorem 6)	exact + finite sample	exact in population; pooled finite-sample target narrowly not met at 10^4 draws per state (0.944)
Rank gain (Proposition 2), mixtures (Proposition 5), repulsion (Theorem 7)	exact	hold to 10^{-15}

E.1 E1: WHAT GOVERNS IDENTIFICATION?

Theorems 2, 3 and 4 say reward is identified only up to $\mathcal{N}$, that control can still be certified, and that the identifiable component converges at rate $n_-^{-1/2}$. When the true reward has a component in $\mathcal{N}$, the full parameter error retains that component as a separate floor.

Setup. Room MDPs with $|\mathcal{S}| = 20$ states, $|\mathcal{A}| = 4$ actions, and horizon $H = 30$, in which each room reached by the negative selector activates a declared subset of feature coordinates and thereby fixes the centered rank, plus a gridworld replicate. We cross feature dimension $d \in \{4, 8, 16\}$, rank fraction $r_M/d \in \{0.25, 0.5, 0.75, 1\}$, and whether the unidentified directions are policy-relevant, with n_- from 20 to 2,500 and 50 seeds per cell. A separate axis thins the transition log.

Results. The exact identities hold on every record: 8,400 rows for the subspace, decomposition and kernel checks, and the 6,300 rank-deficient rows for the null-space flatness check. Identifiable-subspace error decays at the $n_-^{-1/2}$ rate in all 18 rank-deficient cells, while full error retains the component in $\mathcal{N}$. When the unidentified directions are policy-irrelevant, recovery reaches 0.94–1.00 at $n_- = 2500$ at every rank fraction; when they are policy-relevant, recovery is exactly 0 at every rank fraction below 1 and every sample size, and the certificate abstains in exactly those cells. Thinning the log leaves the reward estimate, the held-out likelihood, and the restricted Fisher eigenvalue unchanged at stored floating-point precision while raising support loss by 0.18. Recovery at $n_- = 2500$ is 0.32, 0.56, 0.88, 0.94, 1.00 for $\beta = 0.5, 1, 2, 4, 8$.

Reading. The controlling variable for policy recovery is not the rank fraction but whether the unidentified directions matter for control, which is what Theorem 3 predicts and what the two recovery rows of Figure 2 show: near-complete recovery at every rank fraction when the unidentified directions are policy-irrelevant, and none at any deficient rank fraction when they are not. Rank fraction still governs the *full* reward error (Figure 3): its identifiable component shrinks with more data, but its null-space component cannot. Recovery also requires informative selection: at $\beta = 2$, only one of nine cells reaches the 0.9 recovery-rate bar based on $\varepsilon_{\mathrm{pol}}$, indicating limited information at that sample size rather than a change in population rank.

E.2 E2: WHEN IS BAD DATA MORE INFORMATIVE THAN GOOD DATA?

Proposition 3 says the answer is decided by Fisher information in the reward-relevant directions, not by the sign of the label.

Setup. Complete trees of depth 3 and branching 2 with $d = 8$, paired seeds, equal sample counts, one reference law; the positive channel is the negative channel of $-\Phi$, so the two are exactly matched. Separation $\in \{1, 4, 8\}$, $\beta \in \{0.5, 1, 2, 4\}$, n from 50 to 2,500, 50 seeds. A cell is *decided* when at least one channel reaches the pre-specified 0.9 recovery-rate bar within the sampled n-grid; "reaches first" compares the interpolated crossing sample sizes. Two families were pre-specified: *mirror* (diverse successes, concentrated failures) and *collapse* (the reverse). A third, *avoidance*, was added afterwards and is post hoc: the initial state lands with equal probability in a subtree holding the single high-return mode or in a subtree holding only distinct failure modes, so ordering the failures is what the optimal policy requires.

Results. The analytic construction reproduces the predicted information ratio (fitted slope 8.01–8.15 against 8). The designed per-sample ordering holds in $3{,}598$ of $3{,}600$ directions; the two exceptions are reversed by 2.2% and 2.5%, beyond the declared 1% resolution, so this construction-level check is recorded as not met. Both lie in near-uniform mirror environments at $\beta \leq 1$; the collapse family holds in all $1{,}800$ of its directions. Downstream, one pre-specified prediction holds and the other does not: positives reach the recovery bar first in the mirror family, as predicted (7/7 decided cells), but they also lead in the collapse family, where negatives were predicted to lead (the prediction holds in 1 of 8 decided cells). In the avoidance family negatives lead in 7/7 decided cells and have the lower identifiable-subspace error in 69 of the 72 combinations of β, separation and n.

Reading. The per-sample information claim is supported in 3,598 of 3,600 tested directions, while the downstream comparison refines its interpretation: in both pre-specified families negatives carry

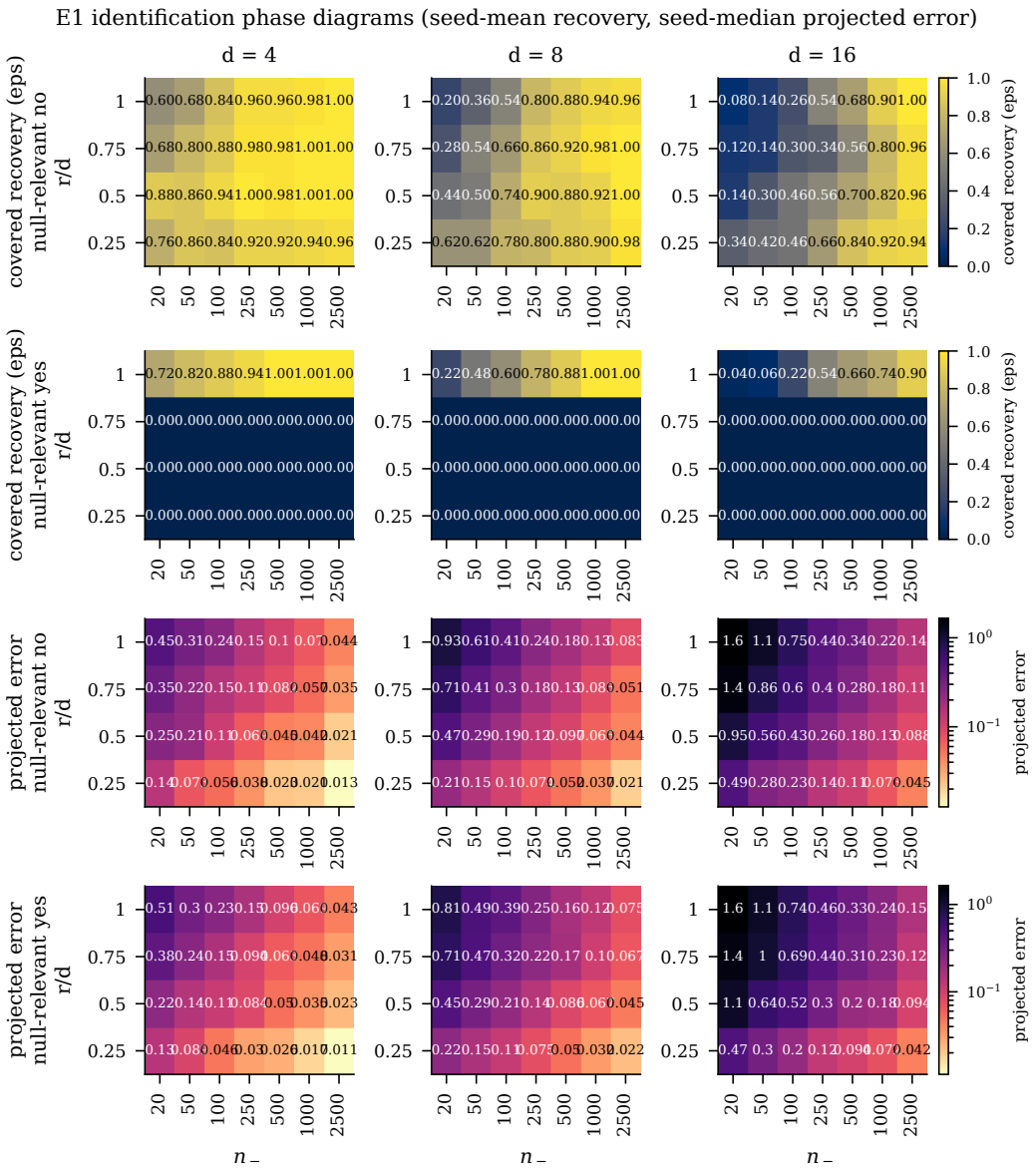

Figure 2: E1 at $\beta = 8$, 50 seeds, over rank fraction and sample size for each feature dimension. The four row groups are recovery and identifiable-subspace reward error, each with the unidentified directions policy-irrelevant (rows one and three) and policy-relevant (rows two and four). Recovery is high at every rank fraction in the first case and exactly zero at every deficient rank fraction in the second; the reward error behaves the same way in both.

more information about contrasts among failures, yet positives recover the policy first, because the optimal policy needs the success contrast. The avoidance family shows the converse, and being post hoc it demonstrates rather than tests the mechanism. Together they bound the claim: what matters is whether a channel's information lies in the directions the policy needs, and neither label is universally better.

E.3 E3: DOES CHOOSING WHICH FAILURES TO COLLECT HELP?

Theorem 5 and Corollary 2 say a budget spent on failure sources buys identified rank, that complementary sources add rank while redundant ones do not, and that greedy selection is within $1 - 1/e$ of optimal.

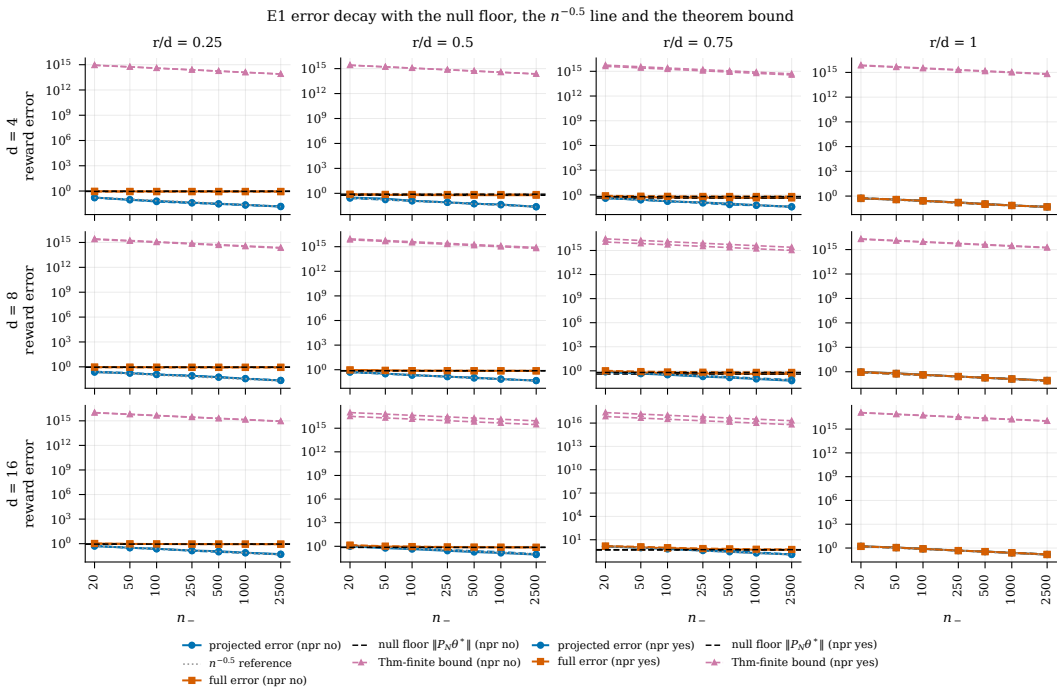

Figure 3: E1 identifiable-subspace reward error against n_-, shown with the norm of the true null-space component and an $n_-^{-1/2}$ reference line. The identifiable-subspace error decreases with sample size; the separately plotted null-space component is zero at full rank and remains irreducible at deficient rank.

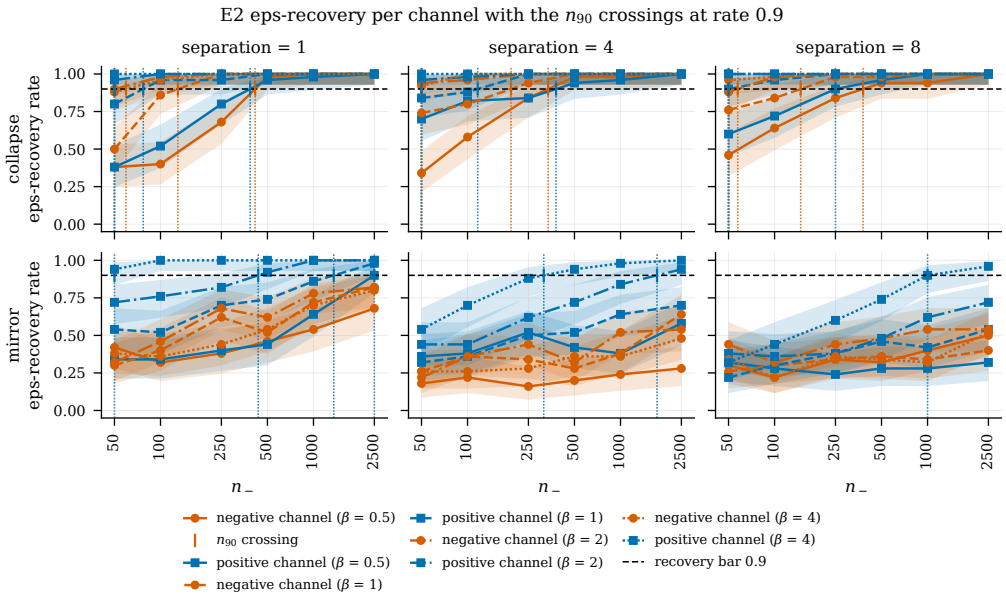

Figure 4: E2 recovery against sample size, negative and positive channels, mirror and collapse families. The success contrast is what the policy needs and positives lead.

Setup. Eight closed source rooms have declared sets of active feature coordinates—four complementary sets spanning the target subspace, two redundant sets, and two overlapping sets—so each source's rank contribution is fixed by construction. A small equal pilot estimates each F_j, the rule

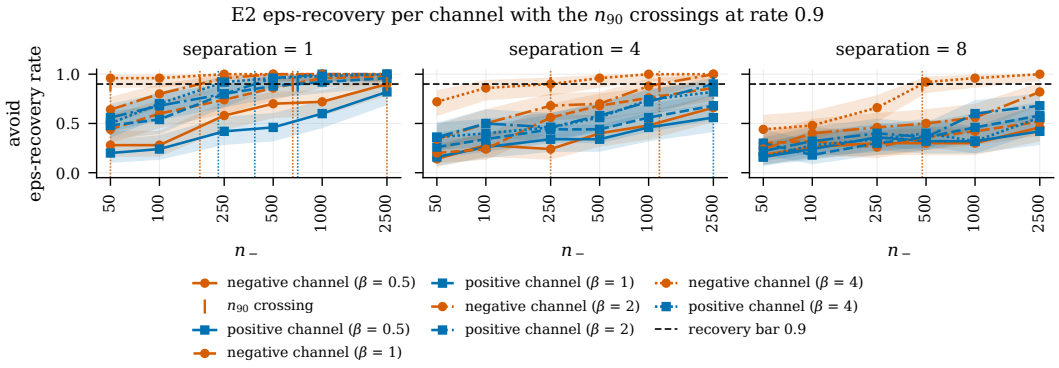

Figure 5: E2 avoidance family (post hoc): ordering the failure modes is what the policy needs, and the ordering of the channels reverses.

is frozen, the rest of the budget is spent. Budgets $\{200, 500, 1000, 2500\}$, pilot fractions $\{0.1, 0.2\}$, oracle and pilot F_j, 50 seeds and a 200-seed replicate. Rules: E-optimal, greedy rank gain then uniform within the selected set, uniform, random by source frequency, largest trace, most-common source.

Results. Every exact statement holds on every record: greedy gains equal the values implied by the active-coordinate sets, greedy equals the brute-force optimum at every mechanism budget and satisfies $1 - 1/e$, and $\ker F(w;\theta) = \bigcap_j \ker F_j(\theta)$ on the support of w. The E-optimal program attains the largest $\lambda_{\min}(F(w;\theta)|_{\mathcal{U}})$ among the rules—0.112 against 0.070 greedy, 0.058 uniform, 0.027 random, and 0 for the two single-source rules—with a certified duality gap below 10^{-8}. Downstream it reaches higher recovery than greedy, uniform and random at $n = 500$ ($+0.11$, $+0.10$, $+0.13$, intervals excluding zero) and is indistinguishable from them once recovery saturates. Along the direction uniform sampling covers least it is below uniform and random in every one of 32 point estimates, and with an interval excluding zero in 29 of them, across the core grid and a variant with θ_* aligned to that direction; the pre-specified form of the prediction demanded all 32 and is recorded as not met. Sweeping the mechanism budget gives identified rank $2, 4, 6, 8$ for $k = 1, \ldots, 4$ and flat thereafter; at $d = 16$ with twelve sources the budget binds, and greedy identifies half of $\mathcal{U}$ while the fractional design identifies all of it.

Reading. On a construction where each source's rank contribution is known by design, greedy recovers the rank-maximizing source set and redundant sources contribute exactly nothing. Design matters most where it changes rank: the two single-source rules never leave rank 2 and never reach the recovery bar, ending at 0.53–0.55 at the largest budget. Once every rule already spans $\mathcal{U}$, uniform allocation over informative sources is a strong baseline, so the advantage of E-optimal is concentrated in the worst-identified direction and at small budgets. The cardinality-constrained greedy rule should not be preferred when a fractional allocation is available: at $d = 16$, pilot fraction 0.1 and $n = 2500$ it reaches 0.18 recovery against random selection's 0.50.

E.4 E4a: Is semantic corruption the same failure as missing coverage?

The theory treats these as independent axes. E4a implements separate interventions on a full-rank room generator, crossing four semantic conditions—correct tilt, perturbed reference, temperature mixture, bounded contamination—with three coverage-thinning levels $c \in \{0, 0.25, 0.5\}$, where c is the targeted fraction of policy-relevant transition support removed from the planning log; 50 seeds, 600 records.

Results and reading. As an implementation check, under thinning the reward estimate, its norm, the held-out likelihood and the misspecification radius have identical stored floating-point values across c in all 200 matched seed–semantic groups, while support loss rises to 0.42 $[0.38, 0.47]$ and total regret with it. Under semantic corruption the estimate moves and the support does not: $+0.357$ $[0.337, 0.373]$ identifiable-subspace reward error for the temperature mixture and $+0.342$

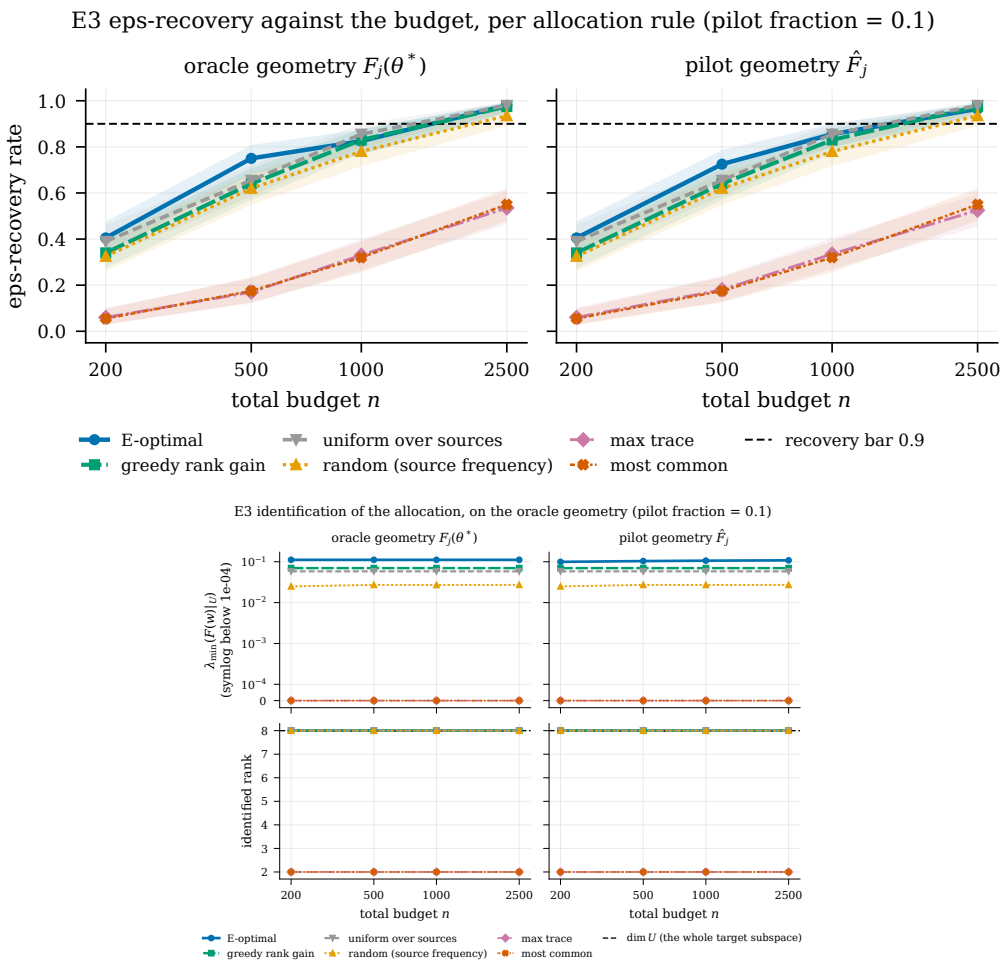

Figure 6: E3, 200 seeds, pilot fraction 0.1. Top: recovery against budget per rule. Bottom: identified rank and restricted minimum eigenvalue of each allocation; the single-source rules never leave rank 2.

$[0.313, 0.371]$ for contamination. The bit-identity is stronger than a statistical result—deleting transition records in this intervention cannot change the estimate, because it does not touch the negative channel. The perturbed reference raises reward error by only 0.006 $[0.002, 0.011]$; we report this without an explanation, since the condition that would explain it, an unrestricted jointly learned reference, is the one pre-specified condition we did not run.

E.5 E4b: Does this work on a sequential control task?

E1–E4a plan on true dynamics. E4b asks whether the estimator still yields a good policy when the model must also be estimated and the selection model is only approximate.

Setup. The task, data source, exact-tilt protocol, and complete baseline comparison are now stated in Section 5 and Table 1. Here we report the additional misspecified-channel result and interpretation.

Results. Table 1 gives the full comparison in the tilt channel. The estimator is ahead of every baseline at $n_- \geq 500$ and behind occupancy repulsion at $n_- = 100$. On the violation rate it is level with repulsion and well ahead of everything else. The bottom-quantile channel, fitted as though it were an exact tilt, gives selected-ridge within-support regret 0.027, 0.025 and 0.023 at the three sample sizes.

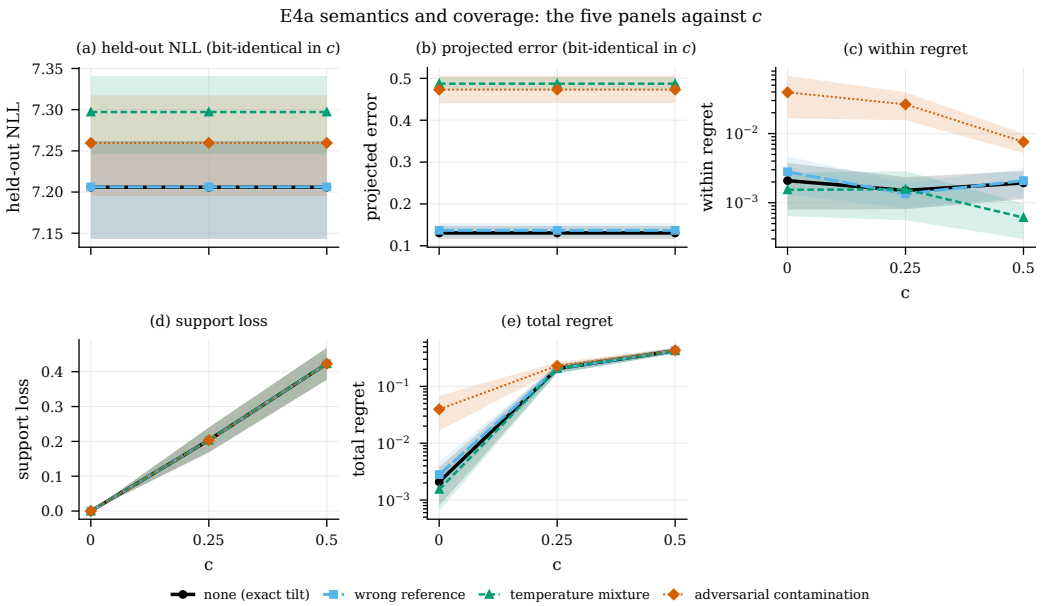

Figure 7: E4a: five metrics against the coverage level, one curve per semantic condition. Thinning moves support loss and total regret; semantic corruption moves the reward estimate.

Reading. In this synthetic task, negatives alone identify enough of the reward to control under model error: with 2,000 of them the estimator is at 0.026 against a model-error floor of 0.017, while behavior cloning on the same log is at 0.147 and the label-only heuristics near 0.48. Two caveats belong with it. Occupancy repulsion is competitive—better at the smallest sample and level on safety—so the gain over a well-chosen avoidance baseline is a matter of sample size, not of kind. And the misspecified-channel result shows robustness to the particular bottom-quantile mechanism tested here. It does not establish that an arbitrary mechanism that favors low-return trajectories suffices; the formal ordering guarantee in Proposition 4 additionally requires uniformly bounded log-odds error and a sufficiently large return gap.

E.6 Diagnostics

The remaining algebraic statements are checked directly as implementation diagnostics. Proposition 2: $\ker(I_0 + nF) = \ker I_0 \cap \mathcal{N}$ with zero principal angle in 300 of 300 instances, and the unidentified dimension drops by exactly $\dim(\ker I_0) - \dim(\ker I_0 \cap \mathcal{N})$, equivalently the rank added by F on directions left unresolved by I_0. Theorem 1: over 200 instances with permuted action labels, let r_1 and r_2 denote the two observationally equivalent rewards that swap the admissible alternatives. Neither label-only heuristic ever returns the bad action, and each recovers the r_1-optimum in 0.525 $[0.45, 0.60]$ of them—the coin flip the theorem predicts. Proposition 5: the population identities hold to 10^{-16}; with an anchor set the mixing proportion is recovered on the small pools (0.296–0.313 against 0.3), while without one the plug-in is inconsistent, collapsing to 0.10 on the large pool. Theorem 7: the one-state Euclidean tie is exact and the reweighted separation is at least 1.5; on enumerable MDPs the farthest-occupancy policy is optimal in 36–38% of instances on one construction and 66% on the other under a non-uniform occupancy metric, though in 100% of the latter under the Euclidean metric, so "farthest need not be best" is realized rather than assumed.

Two results are weaker than the theory suggests. Theorem 4's radius covers in 100% of 2,000 data seeds, but arithmetically: the radii are 63–404 and 1.6×10^5–2×10^{16} against a maximum identifiable-subspace error of 2.61. The asymptotic Wald radius (0.34–2.88) is the informative one, and it over-covers at 0.985–1.00 against a nominal 0.95. Theorem 6 is exact in population and its reference correction is necessary (+0.16 recovery under a behavior reference, exactly 0 under a uniform one), but its finite-sample behavior depends on selection strength: with per-state gaps of

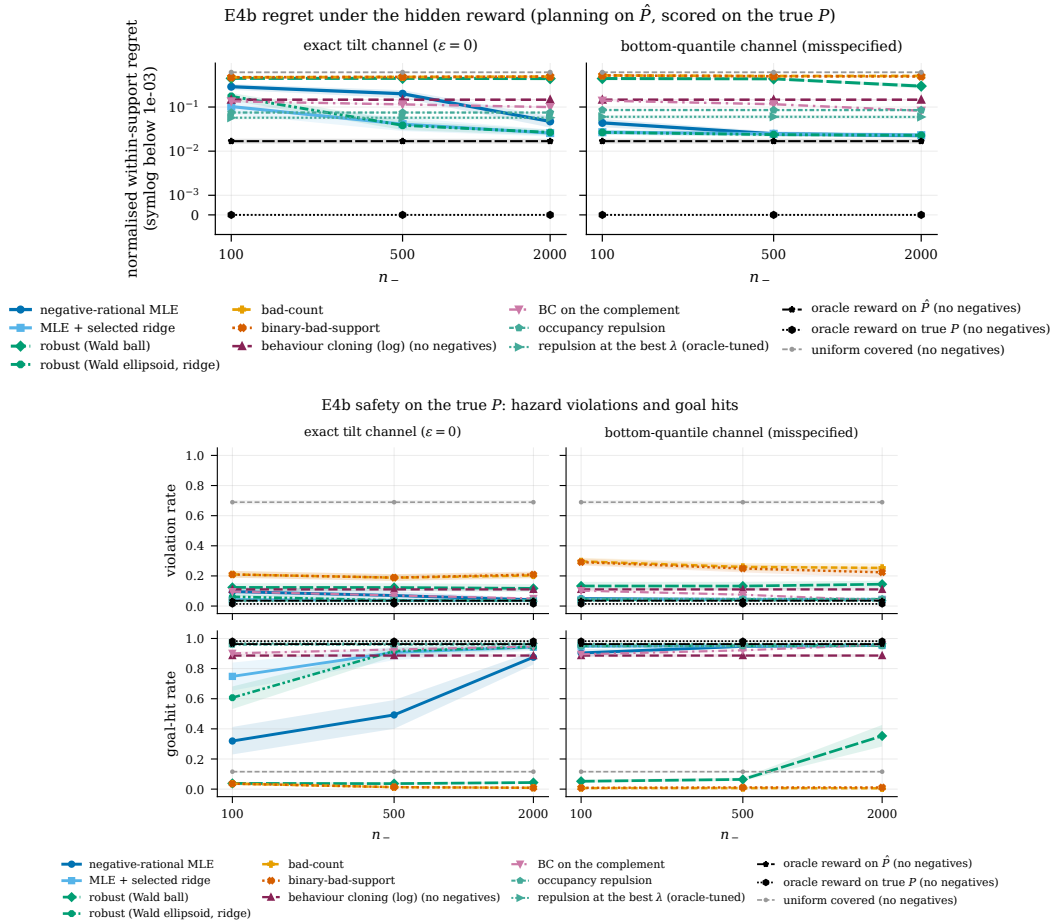

Figure 8: E4b: regret (top) and violation and goal-reaching rates (bottom) against n_-, both channels. All methods plan on $\widehat{P}$ and are scored on the true model.

0.006, the 0.95 bar is met at 10^4 draws per state once $\beta \geq 4$ (0.97) and missed at $\beta = 1$ (0.90), so pooled over the three temperatures it is missed (0.944 [0.933, 0.953]).

Finally, the control certificate on 200 random unstructured MDPs never certifies an incorrect policy (67 positive verdicts, 0 false positives), but it certifies only 33.5% of instances against a 95% recovery rate, and 123 of its 133 negative verdicts recover anyway. It is a sound but heavily abstaining test, not a sharp one.

Implication for the main claim. These diagnostics separate an exact population boundary from a practically useful finite-sample decision. The algebraic checks confirm the predicted ambiguity and failure modes, but the loose finite-sample radius and abstaining affine certificate do not yet provide a sharp rule for deciding from a small real failure log. Likewise, statewise identification is exact only where the relevant actions and states are observed; its weak-selection finite-sample result warns against treating population identifiability as automatic recovery.

E.7 Ablations

Five ablations replace one assumption at a time, 50 seeds each; the certificate study above is the sixth. With β treated as unknown, the likelihood identifies only the projection of $\vartheta = \beta\theta$ onto $U = \mathcal{N}^{\perp}$; held-out likelihood cannot select β separately without an additional scale normalization. We therefore evaluate the direction of $\mathrm{Proj}_U \widehat{\vartheta}$ and its induced plug-in policy, both of which are unchanged relative to a known-β parameterization on U. We do not interpret the fitted scale as

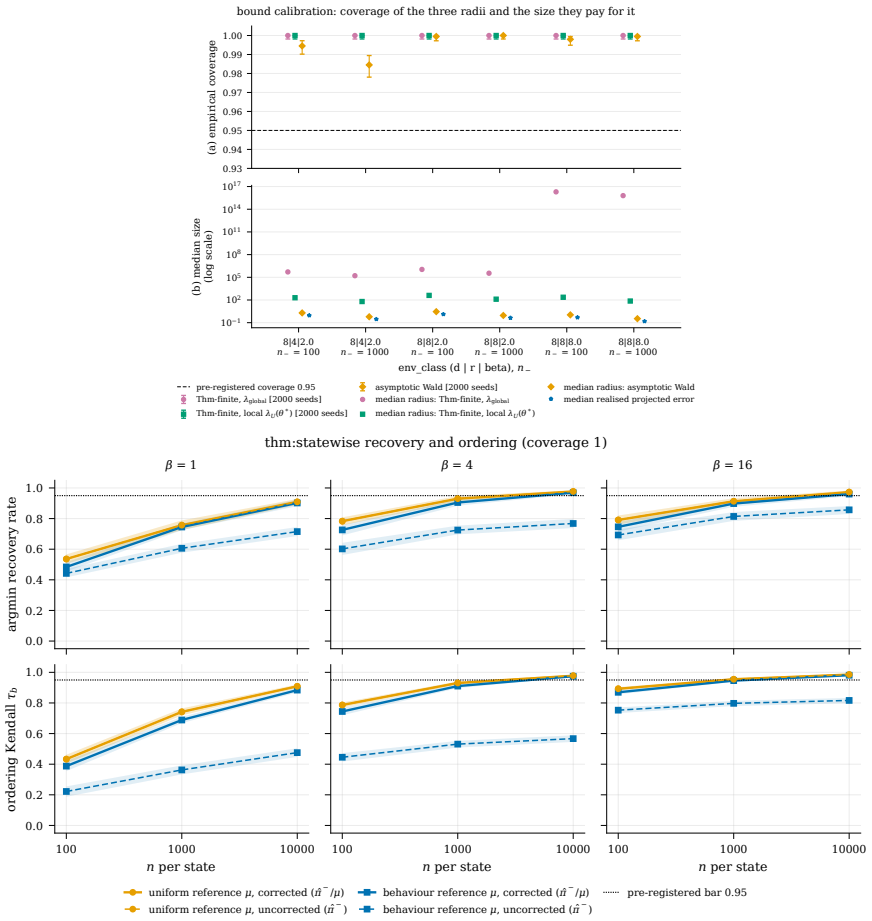

Figure 9: Top: coverage of the finite-sample radii over 2,000 data seeds, with the median identifiable-subspace error for scale. Bottom: statewise argmin recovery and rank correlation against draws per state, with and without the reference correction.

an estimate of β. Under label flips at rates 0.1 and 0.3 an unaware learner degrades as predicted ($+0.28$ to $+0.42$ identifiable-subspace error), while a learner that knows the inversion matches a clean negative-only reference in five of six groups—the sixth misses the declared equivalence band, so the pre-specified equivalence criterion is not met. Under nonlinear selection logits $g(z) = z + \kappa z^2$, where z is trajectory return and κ is the quadratic-link coefficient, the error floor grows with κ ($+0.30$ $[0.25, 0.36]$) while the policy still beats the uniform reference by 0.43–0.48. Under temporally correlated negatives the identifiable-subspace error still decreases with sample size at both correlation levels, with a gap to an independent reference that does not close. Varying the rank tolerance from 10^{-12} to 10^{-6} leaves every metric unchanged at stored precision, and a deliberately coarse control tolerance moves the identified dimension in all 150 comparisons, so the insensitivity is a property of the spectrum—exactly zero beyond the rank in these generators —rather than of an uninformative check. Separately, E3's pilot-versus-oracle design test responds as intended: covariance shrinkage of the pilot information collapses the rank-based rule from rank 8 to rank 4 and leaves E-optimal untouched.

F Scope, out-of-scope questions, and future directions

This appendix separates limits of the reported experiments from limits of the observation model itself. It then turns those boundaries into concrete directions for extending the identification and planning framework.

F.1 Empirical scope

E1, E3, E4a and E4b run at strong selection ($\beta = 8$); at $\beta = 2$ the E1 recovery bar is met in one of nine cells. E2 spans $\beta \in \{0.5, 1, 2, 4\}$ and the statewise diagnostic $\beta \in \{1, 4, 16\}$. The constant in Theorem 4 is loose at these scales, so the finite-sample result should be read as a rate, and a planner adopting that radius is correspondingly conservative. Design helps most where it changes identified rank; once every source is informative, uniform allocation matches E-optimal. The statewise ordering needs 10^4 draws per state at $\beta \geq 4$ and more at weaker selection, at gaps of 0.006. The certificate is sound but abstains on two thirds of instances. All environments are tabular and synthetic with an exactly known feature map, and in the room and hazard families the spectrum is exactly zero beyond the identified rank; the sequential-control study checks that the estimator survives model error and an approximate selection model, and is not evidence about real logged data. One pre-specified condition, an unrestricted jointly learned reference, was not run.

F.2 Modeling scope and out-of-scope questions

The negative-rational model is a model of logging or curation, not a general model of human intent. The positive identification results condition on a known or otherwise restricted reference law, a known feature map, and a linear reward. A poor terminal outcome also does not identify which preceding action caused it. Causal attribution is therefore outside the present identification model unless the dynamics, intervention, or sequential assignment mechanism supplies the necessary counterfactual information. Likewise, the paper does not address strategic labelers, unrestricted mixtures of annotator types, continuous control with learned representations, or deployment outside the support of the transition log.

These exclusions sharpen the interpretation of the results. The paper establishes what negative selection identifies under explicit semantics; it does not claim that every naturally occurring failure log satisfies those semantics. Mixtures may contain more than two behavior types, occupancy distances depend on representation, and an unrestricted learned reference can absorb the reward signal. The finite-sample theorem assumes independent trajectories and uniform likelihood curvature, while the affine control certificate permits every invisible reward direction and consequently abstains often. These are targets for extensions rather than evidence against the population characterization proved here.

F.3 Directions for future work

Several directions follow directly from these boundaries. First, jointly estimating a structured reference law and reward would replace known q_0 with a confidence set that propagates reference uncertainty into reward and policy ambiguity. Second, bounded, sparse, or otherwise structured reward version sets could produce sharper control certificates than the affine test. Third, adaptive collection could estimate F_j online while accounting for pilot uncertainty, nonstationary failure mechanisms, and unequal collection costs. Fourth, nonlinear rewards and learned representations require identifiability conditions that prevent the representation itself from manufacturing apparent failure separation. Fifth, dependent trajectories, heterogeneous annotators, and delayed terminal failures motivate sequential partial-identification and causal-attribution analyses. Finally, jointly robust reward-and-dynamics planning with principled abstention would connect the present tabular analysis to real logged-control domains.

G Checklist for applying the theory

Before treating bad demonstrations as reward information, a practitioner should specify: (i) what event caused the negative label; (ii) whether selection frequencies, rankings, or only labels are observed; (iii) the reference behavior q_0 and its uncertainty; (iv) which policy feature contrasts vary in the negative channel; (v) any mixture proportions or anchors; (vi) transition-log support for alternative actions; and (vii) a bound or estimate for the return gap between competing policies. Failure of an identification condition should result in a version set or abstention, not a point reward estimate.

H DETAILED RELATED WORK

This appendix expands the positioning in Section 2 along four connections: reward identifiability, behavioral observation models, learning from imperfect or negative demonstrations, and data-collection design for offline control. The organizing distinction is between what the observation channel identifies and what a downstream optimization method selects.

Reward and policy identification. Inverse reinforcement learning is generally ill-posed because multiple rewards can explain the same behavior or induce the same policy (Ng & Russell, 2000; Ziebart et al., 2008). Prior work characterizes identifiability under maximum-entropy behavior, reward invariances, and observations from multiple environments or experts (Kim et al., 2021; Cao et al., 2021; Rolland et al., 2022; Shehab et al., 2024). Constrained variants show that the admissible policy set and the chosen regularizer can alter both identifiability and transfer (Schlaginhaufen & Kamgarpour, 2023), while recent finite-sample theory separates the statistical difficulty of reward estimation from that of control (Zhao et al., 2024). Our observation channel is different: it preferentially retains trajectories with low latent return. The contribution is therefore not a sign reversal of maximum-entropy IRL. The negative-selection reference law determines which frequency ratios have reward meaning, centered failure features determine the observable reward subspace, and the covered policy class determines whether the remaining ambiguity matters for control.

Behavioral and observation-model assumptions. Reward inference is inseparable from a model of how demonstrations become observations. Models that assume noisily optimal behavior can confound systematic demonstrator biases with reward, motivating work that learns or relaxes the behavioral model (Shah et al., 2019). Other work explicitly models occlusion or observer noise (Suresh & Doshi, 2022), or jointly estimates reward and transition dynamics when the latter are uncertain (Herman et al., 2016). These papers concern different observation mechanisms, but they support the same methodological lesson: frequencies become reward evidence only relative to a specified data-generating channel. Our known proposal law q_0 isolates availability from negative selection; Proposition 1 shows that without such a restriction the proposal can absorb any candidate reward.

Imperfect and negative demonstrations. Confidence- and discriminator-weighted methods learn from demonstrations of varying quality (Wu et al., 2019; Xu et al., 2022). Trajectory rankings and human comparisons supply comparisons that can reduce reward ambiguity or support better-than-demonstrator behavior (Christiano et al., 2017; Brown et al., 2019; 2020; Kang et al., 2023). Those comparisons contain information absent from an isolated negative label. Inverse constrained RL instead recovers constraints from constraint-abiding demonstrations (Malik et al., 2021), and demonstrations collected under multiple unknown rewards can reveal shared safety constraints (Lindner et al., 2024). SafeDICE removes non-preferred occupancy from an unlabeled distribution (Jang et al., 2023), whereas UNIQ learns avoidance directly from undesirable demonstrations (Hoang et al., 2025). These approaches optimize under their respective supervision or objectives. We characterize what a negative observation law identifies before optimization and distinguish reward or policy evidence from avoidance alone.

Collection design and offline RL. Active IRL chooses expert queries or exploration policies (Brown et al., 2018; Lindner et al., 2022); inverse-bandit results likewise show that exploration along the learning path can be essential for reward estimation (Guo et al., 2022). Our fixed-budget design instead allocates samples among available failure-selection mechanisms, such as stress tests or initial-state strata, without assuming access to an expert query oracle. Offline RL must separately address distribution shift and transition coverage (Levine et al., 2020; Foster et al., 2022). Unlabeled transitions can reduce dynamics error but introduce reward bias (Yu et al., 2022), and offline learning from observations studies other incomplete-reward regimes (Li et al., 2023). We separate the source of reward information from the transition log's ability to evaluate a resulting policy: identifying reward contrasts does not create missing transitions, and additional transition data do not by themselves determine how negative selection relates to reward.