Information-Theoretic Limits of Value Learning under Distribution Shift

CΛ / Lightman Chang
Independent Researcher
lightman.chang@gmail.com

May 7, 2026

Abstract

We study the sample complexity of value learning when the algorithm is evaluated under a deployment distribution that differs from the training distribution. Within the inverse reinforcement learning (IRL) setting in which only expert demonstrations are observed, we prove a deploy-shift floor: there exist Markov decision processes carrying two reward functions whose demonstration-induced trajectory distributions are identical, yet whose optimal actions differ on a state of deployment mass $\beta > 0$ that is never visited under training. For any learning algorithm that consumes only demonstrations, the worst-case expected deployment loss is at least $\beta\Delta/2$, where Δ is the value gap on the out-of-distribution state. Crucially, the bound is independent of the number of training samples m. This is a strict refinement of, rather than a replacement for, in-distribution sample complexity bounds (Komanduru and Honorio, 2021; Metelli et al., 2023). We complement this floor with three propositions: a phase transition for Goodhart's law at the threshold $D_\infty \cdot \varepsilon = \Theta(1)$, where D_∞ is the Rényi-∞ ratio between the optimization-induced and reference state occupancies; a Fano lower bound $m > \sigma^2 \log(N/4)/(4\varepsilon^2)$ for the Gaussian-noise reward channel; and a comparison showing that pure demonstration channels carry zero information on argmax-coincident hypothesis pairs whereas preference and reward channels carry $\Theta(\log N/\varepsilon^2)$. Active demonstration querying does not break the floor. The results give a quantitative information-theoretic vocabulary for the channel-versus-algorithm distinction in alignment-relevant value learning.

Keywords. Inverse reinforcement learning; value learning; sample complexity; Goodhart's law; distribution shift; information theory.
MSC 2020. 68T05, 94A17, 62B10.

1 Introduction

Problem. Value learning seeks a policy whose realized behavior approximates the expectations of an underlying reward V^* that is observed only through a low-bandwidth signal — expert demonstrations, pairwise preferences, or scalar reward labels. In the demonstration-only setting, two distinct rewards may induce identical optimal policies and hence identical demonstration distributions. The resulting underdetermination has been informally invoked since the inception of IRL [1, 8], but it is sensitive to the choice of evaluation metric: under an expert-supported metric, a mimic-the-expert algorithm achieves zero loss and the underdetermination is vacuous. The present paper isolates the regime in which underdetermination produces a quantitative, sample-independent obstruction — namely, when the deployment distribution places nontrivial mass on out-of-distribution (OOD) states.

Prior art. Two recent lines bear directly on our results. Komanduru and Honorio [3] prove a Fano-style sample complexity lower bound of order $\Omega(n \log n)$ for IRL on n-state MDPs in the in-distribution regime, where the algorithm is judged on the same distribution from which demonstrations are drawn. Metelli et al. [7] establish a tight $\Omega\big(H^3SA\big(\log(1/\delta) + S\big)/\varepsilon^2\big)$ PAC bound for IRL with generative-model access. Both bounds are asymptotic in m and assume in-distribution evaluation; they do not address the deployment-shift floor we identify. On the Goodhart side, Karwowski et al. [9] characterize Goodhart phenomena via projected angle distance between reward functions, and El-Mhamdi and Hoang [6] compare light- and heavy-tailed regimes through tail dominance. Our framework employs the Rényi-∞ divergence D_∞ between occupancy measures, yielding a sharp threshold $D_\infty \cdot \varepsilon = \Theta(1)$ that operates at a different abstraction layer; we expand on the deltas in Section 6. Manheim and Garrabrant [5] provide a four-class taxonomy of Goodhart variants, which we map onto distinct D_∞ regimes.

Contribution. We prove that for any learning algorithm consuming only demonstrations, the supremum over a two-point family of rewards of the expected deployment loss is at least $\beta\Delta/2$, where $\beta > 0$ is the deployment mass on an OOD state and Δ is the per-state value gap (Theorem 4). The bound holds for all $m \geq 0$. We then exhibit a spike construction that produces an order-one optimization gap with controlled L^1 misspecification (Proposition 12), prove a Fano lower bound for the Gaussian reward channel with explicit constants (Proposition 17), compare the information content of demonstration, preference, and reward channels (Proposition 19), and show that active demonstration querying cannot escape the deploy-shift floor (Proposition 22). The contribution is conceptual as well as technical: the floor isolates a regime in which the choice of *channel* is more decisive than the choice of *algorithm*.

Organization. Section 2 fixes notation and recalls the information-theoretic tools. Section 3 develops the deploy-shift IRL floor. Section 4 treats the Goodhart phase transition. Section 5 gives the Fano sample complexity bound, the channel comparison, and the active-learning result. Section 6 discusses relations to prior work, limitations, and open problems.

2 Preliminaries

2.1 Markov decision processes and policies

A finite MDP is a tuple $\mathcal{M} = (S, A, P, \rho_0, \gamma, H)$ where S and A are finite state and action spaces, $P : S \times A \to \Delta(S)$ is a sub-stochastic transition kernel, $\rho_0 \in \Delta(S)$ is the initial-state distribution, $\gamma \in (0, 1]$ is a discount factor, and $H \in \mathbb{N} \cup \{\infty\}$ is the horizon. A reward function is a map $V : S \times A \to \mathbb{R}$ with bounded range. A (stationary) policy is a map $\pi : S \to \Delta(A)$. The occupancy measure of π is

$$\rho^\pi(s) \;=\; (1-\gamma) \sum_{t=0}^{H-1} \gamma^t \mathbb{P}_\pi[s_t = s],$$

with the convention $(1-\gamma)\sum_t = 1/H$ when $\gamma = 1$ and $H < \infty$. We write $V(\pi) := \mathbb{E}_{s\sim\rho^\pi}[V(s, \pi(s))]$ for the expected return; for stochastic π, the inner expectation is over $a \sim \pi(\cdot|s)$. An optimal policy for V is denoted $\pi^*_V \in \arg\max_\pi V(\pi)$.

2.2 IRL and information channels

Let $\mathcal{V}$ be a hypothesis class of reward functions and let $V^* \in \mathcal{V}$ denote the unknown ground truth. We consider four single-sample information channels:

- **(I-demo)** demonstrations (s, a) with $s \sim \mu$ and $a \sim \pi^*_{V^*}(\cdot|s)$;
- **(I-pref)** pairwise preferences (τ_A, τ_B, b) where $b \sim \mathrm{Bernoulli}\big(\sigma_{\mathrm{BT}}(V^*(\tau_A) - V^*(\tau_B))\big)$ under the Bradley–Terry model;
- **(I-rew)** noisy reward labels $r = V^*(s, a) + \xi$ with $\xi \sim \mathcal{N}(0, \sigma^2)$;
- **(I-act)** active demonstration queries: a sequential protocol in which the learner chooses s_t adaptively and receives $a_t \sim \pi^*_{V^*}(\cdot|s_t)$.

A learning algorithm is a (possibly randomized) map $A : \mathcal{D} \to \Pi$, where $\mathcal{D}$ is the dataset domain.

2.3 Loss functionals

Throughout, we write

$$\mathcal{L}_\mu(\hat{\pi}; V^*) \;:=\; \mathbb{E}_{s\sim\mu}\big[V^*(s, \pi^*_{V^*}(s)) - V^*(s, \hat{\pi}(s))\big]$$

for the value-loss of $\hat{\pi}$ relative to the optimal policy of V^* under reference distribution μ. The deploy-shift loss in Section 3 is the special case $\mu = \rho_{\mathrm{deploy}}$ which differs from the demonstration distribution.

2.4 Information-theoretic tools

We use Fano's inequality and Le Cam's two-point method in their standard forms.

Lemma 1 (Fano's inequality)**.** *Let V^* be uniform on a finite set $\{V_1, \dots, V_N\}$ and let D be observed data. For any estimator $\hat{V} = \hat{V}(D)$,*

$$\mathbb{P}\big[\hat{V} \neq V^*\big] \;\geq\; 1 - \frac{I(V^*; D) + \log 2}{\log N}.$$

Lemma 2 (Le Cam two-point)**.** *Let V_0, V_1 have separation Δ in the loss functional, in the sense that any estimator $\hat{V}$ satisfies $\mathcal{L}(\hat{V}; V_0) + \mathcal{L}(\hat{V}; V_1) \geq \Delta$. Then*

$$\inf_{\hat{V}} \sup_{V\in\{V_0,V_1\}} \mathbb{E}\big[\mathcal{L}(\hat{V}; V)\big] \;\geq\; \tfrac{\Delta}{2}\big(1 - \mathrm{TV}(P_0, P_1)\big),$$

where P_i is the data distribution under V_i.

2.5 Rényi divergence

For probability measures $\nu \ll \rho$ on a measurable space, the Rényi-∞ divergence is

$$D_\infty(\nu\|\rho) \;:=\; \log \operatorname{ess\,sup}_{x\sim\rho} \frac{d\nu}{d\rho}(x).$$

With a slight abuse of notation we will also write $D_\infty(\nu\|\rho)$ for the unlogged ratio $\operatorname{ess\,sup} d\nu/d\rho$ when this is clearer; the usage is signposted at each occurrence. Either reading produces the same phase transition up to a logarithm.

2.6 IRL setup and underdetermination

We adopt the formulation of Ng and Russell [8]: the unknown reward V^* is observed only through the optimal policy $\pi^*_{V^*}$, and the IRL learner outputs a policy that should match the optimal behavior of V^* on a target distribution. *In-distribution* IRL evaluates the learner on the demonstration distribution; this is the regime studied by Komanduru and Honorio [3], Metelli et al. [7]. *Deploy-shift* IRL — our focus — evaluates the learner on a distribution that may place mass on states never visited by the demonstrating expert. The transition between these two regimes governs whether classical underdetermination is vacuous or operative.

3 Main Result: the Deploy-Shift IRL Floor

3.1 The MDP construction

We construct a four-state, two-action MDP that exhibits demonstration-indistinguishability between two reward hypotheses while assigning conflicting OOD optima.

Let $S = \{s_0, s_1, s', s_T\}$ and $A = \{a_1, a_2\}$. Transitions:

$$\begin{aligned} s_0 &\xrightarrow{a_2} s_1, & s_0 &\xrightarrow{a_1} s_T, \\ s_1 &\xrightarrow{a_1} s_T, & s_1 &\xrightarrow{a_2} s_T, \\ s' &\xrightarrow{a_1} s_T, & s' &\xrightarrow{a_2} s_T, \end{aligned}$$

with s_T absorbing. The training initial distribution is $\rho_0 = \delta_{s_0}$.

Reward hypotheses. The two hypotheses agree on the demonstration trajectory and disagree on the OOD state s'. Specifically, V_1 and V_2 both assign reward 0 to (s_0, a_2) and reward 3 to (s_1, a_1), and both assign reward 0 at s_T. They differ only at s':

	V_1	V_2
$V(s', a_1)$	$+1$	-1
$V(s', a_2)$	-1	$+1$

We set $\Delta := V_1(s', a_1) - V_1(s', a_2) = 2 = V_2(s', a_2) - V_2(s', a_1)$.

Remark 3 (Why the OOD state is required)**.** A construction without s' — in which V_1, V_2 disagree only on actions never optimal under π^* — does *not* produce a floor: a mimic-the-expert algorithm achieves zero loss on ρ^{π^*}. The floor materializes precisely because the deployment distribution forces visitation of a state on which the two hypotheses prescribe conflicting optima. Without OOD mass $\beta > 0$, the bound in Theorem 4 is vacuous.

3.2 The demonstrated policy

The expert policy π^* is identical under V_1 and V_2 along the demonstration trajectory: $\pi^*(s_0) = a_2$, $\pi^*(s_1) = a_1$, terminating at s_T. The demonstration trajectory is $(s_0, a_2, s_1, a_1, s_T)$ with probability one. The behavior of π^* at s' is unspecified by the demonstrations because s' is never visited.

3.3 Disagreement on the OOD state

By the table in Section 3.1,

$$\pi^*_{V_1}(s') = a_1, \qquad \pi^*_{V_2}(s') = a_2,$$

and the value gap incurred by choosing the wrong action is exactly $\Delta = 2$ in either case.

3.4 Indistinguishability of demonstrations

Let $D = \big((s_0^{(i)}, a_0^{(i)}), (s_1^{(i)}, a_1^{(i)})\big)_{i=1}^m$ be a dataset of m demonstrations. Because both hypotheses share the same expert policy along the demonstrated states and because s' is never reached under ρ^{π^*}, the dataset distributions coincide:

$$P(D \,|\, V_1) \;=\; P(D \,|\, V_2).$$

In particular, $I(V^*; D) = 0$ for V^* uniform on $\{V_1, V_2\}$.

3.5 Deployment loss and minimax

The deployment distribution places mass β on s' and mass $1-\beta$ on the demonstration support, on which V_1 and V_2 agree. Hence the deployment loss reduces to

$$\mathcal{L}_{\text{deploy}}(\hat{\pi}; V^*) \;=\; \beta\big[V^*(s', \pi^*_{V^*}(s')) - V^*(s', \hat{\pi}(s'))\big].$$

Let $q := \mathbb{P}_{\hat{\pi}\sim A(D)}[\hat{\pi}(s') = a_2]$. Because $P(D|V_1) = P(D|V_2)$, the marginal distribution of $A(D)$, and hence q, is the same under V_1 and V_2.

Theorem 4 (Deploy-Shift IRL Floor)**.** *For any (I-demo)-only learning algorithm A and any number of training demonstrations $m \geq 0$,*

$$\sup_{V^*\in\{V_1,V_2\}} \mathbb{E}_{D\sim P(\cdot|V^*)}\big[\mathcal{L}_{\text{deploy}}(A(D); V^*)\big] \;\geq\; \frac{\beta\Delta}{2}.$$

The bound is independent of m.

Proof. Fix A and let q be defined as above. Conditional on $V^* = V_1$ the deployment loss is $\beta\Delta \cdot q$ in expectation, since the algorithm is wrong at s' exactly when it outputs a_2. Conditional on $V^* = V_2$ the loss is $\beta\Delta \cdot (1-q)$. Therefore

$$\sup_{V^*} \mathbb{E}\big[\mathcal{L}_{\text{deploy}}\big] \;=\; \beta\Delta \cdot \max(q, 1-q) \;\geq\; \frac{\beta\Delta}{2},$$

with equality at $q = 1/2$. Because $P(D|V_1) = P(D|V_2)$, the algorithm cannot vary q with V^*, and the bound holds independently of m. □

Remark 5 (Relation to in-distribution bounds)**.** Theorem 4 is a strict refinement of, not a replacement for, in-distribution bounds such as those of Komanduru and Honorio [3] and Metelli et al. [7]. With $\beta = 0$ the floor degenerates to zero and the in-distribution sample complexity governs the rate. The deploy-shift floor isolates the irreducible component arising from the unobserved OOD region; this component is structurally orthogonal to the in-distribution rate.

Corollary 6 (Two-point Le Cam form)**.** *With the construction of Section 3.1,* $\mathrm{TV}(P(\cdot|V_1), P(\cdot|V_2)) = 0$ *and Le Cam's two-point inequality reproduces the bound of Theorem 4.*

Corollary 7 (General floor under demonstration coincidence)**.** *The conclusion of Theorem 4 extends to any MDP and any value-function pair (V_1, V_2) satisfying the hypothesis of Lemma 9 with deployment-OOD disagreement of mass $\beta > 0$ and value gap $\Delta > 0$. Specifically, for any (I-demo)-only algorithm,*

$$\sup_{V^*\in\{V_1,V_2\}} \mathbb{E}\big[\mathcal{L}_{\text{deploy}}(A(D); V^*)\big] \;\geq\; \frac{\beta\Delta}{2},$$

whenever such a configuration exists. The four-state construction of Section 3.1 is one witness; the floor is a structural consequence of demonstration coincidence plus deployment-OOD value gap, not of the specific MDP.

Proof. The argument of Theorem 4 uses only (a) $P(D|V_1) = P(D|V_2)$, supplied by Lemma 9, and (b) the existence of a deployment state on which V_1, V_2 prescribe disagreeing optima with value gap Δ, of mass β under ρ_{deploy}. Both hypotheses transfer verbatim, and the minimax computation $\beta\Delta \cdot \max(q, 1-q) \geq \beta\Delta/2$ is unchanged. □

Remark 8 (Stochastic policies and randomized algorithms)**.** Theorem 4 is stated for the policy distribution induced by the algorithm and demonstration data; randomization of A is absorbed into the marginal q. Stochastic expert policies π^* that randomize on s_0, s_1 but produce trajectories with the same distribution under V_1 and V_2 leave the proof unchanged.

3.6 Why the demonstration distribution is uninformative

The technical heart of Theorem 4 is that two distinct rewards can induce identical demonstration distributions. We isolate this point in a self-contained statement.

Lemma 9 (Demonstration kernel coincidence)**.** *Let V_1, V_2 be two reward functions on an MDP, and suppose there is a state set $S_0 \subseteq S$ with $\rho^{\pi^*_{V_1}}(S \setminus S_0) = \rho^{\pi^*_{V_2}}(S \setminus S_0) = 0$ such that*

$$\pi^*_{V_1}(s) = \pi^*_{V_2}(s) \qquad \text{for all } s \in S_0.$$

Then for every $m \geq 0$ the joint demonstration distributions coincide: $P(D|V_1) = P(D|V_2)$ on $D = ((s_t, a_t))_{t<m}$.

Proof. By induction on m. At $m = 0$ the empty dataset has trivial coincident distributions. For the inductive step, the conditional distribution of (s_m, a_m) given $(s_{<m}, a_{<m})$ depends only on the transition kernel P and on $\pi^*_{V^*}$ at the visited state, both of which are by hypothesis identical under V_1 and V_2 on the support S_0. □

In our construction, $S_0 = \{s_0, s_1, s_T\}$, the OOD state s' is excluded, and Lemma 9 immediately yields the indistinguishability used in the proof of Theorem 4. The lemma also explains why active demonstration querying does not break the bound on argmax-shared pairs: the queryable expert maps every state to the same action under both hypotheses, so the lemma applies state-by-state to the active transcript.

Example 10 (A concrete instantiation)**.** Take $\beta = 1/4$, $\Delta = 2$, and $m = 10^6$. Theorem 4 gives a worst-case deployment loss of at least $1/4$, regardless of the algorithm. By contrast, the in-distribution loss on ρ^{π^*} can be driven arbitrarily close to zero by mimic-the-expert.

4 Goodhart Optimization Gap: a Phase Transition

We now show that even when a proxy reward $\hat{V}$ is close to V^* in expectation under a reference distribution ρ_0, optimization of $\hat{V}$ can produce an order-one gap. The threshold is governed by the Rényi-∞ ratio between the optimization-induced occupancy and ρ_0.

4.1 Spike construction

Let $S = [0, 1]$, $\rho_0 = \text{Uniform}[0, 1]$, and $V^*(s) = s$. Fix $\varepsilon \in (0, 1/2)$ and let $\eta := \varepsilon^2$. Define

$$g(s) \;:=\; \frac{1}{\eta}\mathbb{1}[s \in [0, \eta]] \;-\; 1, \qquad \hat{V}(s) \;:=\; s + \varepsilon \cdot g(s).$$

The proxy $\hat{V}$ exceeds V^* by approximately $1/\varepsilon$ on the spike region $[0, \varepsilon^2]$ and is below V^* by ε elsewhere.

4.2 Closeness

Lemma 11. $\mathbb{E}_{\rho_0}|\hat{V}(s) - V^*(s)| = 2\varepsilon(1-\eta) \le 2\varepsilon$.

Proof. $|\hat{V} - V^*| = \varepsilon|g|$, and

$$\int_0^1 |g(s)|\,ds \;=\; \int_0^\eta \Big(\tfrac{1}{\eta} - 1\Big)ds + \int_\eta^1 1\,ds \;=\; (1-\eta) + (1-\eta) \;=\; 2(1-\eta). \qquad \square$$

4.3 Optimization gap

We take $\hat{\pi}$ to be the uniform distribution on the spike region $[0, \varepsilon^2]$, an ε^2-smoothed argmax of $\hat{V}$ (the natural smoothing, since the spike has ρ_0-measure ε^2). With this smoothing, $d\rho^{\hat{\pi}}/d\rho_0 = 1/\varepsilon^2$ on the spike and 0 elsewhere, so $\operatorname{ess\,sup} d\rho^{\hat{\pi}}/d\rho_0 = \varepsilon^{-2}$ exactly. The expected return under V^* is $V^*(\hat{\pi}) = \mathbb{E}_{s\sim\rho^{\hat{\pi}}}[s] = \varepsilon^2/2$, while $V^*(\pi^*) = \sup_s s = 1$, giving a gap of $1 - \varepsilon^2/2$.

Proposition 12 (Goodhart phase transition: spike attains the upper bound)**.** *The Goodhart gap admits a two-sided characterization. Throughout, write* $D_\infty^{\hat{\pi}} := \operatorname{ess\,sup} d\rho^{\hat{\pi}}/d\rho_0$ *and* $D_\infty^{\pi^*} := \operatorname{ess\,sup} d\rho^{\pi^*}/d\rho_0$ *(unlogged Rényi-∞ ratios).*

(i) (Spike lower bound, unconditional.) *Under the spike construction with* $\eta = \varepsilon^2$ *and the smoothed policy* $\hat{\pi} = \mathrm{Uniform}[0, \varepsilon^2]$ *of §3.3,*

$$\mathbb{E}_{\rho_0}|\hat{V} - V^*| \;\le\; 2\varepsilon, \qquad V^*(\pi^*) - V^*(\hat{\pi}) \;\ge\; 1 - \tfrac{\varepsilon^2}{2},$$

and $D_\infty^{\hat{\pi}} = \varepsilon^{-2}$ *exactly.*

(ii) (General upper bound.) *For any* $V^*, \hat{V}$ *with* $\mathbb{E}_{\rho_0}|\hat{V} - V^*| \le \varepsilon$*, any reference* ρ_0*, and any* $\hat{\pi}$ *chosen by optimizing* $\hat{V}$*,*

$$V^*(\pi^*) - V^*(\hat{\pi}) \;\le\; \mathbb{E}_{\rho^{\hat{\pi}}}|V^* - \hat{V}| + \mathbb{E}_{\rho^{\pi^*}}|V^* - \hat{V}| \;\le\; \varepsilon \cdot \big(D_\infty^{\hat{\pi}} + D_\infty^{\pi^*}\big).$$

(iii) (Phase transition.) *When* $D_\infty^{\hat{\pi}} = O(1)$ *(regressional regime), the gap is* $O(\varepsilon)$*. When* $D_\infty^{\hat{\pi}} = \Theta(1/\varepsilon)$ *or larger (extremal regime), the upper bound permits a* $\Theta(1)$ *gap, and the spike construction in (i) attains it (with* $D_\infty^{\hat{\pi}} = 1/\varepsilon^2$ *and gap* $1 - \varepsilon^2/2$*). The product* $D_\infty^{\hat{\pi}} \cdot \varepsilon$ *is the phase-transition order parameter.*

Proof. **Part (i).** Lemma 11 gives the closeness statement (in fact $\mathbb{E}_{\rho_0}|\hat{V} - V^*| = 2\varepsilon(1-\varepsilon^2) \le 2\varepsilon$). Under the smoothed $\hat{\pi} = \mathrm{Uniform}[0, \varepsilon^2]$, $V^*(\hat{\pi}) = \mathbb{E}_{s\sim\hat{\pi}}[s] = \varepsilon^2/2$, while $V^*(\pi^*) = \sup_{s\in[0,1]} s = 1$, so the gap is $1 - \varepsilon^2/2$. The Radon–Nikodym derivative $d\rho^{\hat{\pi}}/d\rho_0$ equals $1/\varepsilon^2$ on $[0, \varepsilon^2]$ and 0 elsewhere, hence $D_\infty^{\hat{\pi}} = \varepsilon^{-2}$.

Part (ii). Decompose

$$V^*(\pi^*) - V^*(\hat{\pi}) = \underbrace{\big[V^*(\pi^*) - \hat{V}(\pi^*)\big]}_{A_1} + \underbrace{\big[\hat{V}(\pi^*) - \hat{V}(\hat{\pi})\big]}_{A_2} + \underbrace{\big[\hat{V}(\hat{\pi}) - V^*(\hat{\pi})\big]}_{A_3}.$$

By the optimality of $\hat{\pi}$ for $\hat{V}$, $A_2 \le 0$. Hence

$$V^*(\pi^*) - V^*(\hat{\pi}) \;\le\; A_1 + A_3 \;\le\; |A_1| + |A_3| \;\le\; \mathbb{E}_{\rho^{\pi^*}}|V^* - \hat{V}| + \mathbb{E}_{\rho^{\hat{\pi}}}|V^* - \hat{V}|.$$

Each term admits an importance-ratio bound:

$$\mathbb{E}_{\rho^{\hat{\pi}}}|V^* - \hat{V}| \;\le\; D_\infty^{\hat{\pi}} \cdot \mathbb{E}_{\rho_0}|V^* - \hat{V}| \;\le\; \varepsilon \cdot D_\infty^{\hat{\pi}},$$

and analogously for ρ^{π^*}. Summing yields the claim.

Part (iii). The phase-transition reading follows from (ii) and (i). When $D_\infty^{\hat{\pi}} = O(1)$, (ii) gives gap $\leq \varepsilon \cdot O(1) = O(\varepsilon)$. When $D_\infty^{\hat{\pi}} \geq c/\varepsilon$ for some constant c, (ii) permits gap up to $\Theta(1)$, and (i) exhibits a concrete construction with $D_\infty^{\hat{\pi}} = 1/\varepsilon^2$ and gap $1 - \varepsilon^2/2 = \Theta(1)$, witnessing tightness of the upper bound in the extremal regime. □

Remark 13 (Phase transition threshold)**.** The product $\varepsilon \cdot D_\infty^{\hat{\pi}}$ is the order parameter: when it is bounded, the gap is $O(\varepsilon \cdot D_\infty^{\hat{\pi}})$ (regressional); when it crosses an order-one threshold, the upper bound permits and the spike attains a $\Theta(1)$ gap (extremal).

Remark 14 (Policy admissibility)**.** Part (ii) requires $\hat{\pi}$ to be a ρ_0-absolutely continuous policy so that $D_\infty^{\hat{\pi}}$ is finite; pure point-mass policies give $D_\infty^{\hat{\pi}} = \infty$ in the continuous setting. The smoothed argmax of §3.3 is the natural admissible choice.

4.4 Worked example of the gap

To illustrate the phase transition, take $\varepsilon = 1/10$. Then $\eta = 1/100$, the proxy $\hat{V}$ exceeds V^* by 9.9 on the spike region $[0, 1/100]$, and is below V^* by $1/10$ elsewhere. The L^1 misspecification is $\mathbb{E}_{\rho_0}|\hat{V} - V^*| \approx 0.198$, well within 2ε. The smoothed optimization policy is uniform on $[0, 1/100]$, giving $V^*(\hat{\pi}) = \varepsilon^2/2 = 1/200$ and gap $1 - 1/200 = 0.995$. The Rényi-∞ ratio is exactly $D_\infty^{\hat{\pi}} = 100$, and the product $\varepsilon \cdot D_\infty^{\hat{\pi}} = 10$ is well above the order-one threshold, consistent with Proposition 12(iii).

4.5 Mapping the Manheim–Garrabrant taxonomy

Manheim and Garrabrant [5] distinguish four Goodhart variants. We map each to a D_∞ regime in our framework.

Proposition 15 (Taxonomy mapping)**.** *Under Proposition 12,*

- ***Regressional Goodhart*** *($D_\infty = O(1)$, optimization stays close to ρ_0): gap $= O(\varepsilon)$.*
- ***Extremal Goodhart*** *($D_\infty = \Theta(1/\varepsilon)$ or larger): gap $= \Theta(1)$, the phase-transition regime.*
- ***Causal Goodhart*** *(optimization induces feedback on ρ_0): D_∞ is itself a function of $\hat{V}$; the framework requires extension.*
- ***Adversarial Goodhart*** *($\hat{V}$ chosen by an adversary): the spike construction realizes the lower bound up to constants.*

Remark 16 (Comparison with prior Goodhart frameworks)**.** Karwowski et al. [9] measure proxy–target deviation through projected angle distance and obtain an early-stopping bound (their Proposition 5). Our framework replaces the geometric distance with the information-theoretic Rényi-∞ ratio and produces a sharp phase-transition threshold rather than a smooth bound. El-Mhamdi and Hoang [6] compare tail behavior between target and proxy and obtain dominance results in heavy-tailed regimes; our D_∞-based threshold is one (information-theoretic) view among several, and it is the appropriate one when the relevant pathology is concentration of the proxy-induced occupancy.

5 Sample Complexity Lower Bounds

5.1 Fano lower bound for the Gaussian reward channel

Suppose the hypothesis class $\mathcal{V}$ contains an N-element *local* packing $\{V_1, \dots, V_N\}$, i.e. pairwise $L^2(\mu)$ distance $\|V_i - V_j\|_{L^2(\mu)} \in [\varepsilon, 2\varepsilon]$ for all $i \neq j$ (the standard local-packing setup; cf. 10, Theorem 2.5). The channel is (I-rew) with i.i.d. Gaussian noise of variance σ^2.

Proposition 17 (PAC Fano lower bound)**.** *Let V^* be uniform on $\{V_1, \dots, V_N\}$. For any algorithm A that observes m i.i.d. noisy reward samples and outputs $\hat{V}$, the probability of error $P_e := \mathbb{P}[\hat{V} \neq V^*]$ satisfies*

$$P_e \;\geq\; 1 - \frac{2m\varepsilon^2/\sigma^2 + \log 2}{\log N}.$$

In particular, the necessary condition $P_e < 1/2$ requires

$$\boxed{m \;>\; \frac{\sigma^2 \log(N/4)}{4\varepsilon^2}.}$$

Proof. By Fano,

$$P_e \;\geq\; 1 - \frac{I(V^*; D) + \log 2}{\log N}.$$

For Gaussian channels with shared variance, integrating against μ, $\mathrm{KL}(P_i \,\|\, P_j) = \|V_i - V_j\|^2_{L^2(\mu)}/(2\sigma^2)$. By the local-packing assumption, $\|V_i - V_j\|_{L^2(\mu)} \leq 2\varepsilon$ for all $i \neq j$, hence

$$\bar{K} \;:=\; \frac{1}{N^2} \sum_{i,j} \mathrm{KL}(P_i \| P_j) \;\leq\; \frac{(2\varepsilon)^2}{2\sigma^2} \;=\; \frac{2\varepsilon^2}{\sigma^2}.$$

For m i.i.d. samples this scales as $I(V^*; D) \leq m\bar{K} \leq 2m\varepsilon^2/\sigma^2$. Substituting,

$$P_e \;\geq\; 1 - \frac{2m\varepsilon^2/\sigma^2 + \log 2}{\log N}.$$

Requiring $P_e < 1/2$ yields $2m\varepsilon^2/\sigma^2 + \log 2 > (\log N)/2$, i.e. $m > \sigma^2 \log(N/4)/(4\varepsilon^2)$. □

Remark 18 (Constants)**.** The constant in front of $\log N$ depends on the precise success probability targeted. The bound stated here is the necessary condition for $P_e < 1/2$; analogous bounds with different multiplicative constants follow for $P_e < \delta$ with general δ. The asymptotic statement $m = \Omega(\sigma^2 \log N/\varepsilon^2)$ holds in all variants.

5.2 Channel comparison

Proposition 19 (Channel comparison on argmax-shared pairs)**.** *Let V_1, V_2 be a pair sharing the same* $\arg\max$ *at every state in* $\mathrm{supp}(\mu)$*. Let D be a single observation drawn under $V^* \in \{V_1, V_2\}$.*

(i) ***(I-demo)****: $I(V^*; D) = 0$ on the argmax-shared pair, so the Fano bound certifies failure for any m.*

(ii) ***(I-rew)*** *with Gaussian noise: $I(V^*; D) = \Theta(\varepsilon^2/\sigma^2)$ per sample for an ε-separated pair, hence $m = \Theta(\log N/\varepsilon^2)$ suffices for PAC-success.*

(iii) ***(I-pref)*** *under Bradley–Terry with separation ε: $I(V^*; D) = \Theta(\varepsilon^2)$ per sample, hence $m = \Theta(\log N/\varepsilon^2)$.*

Proof. (i) Demonstrations are sampled from $\pi^*_{V^*} = \arg\max V^*$, which by hypothesis coincides under V_1 and V_2. Hence $P(D|V_1) = P(D|V_2)$ and $I(V^*; D) = 0$.

(ii) Standard for Gaussian channels: $I(V^*; D) \leq$ KL-divergence between the two reward distributions, which is $(V_1 - V_2)^2/(2\sigma^2) \leq 2\varepsilon^2/\sigma^2$. The matching upper bound follows from the standard maximum likelihood estimator for the Gaussian channel.

(iii) For the Bradley–Terry model the Bernoulli parameter is $p_V = \sigma_{\mathrm{BT}}(V(\tau_A) - V(\tau_B))$ where $\sigma_{\mathrm{BT}}(x) = 1/(1 + e^{-x})$. The per-sample KL between the two Bernoulli laws is

$$\mathrm{KL}\big(p_{V_1} \,\|\, p_{V_2}\big) \;=\; p_{V_1} \log \frac{p_{V_1}}{p_{V_2}} + (1 - p_{V_1}) \log \frac{1 - p_{V_1}}{1 - p_{V_2}}.$$

A Taylor expansion at $V_1 = V_2$ uses $\sigma'_{\mathrm{BT}}(0) = 1/4$ together with $\mathrm{KL}(p\|q) = (p-q)^2/(2q(1-q)) + O((p-q)^3) \to 2(p-q)^2$ as $p, q \to 1/2$, giving $\mathrm{KL} = 2 \cdot (V_1 - V_2)^2/16 = \frac{1}{8}(V_1 - V_2)^2 + O((V_1 - V_2)^3)$, which is $\Theta(\varepsilon^2)$ for separation $\varepsilon = O(1)$. The matching upper bound for $m = \Theta(\log N/\varepsilon^2)$ is the standard pairwise-comparison MLE rate. □

Remark 20 (Channel versus algorithm)**.** Proposition 19(i) is the source of the deploy-shift floor of Theorem 4: on argmax-shared hypothesis pairs, the demonstration channel is uninformative no matter how the algorithm processes the data. Switching to (I-rew) or (I-pref) restores positive information rate, but Theorem 4 continues to apply within the (I-demo) channel.

Remark 21 (Softmax demonstrations)**.** A common refinement of (I-demo) replaces deterministic argmax expert play with a softmax (Boltzmann) policy of inverse temperature β_{exp}. Under this model, the per-sample mutual information on an ε-separated argmax-shared pair is $\Theta(\beta^2_{\mathrm{exp}}\varepsilon^2)$ in the regime $\beta_{\mathrm{exp}}\varepsilon \ll 1$. The required sample size is $m = \Theta(\log N/(\beta^2_{\mathrm{exp}}\varepsilon^2))$, which diverges as $\beta_{\mathrm{exp}} \to \infty$ (perfectly rational expert). Theorem 4 captures the limit $\beta_{\mathrm{exp}} = \infty$: the floor is the obstruction that survives even when expert randomness vanishes.

5.3 Active demonstration querying

A natural counter to Theorem 4 is to allow the learner to query the expert at chosen states, including the OOD state s'. We show that under the demonstration channel this does not lift the floor: the expert can only report her optimal action, which on argmax-shared hypothesis pairs is itself uninformative.

Proposition 22 (Active demonstration does not break the floor)**.** *Suppose the learner adaptively queries an expert at any state and receives the action $\hat{a} \sim \pi^*_{V^*}(\cdot|s)$. If V_1, V_2 share the same* arg max *at every state, then for every transcript T of q adaptive queries,*

$$P(T|V_1) \;=\; P(T|V_2),$$

and $I(V^*; T) = 0$. Consequently, the Fano bound and the deploy-shift floor of Theorem 4 apply unchanged.

Proof. By induction on q. For $q = 1$ the claim is Proposition 19(i). For the inductive step, suppose the transcript T_{q-1} has identical distributions under V_1 and V_2. The next query s_q is determined by T_{q-1} (deterministically or via a fixed randomization), and the response distribution at s_q is

identical under both hypotheses by the argmax-shared assumption. Hence the joint distribution of T_q is identical under V_1 and V_2.

The bound on Theorem 4 now follows: for any algorithm A that uses the active transcript T to produce a deployment policy, the marginal $q = \mathbb{P}[A(T)(s') = a_2]$ is identical under V_1 and V_2, and the minimax argument yields $\sup_{V^*} \mathbb{E}[\mathcal{L}_{\text{deploy}}] \geq \beta\Delta/2$. □

Remark 23 (Sequential Fano)**.** Proposition 22 can be strengthened to a sequential Fano statement using the data-processing inequality on stopping-time-bounded transcripts. We do not pursue the most general version here; see Section 6 for open problems.

5.4 Implications for channel design

The combination of Theorem 4, Proposition 19, and Proposition 22 yields a coherent picture of channel-induced information limits.

Corollary 24 (Channel hierarchy on argmax-shared pairs)**.** *On any pair V_1, V_2 that share* arg max *on the demonstration support and disagree on an OOD state of deployment mass $\beta > 0$:*

(i) ***(I-demo)**, including active demonstration querying: deployment loss is bounded below by $\beta\Delta/2$ for every m.*

(ii) ***(I-rew)** with Gaussian noise of variance σ^2 and L^2-separation ε: deployment loss can be driven below any $\varepsilon' > 0$ with $m = \Theta(\sigma^2 \log N/(\varepsilon')^2)$ samples.*

(iii) ***(I-pref)** under Bradley–Terry: deployment loss can be driven below any $\varepsilon' > 0$ with $m = \Theta(\log N/(\varepsilon')^2)$ samples.*

Proof. (i) is Theorem 4 together with Proposition 22. (ii) and (iii) follow from Proposition 19 together with the upper bound provided by maximum-likelihood-style estimators for the respective channels. The qualitative point is that the demonstration channel has zero information rate on the relevant hypothesis pair, whereas the reward and preference channels have positive rate that allows arbitrary accuracy at $\Theta(\log N/\varepsilon^2)$ rate. □

The corollary justifies a maxim that has been folkloric in alignment-relevant value learning: *when the deployment distribution differs from the demonstration distribution, the practical question is whether the channel disambiguates argmax-shared rewards on the OOD region.* Demonstrations alone cannot; a small admixture of preference or reward signal restores identifiability.

6 Discussion

6.1 Comparison with prior IRL bounds

Komanduru and Honorio [3] prove a lower bound of order $\Omega(n \log n)$ on the number of expert trajectories needed to recover an IRL solution on n-state MDPs in-distribution. Their bound is asymptotic in m for given accuracy. Theorem 4 is fundamentally different: it asserts an irreducible floor $\beta\Delta/2$ on the deployment loss that no m can overcome, valid even when in-distribution recovery is perfect. The two results are complementary and address distinct regimes.

Metelli et al. [7] establish a tight $\Omega(H^3SA(\log(1/\delta)+S)/\varepsilon^2)$ PAC bound for IRL with generative-model access. As they note, their analysis is in-distribution: the deployment evaluation matches the sampling distribution. They do not analyze deployment shift. Our Theorem 4 isolates the OOD ambiguity that their framework, by construction, cannot capture. The two results are consistent: in the regime $\beta = 0$ the deploy-shift floor degenerates and the in-distribution Metelli rate governs.

6.2 Comparison with prior Goodhart frameworks

Karwowski et al. [9] introduce projected angle distance between reward functions and prove early-stopping bounds (their Proposition 5). Their measurement is geometric and produces smooth (rather than threshold) bounds. Our framework uses the Rényi-∞ divergence and yields a phase-transition threshold $D_\infty \cdot \varepsilon = \Theta(1)$ that sharply distinguishes regressional from extremal regimes. The frameworks are complementary; we view neither as subsuming the other.

El-Mhamdi and Hoang [6] compare light- and heavy-tailed proxy distributions, obtaining tail-dominance results principally in the in-distribution setting (their §4.2 contains limited deployment-time discussion). Our deploy-shift framing and D_∞ threshold provide a different lens; for proxies whose pathology is concentration of induced occupancy rather than tail mismatch, the D_∞ view is the natural one.

Manheim and Garrabrant [5] provide an informal four-class taxonomy. Proposition 15 maps each class onto a quantitative D_∞ regime. The mapping is exact for regressional, extremal, and adversarial Goodhart; for causal Goodhart the framework requires an extension to handle policy-induced changes in ρ_0.

6.3 Limitations and scope

The deploy-shift floor is a strict refinement of, not a substitute for, in-distribution PAC bounds. With $\beta = 0$ it produces no obstruction. The two-point construction gives a lower bound for the worst-case-over-rewards minimax loss but says nothing about average-case performance under priors that put low mass on argmax-coincident pairs. The Goodhart phase transition is stated for the Rényi-∞ divergence; analogous results for D_α at finite α require a separate argument. The Fano bound assumes the Gaussian reward channel; the constants for sub-Gaussian or Bernoulli channels differ but the asymptotic rate $m = \Omega(\log N/\varepsilon^2)$ is preserved. The active-learning result assumes the expert reports only optimal actions; richer expert protocols (preference labels, scalar feedback, demonstrations augmented by counterfactual annotations) circumvent the result, consistent with Proposition 19.

6.4 Open problems

- **Uniform sharp bound across regressional and extremal regimes.** A unified bound that smoothly interpolates between the $O(\varepsilon)$ rate at $D_\infty = O(1)$ and the $\Theta(1)$ rate at $D_\infty = \Theta(1/\varepsilon)$ would clarify the transition geometry. The current statement gives the two endpoints but not the interpolation.

- **Sequential Fano for active learning.** A general lower bound on the sample complexity of active demonstration querying as a function of the channel and the hypothesis structure would extend Proposition 22 beyond the argmax-shared regime.

- **Beyond demonstrations.** The deploy-shift floor mechanism applies to any channel that is invariant under hypothesis transformations preserving the demonstration distribution. A taxonomy of channel-induced floors would unify the demonstration-only result with analogous obstructions for partial preferences and partial reward signals.

- **Continuous state and action spaces.** The construction in Section 3.1 is finite. A continuous version requires care with measurability of ρ^π on the OOD region; we conjecture the floor extends without modification.

6.5 Conclusion

The principal lesson is that when value learning is evaluated under deployment shift, the choice of *information channel* is more decisive than the choice of *algorithm*. The demonstration channel admits an irreducible floor on a class of MDPs that is non-trivial under almost any deployment distribution that places mass on out-of-distribution states. Increasing m does not help. The remedy is channel upgrade — preference or reward signals — coupled with control of D_∞ to avoid the extremal Goodhart regime.

Acknowledgments

The author thanks colleagues in the alignment-theory community for discussions that sharpened the comparisons with prior work. All errors are the author's.

References

[1] S. Armstrong and S. Mindermann. Impossibility of deducing preferences and rationality from human policy. In *Advances in Neural Information Processing Systems*, 2017.

[2] R. M. Fano. *Transmission of Information*. MIT Press, 1961.

[3] A. Komanduru and J. Honorio. A lower bound for the sample complexity of inverse reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning*, 2021. arXiv:2103.04446.

[4] L. Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53, 1973.

[5] D. Manheim and S. Garrabrant. Categorizing variants of Goodhart's law. arXiv:1803.04585, 2018.

[6] E.-M. El-Mhamdi and L.-N. Hoang. On Goodhart's law, with an application to value alignment. arXiv:2410.09638, 2024.

[7] A. M. Metelli et al. Towards theoretical understanding of inverse reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. arXiv:2304.12966.

[8] A. Y. Ng and S. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning*, 2000.

[9] J. Karwowski, O. Hayman, X. Bai, K. Kiendlhofer, C. Griffin, and J. Skalse. Goodhart's law in reinforcement learning. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2310.09144.

[10] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.