Inverse Reinforcement Learning with Just Classification and a Few Regressions

Lars van der Laan
Department of Statistics, University of Washington
lvdlaan@uw.edu

Nathan Kallus
Netflix
Cornell University
nkallus@cornell.edu

Aurelien Bibaut
Netflix
abibaut@netflix.com

Abstract

Inverse reinforcement learning (IRL) seeks to recover a reward function from observed behavior, but rewards are typically only partially identified: many reward–value pairs can induce the same behavior policy. We study this problem in the maximum-entropy, or Gumbel-shock, model under a broad class of statewise affine normalization constraints, with anchor-action constraints as a special case. This leads to Generalized Policy-to-Q-to-Reward (GenPQR), a modular approach to normalized reward recovery based on policy estimation and Q-evaluation via the Bellman equation; both components can be instantiated with off-the-shelf classification and regression methods. We establish modular finite-sample guarantees under general function approximation, with separate terms for policy-estimation and Q-function estimation error. As a concrete instantiation, we study GenPQR with fitted Q-evaluation, reducing IRL to policy estimation followed by regression. In experiments, GenPQR improves on DeepPQR in reward recovery while remaining simpler and more modular. Relative to DeepPQR, our theory is broader: it goes beyond anchor actions, accommodates large action spaces, and is not tied to a specific neural-network architecture or training procedure.

1 Introduction

Behavioral data are abundant in robotics, economics, healthcare, and human–computer interaction. Inverse reinforcement learning (IRL) seeks to explain such behavior by recovering a reward under which the observed policy is optimal. Classical IRL often treats agents as exactly optimal [29, 1], but this deterministic view can miss the variability in real behavior. A common alternative is stochastic choice, for example through entropy regularization, which yields softmax-type policies [55, 56, 14]. In maximum-entropy (MaxEnt) IRL [55], closely related to dynamic discrete-choice (DDC) models with i.i.d. Gumbel shocks [33], the observed policy has a softmax form induced by an unknown reward and continuation value.

Even in the softmax setting, rewards are only partially identified: different reward–value pairs can induce the same behavior policy through potential-based shaping transformations [30, 5, 36, 35]. Thus policy fit alone does not imply reward recovery. Existing methods obtain unique rewards only by adding restrictions, often implicitly: MaxEnt IRL identifies an equivalence class unless the reward class is restricted, for example to be linear [55]; adversarial IRL requires strong conditions for reward recovery, such as state-only rewards and deterministic transitions [11]; and other neural IRL objectives depend on their chosen reward parameterization [24, 48, 16, 37]. This motivates

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

normalized reward recovery, a classical idea in DDC/econometrics: impose an explicit identifying normalization and recover the unique reward–value pair satisfying it [33, 17, 3, 13].

Building on classical DDC for discrete actions, Deep Policy-to-Q-to-Reward (DeepPQR) [13] implements this idea under anchor-action normalization. Given an observed behavior policy, or a policy estimated by an upstream method such as imitation learning or adversarial IRL, DeepPQR estimates the associated Q-function using value-based offline RL tools and then recovers rewards from the normalized Bellman equation. Importantly, the procedure uses only the upstream policy, not the upstream reward model, so reward recovery does not rely on the correctness of that model. Its identification strategy, however, is tied to a fixed anchor action: one must specify a well-supported reference action, typically a do-nothing action, whose reward is known. This can be restrictive when actions are continuous, weakly supported, or lack a canonical reference option. Other settings may instead motivate normalizations based on mean rewards, state-dependent anchors, or value-based constraints.

We extend this policy-to-Q-to-reward perspective to statewise affine normalizations. In this formulation, a problem-informed normalization identifies a unique reward without requiring the normalization to take the form of a fixed anchor action. Given an estimated behavior policy, GenPQR solves the corresponding Bellman fixed point for Q and recovers the normalized reward from the Bellman equation. This retains the modular structure of DeepPQR while allowing normalization choices beyond a single reference action.

Our contributions. First, we characterize reward identification in maximum-entropy IRL, equivalently the Gumbel-shock discrete-choice model. We show that behavior identifies only an equivalence class of reward–value pairs, and introduce statewise affine normalizations that select a unique representative. This generalizes the anchor-action normalization of DeepPQR [13] and clarifies when exact reward recovery is possible, and when it is unnecessary for policy comparison.

Second, this characterization yields a modular identification and estimation strategy. Under any statewise affine normalization, reward recovery reduces to estimating the behavior policy, solving a linear Bellman fixed point for an associated Q-function, and applying the normalized Bellman equation. This gives *Generalized Policy-to-Q-to-Reward* (GenPQR), which treats reward recovery as Q-evaluation after policy estimation rather than as a specialized joint IRL objective. In the fixed-anchor setting, GenPQR recovers rewards directly from Q, avoiding the additional reward-regression step in DeepPQR; DeepPQR can be viewed as one neural-network implementation of the Q-estimation step. More generally, the same reduction covers state-dependent anchors, mean-reward and value normalizations, and large or continuous action spaces through the choice of policy and Q-estimation methods.

Third, we prove finite-sample guarantees for GenPQR under function approximation. The bounds apply to generic policy and Q-function estimators; for FQE, they separate policy-estimation, Bellman-approximation, statistical, and iteration errors. Relative to the DeepPQR analysis, our theory replaces sup-norm policy-error conditions with weaker error measures, makes coverage explicit, allows approximate rather than exact Bellman completeness, and is not tied to a particular neural-network architecture or training procedure.

Fourth, controlled experiments isolate the practical consequences of this generality. GenPQR remains accurate as the number of actions grows and, in a 20-action data-fusion setting, combines state–action transitions with an independent state–outcome sample to improve reward recovery beyond behavior-only anchor identification.

1.1 Related Work

Identifiability, shaping, and anchor-action methods. Rewards in MaxEnt IRL are only partially identified because behavior is invariant under potential-based shaping [30, 11, 5, 36, 35]. Our work is closest to DeepPQR [13], which studies anchor-action normalization and gives finite-sample guarantees for a specific neural-network procedure. We extend the anchor-action view to general affine normalizations, clarify the identification structure, simplify the recovery step (Section 3.3), and develop theory under general function approximation.

Behavioral and statistical misspecification. Behavioral misspecification arises when the observed policy is not generated by the assumed decision model, whereas statistical misspecification arises

when the policy or Q-function classes cannot represent their population targets. Any full-support policy admits a MaxEnt rationalization, and our normalization selects one reward from the resulting shaping class. Under behavioral misspecification, this model-implied reward remains well defined but need not equal the agent's structural reward; robustness depends on the form of the departure from the assumed choice model [34, 35]. Our finite-sample bounds separately quantify policy-estimation and Bellman-approximation errors within the MaxEnt/Gumbel-shock model.

MaxEnt IRL and adversarial imitation learning. Maximum-entropy IRL fits stochastic policies induced by soft Bellman equations, often under structured reward parameterizations [55, 56, 24, 48, 53]. Adversarial methods such as GAIL and AIRL are effective for imitation learning and behavior-policy estimation through joint reward-policy optimization [16, 11]. As noted by Geng et al. [13], however, they generally do not resolve reward nonidentifiability without stronger assumptions, such as state-only rewards. We instead separate behavior-policy estimation from normalized reward recovery, allowing action-dependent rewards. Thus, methods designed primarily to reproduce behavior can still serve as the first stage of a modular reward-recovery procedure.

Entropy-regularized RL and control as inference. Our analysis is also connected to entropy-regularized control and the control-as-inference perspective [21, 38, 23], including path-consistency learning [28], soft actor-critic [15], and entropy-regularized offline RL [14, 43]. We use the same soft Bellman structure, but for the inverse problem: the behavior policy is assumed to solve an entropy-regularized control problem for an unknown reward, and the goal is to recover that reward.

Value-based offline RL. The recovery step is closely connected to value-based offline RL. Fitted Q-iteration and fitted Q-evaluation estimate Bellman fixed points by regression [7, 27, 25, 46], while minimax and critic-based methods relax completeness assumptions [41–43, 49, 50, 54]. We do not introduce a new RL update rule; instead, we show that normalized reward recovery reduces to solving a linear fixed-point equation with existing tools. Our finite-sample analysis separates first-stage policy-estimation error from second-stage value-learning error.

2 Problem Setup

We consider a discounted MDP with state space $\mathcal{S}$, finite or continuous action space $\mathcal{A}$, transition kernel P, reward $r^\dagger : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and discount $\gamma \in [0, 1)$. Let $\pi(a \mid s)$ be the behavior policy and ρ the sampling distribution over states. We observe transitions $\{(s_i, a_i, s_i')\}_{i=1}^n$ with

$$s_i \sim \rho, \qquad a_i \sim \pi(\cdot \mid s_i), \qquad s_i' \sim P(\cdot \mid s_i, a_i).$$

Equivalently, $(s_i, a_i) \sim \nu_\pi$, where $\nu_\pi(ds, a) := \rho(ds)\pi(a \mid s)$. Thus P is identified from the observed dynamics, whereas the reward is not. We use finite-action notation in the main text, following Rust [33], Geng et al. [13]; continuous actions replace sums and softmax normalizers by integrals and Boltzmann densities with respect to a reference measure (Appendix G). For any state-action function f, write

$$(\mu f)(s) := \sum_a \mu(a \mid s) f(s, a).$$

Throughout the identification analysis, $\pi(a \mid s) > 0$ for every action and ρ-almost every state, $u^\star = \log \pi$ is bounded, and the one-step coverage condition $\nu_\pi P \ll \rho$ holds. Statewise conclusions are understood ρ-almost everywhere and state–action conclusions ν_π-almost everywhere. The **goal** of IRL is to recover a reward r for which π is optimal in the MDP $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ under the entropy-regularized criterion below. We review the MaxEnt IRL setting from structural discrete-choice and maximum-entropy perspectives, then formulate the optimization problem central to our analysis.

From dynamic discrete choice to MaxEnt IRL. We adopt the dynamic discrete-choice formulation [33, 17, 3]. At time t, an agent in state s_t who takes action a_t receives utility $r^\dagger(s_t, a_t) + \varepsilon_t(a_t)$, where $r^\dagger$ is the unknown mean reward and $\varepsilon_t(a)$ is an idiosyncratic shock with known distribution. Let $V^\dagger(s)$ be the optimal ex ante value, define $Pf(s, a) := E\{f(s') \mid s, a\}$, and set

$$Q^\dagger(s, a) := r^\dagger(s, a) + \gamma P V^\dagger(s, a).$$

For shock scale $\tau > 0$, define

$$\mathrm{LSE}_\tau\, f(s) := \tau \log \sum_a \exp\{f(s, a)/\tau\}.$$

Under i.i.d. Gumbel type-I extreme-value shocks with scale τ, the optimal policy is softmax:

$$\pi^\dagger(a \mid s) \propto \exp\{Q^\dagger(s,a)/\tau\},$$

and

$$Q^\dagger = r^\dagger + \gamma P \operatorname{LSE}_\tau Q^\dagger, \qquad V^\dagger(s) = \operatorname{LSE}_\tau Q^\dagger(s).$$

Equivalently, with state-action continuation value $v^\dagger := P \operatorname{LSE}_\tau Q^\dagger$,

$$\pi^\dagger(a \mid s) \propto \exp\{(r^\dagger(s,a) + \gamma v^\dagger(s,a))/\tau\}, \qquad v^\dagger = P \operatorname{LSE}_\tau(r^\dagger + \gamma v^\dagger).$$

We call the latter the soft Bellman equation. We set $\tau = 1$, write $\operatorname{LSE} := \operatorname{LSE}_1$, and absorb the known scale into $r^\dagger$.

An equivalent perspective comes from maximum-entropy IRL [55, 56], where the agent maximizes expected discounted reward plus an entropy bonus for stochastic action selection. This yields the same optimal policy $\pi^\dagger$ and soft Bellman equation, with $V^\dagger$ and $Q^\dagger$ the entropy-regularized value and Q-functions [14]. Thus, despite different motivations, dynamic discrete choice and MaxEnt IRL reduce to the same mathematical object: a soft Bellman system with a softmax policy.

Partial identification and the role of normalization. Our goal is therefore to recover a reward function whose induced soft-optimal policy best matches the observed behavior, for example by minimizing the state-averaged Kullback–Leibler divergence from $\pi(\cdot \mid s)$ to $\pi^\star(\cdot \mid s)$. However, matching the policy does not in general identify the reward. Under a softmax policy, adding a state-dependent offset to all action values leaves the policy unchanged, so many rewards induce the same behavior policy [56, 5]. The reward is therefore only *partially identified*, up to an equivalence class, and selecting a unique representative requires a normalization constraint [33, 13].

We use **statewise affine normalizations**, which generalize standard choices such as anchor-action, outside-option, and sum-to-zero constraints. Let $\mu(\cdot \mid s)$ be a reference distribution over actions and $g : \mathcal{S} \to \mathbb{R}$ a specified anchor function. We impose

$$\sum_a \mu(a \mid s) r(s,a) = g(s) \qquad \text{for } \rho\text{-almost every } s, \tag{1}$$

or equivalently $\mu r = g$, with $\sup_s |g(s)| < \infty$. Thus, the μ-average reward at each state is fixed at $g(s)$, with μ and g typically chosen from domain knowledge.

Such constraints are standard in economics and are often substantively meaningful [17, 4]. The fixed anchor-action constraint $r(s, a^\dagger) = g(s)$ is the special case $\mu(a \mid s) = 1\{a = a^\dagger\}$ [13]; when $g \equiv 0$, it reduces to the classical zero-reward normalization $r(s, a^\dagger) = 0$ used in Rust's engine-replacement model [33]. More generally, taking $\mu(a \mid s) = 1\{a = a^\dagger(s)\}$ allows state-specific anchors, where the reference action may vary with the state. DDC models often use such anchors to normalize the payoff of a no-action, status-quo, or outside option to be zero or otherwise known [33, 17, 3, 13]. If $\mu(a \mid s) = 1/|\mathcal{A}|$, the constraint fixes the statewise average reward and, when $g \equiv 0$, becomes a sum-to-zero constraint [19].

The framework also permits data-driven choices of (μ, g). A statewise value constraint is a special case: if V_r^μ is the value function of policy μ under reward r, then imposing $V_r^\mu = h$ is equivalent, by the Bellman equation $V_r^\mu = \mu r + \gamma \mu P V_r^\mu$, to taking $g := h - \gamma \mu P h$. For example, in medical decision problems, one may wish to normalize rewards relative to standard care using historical outcome data [20]. If an auxiliary dataset records states and realized outcomes (s_i, y_i) from a population with behavior policy π, where $E[y \mid s, a] = r(s,a)$, then $g(s) = E[y \mid s] = \sum_a \pi(a \mid s) r(s,a)$. Taking $\mu = \pi$ therefore yields a natural data-driven normalization.

Main problem: Putting these pieces together, we obtain the following constrained maximum-likelihood problem for recovering $(r^\star, v^\star)$:

$$\begin{aligned} \arg\max_{r,v} \quad & \mathbb{E}_{(s,a)\sim\nu_\pi}[r(s,a) + \gamma v(s,a) - \operatorname{LSE}(r + \gamma v)(s)] \\ \text{s.t.} \quad & v = P \operatorname{LSE}(r + \gamma v), \qquad \text{(soft Bellman)} \\ & \mu r = g. \qquad \text{(affine normalization)} \end{aligned} \tag{2}$$

That is, we maximize the conditional log-likelihood over state–action functions r and v subject to the soft Bellman equation and a statewise affine normalization constraint.

The remainder of the paper shows that solving (2) reduces to two steps: estimate π, then solve a linear Bellman evaluation equation to recover the unique pair $(r^\star, v^\star)$. We then turn this characterization into a simple algorithm.

3 From partial to point identification

To expose the core structure of the problem, we first drop the normalization constraint and study the resulting relaxed optimization problem.

Relaxed problem: We remove the reward normalization in (2) but keep soft Bellman consistency:

$$\begin{aligned} \arg\max_{r,v} \quad & \mathbb{E}_{(s,a)\sim\nu_\pi}[r(s,a) + \gamma v(s,a) - \mathrm{LSE}(r+\gamma v)(s)] \\ \text{s.t.} \quad & v = P\,\mathrm{LSE}(r+\gamma v). \qquad \text{(soft Bellman)} \end{aligned} \tag{3}$$

This problem is highly non-unique. Because the objective is invariant under potential-based transformations of the reward, the relaxed problem admits an entire equivalence class of solutions. We will exploit this invariance later to recover a solution to the normalized problem (2) from a single convenient representative.

3.1 Behavior cloning solves the relaxed problem

Without normalization, the relaxed problem admits a particularly simple solution: estimate the behavior policy π and set $r(s,a) = \log \pi(a \mid s)$ and $v(s) = 0$. This pair maximizes the conditional log-likelihood and satisfies the soft Bellman equation. Indeed, $P\,\mathrm{LSE}(r+\gamma v) = 0$ since $\exp\{\mathrm{LSE}(r+\gamma v)(s)\} = \sum_a \exp\{r(s,a)\} = \sum_a \pi(a \mid s) = 1$. We will use the shorthand

$$u^\star(s,a) := \log \pi(a \mid s).$$

Lemma 1 (Trivial optimum of the relaxed problem)**.** *The pair $(r,v) := (u^\star, 0)$ solves (3).*

A related observation appears in Section 4 of Fu et al. [11], but not as a tool for identification or estimation. Here, by contrast, it is the starting point for normalization-based identification. As shown in Section 3.4, this trivial solution already suffices for policy-value comparisons.

3.2 An invariance among solutions

The relaxed problem is invariant to *potential-based shaping*: adding a state-only potential $c : \mathcal{S} \to \mathbb{R}$ shifts all logits in the softmax by the same amount per state, leaving both feasibility and likelihood unchanged. This is the entropy-regularized analogue of reward shaping in classical RL [30] and explains why the relaxed objective is flat along an affine subspace.

Lemma 2 (Potential-based shaping invariance)**.** *Let (r,v) be feasible for (3), and let $c : \mathcal{S} \to \mathbb{R}$ be bounded. Define $\tilde{r} = r + c - \gamma Pc$ and $\tilde{v} = v + Pc$. Then $(\tilde{r}, \tilde{v})$ is also feasible for (3) and attains the same objective value as (r,v). In particular, $r + \gamma v - \mathrm{LSE}(r+\gamma v)$, the induced log-policy, is unchanged.*

Related partial-identification results appear in Cao et al. [5, Theorem 1], Fu et al. [11], and Geng et al. [13, Lemma 2]. We next show that our normalization selects a unique representative from this class.

3.3 Solving the original normalized problem

Lemma 1 gives one relaxed optimum, $(u^\star, 0)$, and Lemma 2 characterizes all others via potential-based transformations of the form $(r,v) = (u^\star + c - \gamma Pc, Pc)$. Since the constrained and relaxed problems have the same optimal value, solving (2) amounts to finding the shaping function c such that the corresponding pair (r,v) satisfies the desired constraint. For our normalization $\mu r = g$, c is uniquely determined by a Bellman equation.

The next result gives the corresponding solution in terms of $Q^\mu_{u^\star - g}$, the Q-function under reward $u^\star - g$ and policy μ, where $u^\star = \log \pi$. Recall that

$$(P\mu Q)(s,a) := \mathbb{E}_{s'\sim P(\cdot|s,a),\, a'\sim\mu(\cdot|s')}\big[Q(s',a')\big].$$

Theorem 1 (IRL via a Bellman equation)**.** *Let $Q^\mu_{u^\star - g}$ be the unique bounded solution in $L^\infty(\nu_\pi)$ to*

$$Q^\mu_{u^\star - g}(s,a) = u^\star(s,a) - g(s) + \gamma(P\mu Q^\mu_{u^\star - g})(s,a).$$

Then (2) *admits a unique optimal solution in* $L^\infty(\nu_\pi)$*. For* ρ*-almost every* s *and every action* a*, it is given by*

$$r^\star(s,a) = Q^\mu_{u^\star - g}(s,a) - (\mu Q^\mu_{u^\star - g})(s) + g(s),$$
$$v^\star(s,a) = -(P\mu Q^\mu_{u^\star - g})(s,a).$$

When $\gamma > 0$*, equivalently* $v^\star = \gamma^{-1}(u^\star - g - Q^\mu_{u^\star - g})$*.*

This theorem is the main result of the paper. It shows that normalized reward recovery reduces to two steps: first estimate $u^\star = \log \pi$, then solve the linear Bellman equation for $Q^\mu_{u^\star - g}$. The normalized reward is then obtained in closed form as the advantage function $Q^\mu_{u^\star - g} - \mu Q^\mu_{u^\star - g}$, shifted by g. By Lemma 2, every feasible representative in the shaping class induces the same $u^\star$. Therefore, once one feasible representative is known, we can impose a different normalization without re-solving the original IRL problem.

Comparison to DeepPQR. In the anchor-action setting $\mu(a \mid s) = 1\{a = a^\dagger\}$, DeepPQR [13] first estimates the behavior policy and then learns the entropy-regularized soft anchor Q-function, denoted here by Q_A, which satisfies

$$Q_A(s) = g(s) + \gamma \mathbb{E}\left[Q_A(s') - u^\star(s', a^\dagger) \mid s, a^\dagger\right].$$

This quantity is related to GenPQR's anchor-policy value by $W(s) = u^\star(s, a^\dagger) - Q_A(s)$, which obeys

$$W(s) = u^\star(s, a^\dagger) - g(s) + \gamma \mathbb{E}[W(s') \mid s, a^\dagger].$$

Thus, DeepPQR estimates Q_A, which can be transformed into W, rather than estimating W directly. GenPQR instead works with $Q^\mu_{u^\star - g}$ itself, from which the normalized reward and continuation value $v^\star$ follow by the normalized Bellman equation. This avoids separately estimating PW and gives a modular second stage based on standard Q-evaluation.

The comparison also highlights a practical tradeoff. DeepPQR relies on a well-supported anchor action $a^\dagger$. When such observations are limited, GenPQR can estimate the relevant Q-function using methods that share information across actions, and it extends beyond fixed-anchor discrete-action settings. This is consistent with our experiments: GenPQR is more stable when anchor actions are rare or the action space is large, whereas DeepPQR's anchor-action regression degrades in these settings and does not directly extend to continuous actions.

3.4 Behavior cloning suffices for policy comparison

For policy comparison, exact reward recovery is unnecessary once the transition kernel P and discount factor γ are fixed. Let $V^\pi_r = \pi Q^\pi_r$ denote the value function under reward r and policy π, where Q^π_r is the corresponding Q-function.

Theorem 2 (Identification of policy value differences)**.** *Let* (r, v) *solve* (3)*, for example* $(u^\star, 0)$*. Then, for any two policies* π_1, π_2*,*

$$V^{\pi_1}_{r^\star}(s) - V^{\pi_2}_{r^\star}(s) = V^{\pi_1}_r(s) - V^{\pi_2}_r(s) \qquad \textit{for } \rho\textit{-almost every } s.$$

Thus, exact reward recovery is needed only for targets that depend on the normalization itself, such as evaluation under counterfactual transition dynamics or discount factors, or interpretation of structural features of r. Otherwise, $u^\star$, or any reward solving the relaxed problem, suffices for policy comparison. This parallels Hotz and Miller [17], where under Gumbel shocks value differences are identified from log-odds of observed choices.

4 A generic algorithm

Algorithm 1 presents Generalized Policy-to-Q-to-Reward (GenPQR) for reward recovery. The method has two standard steps. First, we estimate the behavior policy $\pi(a \mid s)$, equivalently $u^\star(s,a) = \log \pi(a \mid s)$, yielding $\hat{u}$. Second, we recover $(r^\star, v^\star)$ by estimating the Q-function of policy μ under reward $\hat{u} - g$, then applying the plug-in formula in Theorem 1. Appendix G gives the continuous-action version: estimate a conditional log-density, solve the same Q-fixed point with μ-integrals, and apply the same normalized advantage formula. When $\hat{\pi}(a \mid s)$ is close to 0, $\hat{u}$ may

Algorithm 1 GENERALIZED POLICY-TO-Q-TO-REWARD (GENPQR)

input Data $\{(s_i, a_i, s_i')\}_{i=1}^n$, normalization (μ, g), discount γ

1: **Policy estimation:** fit $\hat{u}(s,a) \approx \log \pi(a \mid s)$

2: Q**-evaluation:** estimate the fixed point

3: $\hat{Q}(s,a) \approx \hat{u}(s,a) - g(s) + \gamma(P\mu\hat{Q})(s,a)$

output $\hat{r}(s,a) = \hat{Q}(s,a) - (\mu\hat{Q})(s) + g(s)$

Algorithm 2 FITTED Q EVALUATION for $Q^{\mu}_{\hat{u}-g}$

input Data $\{(s_i, a_i, s_i')\}_{i=1}^n$, log-policy $\hat{u}$, normalization (μ, g)

input Discount γ, class $\mathcal{F}$, iterations K

1: Initialize $\hat{Q}^{(0)}(s,a) \leftarrow 0$

2: **for** $k = 1, \ldots, K$ **do**

3: For each i, set

$$y_i \leftarrow \hat{u}(s_i, a_i) - g(s_i) + \gamma \sum_{a'} \mu(a' \mid s_i') \times \hat{Q}^{(k-1)}(s_i', a')$$

4: Fit $\hat{Q}^{(k)} \in \mathcal{F}$ by regressing y_i on (s_i, a_i)

5: **end for**

output $\hat{Q} \leftarrow \hat{Q}^{(K)}$

be unstable; in practice, one may clip $\hat{\pi}$ away from zero [18]. Thus, in our setting, IRL reduces to behavior-policy estimation plus offline policy evaluation, rather than a fundamentally new estimation problem. The unit-scale Gumbel model fixes the softmax temperature. With shock scale $\tau > 0$, behavior identifies r/τ; a known τ rescales the recovered reward, whereas choices alone do not identify an unknown τ [13, Appendix C.2.3]. Changing τ rescales reward and value differences but preserves their signs and orderings, policy rankings, and the optimal policy; cardinal interpretation requires an external scale normalization.

The procedure is modular and black-box. The first stage can use any probabilistic classifier, behavior-cloning, imitation-learning, or IRL method trained on the observed data, including MaxEnt IRL and adversarial IRL. The second stage can use a range of existing offline RL methods, including temporal-difference learning [40], fitted Q-evaluation (FQE) [27, 46], and minimax or saddle-point Q-learning methods [41, 42, 49, 50]. This separation lets each stage leverage existing methods and remain compatible with techniques for handling misspecification or distribution shift [12, 2, 6, 9]. In particular, if an upstream IRL method returns a feasible reward–value pair under one normalization, GenPQR can use the induced policy as input and recover the corresponding reward under another normalization, without re-solving the original IRL objective.

Algorithm 2 instantiates the second stage of GenPQR using FQE, thereby generalizing DeepPQR [13]. It requires only standard tools for policy estimation and regression. Moreover, the regression step in Algorithm 2 need not be solved exactly at each iteration: one may instead parametrize Q by a neural network and take one or a few stochastic-gradient steps per iteration, as in deep fitted Q-learning or boosting [32, 39]. The expectation under μ can also be approximated by Monte Carlo sampling to avoid explicit integration.

5 Theoretical Guarantees

5.1 A generic and modular deterministic bound

We now turn to finite-sample guarantees for $\hat{r} = \hat{Q} - \mu\hat{Q} + g$ obtained via Algorithm 1. The following result shows how errors in the estimated log-policy $\hat{u}$ and anchored Q-function $\hat{Q}$ propagate to the recovered reward. We measure errors in the behavior norm $L^2(\nu_\pi)$, but the Bellman operator is most naturally analyzed in $L^2(d_\mu)$, where d_μ is any stationary state distribution under (μ, P) [31, 46], that is,

$$d_\mu = d_\mu P_\mu, \qquad P_\mu(s' \mid s) := \int_{\mathcal{A}} P(s' \mid s, a)\, \mu(a \mid s)\, da.$$

Recalling that $\nu_\pi := \rho \otimes \pi$, we write

$$\|f\|_{2,\mathrm{beh}} := \left(\mathbb{E}_{(s,a)\sim\nu_\pi}[f(s,a)^2]\right)^{1/2}, \qquad \|f\|_\infty := \operatorname*{ess\,sup}_{s,a} |f(s,a)|.$$

We impose a coverage assumption requiring the observed system (ρ, π, P) to provide sufficient support relative to the normalization policy μ under the same dynamics P: π must cover μ, the behavior state distribution ρ must cover the stationary distribution d_μ, and d_μ must cover the one-step state distribution $\nu_\pi P$.

Assumption 1 (Policy mismatch, stationary-state coverage, and one-step state coverage)**.** *Assume*

$$C_{\mathrm{cov}} := C_{\nu_\pi P/d_\mu}\, C_{d_\mu/\rho}\, C_{\mu/\pi} < \infty,$$

where

$$C_{\mu/\pi} := \sup_{s,a} \frac{\mu(a \mid s)}{\pi(a \mid s)}, \quad C_{d_\mu/\rho} := \left\| \frac{dd_\mu}{d\rho} \right\|_\infty, \quad C_{\nu_\pi P/d_\mu} := \left\| \frac{d(\nu_\pi P)}{dd_\mu} \right\|_\infty.$$

The coverage constant C_{cov} is analogous to standard coverage and concentrability coefficients in offline RL [51, 54]. It equals 1 under stationary sampling from (P, μ). Since the normalization policy μ is typically chosen with knowledge of the behavior policy π, this condition is often less restrictive here than in standard offline policy evaluation. For example, in the anchor-action case, one would not anchor on an action that is rarely or never observed.

Theorem 3 (Reward recovery bound)**.** *Under Assumption 1,*

$$\|r^\star - \hat{r}\|_{2,\mathrm{beh}} \le \sqrt{C_{\mu/\pi}} \left\{ \|Q^\mu_{\hat{u}-g} - \hat{Q}\|_{2,\mathrm{beh}} + \left(1 + \frac{\gamma\sqrt{C_{\mathrm{cov}}}}{1-\gamma}\right) \|u^\star - \hat{u}\|_{2,\mathrm{beh}} \right\}.$$

The reward-estimation error decomposes into a Q-estimation term, $\|Q^\mu_{\hat{u}-g} - \hat{Q}\|_{2,\mathrm{beh}}$, and a policy-estimation term, $\|\hat{u} - u^\star\|_{2,\mathrm{beh}}$. This makes the bound modular: any guarantees for Q-learning and policy estimation translate directly into a reward-recovery guarantee, with the latter scaled by $(1-\gamma)^{-1}$ and the coverage coefficients in Assumption 1. Under sample splitting [8], for example, existing results apply directly to $\hat{u}$ and $\hat{Q}$, including bounds for FQE [27, 46] and minimax Q-learning [41]. In the next section, we apply this template to derive finite-sample bounds for PQR with FQE.

5.2 Generalization conditions for GenPQR with FQE

We now specialize Theorem 3 to the case in which $\hat{Q}$ is obtained by running K steps of FQE over a regression class $\mathcal{F}$, initialized at $\hat{Q}^{(0)}$, as in Algorithm 2. To state the resulting finite-sample bound, we assume PAC-style generalization guarantees for the policy-estimation and regression steps.

Assumption 2 (Policy generalization)**.** *There exists $\underline{p} > 0$ such that, for all $\delta \in (0, 0.5)$, with probability at least $1-\delta$,*

$$\min\{\pi(a \mid s), \hat{\pi}(a \mid s)\} \ge \underline{p} \quad \textit{for all } (s,a), \qquad \{\mathbb{E}_{s\sim\rho}[\mathrm{KL}(\pi(\cdot \mid s) \,\|\, \hat{\pi}(\cdot \mid s))]\}^{1/2} \le \rho_\pi(n,\delta).$$

Lower bounds on $\hat{\pi}$ and π ensure that KL divergence controls $\|\hat{u} - u^\star\|_{2,\mathrm{beh}}$. For maximum likelihood estimation over a class $\mathcal{U}$, one typically has $\rho_\pi(n,\delta) \lesssim r_{\mathcal{U}}(n) + \sqrt{\log(1/\delta)/n}$, where $r_{\mathcal{U}}(n)$ is a local complexity measure of $\mathcal{U}$ (e.g., VC dimension) [44, 47].

We require each regression step to achieve small excess risk relative to the best approximation in $\mathcal{F}$ to the Bellman target $\mathcal{T}^\mu_{\hat{u}}(\hat{Q}^{(k-1)})$, where $\mathcal{T}^\mu_u f := u - g + \gamma P \mu f$. Define

$$\mathrm{reg}(Q \mid u, Q') := \|\mathcal{T}^\mu_u(Q') - Q\|^2_{2,\mathrm{beh}} - \inf_{f\in\mathcal{F}} \|\mathcal{T}^\mu_u(Q') - f\|^2_{2,\mathrm{beh}}.$$

Assumption 3 (One-step regression generalization)**.** *Assume $0 \in \mathcal{F}$ and $\sup_{f\in\mathcal{F}} \|f\|_\infty < \infty$. For every $k \in [K]$ and $\delta \in (0, 0.5)$, with probability at least $1-\delta$, $\{\mathrm{reg}(\hat{Q}^{(k)} \mid \hat{u}, \hat{Q}^{(k-1)})\}^{1/2} \le \rho_Q(n,\delta)$.*

Such bounds for nuisance-dependent regression targets follow from Foster and Syrgkanis [8], van der Laan [45]. Under a fresh-sample analysis [27], one typically has $\rho_Q(n,\delta) \lesssim r_{\mathcal{F}}(n/K) + \sqrt{K\log(1/\delta)/n}$, where $r_{\mathcal{F}}(n/K)$ is the local Rademacher critical radius of $\mathcal{F}$. Similar bounds hold for weakly dependent trajectory data under mixing conditions [52, 26].

5.3 Finite-sample bound for GenPQR with FQE

We now state our main result for GenPQR with FQE under approximate Bellman-completeness. To account for misspecification in the regression steps, we introduce the *inherent Bellman error* of the regression class [27]:

$$\varepsilon_{\mathcal{F}} := \sup_{f\in\mathcal{F}} \inf_{g\in\mathcal{F}} \|\mathcal{T}^{\mu}_{u^\star} f - g\|_\infty.$$

This quantity is zero when $\mathcal{F}$ is Bellman complete, that is, when $\mathcal{T}^{\mu}_{u^\star} f \in \mathcal{F}$ for every $f \in \mathcal{F}$ [6]. In that case, each regression step in Algorithm 2 is correctly specified, since its population target lies in $\mathcal{F}$. We define the standard concentrability coefficient

$$C_{\mathrm{conc}} := \sup_{m\geq 0} \left\| \frac{d\big(\nu_\pi (P\mu)^m\big)}{d\nu_\pi} \right\|_\infty < \infty,$$

which is finite under Assumption 1 with $C_{\mathrm{conc}} \leq C_{\mathrm{cov}}$ by Lemma 5 in Appendix F.2. Set

$$\kappa_{\mathrm{conc}} := 1 + \frac{\gamma\sqrt{C_{\mathrm{conc}}}}{1-\gamma}.$$

Theorem 4 (Finite-sample bound for GenPQR with FQE)**.** *Assume Assumptions 1, 2, and 3. Then with probability at least* $1-\delta$*,*

$$\|r^\star - \hat{r}\|_{2,\mathrm{beh}} \leq \sqrt{C_{\mu/\pi}} \Bigg\{ \sqrt{C_{\mathrm{conc}}}\,\gamma^K \|\hat{Q}^{(0)} - Q^{\mu}_{\hat{u}-g}\|_\infty + \kappa_{\mathrm{conc}} \Big(\varepsilon_{\mathcal{F}} + \rho_Q\big(n, \tfrac{\delta}{2K}\big)\Big) + \sqrt{\frac{2}{\underline{p}}} \left(1 + \kappa_{\mathrm{conc}} + \frac{\gamma\sqrt{C_{\mathrm{cov}}}}{1-\gamma}\right) \rho_\pi\big(n, \tfrac{\delta}{2}\big) \Bigg\}.$$

Proof sketch. Apply Theorem 3 and bound $\|Q^{\mu}_{\hat{u}-g} - \hat{Q}\|_{2,\mathrm{beh}}$ by the inexact Picard argument of Munos and Szepesvári [27]. The only change is that inherent Bellman error is measured for $\mathcal{T}^{\mu}_{u^\star}$, not the data-dependent $\mathcal{T}^{\mu}_{\hat{u}}$; Lemma 4 shows this suffices up to policy-estimation error. Approximate Bellman completeness could be relaxed using suitable weighting [46] or minimax formulations [43].

Discussion. The bound matches the usual FQE structure [27, 46], with additional terms due to estimation of $u^\star$. The three terms respectively capture finite-iteration error, approximation and statistical error from the fitted Bellman updates, and the IRL-specific error from estimating $u^\star$. The last term can be large when the behavior policy is nearly deterministic, for example in low-temperature softmax regimes, because KL divergence controls L^2 log-policy error with a factor of $\underline{p}^{-1/2}$. In particular, if $\hat{u}$ and $\hat{Q}$ are learned by ERM over parametric classes $\mathcal{U}$ and $\mathcal{F}$ with pseudo-dimensions $d_{\mathcal{U}}$ and $d_{\mathcal{F}}$, and $K \asymp \log n$, then one typically obtains, up to $O(\sqrt{\log\log n/n})$ terms,

$$\|\hat{r} - r^\star\|_{2,\mathrm{beh}} \lesssim \frac{\sqrt{C_{\mu/\pi} C_{\mathrm{cov}}}}{1-\gamma} \left(\varepsilon_{\mathcal{F}} + \sqrt{\frac{d_{\mathcal{F}}}{n}} + \frac{1}{\sqrt{\underline{p}}}\sqrt{\frac{d_{\mathcal{U}}}{n}}\right).$$

Comparison to DeepPQR. Theorem 2 of Geng et al. [13] gives a related FQE bound for anchor-action reward recovery under a specific neural-network architecture. Our result is more modular: it is not tied to a particular function class, allows approximate Bellman completeness for $\mathcal{T}^{\mu}_{u^\star}$, makes coverage explicit, and replaces sup-norm policy-error control with Kullback–Leibler or L^2-type control, which is natural for likelihood-based policy estimators, multiclass classification, MaxEnt IRL [55], and adversarial IRL [11, 37, 22, 10].

6 Experimental investigation

We adapt DeepPQR's synthetic study to isolate identification rather than imitation performance, using its infinite-horizon environment with continuous states, deterministic transitions, and a zero-reward anchor action. We first compare DeepPQR and GenPQR under matched policy estimation, then vary the action count, and finally study a complementary-source normalization unavailable to anchor-only

Setting	Method	Anchor ct. (frac.)	MSE ↓	Corr. ↑	Time ↓
200 / rare	DeepPQR	311 (0.16)	2.60 ± 0.15	0.46 ± 0.03	7.33 ± 0.10
200 / rare	GenPQR	311 (0.16)	$\mathbf{0.91 \pm 0.11}$	$\mathbf{0.66 \pm 0.03}$	$\mathbf{5.22 \pm 0.08}$
1000 / rare	DeepPQR	1559 (0.16)	1.48 ± 0.10	0.61 ± 0.03	38.84 ± 0.60
1000 / rare	GenPQR	1559 (0.16)	$\mathbf{0.76 \pm 0.08}$	$\mathbf{0.73 \pm 0.02}$	$\mathbf{27.76 \pm 0.42}$
2500 / common	DeepPQR	6280 (0.25)	0.70 ± 0.07	0.73 ± 0.02	92.82 ± 1.28
2500 / common	GenPQR	6280 (0.25)	$\mathbf{0.60 \pm 0.06}$	$\mathbf{0.73 \pm 0.02}$	$\mathbf{64.09 \pm 0.90}$

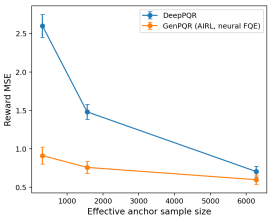


Figure 1: Matched DeepPQR vs. GenPQR comparison. Both use the same AIRL policy estimate and neural FQE. Entries are mean ± 95% CI over 100 seeds. Anchor count is the number of anchor-action transitions used by DeepPQR's anchor-Q step.

methods. We evaluate the noise-free reward matrix on independent test states and report means with 95% confidence intervals over 100 seeds; complete design and estimator details are in Appendix C.

Matched comparison to DeepPQR. Both methods use the same AIRL policy estimate and neural downstream approximation, so differences reflect the identification step (Section 3.3). We vary trajectory count and anchor frequency, which determine DeepPQR's effective anchor sample size. Figure 1 summarizes three regimes; Appendix D reports the full estimator comparison.

Scaling the action set. Holding the behavior sample fixed at 2500 trajectories, we vary $|\mathcal{A}| \in \{5, 10, 20, 40\}$. As the action set grows, the realized anchor count falls from about 6400 to 800. Figure 3 in Appendix D shows the resulting separation: DeepPQR's reward MSE rises above 4 at 40 actions, whereas both GenPQR variants remain below 0.5 because their FQE stage pools the full behavior sample across actions.

Complementary-source data fusion. We next use 20 actions and two independent samples from a shared MDP. The behavior source contains 2500 trajectories of (S, A, S'), but no rewards. An auxiliary source contains $n_g \in \{100, 250, 500, 1000, 2500, 5000\}$ trajectories of (S, Y), but with actions withheld, under the same reference and behavior policy $\mu = \pi$, with $\mathbb{E}[Y \mid S = s, A = a] = r(s, a)$. It identifies $g(s) = \mathbb{E}[Y \mid S = s] = \sum_a \pi(a \mid s) r(s, a)$, while the behavior source identifies π. Neither source identifies the data-driven normalization pair (π, g) alone, but together they do. We estimate π by behavior cloning and g by regression, then apply GenPQR. The true reward also obeys $r(s, 0) = 0$, so DeepPQR and anchor GenPQR remain identified from the behavior source alone; however, they cannot exploit the auxiliary outcomes.

At 100 auxiliary trajectories, fusion and anchor GenPQR are statistically indistinguishable (paired MSE difference 0.008 ± 0.026). From 250 trajectories onward, fusion is more accurate. At 5000, its behavior-weighted reward MSE is 0.210 ± 0.026, versus 0.443 ± 0.064 for anchor GenPQR and 2.253 ± 0.089 for DeepPQR: a 52.6% reduction relative to anchor GenPQR, with improvement in all 100 paired seeds.

Conclusion. GenPQR reduces normalized reward recovery to policy estimation and Bellman evaluation. The target is determined by the behavioral model and normalization; estimation requires adequate coverage and accurate policy and Q-function fits.

Acknowledgments and Disclosure of Funding

The authors have no funding to disclose. The authors declare no competing interests.

References

[1] Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *ICML*, 2004.

[2] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems*, 34:29304–29320, 2021.

[3] Victor Aguirregabiria and Pedro Mira. Dynamic discrete choice structural models: A survey. *Journal of Econometrics*, 156(1):38–67, 2010.

[4] Patrick Bajari, Han Hong, and Stephen P Ryan. Identification and estimation of a discrete game of complete information. *Econometrica*, 78(5):1529–1568, 2010.

[5] Haoyang Cao, Samuel Cohen, and Lukasz Szpruch. Identifiability in inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12362–12373, 2021.

[6] Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International conference on machine learning*, pages 1042–1051. PMLR, 2019.

[7] Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6, 2005.

[8] Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3):879–908, 2023.

[9] Dylan J Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. *arXiv preprint arXiv:2111.10919*, 2021.

[10] Dylan J Foster, Adam Block, and Dipendra Misra. Is behavior cloning all you need? understanding horizon in imitation learning. *Advances in Neural Information Processing Systems*, 37: 120602–120666, 2024.

[11] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. In *ICLR*, 2018.

[12] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pages 2052–2062. PMLR, 2019.

[13] Sinong Geng, Houssam Nassif, Carlos Manzanares, Max Reppen, and Ronnie Sircar. Deep pqr: Solving inverse reinforcement learning using anchor actions. In *International Conference on Machine Learning*, pages 3431–3441. PMLR, 2020.

[14] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International conference on machine learning*, pages 1352–1361. PMLR, 2017.

[15] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *ICML*, 2018.

[16] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *NeurIPS*, 2016.

[17] V. Joseph Hotz and Robert A. Miller. Conditional choice probabilities and the estimation of dynamic models. *The Review of Economic Studies*, 60(3):497–529, 1993.

[18] Edward L Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

[19] Nathan Kallus and Madeleine Udell. Revealed preference at scale: Learning personalized preferences from assortment choices. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 821–837, 2016.

[20] Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. *Advances in neural information processing systems*, 31, 2018.

[21] Hilbert J. Kappen. Linear theory for control of nonlinear stochastic systems. *Physical Review Letters*, 95(20):200201, 2005.

[22] Liyiming Ke, Sanjiban Choudhury, Matt Barnes, Wen Sun, Gilwoo Lee, and Siddhartha Srinivasa. Imitation learning as f-divergence minimization. In *International workshop on the algorithmic foundations of robotics*, pages 313–329. Springer, 2020.

[23] Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

[24] Sergey Levine, Zoran Popović, and Vladlen Koltun. Nonlinear inverse reinforcement learning with gaussian processes. In *NeurIPS*, 2011.

[25] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[26] Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary φ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11(2), 2010.

[27] Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

[28] Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Bridging the gap between value and policy based reinforcement learning. In *NeurIPS*, 2017.

[29] Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *ICML*, 2000.

[30] Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pages 278–287. Citeseer, 1999.

[31] Andrew Patterson, Adam White, and Martha White. A generalized projected bellman error for off-policy value estimation in reinforcement learning. *Journal of Machine Learning Research*, 23(145):1–61, 2022.

[32] Martin Riedmiller. Neural fitted q iteration–first experiences with a data efficient neural reinforcement learning method. In *European conference on machine learning*, pages 317–328. Springer, 2005.

[33] John Rust. Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica: Journal of the Econometric Society*, pages 999–1033, 1987.

[34] Joar Skalse and Alessandro Abate. Misspecification in inverse reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15136–15143, 2023. doi: 10.1609/aaai.v37i12.26766.

[35] Joar Skalse and Alessandro Abate. Partial identifiability and misspecification in inverse reinforcement learning. *Artificial Intelligence*, 356:104525, 2026. doi: 10.1016/j.artint.2026.104525.

[36] Joar Max Viktor Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In *International Conference on Machine Learning*, pages 32033–32058. PMLR, 2023.

[37] Aaron J Snoswell, Surya PN Singh, and Nan Ye. Revisiting maximum entropy inverse reinforcement learning: New perspectives and algorithms. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 241–249. IEEE, 2020.

[38] Emanuel Todorov. Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences*, 106(28):11478–11483, 2009.

[39] Samuele Tosatto, Matteo Pirotta, Carlo d'Eramo, and Marcello Restelli. Boosted fitted q-iteration. In *International Conference on Machine Learning*, pages 3434–3443. PMLR, 2017.

[40] John Tsitsiklis and Benjamin Van Roy. Analysis of temporal-difference learning with function approximation. *Advances in neural information processing systems*, 9, 1996.

[41] Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pages 9659–9668. PMLR, 2020.

[42] Masatoshi Uehara, Masaaki Imaizumi, Nan Jiang, Nathan Kallus, Wen Sun, and Tengyang Xie. Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981*, 2021.

[43] Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. Offline minimax soft-q-learning under realizability and partial coverage. *Advances in Neural Information Processing Systems*, 36:12797–12809, 2023.

[44] Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

[45] Lars van der Laan. A researcher's guide to empirical risk minimization. *arXiv preprint arXiv:2602.21501*, 2026.

[46] Lars van der Laan and Nathan Kallus. Fitted q evaluation without bellman completeness via stationary weighting. *arXiv preprint arXiv:2512.23805*, 2025.

[47] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

[48] Markus Wulfmeier, Peter Ondruska, and Ingmar Posner. Maximum entropy deep inverse reinforcement learning. In *AAAI*, 2016.

[49] Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, pages 550–559. PMLR, 2020.

[50] Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pages 11404–11413. PMLR, 2021.

[51] Tengyang Xie, Dylan J Foster, Yu Bai, Nan Jiang, and Sham M Kakade. The role of coverage in online reinforcement learning. *arXiv preprint arXiv:2210.04157*, 2022.

[52] Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.

[53] Siliang Zeng, Chenliang Li, Alfredo Garcia, and Mingyi Hong. Maximum-likelihood inverse reinforcement learning with finite-time guarantees. *Advances in Neural Information Processing Systems*, 35:10122–10135, 2022.

[54] Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pages 2730–2775. PMLR, 2022.

[55] Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, 2008.

[56] Brian D Ziebart, J Andrew Bagnell, and Anind K Dey. Modeling interaction via the principle of maximum causal entropy. In *ICML*, 2010.

Lay summary

People's choices can reveal what they value, but behavior alone does not uniquely determine the underlying reward. We show how to recover a well-defined reward after specifying a transparent normalization. Our method, GenPQR, reduces the problem to two familiar statistical tasks: classifying which action an agent takes and using regression to evaluate future outcomes. This makes the approach compatible with standard machine-learning tools and auxiliary outcome data. We establish when the normalized reward is identified and bound the estimation error from each stage. In simulations, GenPQR is more accurate than an anchor-action method when the number of possible actions grows and when complementary outcome data are available. The recovered reward depends on the behavioral model and normalization and should be interpreted accordingly.

Method	MSE ↓	Corr. ↑	NLL ↓	Time ↓
GenPQR (BC, NN)	**0.37 ± 0.06**	**0.81 ± 0.01**	**1.546 ± 0.004**	4.74 ± 0.06
GenPQR (AIRL, NN)	0.63 ± 0.06	0.72 ± 0.03	1.608 ± 0.006	26.43 ± 0.34
GenPQR (BC, GBT)	1.07 ± 0.19	0.49 ± 0.03	**1.546 ± 0.004**	**2.25 ± 0.03**
GenPQR (AIRL, GBT)	1.07 ± 0.19	0.41 ± 0.05	1.608 ± 0.006	23.97 ± 0.32
DeepPQR	0.79 ± 0.07	0.69 ± 0.02	1.608 ± 0.006	38.24 ± 0.42
MaxEnt-IRL	1.17 ± 0.18	0.35 ± 0.03	1.624 ± 0.005	8.57 ± 0.11
AIRL state reward	2.17 ± 0.21	-0.01 ± 0.03	1.608 ± 0.006	23.83 ± 0.32
Log-policy pseudo	3.40 ± 0.18	0.37 ± 0.05	1.608 ± 0.006	23.83 ± 0.32

Figure 2: Higher-sample method comparison. Entries are mean ± 95% confidence interval over 100 seeds. NN = neural FQE; GBT = boosted FQE.

A Experiment figure: high sample comparison

B Abstract linear normalizations.

The same argument used to derive Theorem 1 extends to general linear normalizations. Let A be a bounded linear operator on reward functions, and impose $Ar = g$. Every relaxed optimum inducing the observed policy has the form

$$r_c = u^\star + c - \gamma Pc, \qquad v_c = Pc,$$

for some state function c. Thus the normalization $Ar_c = g$ is equivalent to

$$A(c - \gamma Pc) = g - Au^\star.$$

If $g - Au^\star$ lies in the range of the map $c \mapsto A(c - \gamma Pc)$, and this map is injective on state functions, then the normalization selects a unique representative of the reward class:

$$c^\star = \{A(I - \gamma P)\}^{-1}(g - Au^\star), \qquad r^\star = u^\star + c^\star - \gamma Pc^\star, \qquad v^\star = Pc^\star.$$

The statewise normalization above is the special case $A = \mu$. In this case, the map becomes $c \mapsto c - \gamma P_\mu c = (I - \gamma P_\mu)c$, so the abstract equation reduces to the Bellman resolvent equation in Theorem 1.

C Simulation Details

This appendix summarizes the simulation design and implementation choices used in Section 6. The experiment code, pinned environment, exact commands, tests, and plotting scripts are available in the public reproduction repository.

Environment. We adapt the synthetic environment of Geng et al. [13]. States are continuous with dimension $p = 5$, the action set has $|\mathcal{A}| = 5$ actions, and the discount factor is $\gamma = 0.9$. We use an infinite-horizon stationary data-generating process and generate finite trajectories of length 10 for offline training and evaluation. The normalization is the anchor constraint $r(s, a^\dagger) = 0$, implemented as anchor-action normalization with $a^\dagger = 0$ and $g(s) = 0$. Train and test sets within each seed are generated from the same underlying MDP parameters, with independent rollouts.

Behavior policy and transitions. The logged policy is generated from a soft Q-planner with $\alpha = 1$, so that the DeepPQR log-policy identity is correctly matched. The planner induces a heterogeneous but non-degenerate behavior policy, and we vary anchor support by shifting the anchor-action logit. Transitions are deterministic conditional on state and action up to boundary handling: actions induce action-specific state shifts, and trajectories that leave the bounded state region are reset uniformly within that region. We clip estimated action probabilities to $[0.01, 0.99]$ and renormalize before using them in either GenPQR or DeepPQR.

Experiment 1. The matched DeepPQR-vs.-GenPQR comparison uses shared AIRL policy estimation and neural downstream approximation. We consider three settings: `(200, -1.0)`, `(1000, -1.0)`, and `(2500, 0.0)`, where the first entry is the number of training trajectories and the second is the anchor-logit shift. These correspond to low-sample rare-anchor, mid-sample rare-anchor, and high-sample common-anchor regimes. Each setting uses 300 test trajectories and 100 random seeds.

Experiment 2. The higher-sample comparison fixes 1000 training trajectories, 300 test trajectories, and anchor-logit shift 0.0, again over 100 seeds. We compare DeepPQR, GenPQR with two policy estimators and two Q-estimators, and several reward-recovery baselines that are standard in the IRL literature.

Experiment 3: action scaling. We fix 2500 behavior trajectories and 300 test trajectories and vary $|\mathcal{A}| \in \{5, 10, 20, 40\}$, using 100 seeds at each action count. The anchor-logit shift remains 0. We compare DeepPQR and GenPQR using both behavior cloning and AIRL policy estimates, with neural downstream estimators.

Experiment 4: complementary-source data fusion. We fix 20 actions, 2500 behavior trajectories, and 300 independent test trajectories. Within each seed, an independent auxiliary sample contains $n_g \in \{100, 250, 500, 1000, 2500, 5000\}$ trajectories of length 10. Only (S, Y) is exposed from this source, while only (S, A, S') is exposed from the behavior source. The reference and behavior policies coincide, so $g(s) = \mathbb{E}[Y \mid S = s] = \sum_a \pi(a \mid s) r(s, a)$. We fit one shared behavior-cloning policy per seed, estimate g by gradient-boosted regression, and use neural FQE for all GenPQR variants. Auxiliary samples are nested by complete trajectory, and all results use 100 fresh confirmatory seeds disjoint from the development runs. The oracle-nuisance variant uses the true (π, g) but still estimates FQE; it is reported only as a diagnostic upper bound.

Policy estimators. AIRL uses a standard action-independent reward network and potential network, both implemented as two-layer ReLU MLPs with hidden widths $(64, 64)$, Adam step size 10^{-3}, 60 behavior-cloning warm-start epochs, and 80 adversarial updates in the paper experiments. Behavior cloning uses the same $(64, 64)$ MLP architecture and 40 epochs in Experiment 2. MaxEnt-IRL is implemented as a neural softmax-Q model with hidden widths $(128, 128)$ and 150 gradient steps. These architectures were chosen to be standard for low-dimensional synthetic control problems and to remain stable across seeds; larger networks did not materially improve performance in pilot runs.

Q-evaluation. Neural FQE uses a dueling-style MLP $Q(s, a) = V(s) + A(s, a) - \bar{A}(s)$ with hidden widths $(128, 128)$, Adam step size 5×10^{-3}, 8 Bellman iterations, and 4 epochs of regression per iteration. Boosted FQE uses LightGBM directly, with 4 outer Bellman iterations and 30 boosting rounds per iteration, learning rate 0.05, 32 leaves, and minimum leaf size 20. We selected these settings to balance Bellman-fit accuracy, runtime, and seed-to-seed stability; the boosted configuration follows the repository's adapted FQE implementation and avoids long inner re-fitting loops.

DeepPQR and baselines. DeepPQR uses the same shared AIRL policy estimate as GenPQR in the matched comparison. It then estimates the anchor Q-function on the anchor-action subset, reconstructs the full Q-function from log-policy ratios, and fits the final reward-regression network. The AIRL state-reward baseline uses the AIRL reward head directly; the log-policy pseudo-reward baseline uses $\log \hat{\pi}(a \mid s) - g(s)$; MaxEnt-IRL uses its learned Q-surrogate as reward; and the linear reward baseline fits a strictly linear state-action reward model with action-specific coefficients and ridge regularization.

Reporting. For each method we report reward mean squared error, reward correlation with the ground-truth normalized reward, held-out policy negative log-likelihood when applicable, and wall-clock runtime. Confidence intervals are normal approximation intervals based on the 100 seed-level estimates. The data-fusion study reports both behavior-policy-weighted and uniform-over-action MSE on the full noise-free test reward matrix; no oracle quantity is used to fit or select an estimated method.

D Additional Experimental Results

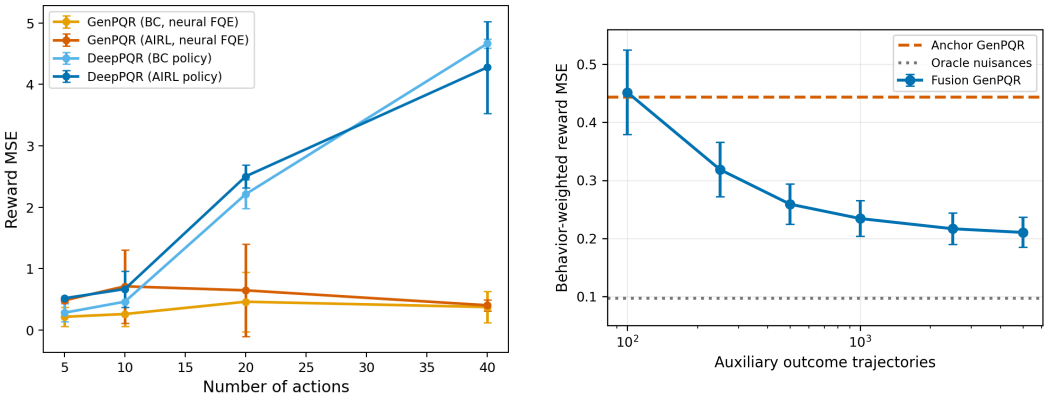


Figure 3: Large-action experiments. Left: reward MSE as the number of actions grows at fixed behavior-sample size. Right: 20-action GenPQR using an independent state–outcome source. Error bars are 95% confidence intervals over 100 seeds.

Method comparison. Figure 4 corresponds to the second experiment in the main text.

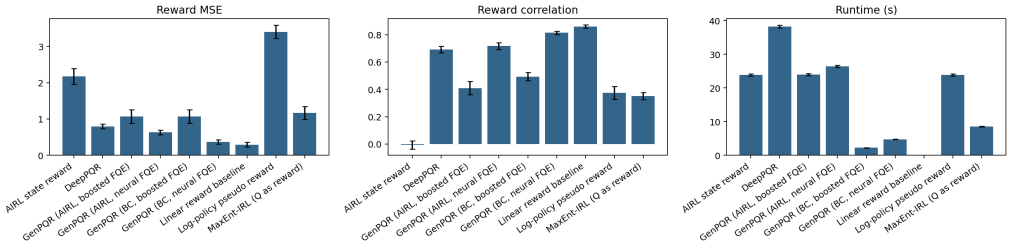


Figure 4: Higher-sample method comparison. Neural GenPQR variants outperform non-identifying baselines, and behavior cloning with neural FQE provides the best accuracy-runtime tradeoff in this setting. Error bars denote 95% confidence intervals over 100 seeds.

Action-support diagnostic. Figure 5 reports the anchor counts underlying the left panel of Figure 3.

Complementary-source results. Table 1 gives the exact results underlying Figure 3 (right). The projected DeepPQR row enforces the known zero anchor on its final predictions; this sensitivity analysis improves its behavior-weighted MSE by only 3.5%, so anchor violation does not explain the gap.

Method	n_g	Behavior MSE	Uniform MSE	Uniform corr.	g RMSE
DeepPQR	–	2.253 ± 0.089	2.247 ± 0.089	0.560 ± 0.025	–
DeepPQR, anchor projected	–	2.174 ± 0.086	2.187 ± 0.087	0.566 ± 0.025	–
Anchor GenPQR	–	0.443 ± 0.064	0.454 ± 0.064	0.790 ± 0.011	–
Fusion GenPQR	100	0.452 ± 0.073	0.471 ± 0.074	0.795 ± 0.009	0.388 ± 0.035
Fusion GenPQR	250	0.319 ± 0.047	0.337 ± 0.048	0.841 ± 0.006	0.291 ± 0.026
Fusion GenPQR	500	0.259 ± 0.035	0.277 ± 0.036	0.864 ± 0.006	0.230 ± 0.021
Fusion GenPQR	1000	0.234 ± 0.031	0.252 ± 0.032	0.876 ± 0.006	0.181 ± 0.017
Fusion GenPQR	2500	0.217 ± 0.027	0.233 ± 0.028	0.884 ± 0.007	0.137 ± 0.013
Fusion GenPQR	5000	0.210 ± 0.026	0.227 ± 0.027	0.888 ± 0.007	0.114 ± 0.011
GenPQR, oracle nuisances	–	0.097 ± 0.017	0.105 ± 0.018	0.963 ± 0.003	0

Table 1: Twenty-action complementary-source experiment. Entries are mean $\pm$ 95% confidence interval over 100 seeds. The oracle-nuisance row is a diagnostic, not a deployable estimator.

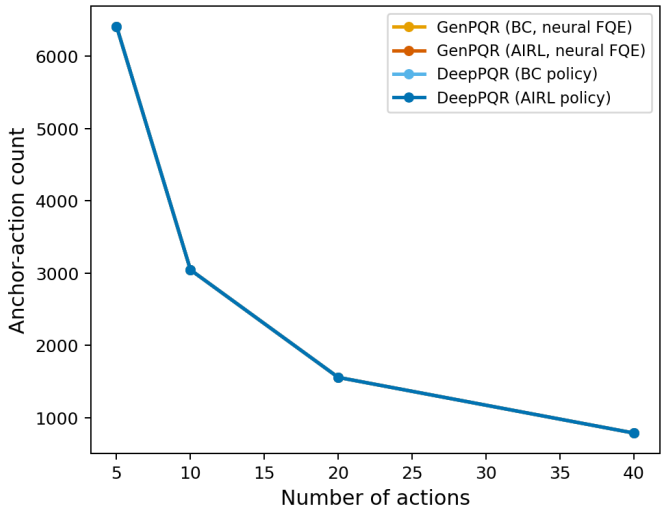


Figure 5: Realized anchor-action count as the action set grows with the total behavior sample held fixed.

E Identification and Representation

E.1 Proof of Lemma 1

Proof. Let $u^\star(s,a) := \log \pi(a \mid s)$. Since

$$\sum_{a'} \exp\{u^\star(s,a')\} = \sum_{a'} \pi(a' \mid s) = 1,$$

we have $\mathrm{LSE}(u^\star)(s) = 0$ for ρ-almost every s. Since $\nu_\pi P \ll \rho$, $P\,\mathrm{LSE}(u^\star) = 0$ ν_π-almost everywhere, so $(u^\star, 0)$ is feasible for (3).

To prove optimality, define $q(s,a) := r(s,a) + \gamma v(s,a)$. For any feasible (r,v), the objective in (3) can be written as

$$\mathbb{E}_{(s,a)\sim\nu_\pi}\big[q(s,a) - \mathrm{LSE}\, q(s)\big].$$

For each state s, the quantity $q(s,a) - \mathrm{LSE}\, q(s)$ is the log-probability of action a under the softmax policy induced by q,

$$\pi_q(a \mid s) := \exp\{q(s,a) - \mathrm{LSE}\, q(s)\}.$$

Therefore

$$\begin{aligned}\mathbb{E}_{(s,a)\sim\nu_\pi}\big[q(s,a) - \mathrm{LSE}\, q(s)\big] &= \mathbb{E}_{s\sim\rho}\left[\sum_a \pi(a \mid s) \log \pi_q(a \mid s)\right] \\ &= -\mathbb{E}_{s\sim\rho}\Big[\mathrm{KL}\big(\pi(\cdot \mid s) \,\|\, \pi_q(\cdot \mid s)\big)\Big] - \mathbb{E}_{s\sim\rho}\big[H(\pi(\cdot \mid s))\big].\end{aligned}$$

This is maximized when $\pi_q(\cdot \mid s) = \pi(\cdot \mid s)$ for ρ-almost every s, which is attained by $q = u^\star$, i.e., by $(r,v) = (u^\star, 0)$. □

E.2 Proof of Lemma 2

Proof. Let $q := r + \gamma v$, and define

$$\tilde{r} = r + c - \gamma P c, \qquad \tilde{v} = v + Pc.$$

Then

$$\tilde{r} + \gamma\tilde{v} = r + \gamma v + c = q + c,$$

where $c(s)$ is added to every action at state s. Consequently,

$$\mathrm{LSE}(\tilde{r} + \gamma\tilde{v})(s) = \mathrm{LSE}(q + c)(s) = c(s) + \mathrm{LSE}\, q(s).$$

Applying P to both sides gives

$$P\,\mathrm{LSE}(\tilde{r}+\gamma\tilde{v}) = Pc + P\,\mathrm{LSE}\,q = Pc + v = \tilde{v},$$

so $(\tilde{r},\tilde{v})$ is feasible for (3).

The objective value is unchanged because

$$\tilde{r}(s,a)+\gamma\tilde{v}(s,a) - \mathrm{LSE}(\tilde{r}+\gamma\tilde{v})(s) = q(s,a)+c(s) - \big(\mathrm{LSE}\,q(s)+c(s)\big) = q(s,a) - \mathrm{LSE}\,q(s).$$

Thus $(\tilde{r},\tilde{v})$ attains the same objective value as (r,v). The induced log-policy,

$$\tilde{r}+\gamma\tilde{v} - \mathrm{LSE}(\tilde{r}+\gamma\tilde{v}) = q - \mathrm{LSE}\,q,$$

is therefore also unchanged. □

E.3 Proof of Theorem 1

Proof. Let $Q^{\mu}_{u^\star - g}$ denote the unique bounded fixed point of

$$Q = u^\star - g + \gamma P\mu Q.$$

Define

$$r^\star := Q^{\mu}_{u^\star-g} - \mu Q^{\mu}_{u^\star-g} + g, \qquad v^\star := -P\mu Q^{\mu}_{u^\star-g}.$$

We first verify feasibility for (2). The normalization is immediate:

$$\mu r^\star = \mu Q^{\mu}_{u^\star-g} - \mu Q^{\mu}_{u^\star-g} + g = g.$$

Next, the Bellman equation for $Q^{\mu}_{u^\star-g}$ implies

$$\gamma v^\star = -\gamma P\mu Q^{\mu}_{u^\star-g} = u^\star - g - Q^{\mu}_{u^\star-g}.$$

Hence

$$r^\star + \gamma v^\star = Q^{\mu}_{u^\star-g} - \mu Q^{\mu}_{u^\star-g} + g + u^\star - g - Q^{\mu}_{u^\star-g} = u^\star - \mu Q^{\mu}_{u^\star-g}.$$

Since $\mu Q^{\mu}_{u^\star-g}$ is state-only and $\mathrm{LSE}(u^\star)=0$,

$$\mathrm{LSE}(r^\star+\gamma v^\star) = \mathrm{LSE}\big(u^\star - \mu Q^{\mu}_{u^\star-g}\big) = -\mu Q^{\mu}_{u^\star-g}.$$

Applying P gives

$$P\,\mathrm{LSE}(r^\star+\gamma v^\star) = -P\mu Q^{\mu}_{u^\star-g} = v^\star.$$

Thus $(r^\star, v^\star)$ is feasible.

To prove optimality, observe that

$$r^\star + \gamma v^\star = u^\star + c^\star, \qquad c^\star(s) := -(\mu Q^{\mu}_{u^\star-g})(s).$$

By the same log-likelihood argument as in Lemma 1, adding a state-only shift $c^\star$ does not change the induced policy, so $(r^\star, v^\star)$ attains the same objective value as $(u^\star, 0)$. Hence it is optimal.

It remains to show uniqueness. Let (r,v) be any optimal feasible pair for (2), and set $q := r+\gamma v$. Since (r,v) is also feasible for the relaxed problem and achieves the same maximal likelihood as $(u^\star,0)$, the induced policy equals π for ρ-almost every state. Therefore, for ρ-almost every s and every action a,

$$u^\star(s,a) = q(s,a) - \mathrm{LSE}\,q(s),$$

which implies

$$q(s,a) = u^\star(s,a) + c(s), \qquad c(s) := \mathrm{LSE}\,q(s).$$

Feasibility then gives

$$v = P\,\mathrm{LSE}\,q = Pc, \qquad r = q - \gamma v = u^\star + c - \gamma Pc.$$

The normalization condition $\mu r = g$ becomes

$$g = \mu u^\star + c - \gamma P_\mu c, \qquad P_\mu c(s) := \int Pc(s,a)\,\mu(da \mid s).$$

Equivalently,

$$(I - \gamma P_\mu)c = g - \mu u^\star.$$

Since $\pi(a \mid s) > 0$ for every action and $\nu_\pi P \ll \rho$, $\rho P_\mu \ll \rho$. Hence P_μ is a Markov contraction on $L^\infty(\rho)$. Since $\gamma < 1$, the operator $I - \gamma P_\mu$ is invertible on $L^\infty(\rho)$ by the Neumann series. Thus c is unique in $L^\infty(\rho)$, and (r,v) is unique in $L^\infty(\nu_\pi)$. □

E.4 An Equivalent Identification

The next result rewrites the identification in terms of the continuation value $v^\star$. It is equivalent to Theorem 1, but it enables direct modeling of $v^\star$, from which the reward $r^\star$ is obtained immediately.

Theorem 5 (An equivalent identification with anchor function)**.** *Equation* (2) *admits a unique optimal solution in* $L^\infty(\nu_\pi)$*, where*

$$r^\star = u^\star + \mu(g + \gamma v^\star - u^\star) - \gamma v^\star,$$

and $v^\star$ *is the unique bounded solution to the fixed-point equation*

$$v^\star = P\mu(g + \gamma v^\star - u^\star)\,.$$

Proof. Let $Q^\mu_{u^\star - g}$ denote the unique bounded solution to

$$Q^\mu_{u^\star - g} = u^\star - g + \gamma P \mu Q^\mu_{u^\star - g}.$$

By Theorem 1,

$$r^\star = Q^\mu_{u^\star - g} - \mu Q^\mu_{u^\star - g} + g, \qquad v^\star = -P\mu Q^\mu_{u^\star - g}.$$

The Bellman equation therefore gives

$$Q^\mu_{u^\star - g} = u^\star - g - \gamma v^\star.$$

Using this identity in $v^\star = -P\mu Q^\mu_{u^\star - g}$ gives

$$v^\star = -P\mu(u^\star - g - \gamma v^\star) = P\mu(g + \gamma v^\star - u^\star),$$

which is the claimed fixed-point equation.

For the reward,

$$\begin{aligned} r^\star &= Q^\mu_{u^\star - g} - \mu Q^\mu_{u^\star - g} + g \\ &= (u^\star - g - \gamma v^\star) - \mu(u^\star - g - \gamma v^\star) + g \\ &= u^\star + \mu(g + \gamma v^\star - u^\star) - \gamma v^\star. \end{aligned}$$

Uniqueness follows directly from Theorem 1. □

E.5 Proof of Theorem 2

Proof. Let (r, v) solve (3), and define

$$q := r + \gamma v, \qquad c(s) := \mathrm{LSE}\, q(s).$$

Since (r, v) is optimal for the relaxed problem, it induces the behavior policy π for ρ-almost every state. Therefore

$$u^\star(s, a) = q(s, a) - \mathrm{LSE}\, q(s) = q(s, a) - c(s),$$

so

$$q(s, a) = u^\star(s, a) + c(s).$$

The Bellman feasibility condition then gives

$$v = P\,\mathrm{LSE}\, q = Pc, \qquad r = q - \gamma v = u^\star + c - \gamma Pc.$$

Fix any policy π_1, and let $Q^{\pi_1}_{u^\star}$ denote the unique bounded solution to

$$Q^{\pi_1}_{u^\star} = u^\star + \gamma P \pi_1 Q^{\pi_1}_{u^\star}.$$

Because c is state-only,

$$P\pi_1 c = Pc.$$

Hence

$$\begin{aligned} r + \gamma P \pi_1 \big(Q^{\pi_1}_{u^\star} + c\big) &= u^\star + c - \gamma Pc + \gamma P \pi_1 Q^{\pi_1}_{u^\star} + \gamma P \pi_1 c \\ &= u^\star + \gamma P \pi_1 Q^{\pi_1}_{u^\star} + c \\ &= Q^{\pi_1}_{u^\star} + c. \end{aligned}$$

So $Q^{\pi_1}_{u^\star} + c$ solves the Bellman equation for reward r under policy π_1. By uniqueness of bounded Bellman fixed points,

$$Q^{\pi_1}_r = Q^{\pi_1}_{u^\star} + c,$$

which is equivalent to

$$Q^{\pi_1}_{u^\star} = Q^{\pi_1}_r - c.$$

Now let $c^\star$ denote the state shift corresponding to the normalized solution $r^\star$. By Theorem 1, $r^\star$ is also a shaped version of $u^\star$, so the same argument gives

$$Q^{\pi_i}_{r^\star} = Q^{\pi_i}_{u^\star} + c^\star, \qquad i \in \{1, 2\}.$$

Taking policy values and using that c and $c^\star$ are state-only,

$$V^{\pi_i}_r = V^{\pi_i}_{u^\star} + c, \qquad V^{\pi_i}_{r^\star} = V^{\pi_i}_{u^\star} + c^\star.$$

Subtracting the identities for $i = 1$ and $i = 2$ proves, for ρ-almost every s,

$$V^{\pi_1}_{r^\star}(s) - V^{\pi_2}_{r^\star}(s) = V^{\pi_1}_r(s) - V^{\pi_2}_r(s). \qquad \square$$

F Reward Recovery and Finite-Sample Analysis

F.1 Proof of Theorem 3

Proof. For any measure d on $\mathcal{S}$, let $\|h\|_{2,d} := \{\mathbb{E}_{s\sim d}[h(s)^2]\}^{1/2}$. Since $r^\star(s,a) = Q^\mu_{u^\star - g}(s,a) - (\mu Q^\mu_{u^\star - g})(s) + g(s)$, write $\widetilde{\Delta}_Q := Q^\mu_{u^\star - g} - \hat{Q}$. Then $r^\star - \hat{r} = \widetilde{\Delta}_Q - \mu\widetilde{\Delta}_Q$. For a state–action function f, define

$$w_s(a) := \frac{\mu(a \mid s)}{\pi(a \mid s)}, \qquad \bar{f}_\pi(s) := \sum_a \pi(a \mid s) f(s,a), \qquad h(s,a) := f(s,a) - \bar{f}_\pi(s).$$

Using the $L^2\{\pi(\cdot \mid s)\}$ inner product, h is orthogonal to constants and

$$f - \mu f = h - \langle w_s, h\rangle_\pi.$$

Moreover, $\langle w_s, 1\rangle_\pi = 1$, so $1 + \|w_s - 1\|^2_{2,\pi(\cdot|s)} = \|w_s\|^2_{2,\pi(\cdot|s)}$. Therefore

$$\begin{aligned}\|f - \mu f\|^2_{2,\pi(\cdot|s)} &= \|h\|^2_{2,\pi(\cdot|s)} + \langle w_s - 1, h\rangle^2_\pi \\ &\le \|w_s\|^2_{2,\pi(\cdot|s)} \|h\|^2_{2,\pi(\cdot|s)} \\ &\le C_{\mu/\pi} \|f\|^2_{2,\pi(\cdot|s)},\end{aligned}$$

where $\|w_s\|^2_{2,\pi(\cdot|s)} = \sum_a \mu(a \mid s)^2/\pi(a \mid s) \le C_{\mu/\pi}$. Integrating over $s \sim \rho$ and taking $f = \widetilde{\Delta}_Q$ yields

$$\|r^\star - \hat{r}\|_{2,\mathrm{beh}} \le \sqrt{C_{\mu/\pi}} \|\widetilde{\Delta}_Q\|_{2,\mathrm{beh}}. \tag{4}$$

Now let $\Delta Q := Q^\mu_{u^\star - g} - Q^\mu_{\hat{u} - g}$ and $\Delta u := u^\star - \hat{u}$. Then

$$\|\widetilde{\Delta}_Q\|_{2,\mathrm{beh}} \le \|Q^\mu_{\hat{u}-g} - \hat{Q}\|_{2,\mathrm{beh}} + \|\Delta Q\|_{2,\mathrm{beh}}. \tag{5}$$

Subtracting the Bellman equations gives $\Delta Q = \Delta u + \gamma P\mu\, \Delta Q$, hence $\|\Delta Q\|_{2,\mathrm{beh}} \le \|\Delta u\|_{2,\mathrm{beh}} + \gamma \|P\mu\, \Delta Q\|_{2,\mathrm{beh}}$.

Using Jensen again, $(P\mu\Delta Q)^2 \le P\mu((\Delta Q)^2)$, so

$$\begin{aligned}\|P\mu\, \Delta Q\|^2_{2,\mathrm{beh}} &\le \int P\mu\big((\Delta Q)^2\big)\, d\nu_\pi = \int (\Delta Q)^2\, d(\nu_\pi P\mu) \\ &= \int \sum_a (\Delta Q(s,a))^2 \mu(a \mid s)\, d(\nu_\pi P)(s) \\ &\le C_{\nu_\pi P/d_\mu} \int \sum_a (\Delta Q(s,a))^2 \mu(a \mid s)\, d d_\mu(s).\end{aligned} \tag{6}$$

Define the mixed norm

$$\|f\|_{2,d_\mu,\mu} := \left\{ \int \sum_a f(s,a)^2 \mu(a \mid s)\, d_\mu(s) \right\}^{1/2}.$$

Then (6) implies

$$\|\Delta Q\|_{2,\mathrm{beh}} \le \|\Delta u\|_{2,\mathrm{beh}} + \gamma \sqrt{C_{\nu_\pi P/d_\mu}}\, \|\Delta Q\|_{2,d_\mu,\mu}. \tag{7}$$

It remains to control $\|\Delta Q\|_{2,d_\mu,\mu}$. Taking $\|\cdot\|_{2,d_\mu,\mu}$ norms in $\Delta Q = \Delta u + \gamma P\mu\, \Delta Q$ gives $\|\Delta Q\|_{2,d_\mu,\mu} \le \|\Delta u\|_{2,d_\mu,\mu} + \gamma \|P\mu\, \Delta Q\|_{2,d_\mu,\mu}$. Because d_μ is stationary for P_μ, Jensen's inequality yields $\|P\mu f\|_{2,d_\mu,\mu} \le \|f\|_{2,d_\mu,\mu}$ for all f. Therefore

$$(1-\gamma)\|\Delta Q\|_{2,d_\mu,\mu} \le \|\Delta u\|_{2,d_\mu,\mu}.$$

Finally,

$$\begin{aligned}
\|\Delta u\|^2_{2,d_\mu,\mu} &= \int (\Delta u(s,a))^2\, d_\mu(s) \mu(da \mid s) \\
&= \int (\Delta u(s,a))^2 \frac{d d_\mu}{d\rho}(s) \frac{\mu(a \mid s)}{\pi(a \mid s)}\, d\nu_\pi(s,a) \\
&\le C_{d_\mu/\rho} C_{\mu/\pi} \|\Delta u\|^2_{2,\mathrm{beh}}.
\end{aligned}$$

Thus

$$\|\Delta Q\|_{2,d_\mu,\mu} \le \frac{\sqrt{C_{d_\mu/\rho} C_{\mu/\pi}}}{1-\gamma} \|u^\star - \hat{u}\|_{2,\mathrm{beh}}. \tag{8}$$

Combining (5), (7), and (8), we obtain

$$\|\widetilde{\Delta}_Q\|_{2,\mathrm{beh}} \le \|Q^\mu_{\hat{u}-g} - \hat{Q}\|_{2,\mathrm{beh}} + \left(1 + \frac{\gamma \sqrt{C_{\nu_\pi P/d_\mu}\, C_{d_\mu/\rho} C_{\mu/\pi}}}{1-\gamma} \right) \|u^\star - \hat{u}\|_{2,\mathrm{beh}}.$$

Substituting this into (4) proves the claim. □

F.2 Technical Lemmas

Lemma 3. *Under Assumption 2, set $C_p := \sqrt{2/\underline{p}}$. Then, with probability at least $1-\delta$,*

$$\|\hat{u} - u^\star\|_{2,\mathrm{beh}} \le C_p\, \rho_\pi(n,\delta).$$

Proof. Let $u^\star = \log \pi$, $\hat{u} = \log \hat{\pi}$, and define the likelihood ratio

$$\vartheta(a \mid s) := \frac{\hat{\pi}(a \mid s)}{\pi(a \mid s)}.$$

Since $\pi(a \mid s) \ge \underline{p}$ and $\hat{\pi}(a \mid s) \ge \underline{p}$, we have

$$\vartheta(a \mid s) \in [\underline{p}, \underline{p}^{-1}]$$

for all (s,a).

Let $x := \log \vartheta$ and $h(x) := e^x - 1 - x$. On $[\log \underline{p}, -\log \underline{p}]$, $h''(x) = e^x \ge \underline{p}$, while $h(0) = h'(0) = 0$. Hence $h(x) \ge \underline{p} x^2/2$. Because $\mathbb{E}_{\pi(\cdot|s)}[\vartheta] = 1$, for each s,

$$\begin{aligned}
\mathrm{KL}\big(\pi(\cdot \mid s) \,\|\, \hat{\pi}(\cdot \mid s)\big) &= \mathbb{E}_{\pi(\cdot|s)}[-x] = \mathbb{E}_{\pi(\cdot|s)}[e^x - 1 - x] \\
&\ge \frac{\underline{p}}{2}\, \mathbb{E}_{\pi(\cdot|s)}[x^2].
\end{aligned}$$

Averaging over s and invoking Assumption 2 yields

$$\|\hat{u} - u^\star\|_{2,\mathrm{beh}} \le \sqrt{\frac{2}{\underline{p}}} \left\{ \mathbb{E}_s \left[\mathrm{KL}\big(\pi(\cdot \mid s) \,\|\, \hat{\pi}(\cdot \mid s)\big) \right] \right\}^{1/2} \le C_p\, \rho_\pi(n,\delta),$$

which proves the claim. □

Lemma 4 (Per-iteration error bound)**.** *Under Assumptions 2–3, fix $\delta_\pi, \delta_Q \in (0, 0.5)$. On the intersection of the policy event at level δ_π and the kth regression event at level δ_Q,*

$$\|\mathcal{T}_{\hat{u}}^{\mu}(\hat{Q}^{(k-1)}) - \hat{Q}^{(k)}\|_{2,\mathrm{beh}} \le \sqrt{2/\underline{p}}\,\rho_\pi(n, \delta_\pi) + \rho_Q(n, \delta_Q) + \varepsilon_{\mathcal{F}}.$$

This event has probability at least $1 - \delta_\pi - \delta_Q$.

Proof. Use $\|\cdot\|$ to denote $\|\cdot\|_{2,\mathrm{beh}}$. Let

$$R_k := \|\mathcal{T}_{\hat{u}}^{\mu}(\hat{Q}^{(k-1)}) - \hat{Q}^{(k)}\|, \qquad R_k^\star := \inf_{f \in \mathcal{F}} \|\mathcal{T}_{\hat{u}}^{\mu}(\hat{Q}^{(k-1)}) - f\|.$$

By definition of the regret and Assumption 3, on the kth regression event at level δ_Q,

$$R_k^2 - (R_k^\star)^2 = \mathrm{reg}(\hat{Q}^{(k)} \mid \hat{u}, \hat{Q}^{(k-1)}) \le \rho_Q(n, \delta_Q)^2.$$

Since

$$R_k^2 - (R_k^\star)^2 = (R_k - R_k^\star)(R_k + R_k^\star) \ge (R_k - R_k^\star)^2,$$

it follows that

$$R_k \le R_k^\star + \rho_Q(n, \delta_Q).$$

Moreover, for any $f \in \mathcal{F}$,

$$\begin{aligned}\|\mathcal{T}_{\hat{u}}^{\mu}(\hat{Q}^{(k-1)}) - f\| &\le \|\mathcal{T}_{\hat{u}}^{\mu}(\hat{Q}^{(k-1)}) - \mathcal{T}_{u^\star}^{\mu}(\hat{Q}^{(k-1)})\| \\ &\quad + \|\mathcal{T}_{u^\star}^{\mu}(\hat{Q}^{(k-1)}) - f\| \\ &= \|\hat{u} - u^\star\| + \|\mathcal{T}_{u^\star}^{\mu}(\hat{Q}^{(k-1)}) - f\|.\end{aligned}$$

Taking the infimum over $f \in \mathcal{F}$ yields

$$R_k^\star \le \|\hat{u} - u^\star\| + \inf_{f \in \mathcal{F}} \|\mathcal{T}_{u^\star}^{\mu}(\hat{Q}^{(k-1)}) - f\|.$$

Since ν_π is a probability measure,

$$\|h\|_{2,\mathrm{beh}} \le \|h\|_\infty \qquad \text{for all } h,$$

so by the definition of $\varepsilon_{\mathcal{F}}$,

$$\inf_{f \in \mathcal{F}} \|\mathcal{T}_{u^\star}^{\mu}(\hat{Q}^{(k-1)}) - f\| \le \inf_{f \in \mathcal{F}} \|\mathcal{T}_{u^\star}^{\mu}(\hat{Q}^{(k-1)}) - f\|_\infty \le \varepsilon_{\mathcal{F}}.$$

Therefore, on the regression event,

$$\|\mathcal{T}_{\hat{u}}^{\mu}(\hat{Q}^{(k-1)}) - \hat{Q}^{(k)}\| \le \rho_Q(n, \delta_Q) + \|\hat{u} - u^\star\| + \varepsilon_{\mathcal{F}}.$$

Finally, on the policy event at level δ_π, Lemma 3 gives

$$\|\hat{u} - u^\star\|_{2,\mathrm{beh}} \le \sqrt{2/\underline{p}}\,\rho_\pi(n, \delta_\pi).$$

Combining the two bounds proves the inequality; the probability statement follows by a union bound. □

Assumption 4 (Concentrability of propagated state-action distributions)**.** *Assume that*

$$C_{\mathrm{conc}} := \sup_{m \ge 0} \left\| \frac{d\big(\nu_\pi (P\mu)^m\big)}{d\nu_\pi} \right\|_\infty < \infty.$$

Lemma 5 (Coverage implies propagated concentrability)**.** *Let $\nu_\pi := \rho \otimes \pi$ and $\nu_\mu := d_\mu \otimes \mu$, where d_μ is a stationary distribution of P_μ. Suppose that*

$$C_{\mu/\pi} := \left\| \frac{\mu(a \mid s)}{\pi(a \mid s)} \right\|_\infty < \infty, \qquad C_{d_\mu/\rho} := \left\| \frac{dd_\mu(s)}{d\rho(s)} \right\|_\infty < \infty,$$

and

$$C_{\nu_\pi P/d_\mu} := \left\| \frac{d(\nu_\pi P)(s)}{dd_\mu(s)} \right\|_\infty < \infty.$$

Then

$$\sup_{m \ge 0} \left\| \frac{d\big(\nu_\pi (P\mu)^m\big)}{d\nu_\pi} \right\|_\infty \le \max\Big\{1,\, C_{\nu_\pi P/d_\mu}\, C_{d_\mu/\rho}\, C_{\mu/\pi}\Big\}.$$

In particular, Assumption 1 implies Assumption 4.

Proof. For $m = 0$,

$$\left\| \frac{d\big(\nu_\pi (P\mu)^0\big)}{d\nu_\pi} \right\|_\infty = \left\| \frac{d\nu_\pi}{d\nu_\pi} \right\|_\infty = 1.$$

Now fix $m \geq 1$. Write

$$\eta_m := (\nu_\pi P)(P_\mu)^{m-1},$$

so that

$$\nu_\pi (P\mu)^m = \eta_m \otimes \mu.$$

Since $\nu_\pi = \rho \otimes \pi$, we have

$$\frac{d\big(\nu_\pi (P\mu)^m\big)}{d\nu_\pi}(s,a) = \frac{d\eta_m}{d\rho}(s)\, \frac{\mu(a \mid s)}{\pi(a \mid s)}.$$

Using $d_\mu \ll \rho$, this becomes

$$\frac{d\big(\nu_\pi (P\mu)^m\big)}{d\nu_\pi}(s,a) = \frac{d\eta_m}{dd_\mu}(s)\, \frac{dd_\mu}{d\rho}(s)\, \frac{\mu(a \mid s)}{\pi(a \mid s)}.$$

Hence

$$\left\| \frac{d\big(\nu_\pi (P\mu)^m\big)}{d\nu_\pi} \right\|_\infty \leq \left\| \frac{d\eta_m}{dd_\mu} \right\|_\infty \left\| \frac{dd_\mu}{d\rho} \right\|_\infty \left\| \frac{\mu}{\pi} \right\|_\infty.$$

It remains to bound the first factor. The one-step coverage condition implies the measure inequality

$$\nu_\pi P \leq C_{\nu_\pi P/d_\mu}\, d_\mu.$$

Applying the Markov kernel P_μ^{m-1} to both measures preserves this inequality. Since $d_\mu P_\mu^{m-1} = d_\mu$ by stationarity,

$$\eta_m = (\nu_\pi P) P_\mu^{m-1} \leq C_{\nu_\pi P/d_\mu}\, d_\mu, \qquad \left\| \frac{d\eta_m}{dd_\mu} \right\|_\infty \leq C_{\nu_\pi P/d_\mu}.$$

Thus, for all $m \geq 1$,

$$\left\| \frac{d\big(\nu_\pi (P\mu)^m\big)}{d\nu_\pi} \right\|_\infty \leq C_{\nu_\pi P/d_\mu}\, C_{d_\mu/\rho}\, C_{\mu/\pi}.$$

Combining this with the case $m = 0$ gives

$$\sup_{m \geq 0} \left\| \frac{d\big(\nu_\pi (P\mu)^m\big)}{d\nu_\pi} \right\|_\infty \leq \max\Big\{1,\, C_{\nu_\pi P/d_\mu}\, C_{d_\mu/\rho}\, C_{\mu/\pi}\Big\}.$$

□

F.3 GenPQR + FQE Bound

Boundedness condition for the FQE recursion. In addition to Assumptions 2–3, assume that the regression class is uniformly bounded:

$$\sup_{f \in \mathcal{F}} \|f\|_\infty \leq B_{\mathcal{F}} < \infty.$$

Assumption 2 already gives $\|\hat{u}\|_\infty \vee \|u^\star\|_\infty \leq B < \infty$, and g is bounded by construction. Hence the Bellman targets $\mathcal{T}_{\hat{u}}^\mu f$ are uniformly bounded whenever $f \in \mathcal{F}$ is. This is the standard boundedness condition used in sup-norm FQE analyses.

Lemma 6 (FQE error bound)**.** *Let $\mathcal{T}_u^\mu f := u - g + \gamma P\mu f$, and let $Q_{\hat{u}-g}^\mu$ denote the unique fixed point of $\mathcal{T}_{\hat{u}}^\mu$. Suppose $\hat{Q}^{(k)}$ is produced by Algorithm 2. Under Assumptions 2–3, Assumption 4, and the boundedness condition above, with probability at least $1 - \delta$,*

$$\|Q_{\hat{u}-g}^\mu - \hat{Q}^{(K)}\|_{2,\mathrm{beh}} \leq \sqrt{C_{\mathrm{conc}}}\, \gamma^K \|Q_{\hat{u}-g}^\mu - \hat{Q}^{(0)}\|_\infty + \kappa_{\mathrm{conc}} \Big(\varepsilon_{\mathcal{F}} + \rho_Q(n, \delta/(2K)) + \sqrt{2/\underline{p}}\, \rho_\pi(n, \delta/2) \Big).$$

Proof. Write $Q_{\hat{u}} := Q^{\mu}_{\hat{u}-g}$, and define

$$\xi_k := \mathcal{T}^{\mu}_{\hat{u}} \hat{Q}^{(k-1)} - \hat{Q}^{(k)}.$$

Since $Q_{\hat{u}}$ is the fixed point of $\mathcal{T}^{\mu}_{\hat{u}}$,

$$Q_{\hat{u}} - \hat{Q}^{(k)} = \mathcal{T}^{\mu}_{\hat{u}}(Q_{\hat{u}}) - \mathcal{T}^{\mu}_{\hat{u}}(\hat{Q}^{(k-1)}) + \xi_k = \gamma P\mu\big(Q_{\hat{u}} - \hat{Q}^{(k-1)}\big) + \xi_k.$$

Iterating this recursion gives

$$Q_{\hat{u}} - \hat{Q}^{(K)} = \gamma^K (P\mu)^K \big(Q_{\hat{u}} - \hat{Q}^{(0)}\big) + \sum_{j=1}^{K} \gamma^{K-j} (P\mu)^{K-j} \xi_j.$$

Therefore,

$$\begin{aligned} \|Q_{\hat{u}} - \hat{Q}^{(K)}\|_{2,\mathrm{beh}} \le \gamma^K \|(P\mu)^K (Q_{\hat{u}} - \hat{Q}^{(0)})\|_{2,\mathrm{beh}} \\ + \sum_{j=1}^{K} \gamma^{K-j} \|(P\mu)^{K-j} \xi_j\|_{2,\mathrm{beh}}. \end{aligned}$$

We now bound each propagated term in behavior norm. For any measurable f and any $m \ge 0$, Jensen's inequality gives

$$\big|(P\mu)^m f\big|^2 \le (P\mu)^m (f^2).$$

Hence

$$\begin{aligned} \|(P\mu)^m f\|^2_{2,\mathrm{beh}} &= \int \big|(P\mu)^m f\big|^2 \, d\nu_\pi \le \int (P\mu)^m (f^2) \, d\nu_\pi \\ &= \int f^2 \, d\big(\nu_\pi (P\mu)^m\big) \le C_{\mathrm{conc}} \int f^2 \, d\nu_\pi = C_{\mathrm{conc}} \|f\|^2_{2,\mathrm{beh}}. \end{aligned}$$

Thus

$$\|(P\mu)^m f\|_{2,\mathrm{beh}} \le \sqrt{C_{\mathrm{conc}}} \, \|f\|_{2,\mathrm{beh}}.$$

Applying this with $f = Q_{\hat{u}} - \hat{Q}^{(0)}$ and using $\|f\|_{2,\mathrm{beh}} \le \|f\|_\infty$, we obtain

$$\|(P\mu)^K (Q_{\hat{u}} - \hat{Q}^{(0)})\|_{2,\mathrm{beh}} \le \sqrt{C_{\mathrm{conc}}} \, \|Q_{\hat{u}} - \hat{Q}^{(0)}\|_\infty.$$

Let $\mathcal{E}_\pi$ be the policy event in Assumption 2 at level $\delta/2$, and let $\mathcal{E}_{Q,j}$ be the jth regression event in Assumption 3 at level $\delta/(2K)$. By a union bound,

$$\mathbb{P}\left(\mathcal{E}_\pi \cap \bigcap_{j=1}^{K} \mathcal{E}_{Q,j}\right) \ge 1 - \delta.$$

On this event, Lemma 4 gives, simultaneously for all $j \in [K]$,

$$\|\xi_j\|_{2,\mathrm{beh}} \le e_K := \varepsilon_{\mathcal{F}} + \rho_Q(n, \delta/(2K)) + \sqrt{2/\underline{p}} \, \rho_\pi(n, \delta/2).$$

The unpropagated term ξ_K has coefficient one. For $j < K$, the concentrability bound with $m = K - j \ge 1$ gives $\|(P\mu)^{K-j} \xi_j\|_{2,\mathrm{beh}} \le \sqrt{C_{\mathrm{conc}}} \|\xi_j\|_{2,\mathrm{beh}}$. Therefore,

$$\begin{aligned} \|Q_{\hat{u}} - \hat{Q}^{(K)}\|_{2,\mathrm{beh}} &\le \sqrt{C_{\mathrm{conc}}} \, \gamma^K \|Q_{\hat{u}} - \hat{Q}^{(0)}\|_\infty \\ &\quad + \left(1 + \sqrt{C_{\mathrm{conc}}} \sum_{m=1}^{K-1} \gamma^m\right) e_K \\ &\le \sqrt{C_{\mathrm{conc}}} \, \gamma^K \|Q_{\hat{u}} - \hat{Q}^{(0)}\|_\infty + \kappa_{\mathrm{conc}} e_K, \end{aligned}$$

which proves the claim. □

Proof of Theorem 4. Apply Theorem 3 with $\hat{Q} = \hat{Q}^{(K)}$:

$$\|r^\star - \hat{r}\|_{2,\mathrm{beh}} \le \sqrt{C_{\mu/\pi}} \Bigg\{ \|Q^\mu_{\hat{u}-g} - \hat{Q}^{(K)}\|_{2,\mathrm{beh}} + \left(1 + \frac{\gamma\sqrt{C_{\nu_\pi P/d_\mu} C_{d_\mu/\rho} C_{\mu/\pi}}}{1-\gamma}\right) \|u^\star - \hat{u}\|_{2,\mathrm{beh}} \Bigg\}.$$

Let $\mathcal{E}$ be the intersection of the policy event at level $\delta/2$ and the K regression events at level $\delta/(2K)$. Then $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$. On $\mathcal{E}$, Lemma 6 gives

$$\|Q^\mu_{\hat{u}-g} - \hat{Q}^{(K)}\|_{2,\mathrm{beh}} \le \sqrt{C_{\mathrm{conc}}}\,\gamma^K \|Q^\mu_{\hat{u}-g} - \hat{Q}^{(0)}\|_\infty + \kappa_{\mathrm{conc}}\Big(\varepsilon_{\mathcal{F}} + \rho_Q(n, \delta/(2K)) + C_p\,\rho_\pi(n, \delta/2)\Big),$$

where $C_p := \sqrt{2/\underline{p}}$. On the same event, Lemma 3 gives

$$\|u^\star - \hat{u}\|_{2,\mathrm{beh}} \le C_p\,\rho_\pi(n, \delta/2).$$

Substituting these two inequalities into the deterministic reward-recovery bound gives the bound in Theorem 4 on $\mathcal{E}$. □

G Continuous-Action Extension

This appendix gives the continuous-action analogue of the identification result and algorithm. Finite-action sums are replaced by integrals, and the softmax policy is replaced by a Boltzmann density.

Let $(\mathcal{A}, \mathcal{B}_\mathcal{A})$ be an action space equipped with a finite reference measure λ. For a bounded measurable state–action function f, define the log-partition operator

$$\mathrm{LSE}\, f(s) := \log \int_\mathcal{A} \exp\{f(s,a)\}\, \lambda(da),$$

and the associated Boltzmann policy

$$\pi_f(da \mid s) = \exp\{f(s,a) - \mathrm{LSE}\, f(s)\}\, \lambda(da).$$

For any policy or reference kernel $\eta(\cdot \mid s)$, write

$$(\eta f)(s) := \int_\mathcal{A} f(s,a)\, \eta(da \mid s), \qquad (P\eta f)(s,a) := \mathbb{E}\left[\int_\mathcal{A} f(s',a')\, \eta(da' \mid s') \,\middle|\, s, a\right].$$

The normalization is the same statewise affine constraint as in the main text,

$$(\mu r)(s) = g(s),$$

now with integrals in place of sums.

Assumption 5 (Continuous-action regularity)**.** *The discount satisfies $\gamma \in [0,1)$, the anchor g is bounded, and P maps bounded measurable state functions to bounded measurable state–action functions. The behavior policy admits a strictly positive density $p_\pi(a \mid s)$ with respect to λ, and*

$$u^\star(s,a) := \log p_\pi(a \mid s)$$

is bounded with $\mathrm{LSE}\, u^\star(s) = 0$ for all s. The normalization kernel $\mu(\cdot \mid s)$ is a Markov kernel on $\mathcal{A}$, dominated by λ, and is used only to define the normalization and the evaluation operator $P\mu$.

Theorem 6 (Continuous-action fixed-point identification)**.** *Under Assumption 5, let $Q^\star$ be the unique bounded solution of*

$$Q^\star(s,a) = u^\star(s,a) - g(s) + \gamma(P\mu Q^\star)(s,a).$$

Then, modulo λ-null action sets at each state, the unique bounded normalized reward representative inducing the Boltzmann policy $\pi(da \mid s) = \exp\{u^\star(s,a)\}\lambda(da)$ is

$$r^\star(s,a) = Q^\star(s,a) - (\mu Q^\star)(s) + g(s), \qquad v^\star(s,a) = -(P\mu Q^\star)(s,a).$$

Equivalently, $(r^\star, v^\star)$ satisfies the continuous soft Bellman equation $v^\star = P\,\mathrm{LSE}(r^\star + \gamma v^\star)$, the normalization $\mu r^\star = g$, and induces the observed Boltzmann density.

Proof. The operator

$$(\mathcal{T}Q)(s,a) := u^\star(s,a) - g(s) + \gamma(P\mu Q)(s,a)$$

is a γ-contraction in sup norm on bounded measurable state–action functions, since $\|P\mu Q - P\mu Q'\|_\infty \le \|Q - Q'\|_\infty$. Thus $Q^\star$ exists and is unique.

Define $r^\star$ and $v^\star$ as in the theorem. The normalization is immediate:

$$\mu r^\star = \mu Q^\star - \mu Q^\star + g = g.$$

Using the fixed-point equation,

$$r^\star + \gamma v^\star = Q^\star - \mu Q^\star + g - \gamma P\mu Q^\star = u^\star - \mu Q^\star.$$

Because $\mu Q^\star$ is state-only and $\mathrm{LSE}\, u^\star = 0$,

$$\mathrm{LSE}(r^\star + \gamma v^\star)(s) = \mathrm{LSE}\{u^\star - \mu Q^\star\}(s) = -(\mu Q^\star)(s).$$

Applying P gives

$$P\,\mathrm{LSE}(r^\star + \gamma v^\star) = -P\mu Q^\star = v^\star,$$

so the continuous soft Bellman equation holds. Moreover,

$$(r^\star + \gamma v^\star) - \mathrm{LSE}(r^\star + \gamma v^\star) = u^\star,$$

and therefore the induced Boltzmann density is the observed density.

It remains to show uniqueness. Let (r, v) be any bounded normalized pair that satisfies the continuous soft Bellman equation and induces the same Boltzmann policy. With $q := r + \gamma v$, equality of Boltzmann densities implies

$$q(s,a) - \mathrm{LSE}\, q(s) = u^\star(s,a) \qquad \text{for each } s \text{ and } \lambda\text{-almost every } a.$$

Thus $q = u^\star + c$ for the state-only function $c(s) := \mathrm{LSE}\, q(s)$. Bellman feasibility gives $v = Pc$ and hence $r = u^\star + c - \gamma Pc$. The normalization becomes

$$g = \mu u^\star + c - \gamma P_\mu c, \qquad P_\mu c(s) := \int_{\mathcal{A}} Pc(s,a)\,\mu(da \mid s).$$

Since P_μ is a Markov operator with $\|P_\mu\|_\infty \le 1$, the operator $I - \gamma P_\mu$ is invertible on bounded state functions by the Neumann series. Hence c, and therefore (r, v), is unique modulo λ-null action sets. □

Algorithm 3 CONTINUOUS-ACTION GENPQR

input Transitions $\{(s_i, a_i, s_i')\}_{i=1}^n$, normalization kernel $\mu(\cdot \mid s)$, anchor $g(s)$, discount γ, class $\mathcal{F}$, iterations K

1: Estimate a conditional log-density $\hat{u}(s,a) \approx \log p_\pi(a \mid s)$ with respect to the reference measure λ

2: Initialize $\hat{Q}^{(0)}(s,a) \leftarrow 0$

3: **for** $k = 1, \dots, K$ **do**

4: For each i, compute

$$m_i^{(k-1)} \leftarrow \int_{\mathcal{A}} \hat{Q}^{(k-1)}(s_i', a')\,\mu(da' \mid s_i')$$

by quadrature or Monte Carlo samples from $\mu(\cdot \mid s_i')$

5: Set

$$y_i \leftarrow \hat{u}(s_i, a_i) - g(s_i) + \gamma m_i^{(k-1)}$$

6: Fit $\hat{Q}^{(k)} \in \mathcal{F}$ by regressing y_i on (s_i, a_i)

7: **end for**

8: Set $\hat{Q} \leftarrow \hat{Q}^{(K)}$

output Reward:

$$\hat{r}(s,a) = \hat{Q}(s,a) - \int_{\mathcal{A}} \hat{Q}(s,a')\,\mu(da' \mid s) + g(s),$$

with the final integral evaluated by the same quadrature or Monte Carlo rule.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction state the main claims, and the paper separates identification results, algorithmic reductions, finite-sample guarantees, and empirical comparisons; see sections 1 and 3 to 6.

   Guidelines:

   - The answer should be yes, no, or n/a.
   - Claims in the abstract and introduction should match the paper's contributions and scope.
   - Important assumptions and limitations should be clearly stated.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The assumptions and discussion identify the main scope restrictions, including coverage, bounded log-policy conditions, approximation error, low-temperature behavior policies, and the synthetic nature of the experiments; see assumption 1 and sections C, D, 5.2 and 5.3.

   Guidelines:

   - The answer should be yes, no, or n/a.
   - Authors are encouraged to discuss strong assumptions, robustness to violations of those assumptions, and the practical scope of the claims.
   - The answer n/a means the paper has no limitations; the answer no means limitations exist but are not discussed.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete proof?

   Answer: [Yes].

   Justification: The main assumptions and results are stated in sections 3 and 5, with proofs, technical lemmas, and the continuous-action extension in sections E to G.

   Guidelines:

   - The answer should be yes, no, or n/a.
   - All assumptions should be stated or referenced in theorem and lemma statements.
   - Complete proofs may appear in the main paper or appendix; informal proof sketches in the main text should be backed by formal proofs.
   - Theorems, lemmas, and external results used in proofs should be properly referenced.

4. **Experimental result reproducibility**

   Question: If the contribution is a dataset, model, algorithm, or empirical evaluation, what steps were taken to make the results reproducible or verifiable?

   Answer: [Yes].

   Justification: The main experimental results are reported in section 6, with simulation details, estimator choices, hyperparameters, repetition counts, and additional diagnostics in sections C and D.

   Guidelines:

   - The answer should be yes, no, or n/a.
   - Reproducibility can be supported by detailed algorithm descriptions, instructions, code, data, checkpoints, hosted access, or other appropriate avenues.
   - If reproducibility is limited, the paper should explain the available path for verifying or reproducing the main claims.

5. **Open access to data and code**

   Question: If the paper includes experiments, does it include the code, data, and instructions needed to reproduce the main experimental results, either in the supplemental material or as a URL?

Answer: [Yes].

Justification: The experiments use a synthetic benchmark adapted from Geng et al. [13]. The public reproduction repository provides a pinned environment, exact setup and execution commands, experiment scripts, tests, and plotting code; see section C.

Guidelines:

- The answer should be yes, no, or n/a.
- Instructions should include the exact commands and environment needed to reproduce the main experimental results, including the new method and baselines.
- If only a subset of experiments is reproducible, the paper should say which subset.
- At submission time, released links or archives should preserve anonymity.

6. **Experimental setting/details**

   Question: If the paper includes experiments, does it specify all training and test details, including data splits, hyperparameters, and how they were chosen?

   Answer: [Yes].

   Justification: The environment, trajectory counts, random-seed protocol, policy estimators, Q-evaluation procedures, baselines, hyperparameters, and reporting metrics are described in section C.

   Guidelines:

   - The answer should be yes, no, or n/a.
   - The answer n/a means the paper does not include experiments.
   - Important experimental details should appear in the main paper; full details may also appear in the appendix, code, or supplemental material.
   - Hyperparameter choices and selection procedures should be stated.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

   Justification: The paper reports means with 95% confidence intervals over 100 random seeds in the main experiments, and describes the interval construction in section C.

   Guidelines:

   - The answer should be yes, no, or n/a.
   - The answer n/a means the paper does not include experiments.
   - Error bars, confidence intervals, or significance tests should be reported for experiments supporting the main claims when appropriate.
   - The paper should state what source of variability is shown and how error bars or intervals are computed.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources needed to reproduce the experiments?

   Answer: [No].

   Justification: The paper reports method runtimes, architectures, and iteration counts, but not the hardware, memory, or total project compute.

   Guidelines:

   - The answer should be yes, no, or n/a.
   - The answer n/a means the paper does not include experiments.
   - The paper should describe the compute type, memory or hardware when relevant, per-run compute, and estimated total compute.
   - The paper should disclose whether substantially more compute was used for the project than for the reported experiments.

9. **Code of ethics**

   Question: Have the authors read the NeurIPS Code of Ethics and ensured that the research conforms to it?

   Answer: [Yes].

Justification: The paper is a methodological contribution to inverse reinforcement learning and offline reward recovery, and we are not aware of any aspect of the work that conflicts with the NeurIPS Code of Ethics.

Guidelines:

- The answer should be yes or no, with explanation if needed.
- Any special circumstances requiring deviation from the NeurIPS Code of Ethics should be explained while preserving anonymity.

10. **Broader impacts**

    Question: If appropriate for the scope and focus of the paper, does it discuss potential negative societal impacts of the work performed?

    Answer: [N/A].

    Justification: The paper develops statistical and algorithmic methodology rather than a deployed decision system.

    Guidelines:

    - The answer should be yes, no, or n/a.
    - Authors should discuss plausible negative impacts when there is a direct path to misuse, fairness, privacy, security, or other societal harms.
    - Foundational or generic methods may answer n/a when no direct negative application is apparent, but limitations and deployment risks should still be handled honestly.
    - Mitigations may be discussed when negative impacts are identified.

11. **Safeguards**

    Question: Does the paper describe safeguards for responsible release of data or models that have a high risk for misuse?

    Answer: [N/A].

    Justification: The paper does not introduce a deployed system, content-generation model, or human-facing interactive tool requiring additional safeguards.

    Guidelines:

    - The answer should be yes, no, or n/a.
    - High-risk or dual-use model releases should include appropriate access, usage, or release safeguards.
    - Dataset releases that may contain unsafe content should describe how unsafe material was avoided or controlled.

12. **Licenses for existing assets**

    Question: If existing assets are used, are the creators or original owners properly credited and are the license and terms of use respected?

    Answer: [No].

    Justification: The experimental environment and methodological components are credited, but the accompanying repository does not state license information for adapted components.

    Guidelines:

    - The answer should be yes, no, or n/a.
    - Existing code, data, models, and benchmarks should be cited and their licenses or terms of use should be respected.
    - Versions and URLs should be provided when possible.
    - If releasing derived assets, the original and derived licenses should be stated where applicable.

13. **New assets**

    Question: If new assets are released, are they documented and are relevant details provided alongside the assets?

    Answer: [No].

    Justification: The accompanying repository documents the experiment scripts, configurations, and plotting code, but does not state a license or intended-use terms for the released code.

    Guidelines:

    - The answer should be yes, no, or n/a.
    - The answer n/a means the paper does not release new assets.

- New datasets, code, or models should be documented with training, license, limitations, and intended-use details as relevant.
- Released assets should be anonymized for initial submission when applicable.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper uses synthetic simulation data and does not conduct crowdsourcing, user studies, or other human-subject experiments.

    Guidelines:

    - The answer should be yes, no, or n/a.
    - The full participant instructions, screenshots, and compensation details should be included when applicable, with more detail in the main paper when human subjects are central to the contribution.
    - Work involving data collection, curation, or other labor should follow the NeurIPS Code of Ethics, including compensation requirements.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: If human-subject research was conducted, does the paper describe participant risks and whether IRB or equivalent approval was obtained?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing or human-subject research, so IRB or equivalent approval is not applicable.

    Guidelines:

    - The answer should be yes, no, or n/a.
    - IRB or equivalent review should be described when required by the relevant institution or jurisdiction.
    - Initial submissions should not include information that breaks anonymity, such as the reviewing institution, when applicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

    Answer: [N/A].

    Justification: LLMs are not part of the proposed method or experiments.

    Guidelines:

    - The answer should be yes, no, or n/a.
    - The answer n/a means LLMs are not important, original, or non-standard components of the core method.
    - LLM use only for writing, editing, or formatting does not need to be declared when it does not affect methodology, rigor, or originality.