Journal of Machine Learning Research 24 (2023) 1-64 Submitted 10/20; Revised 12/22; Published 2/23

Necessary and Sufficient Conditions for Inverse Reinforcement Learning of Bayesian Stopping Time Problems *

Kunal Pattanayak KP487@CORNELL.EDU
Electrical and Computer Engineering
Cornell University
Ithaca, NY 14853*, USA*

Vikram Krishnamurthy VIKRAMK@CORNELL.EDU
Electrical and Computer Engineering
Cornell University
Ithaca, NY 14853*, USA*

Editor: Andreas Krause

Abstract

This paper presents an inverse reinforcement learning (IRL) framework for Bayesian stopping time problems. By observing the actions of a Bayesian decision maker, we provide a necessary and sufficient condition to identify if these actions are consistent with optimizing a cost function. In a Bayesian (partially observed) setting, the inverse learner can at best identify optimality wrt the observed strategies. Our IRL algorithm identifies optimality and then constructs set-valued estimates of the cost function. To achieve this IRL objective, we use novel ideas from Bayesian revealed preferences stemming from microeconomics. We illustrate the proposed IRL scheme using two important examples of stopping time problems, namely, sequential hypothesis testing and Bayesian search. As a real-world example, we illustrate using a YouTube dataset comprising metadata from 190000 videos how the proposed IRL method predicts user engagement in online multimedia platforms with high accuracy. Finally, for finite datasets, we propose an IRL detection algorithm and give finite sample bounds on its error probabilities.

Keywords: Inverse Reinforcement Learning (IRL), Bayesian Revealed Preferences, Stopping Time Problems, Inverse Detection, Sequential Hypothesis Testing (SHT), Bayesian Search, Finite Sample Complexity

1. Introduction

In a stopping time problem, a decision maker obtains noisy observations of a random variable (state of nature) x sequentially over time. Based on the observation history (sigma-algebra generated by the observations), the decision maker decides at each time whether to continue or stop. If the decision maker chooses the continue action, it pays a continuing cost and obtains the next observation. If the decision maker chooses the stop action at a specific time, then the problem terminates, and the decision maker pays a stopping cost. In a *Bayesian* stopping time problem, the decision maker knows the prior distribution of state of nature x and the observation likelihood (conditional distribution of the observations) $p(y|x)$ given the state x, and uses this information to update its belief and choose

*. A short version of partial results has appeared in the Proceedings of the International Conference on Information Fusion, July 2020.

its continue and stop actions. Finally, in an *optimal* Bayesian stopping time problem, the decision maker chooses its continue and stop actions to minimize an expected cumulative cost function.

Inverse reinforcement learning (IRL) aims to estimate the costs/rewards of a decision maker by observing its actions and was first studied by Ng et al. (2000) and Abbeel and Ng (2004). This paper considers IRL for Bayesian stopping time problems. Suppose an inverse learner observes the actions of a decision maker performing Bayesian sequential stopping in *multiple environments*. The decision maker has a fixed observation likelihood and observation cost, and incurs a different stopping cost in each environment[1]. The inverse learner does not know the realizations of the observation sequence nor the observation likelihood of the decision maker; the inverse learner only knows the true state x and observes the stopping action a of the decision maker. The two main questions we address are:

1. How can the inverse learner check if the actions of a Bayesian decision maker are consistent with optimal stopping?
2. If the decision maker's actions are consistent with optimal stopping, how can the inverse learner estimate the stopping costs of the multiple environments?

1.1 Main results and context

The key results in this paper are summarized as follows:

1. Inverse RL for Bayesian sequential stopping: Theorem 3 in Sec. 2 is our first key IRL result. Theorem 3 specifies a set of convex inequalities that are simultaneously necessary and sufficient for the actions of a Bayesian decision maker in multiple environments to be consistent with optimal stopping. If so, then Theorem 3 provides an algorithm for the inverse learner to generate set-valued estimates of the decision maker's costs in the multiple environments. Theorem 3 is especially useful in scenarios where the inverse learner has no knowledge of the decision maker's observation likelihood or observation sample paths, and yet can construct a set-valued estimate of the costs incurred by the decision maker.

2. Inverse RL for SHT and Search: Sec. 3 and Sec. 4 construct IRL algorithms for two specific examples of Bayesian stopping time problems, namely, Sequential Hypothesis Testing (SHT) and Search. The main results, Theorem 6 and Theorem 9 specify necessary and sufficient conditions for the decision maker's actions to be consistent with optimal SHT and optimal search, respectively. If the conditions hold, Theorems 6 and 9 provide algorithms to estimate the incurred misclassification costs (for SHT) and search costs (for Bayesian search). In Sec. 3 for inverse SHT, we also propose an IRL algorithm to compute a point-estimate of the decision maker's costs. The point-estimate is computed by maximizing the regularized margin of the convex feasibility test for inverse SHT proposed in Theorem 6 and estimates the misclassification costs with up to 95% accuracy. Also, in Sec. 3.6, we compare numerically the performance of the IRL algorithm in Theorem 3 with two existing IRL algorithms (Choi and Kim, 2011) in the literature. This numerical comparison highlights how the IRL approach in this paper complements the results of Choi and Kim (2011). Theorem 6 achieves IRL when the inverse learner has partial information about the decision maker's costs.

3. Illustration of Inverse RL for Bayesian stopping on Real-World Dataset: One important use case of IRL is to extract preferences from expert human agents (Lee et al., 2014; Gombolay et al., 2016). In Sec. 5, we illustrate how our IRL algorithms extend to predicting human-level online multimedia user engagement using a massive YouTube dataset comprising video metadata

1. We refer the reader to Rust (1994, Ch. 3.5) and Rolland et al. (2022) for motivating the need to for multiple environments for identifiability of Markov decision processes (MDPs).

from approximately 190000 videos.[2] From the set of costs that pass the convex feasibility test in Theorem 3 for optimal Bayesian stopping, we chose two point-valued IRL costs for IRL prediction, namely, max-margin IRL and entropy-regularized IRL. The main finding is that both point estimates accurately predict YouTube commenting behavior. Also, we observe that the max-margin estimate yields a more accurate prediction compared to the entropy-regularized estimate (in terms of the chi-square and total variation distance).

4. Sample Complexity for IRL: In Sec. 6, we propose IRL detection tests for optimal stopping, optimal SHT and optimal search under finite sample constraints. Theorems 11, 13 and 15 in Sec. 6 comprise our sample complexity results that characterize the robustness of the detection tests by specifying Type-I and Type-II error bounds for the IRL detection tests. To the best of our knowledge, our finite sample complexity results for the IRL detector, namely, the sample size required to achieve a Type-I or Type-II error probability below a specified value for IRL, are novel.

The proofs of all theorems are provided in the Appendix. For a practitioner's perspective, our key IRL algorithms are Theorems 3, 6 and 9, and finite sample complexity results for IRL error bounds are Theorems 11, 13 and 15. The MATLAB codes and the YouTube dataset for our real-world numerical experiment in Sec. 5 are available on Github and are completely reproducible.

1.2 Identifiability. Why IRL for a decision maker in multiple environments?

An important aspect of our IRL framework is that the inverse learner observes the decision maker in multiple environments.[3] The purpose of this section is to motivate this framework.

We consider a decision maker operating over M environments. In each environment, the decision maker solves a stopping time problem with a distinct stopping cost. The decision maker has a fixed observation likelihood (sensor accuracy) and sensing cost (operating cost), where both variables are invariant across multiple environments. Therefore, there are up to M distinct strategies exhibited by the decision maker, one for each environment. Let $J(\mu_m, s_m)$ denote the expected cost incurred by the decision maker when it chooses stopping strategy μ_m in environment m with stopping cost s_m.

Consider now the inverse learner that observes the decision maker. Assume that the inverse learner does not know the stopping costs s_m, but only observes[4] the set of strategies $\{\mu_m, m = 1, 2, \ldots, M\}$. To achieve IRL, the inverse learner must first establish if the decision maker's strategy in each environment is consistent with minimizing an expected cost. Equivalently, the inverse learner must check if the expected cost incurred by the decision maker in environment m by choosing strategy μ_m is less than that incurred by all other (infinitely many) stopping strategies. However, the inverse learner does not observe infinitely many strategies, but only M strategies. Given the decision maker's strategies in M, each with a distinct stopping cost, the inverse learner's procedure to identify if the decision maker is optimal or not is defined below:

IRL identifiability of optimal stopping agent. *Consider a Bayesian stopping agent that chooses*

2. Although understanding YouTube commenting behavior was the main focus of our previous work (Hoiles et al., 2020), the inference methodology and numerical experiments in this paper are new; see Appendix G and `https://github.com/KunalP117/YouTube-Commenting-Analysis` for details.
3. The inverse learner in this paper can be viewed as a *passive* analyst that does not control the environment variables, that is, the agent's stopping costs. An interesting extension of this paper (for future work) is to consider an *active* inverse learner that purposefully adapts the environment variables to minimize IRL detection errors.
4. We are deliberately simplifying the IRL framework here for explanatory reasons. Our main result assumes the inverse learner only observes the actions of the decision maker, and not the strategy.

strategy μ_m in environment m, over multiple environments $m = 1, 2, \ldots, M$. Then, identifying an optimal Bayesian stopping time agent is equivalent to checking if the following inequalities have a feasible solution:

$$\textit{There exists } s_1, s_2, \ldots, s_M \textit{ such that: } J(\mu_m, s_m) \leq J(\mu_n, s_m),\ \forall\, m, n. \tag{1}$$

Here, $J(\mu_m, s_n)$ is the decision maker's cumulative expected cost when the decision maker chooses strategy μ_m and incurs a stopping cost s_n.

The solution of the feasibility problem in (1) is the set-valued IRL estimate of the stopping costs incurred by the decision maker. The comparison in (1) between the performance of the decision maker's strategy in each environment to the strategies chosen in all other (finitely many) environments is formalized in Lemma 2 and is achieved by the inverse learner via the IRL procedure in Theorem 3. We also refer the reader to the seminal work of Rust (1994, Ch. 3.5) and Kim et al. (2021) on identifiability of MDPs for further justification of multiple environments. The above framework of a Bayesian stopping time agent operating in multiple environments arises in several applications; see Sec. A.1 for details.

1.3 Context. Bayesian revealed preference for IRL

The formalism used in this paper to achieve IRL is *Bayesian revealed preferences* studied in microeconomics by Martin (2014); Caplin and Martin (2015) and Caplin and Dean (2015); see Sec. A.2 for more details. This Bayesian revealed preference-based approach *complements* existing IRL results for partially observed Markov decision processes (POMDP) including Choi and Kim (2011). This paper considers a subset of POMDPs, namely, Bayesian stopping time problems. Due to the problem structure, we show that our IRL algorithms *do not* require knowledge of the observation likelihood of the decision maker and also do not require solving a POMDP.

We now briefly discuss how the Bayesian revealed preference based IRL approach differs from classical IRL.

1. The classical IRL frameworks (Ng et al., 2000; Abbeel and Ng, 2004) assume the observed agent is a reward maximizer (or equivalently, cost minimizer) and then seeks to estimate its cost function. The approach in this paper is more fundamental. We first *identify* if the decisions of a single decision maker in multiple environments are consistent with optimality and if so, we then generate set-valued estimates of the costs that are consistent with the observed decisions.
2. Classical IRL assumes complete knowledge of the decision maker's observation likelihood. We assume the inverse learner only knows the state of nature and the action chosen when the decision maker stops, and does not know its observation likelihoods or the sequence of observation realizations. Two important scenarios where this situation arises are:
   (i) *Multimedia Datasets.* In online multimedia datasets such as the YouTube dataset analyzed in Sec. 5, it is impossible to know the attention span (observation likelihood) of the online user. All that is available are the online user's actions (interactions such as comments and comment ratings) and the underlying state of nature (video metadata such as viewcount, thumbnail and video description); see also Hoiles et al. (2020).
   (ii) *Adversarial Signal Processing.* In adversarial signal processing and sensing applications, it is not realistic for the inverse learner to know the model dynamics of the agent. An important example is IRL for radars (Krishnamurthy, 2020), where the radar is the adversary and so it is impossible to know its sensing modes (observation likelihood); however, the inverse learner

records the electromagnetic waveforms (response) emitted by the radar.
Additional applications where only the agent decisions are available for IRL (and not the observation likelihood) include consumer insights and advertisement design research, interpretable ML in smart healthcare and electronic warfare. These are discussed in Appendix A.

3. *Algorithmic Issues:* In classical IRL (Abbeel and Ng, 2004), the inverse learner solves the Bayesian stopping time problem iteratively for various choices of the cost. This can be computationally prohibitive since it involves stochastic dynamic programming over a belief space which is PSPACE hard (Papadimitriou and Tsitsiklis, 1987). The IRL procedure in this paper does not require solving a POMDP and only requires testing for the feasibility of a set of convex inequalities.

For brevity, we discuss related IRL literature and applications of IRL for Bayesian stopping problems in Appendix A.

2. Identifying optimal Bayesian stopping and reconstructing agent costs

Our IRL framework comprises a decision maker's actions in a stopping time problem over M environments, and an *inverse learner* that observes these actions. This section defines the IRL problem that the inverse learner faces and then presents two results regarding the inverse learner:

1. *Identifying Optimal Stopping*. Theorem 3 below provides a necessary and sufficient condition for the inverse learner to identify if the Bayesian decision maker chooses its actions as the solution of an optimal stopping problem.
2. *IRL for Reconstructing Costs*. Theorem 3 is also constructive. It shows that the continue and stopping costs of the Bayesian decision maker can be reconstructed by solving a convex feasibility problem.

This section provides a complete IRL framework for Bayesian stopping time problems and sets the stage for subsequent sections where we formulate generalizations and examples.

2.1 Bayesian stopping agent

A Bayesian stopping time agent is parametrized by the tuple

$$\Xi = (\mathcal{X}, \pi_0, \mathcal{Y}, \mathcal{A}, B, \mu) \tag{2}$$

where

- $\mathcal{X} = \{1, 2, \dots X\}$ is a finite set of states.
- At time 0, the true state $x^o \in \mathcal{X}$ is sampled from prior distribution π_0. x^o is unknown to the agent.
- $\mathcal{Y} \subset \mathbb{R}$ is the observation space. Given state x^o, the observations $y \in \mathcal{Y}$ have conditional probability density $B(y, x^o) = p(y|x^o)$.
- $\mathcal{A} = \{1, 2, \dots A\}$ is the finite set of stopping actions.
- Finally, μ denotes the agent's stopping strategy. The stopping strategy operates sequentially on a sequence of observations $y_1, y_2, \dots$ as discussed below in Protocol 1.

Protocol 1 *Sequential Decision-making protocol: Assume the agent knows* Ξ.

1. *Generate* $x^o \sim \pi_0$*, at time* $t = 0$*. Here* x^o *is not known to the agent.*

2. *At time $t > 0$, agent records observation $y_t \sim B(\cdot, x^o)$.*

3. Belief Update: *Let $\mathcal{F}_t$ denote the sigma-algebra generated by observations $\{y_1, y_2, \dots y_t\}$. The agent updates its belief (posterior) $\pi_t(x) = \mathbb{P}(x^o = x|\mathcal{F}_t), x \in \mathcal{X}$ using Bayes formula as*

$$\pi_t = \frac{B(y_t)\pi_{t-1}}{\mathbf{1}'B(y_t)\pi_{t-1}}, \tag{3}$$

where $B(y) = \text{diag}(\{B(y, x), x \in \mathcal{X}\})$. The belief π_t is an X-dimensional probability vector in the $X - 1$ dimensional unit simplex

$$\Delta(\mathcal{X}) \stackrel{def.}{=} \{\pi \in \mathbb{R}_+^X : \mathbf{1}'\pi = 1\}. \tag{4}$$

4. *Choose action $a_t = \mu(\pi_t, t)$ from the set $\mathcal{A} \cup \{$continue$\}$. If $a_t \in \mathcal{A}$, then stop, else if $a_t =$ continue, set $t = t + 1$ and go to Step 2.*

The stopping strategy μ is a (possibly randomized) time-dependent mapping from the agent's belief at time $t \in \mathbb{Z}^+$ to the set $\mathcal{A} \cup \{\text{continue}\}$ and belongs to $\boldsymbol{\mu}$, the set of admissible stopping strategies:

$$\boldsymbol{\mu} = \{\mu : \Delta(\mathcal{X}) \times \mathbb{Z}^+ \to \mathcal{A} \cup \{\text{continue}\}\}. \tag{5}$$

We define the random variable τ as the time when the agent stops and takes a stop action from $\mathcal{A}$.

$$\tau = \inf\{t \geq 0 |\ \mu(\pi_t, t) \neq \{\text{continue}\}\}. \tag{6}$$

Clearly, the set $\{\tau = t\}$ is measurable wrt $\mathcal{F}_t$, the sigma-algebra generated by observations $\{y_1, y_2, \dots y_t\}$. Hence, the random variable τ is adapted to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$. In the following sub-section, we will introduce costs for the agent's stop and continue actions. We will use τ for expressing the expected cumulative cost of the agent.

To summarize, a Bayesian stopping agent is parameterized by Ξ and operates according to Protocol 1. Several decision problems such as SHT and sequential search fit this formulation.

2.2 Optimal Bayesian stopping agent in multiple environments

So far we have defined a Bayesian stopping agent. Our main IRL result is to identify if a Bayesian stopping agent's behavior in a set of environments $\mathcal{M}$ is *optimal*. The purpose of this section is to define optimal Bayesian stopping (Bertsekas, 2015) in multiple environments. For identifiability reasons (see assumption (A2) below) we require at least two environments ($M \geq 2$).

An optimal Bayesian stopping agent in multiple environments is defined by the tuple

$$\Xi_{opt} = (\Xi, \mathcal{M}, \mathcal{C}, \boldsymbol{s}, \boldsymbol{\mu}^*). \tag{7}$$

In (7),

- $\mathcal{M}$ is the set of M environments.
- The parameters $\mathcal{X}, \mathcal{Y}, \mathcal{A}, \pi_0, p$ in Ξ (2) and continue cost $\mathcal{C}$ (defined below) are the same for all environments in $\mathcal{M}$.
- $\mathcal{C} = \{c_t\}_{t\geq 0}$, $c_t(x) \in \mathbb{R}^+$ is the continue cost incurred in any environment $m \in \mathcal{M}$ at time t given state $x^o = x$.

- $\boldsymbol{s} = \{s_m(x,a), x \in \mathcal{X}, a \in \mathcal{A}, m \in \mathcal{M}\}$, $s_m(x,a) < \infty$ is the cost for taking stop action a when the state $x^o = x$ in the m^{th} environment.
- $\boldsymbol{\mu}^* = \{\mu^*_m, m \in \mathcal{M}\}$ is the set of **optimal** stopping strategies of the Bayesian stopping agent over the set of environments $\mathcal{M}$, where the optimality is defined in Definition 1 below. In environment m, the Bayesian stopping agent employs its stopping strategy $\mu^*_m, m \in \mathcal{M}$ and operates according to Protocol 1.

Definition 1 (Optimal Stopping Strategy) *For each environment $m \in \mathcal{M}$, strategy μ^*_m is optimal for stopping cost $s_m(x,a)$ iff the following conditions hold:*

$$\mu^*_m(\pi, \tau) = \operatorname*{argmin}_{a \in \mathcal{A}} \pi' \bar{s}_{m,a}, \tag{8}$$

$$J(\mu^*_m, s_m) = \inf_{\mu \in \boldsymbol{\mu}} J(\mu, s_m), \tag{9}$$

Recall $\boldsymbol{\mu}$ (5) denotes the set of all stopping strategies. Also $J(\mu, s_m)$ is the expected cumulative cost defined as:

$$J(\mu, s_m) = G(\mu, s_m) + C(\mu), \text{ where}$$

$$G(\mu, s_m) = \mathbb{E}_\mu \left\{ \pi'_\tau \bar{s}_{m,\mu(\pi_\tau,\tau)} \right\}, \; C(\mu) = \mathbb{E}_\mu \left\{ \sum_{t=0}^{\tau-1} \pi'_t \bar{c}_t \right\}, \; \mu \in \boldsymbol{\mu}. \tag{10}$$

$\mathbb{E}_\mu$ denotes expectation parametrized by μ wrt the probability measure induced by $y_{1:\tau}$. Also, $\bar{s}_a$, $\bar{c}_t$ are the stopping and continue[5] cost vectors, respectively, vectorized over states $x \in \mathcal{X}$.

Definition 1 is standard for the optimal strategy in a sequential stopping problem (Krishnamurthy, 2016). The optimal strategy naturally decomposes into two steps: choosing whether to continue or stop according to (9); and if the decision is to stop, then choose a specific stopping action from $\mathcal{A}$ according to (8). The optimal stopping strategies $\mu_m, m \in \mathcal{M}$ that satisfy the conditions (8), (9) can be obtained by solving a stochastic dynamic programming problem (Krishnamurthy, 2016). It is a well-known result (Lovejoy, 1987) that the set of beliefs for which it is optimal to stop is convex.

RELATION TO BAYESIAN CONTEXTUAL BANDITS

For readers familiar with the multi-armed bandit problem, optimal Bayesian stopping can be viewed as an instance of the *partially-observed regularized contextual Bayesian bandit problem*; *contextual* (Agrawal and Goyal, 2013) since the agent faces multiple ground truths x (context), *partially observed* (Krishnamurthy and Wahlberg, 2009) since the agent observes a sequence of noisy measurements of the underlying context x, *Bayesian* (Hong et al., 2022) since the agent minimizes its expected cumulative cost per context averaged over all contexts sampled from a prior distribution π_0, and *regularized* (Fontaine et al., 2019) since the agent minimizes the sum of expected stopping cost and a regularization term, namely, the expected continue cost. Loosely speaking, this paper addresses the problem of IRL for partially-observed regularized contextual bandits. Although our IRL results are introduced in subsequent sections, we remark here that there is ample scope to extend the results in this paper to typical RL decision frameworks that allow underlying state transitions. At a high level, this can be made possible by constructing feasibility tests in terms of the state-occupancy measure induced by the decision maker's policy in multiple environments.

5. Since the continue cost is a positive real, the stopping time τ (6) is finite a.s.

2.3 IRL for inverse optimal stopping. Main result

We now discuss an inverse learner-centric view of the Bayesian stopping time problem and the main IRL result. Suppose the inverse learner observes the actions of a Bayesian stopping agent in M environments, where each environment is characterized by the stopping costs incurred by the agent. Suppose the agent performs several independent trials of Protocol 1 in all M environments. We make the following assumptions about the inverse learner performing IRL.

(A1) The inverse learner knows the dataset

$$\mathcal{D}_M = (\pi_0, \boldsymbol{p}), \text{ where } \boldsymbol{p} = \{p_m(a|x), x \in \mathcal{X}, a \in \mathcal{A}, m \in \mathcal{M}\}. \tag{11}$$

In (11), $p_m(a|x)$ is the Bayesian stopping agent's conditional probability of choosing stop action a at the stopping time given state $x^o = x$ in the m^{th} environment. We call $p_m(a|x)$ as the agent's *action selection policy*.
Note that:
(i) The inverse learner does not know the stopping times; it only has access to the conditional density of which stop action a was chosen given the true state x^o.
(ii) We assume the decision maker visits all states in the support of the prior pmf π_0 (11) infinitely often. In Sec. 6, we address the case where the decision maker visits the states finitely often and provide IRL performance guarantees via finite sample complexity.

(A2) Dataset $\mathcal{D}_M$ is generated by a Bayesian agent acting in at least $M \geq 2$ environments, where each environment has distinct stopping costs.

Both assumptions are discussed below after the main theorem, but let us make some preliminary remarks at this stage. (A1) implies the inverse learner observes the stopping actions chosen by a Bayesian stopping agent in a finite number (M) of environments, where the agent performs an infinite number of independent trials of Protocol 1 in each environment; see discussion in Sec. 2.5 for asymptotic interpretation. In Sec. 6 we will consider finite sample effects where the inverse learner observes the agent performing a finite number of independent trials of Protocol 1. Assumption (A2) is necessary for the inverse optimal stopping problem to be well-posed.

Let μ_m denote the policy chosen by the agent in the m^{th} environment, and $\boldsymbol{\mu}_{\mathcal{M}} = \{\mu_m, m \in \mathcal{M}\}$ denote the set of chosen strategies.[6] The finite assumption on $|\mathcal{M}|$ in (A1) imposes a restriction on our IRL task of identifying optimality of a Bayesian stopping agent formalized below:

Lemma 2 (IRL identifiability of optimal Bayesian stopping agent.) *Given the dataset $\mathcal{D}_M$* (11)*, the inverse learner can identify an optimal Bayesian stopping agent* (7) *acting in M environments if and only if* (8) *and the following relaxation of* (9) *holds:*

$$G_{m,m} + C_m \leq G_{n,m} + C_n, \ \forall m, \ n \in \mathcal{M}, \ m \neq n. \tag{12}$$

In (12)*, $G_{n,m} = G(\mu_n, s_m)$ is the expected stopping cost and $C_m = C(\mu_m)$ is the expected cumulative continue cost for the policy μ_m chosen in environment m, $m \in \mathcal{M}$.*

6. Recall that μ is a generic variable of a stopping policy, $\boldsymbol{\mu}$ is the space of admissible policies, μ_m^* is the optimal policy in environment m and μ_m is a realization of the agent's policy.

	C unknown	$C \in \mathcal{C}$ convex in $p(a\|x)$	$C \in \mathcal{C}$ non-convex in $p(a\|x)$
Identifiability	Absolute Optimality	Absolute Optimality	Relative Optimality
Conditions	(8), (9) in Def. 1	(8), (9) in Def. 1	(8), (12) in Lemma 2
IRL Example	$--$	Inverse Optimal Stopping with Entropic Running Cost	Inverse SHT (Sec. 3)
Reconstruction	Convex reconstruction (94)	Convex reconstruction (94)	Reconstructed cost for a finite set of strategies/

Table 1: IRL Identifiability of Optimal Bayesian stopping.

The proof of Lemma 2 is in Appendix B. Lemma 2 formalizes the IRL identification procedure of the inverse learner in (1). Since the inverse learner only observes the agent's actions from M strategies chosen by the stopping agent, the best the inverse learner can do is check if μ_m is optimal for environment m out of the finite strategies in $\boldsymbol{\mu}_{\mathcal{M}}$. Indeed, the expected stopping cost $G_{n,m}$ is a function of the policy μ_n. However, in Appendix C, we show how the expected stopping cost can be expressed only in terms of the observed variables in $\mathcal{D}_M$, namely, the action selection policies $\{p_m(a|x)\}_{m=1}^M$ of the agent induced by the stopping strategies $\{\mu_m\}_{m=1}^M$. This is precisely what Theorem 3 below achieves when the inverse learner has access to the agent's action selection policies.

Remarks:
(1) If the analyst does not know *a priori* the structure of the expected continue cost in (10), then the IRL identifiability can be generalized from testing for relative optimality (8), (12) to testing for absolute optimality (8), (9) in Definition 1. Specifically, we show a certain reconstruction of the expected continue cost (see (94) in Appendix C) ensures if relative optimality (12) holds, then absolute optimality (9) holds.
(2) In contrast to remark (1) above, if the analyst does know a functional form of the expected continue cost, IRL identifiability cannot be improved from testing for relative optimality. One example is IRL for inverse SHT discussed in Sec. 3 below where the expected continue cost is known to be the expected stopping time of the agent. On a deeper and more subtle level, knowledge of the structure of the expected continue cost imposes an implicit constraint on the reconstructed cost. Ensuring the reconstructed expected continue cost (94) in Appendix C satisfies this implicit constraint is non-trivial and beyond the scope of this paper.

We now present our first main IRL result. The result specifies a set of inequalities that, given the inverse learner's specifications in assumptions (A1) and (A2), are simultaneously *necessary* and *sufficient* for the inverse learner to identify a Bayesian stopping agent's actions to be optimal in the sense of Lemma 2. For readability, we provide the exact expressions for the feasibility inequalities introduced below after the main theorem.

Theorem 3 (IRL for inverse Bayesian optimal stopping (Caplin and Dean, 2015)) *Consider the inverse learner with dataset $\mathcal{D}_M$ (11) obtained from a Bayesian stopping agent's actions over M environments. Assume (A1) and (A2) hold. Then:*
1. *Identifiability: The inverse learner can identify if the dataset $\mathcal{D}_M$ is generated by an optimal Bayesian stopping agent, i.e.,* (8) *and* (9)*; see Lemma 2.*

2. Existence: *There exists an optimal stopping agent parameterized by tuple* Ξ_{opt} (7) *, if and only if there exists a feasible solution to the following convex (in stopping costs) inequalities:*

$$\begin{aligned}
&\textit{Find } s_m(x,a) \in \mathbb{R}_+ \ \forall m \in \mathcal{M} \textit{ s.t.} \\
&\text{NIAS}(\mathcal{D}_M, \{s_m(x,a), x \in \mathcal{X}, a \in \mathcal{A}, m \in \mathcal{M}\}) \leq 0, \qquad (13)\\
&\text{NIAC}(\mathcal{D}_M, \{s_m(x,a), x \in \mathcal{X}, a \in \mathcal{A}, m \in \mathcal{M}\}) \leq 0. \qquad (14)
\end{aligned}$$

The NIAS (No Improving Action Switches) and NIAC (No Improving Action Cycles) inequalities are defined in (16), (17) below, and are convex in the stopping cost $s_m(x,a), m \in \mathcal{M}$.

3. Reconstruction of costs:

(a) *If the inverse learner knows the agent's expected continue cost* C_m *for all environments* m*, the set-valued IRL estimate of the agent's stopping costs is the set of all feasible costs* $\{s_m(x,a), m \in \mathcal{M}\}$ *that satisfy the NIAS* (13)*, NIAC* (14) *and SUMCOST inequalities below:*

$$\text{SUMCOST}(\mathcal{D}_M, \{s_m(x,a), C_m, m \in \mathcal{M}\}) \leq 0, \qquad (15)$$

where SUMCOST *is defined in (18), and* C_m *is the expected cost of the Bayesian stopping agent in environment* m.

(b) *Suppose the inverse learner knows the agent's stopping costs, and the NIAS* (13) *and NIAC* (14) *inequalities are feasible. Then, the set-valued IRL estimate of the agent's expected continue cost is given by the set of all feasible costs* C_m *that satisfy the SUMCOST inequality* (15)*. Also, if the inverse learner knows the agent's expected continue cost is convex, then the SUMCOST inequality structure permits a convex reconstruction of the cost outlined in Definition 4.* ■

Theorem 3 is proved in Appendix C. It says that identifying if a set $\mathcal{M}$ comprising stopping actions of a Bayesian stopping agent in multiple environments is optimal and then reconstructing the costs incurred in the environments is equivalent to solving a convex feasibility problem. Theorem 3 provides a constructive procedure for the inverse learner to generate set valued estimates of the stopping cost $s_m(x,a)$ and expected cumulative continue cost C_m for all environments $m \in \mathcal{M}$. Algorithms for convex feasibility such as interior points methods (Boyd and Vandenberghe, 2004) can be used to check feasibility of (13) and (14) (defined in (16) and (17) below) and construct a feasible solution.

The inequalities NIAS, NIAC and SUMCOST denoted abstractly in Theorem 3 are defined below:

Definition 4 (NIAS, NIAC and SUMCOST inequalities) *Given dataset* $\mathcal{D}_M$*, stopping costs* $\{s_m(x,a), m \in \mathcal{M}\}$ *and expected continue costs* $\{C_m, m \in \mathcal{M}\}$*:*

$$\text{NIAS}: \sum_{x \in \mathcal{X}} p_m(x|a)(s_m(x,a) - s_m(x,b)) \leq 0, \forall a, m. \qquad (16)$$

$$\text{NIAC}: \sum_{m \in \widehat{\mathcal{M}}} \mathbb{E}_{a \sim \sum_x \pi_0(x) p_m(\cdot|x)} \left\{ \min_{a' \in \mathcal{A}} \mathbb{E}_{x \sim p_m(\cdot|a)} \{s_m(x,a) - s_{m+1}(x,a')\} \right\} \leq 0,$$

$$\textit{for any subset of indices } \widehat{\mathcal{M}} \subseteq \mathcal{M}, \textit{ where } m_k + 1 = m_{k+1} \textit{ if } k < l \textit{ and } m_l + 1 = m_1. \qquad (17)$$

$$\text{SUMCOST}: \ \mathbb{E}_{x \sim \pi_0, a \sim p_m(\cdot|x)} \{s_m(x,a)\} + C_m \leq \mathbb{E}_{a \sim p_n(a)} \{\min_{a' \in \mathcal{A}} \mathbb{E}_{x \sim p_n(\cdot|a)} \{s_m(x,a')\}\} + C_n,$$

$$\forall\, m, n \in \mathcal{M}. \qquad (18)$$

Reconstruction of expected cumulative continue cost. *If NIAS, NIAC and SUMCOST inequalities defined above have a feasible solution, the following convex reconstruction of the agent's expected continue cost is consistent with optimal Bayesian stopping* (8)*,* (9)*, a stronger condition compared to relative optimality* (8)*,* (12)*:*

$$\widehat{C}(\mu) = \max_{m=1,2,\ldots,M} \left\{ C_m + G_{m,m} - \tilde{G}(\mu, s_m) \right\}, \text{ where} \tag{19}$$

$$\tilde{G}(\mu, s_m) = \sum_{a \in \mathcal{A}} \left(\sum_{x \in \mathcal{X}} p_\mu(a|x) \pi_0(x) \right) \min_{b \in \mathcal{A}} \sum_{x \in \mathcal{X}} p_\mu(x|a) s_m(x, b), \text{ and} \tag{20}$$

$$G_{m,m} = \sum_{x \in \mathcal{X}, a \in \mathcal{A}} \pi_0(x) p_m(a|x) s_m(x, a) \tag{21}$$

The above reconstruction assumes the agent's mapping from the sequence of observations $y_{1:\tau(\mu)}$ to the space of actions is one-to-one, and is valid if and only if the agent's expected cumulative continue cost is convex.

Let us now provide an intuitive explanation for the abstract inequalities of Theorem 3.
NIAS (13): NIAS applies to each of the M environments in $\mathcal{M}$. NIAS checks if, for every environment, the agent chooses the optimal stop action given its stopping belief and stopping strategy.
NIAC (14): NIAC checks for optimality of the agent's stopping strategies in M environments. Since the stopping agent chooses its strategies in a finite number (M) environments, NIAC checks if the agent's strategy in the m^{th} environment performs at least as well as the strategies of the agent in all other environments given the environment's stopping cost $s_m(x, a)$, for all $m \in \mathcal{M}$. If so, it constructs a feasible set of stopping costs in the M environments so that the chosen strategies are consistent with an optimal stopping agent.
SUMCOST (15): If the Bayesian agent is an optimal stopping agent (NIAS and NIAC have a feasible solution), SUMCOST constructs a set of feasible expected continue costs incurred by the Bayesian agent in the multiple environments. The feasibility of NIAS and NIAC ensures that the SUMCOST inequalities have a feasible solution. In (18), the RHS term is the expected cumulative cost of the agent in environment n given the stopping costs in environment m. The feasibility inequality (18) checks for feasible expected cumulative continue costs so that the agent's stopping strategies in $\mathcal{M}$ are identified as optimal by the inverse learner, i.e., (12) is satisfied. The reconstructed cost $\widehat{C}$ (19) is a convex interpolation of expected stopping costs and feasible scalars C_m (18) such that conditions (8) and (9) for optimal Bayesian stopping hold; see Appendix C for a detailed discussion. We remark that the reconstruction in (19) is only valid when (a) the inverse learner has no information about the agent's observation likelihood, and (b) the inverse learner does not know the agent's expected continue cost. In Table 1, we highlight the subtle issues underpinning IRL identifiability for optimal Bayesian stopping in more detail. In Sec. 3 below, we discuss IRL for optimal Bayesian stopping when the inverse learner knows the agent's expected continue cost; hence, the reconstruction (19) is no more required for achieving IRL.

2.4 Discussion of Theorem 3

We now discuss the implications of Theorem 3 and contextualize the NIAS and NIAC feasibility inequalities (13), (14) of Theorem 3.
(i) Necessity and Sufficiency.
The NIAS and NIAC conditions (13), (14) are necessary and sufficient for the inverse learner to

identify an optimal stopping agent. This makes Theorem 3 a remarkable result. If no feasible solution exists, then the dataset $\mathcal{D}_M$ cannot be rationalized by an optimal Bayesian stopping agent. Also, if there exists a feasible solution, then the dataset $\mathcal{D}_M$ must be generated by an optimal stopping agent in multiple environments (Lemma 2).

(ii) Set valued estimate vs point estimate.

An important consequence of Theorem 3 is that the reconstructed utilities are set-valued estimates rather than point valued estimates even though the dataset $\mathcal{D}_M$ has $K \to \infty$ samples. Estimating the costs from the solution of a cost minimization problem is an ill-posed problem. Put differently, all points in the feasible set of rationalizing costs explain the dataset $\mathcal{D}_M$ equally well.

(iii) Consistency of Set-Valued Estimate.

The NIAS and NIAC inequalities are both necessary and sufficient for optimal Bayesian stopping. The necessity implies that the true stopping costs and expected continue costs incurred by the agent are feasible wrt the convex NIAS and NIAC inequalities. Hence, the IRL procedure is consistent in that the set-valued estimator contains the true generating model.

(iv) Context: NIAS and NIAC.

The inequalities (8), (12) for the inverse learner to identify an optimal stopping agent can be written in abstract notation as (22), (23), respectively, in terms of the variables $\{s_m, C_m\}_{m=1}^{\mathcal{M}}$:

$$\text{NIAS}(\{\{p(y_{1:\tau(\mu_m)}|x), x \in \mathcal{X}\}, s_m, m \in \mathcal{M}\}, \pi_0) \leq 0, \tag{22}$$

$$\text{NIAC}^*(\{\{p(y_{1:\tau(\mu_m)}|x), x \in \mathcal{X}\}, s_m, C_m, m \in \mathcal{M}\}, \pi_0) \leq 0. \tag{23}$$

The inverse learner in our setup does not know the agent's observation sequences $\{y_{1:\tau}, m \in \mathcal{M}\}$, observation likelihood B or the continue cost $\mathcal{C}$. Hence, as shown in Appendix C, the best the inverse learner can do is check for the feasibility of the NIAC (17) that does not depend on C_m. Otherwise, the IRL task is equivalent to using optimality equations (8), (12) expressed abstractly as NIAS and NIAC* above to reconstruct the costs. Eq. 13 and 14 in Theorem 3 specialize to (22) and (23) by replacing the action selection policy $p_m(a|x)$ with the unknown likelihood $p(y_{1:\tau(\mu_m)}|x)$. Put differently, (13) and (14) defined in (16), (17) can be viewed as surrogates of the feasibility conditions (22) and (23), respectively. However, as discussed in the proof in Appendix C, the action selection policy $p_m(a|x)$ suffices for both necessity and sufficiency of Bayes optimality (22), (23) in spite of being a Blackwell noisy measurement of $p(y_{1:\tau(\mu_m)}|x)$. Also, observe the NIAC inequality (17) is independent of C_m and expressed only in terms of stopping costs s_m. However, as shown in Appendix C, the feasibility of both inequalities (23) and (17) are equivalent. Finally, in some examples of stopping time problems such as SHT discussed in Sec. 3, the inverse learner knows the agent's expected cumulative continue cost and hence, can use the NIAC* inequality as is to identify optimality and achieve IRL.

NIAS and NIAC with ε-feasibility. One trivial solution that satisfies both NIAS and NIAC inequalities in Theorem 3 is the degenerate cost of all zeros. Such degeneracy is common in IRL literature due to the fundamental ill-posedness of the inverse optimization problem. In practice, one can ensure only non-trivial solutions pass the NIAS and NIAC feasibility inequalities by introducing a margin constraint:

$$\text{NIAS}(\cdot) \leq -\varepsilon, \ \text{NIAC}(\cdot) \leq -\varepsilon, \ \varepsilon > 0. \tag{24}$$

Margin constraints for ensuring non-degenerate solutions to feasibility tests are common practices in IRL (Ratliff et al., 2006). In complete analogy, using the ϵ restriction of (24), we can ensure only non-trivial informative costs pass the NIAS and NIAC feasibility test of Theorem 3.

(v) Private and Public Beliefs.
The stopping belief π_τ in (9) can be interpreted as the *private belief* evaluated by the agent after measuring $y_{1:\tau}$ in the sense of Bayesian social learning (Krishnamurthy, 2016; Chamley, 2004). Since π_τ is unavailable to the inverse learner, it uses the *public belief* $p(x|a)$ as a result of the agent's stop action to estimate its incurred costs.
(vi) IRL for stopping agent whose observation likelihood changes with the environment. For notational convenience, we assume the Bayesian agent's observation likelihood is fixed across different environments. However, in Appendix C, we discuss under what conditions the inverse learner can achieve IRL when the Bayesian agent's observation likelihoods change with the environment. We provide a specific example of the agent continue cost, namely, the entropic continue cost that facilitates the inverse learner to achieve IRL for different agent observation likelihoods in different environments. The agent's stopping cost in this case is a logistic function in terms of its action selection policy; the logistic function also arises in Max-Entropy IRL (Ziebart et al., 2008). This resemblance is not surprising; the agent in (Ziebart et al., 2008) maximizes its cumulative expected reward subject to a bound on the mutual information between the prior and the distribution of beliefs induced by its policy. The objective function in (9) where C is the mutual information between the prior and the stopping belief is simply the Lagrangian form of the objective the agent aims to optimize in Ziebart et al. (2008). The IRL problem for agents that Maximize their expected terminal rewards with a mutual information penalty has also been studied in the Bayesian revealed preference literature by Caplin et al. (2019).
(vii) IRL for boundedly-rational forward learner.
For general POMDPs, it is difficult[7] for a Bayesian sequential decision maker to compute the optimal policy μ^* in (8), (9). We say that a strategy $\hat{\mu}$ is ϵ-optimal if the following condition holds:

$$\epsilon\text{-optimal Bayesian stopping: } J(\hat{\mu}) - J(\mu^*) \leq \epsilon, \text{ for some } \epsilon \geq 0. \tag{25}$$

Eq. 25 arises when the forward learner uses sub-optimal procedures for solving the POMDP such as approximate value iteration, open loop feedback and finite state controllers. When both the stopping cost and the expected continue cost are free variables like in Theorem 3, detecting ϵ-optimality is non-identifiable and a difficult task. However, if either the stopping cost or the expected continue cost, (such as in the case of SHT discussed in Sec. 3) is known to the inverse learner, one can identify ϵ-optimality based on the feasibility of the IRL inequalities. We briefly discuss identification of ϵ-optimality after Theorem 6; a general framework is beyond the scope of this paper and the subject of future work. Indeed, more precise knowledge of the agent's sub-optimality allows the inverse learner to achieve IRL; see Brown et al. (2019) for a discussion on how to achieve IRL when the inverse learner has access to a ranked set of forward learner's decision trajectories, ranked according to the extent of sub-optimality in each trajectory.
(viii) **No knowledge of observation likelihood by the inverse learner.** This paper assumes the inverse learner has no knowledge of the agent's observation likelihood. The sufficiency proof of Theorem 3 exploits this zero-knowledge assumption and posits that the inverse learner can thus assume a one-to-one mapping from the space of observation sequences $y_{1:\tau(\mu)}$ to the space of stopping actions. Indeed, one can show that if the instantaneous continue cost has an entropic form, for example, the Shannon-Gibbs entropy, Rényi entropy or Tsallis entropy, the optimal mapping

7. Papadimitriou and Tsitsiklis (1987) show that solving partially observed Markov decision processes are in general PSPACE hard. The SHT and Search problems discussed in this paper are special cases where the optimal stopping strategy is stationary due to the problem structure and characterized as a threshold policy in the belief space.

from observation sequences to stopping actions is one-to-one due to the strongly concave nature of these costs; see Caplin et al. (2019) for a discussion of IRL for entropic costs.

(ix) **Partial knowledge of agent costs.** If the Bayesian agent's instantaneous continue cost is zero, then it is optimal to never stop sensing, i.e., the agent observe infinitely many samples and the posterior belief approaches the Dirac delta function centered at the state x[8]. Hence, the optimal $p_m(a|x)$ has non-zero weights if and only if $a \in \operatorname{argmin}_{a' \in \mathcal{A}} \ s_m(x, a')$. Then checking for optimal Bayesian stopping with zero running cost is equivalent to identifying feasible stopping costs that satisfy the following condition:

$$p_m(a|x) \neq 0 \iff a \in \operatorname*{argmin}_{a' \in \mathcal{A}} \ s_m(x, a'). \tag{26}$$

Sec. 3 considers the case where the instantaneous continue cost is a constant, hence the cumulative expected continue cost is proportional to the expected stopping time of the agent. If the inverse learner knows the expected continue cost, IRL is achieved by checking for the existence of feasible stopping costs that satisfy the NIAS (16) and SUMCOST (18) inequalities with C_m set to the agent's expected continue cost in environment m.

(x) **IRL with ε-feasibility.** If neither the stopping costs nor the expected continue costs are known to the inverse learner, the NIAS, NIAC and SUMCOST inequalities are trivially feasible by choosing the degenerate solution of constant costs. In this case it makes sense to construct the inverse learner's non-trivial IRL cost estimate as the set of feasible costs $\{s_m(x, a), C_m, m \in \mathcal{M}\}$ that are ϵ-feasible wrt the NIAC, NIAC and SUMCOST inequalities:

- Choose feasibility margins $\epsilon_{NIAS},\ \epsilon_{NIAC},\ \epsilon_{SUMCOST} \geq 0,$ not all zero.
- Construct the set-valued IRL estimate as the set of all tuples $\{s_m(x, a), C_m, m \in \mathcal{M}\}$ that satisfy

$$\text{NIAS}(\cdot) \leq \epsilon_{NIAS},\ \text{NIAC}(\cdot) \leq \epsilon_{NIAC} \ \text{ and } \ \text{SUMCOST}(\cdot) \leq \epsilon_{SUMCOST}. \tag{27}$$

2.5 Discussion of (A1) and (A2)

(A1): To motivate (A1), suppose for each environment $m \in \mathcal{M}$, the inverse learner records the Bayesian stopping agent's true state $x^o_{k,m}$, stopping action $a_{k,m}$ and stopping time $\tau_k(\mu_m)$ over $k = 1, 2, \ldots, K$ independent trials. Then the pmf $p_m(a|x)$ in (11) is the limit pmf of the empirical pmf $\hat{p}_m(a|x)$ as the number of trials $K \to \infty$ defined as:

$$\hat{p}_m(a|x) = \frac{\sum_{k=1}^{K} \mathbb{1}\{x^o_{k,m} = x, a_{k,m} = a\}}{\sum_{k=1}^{K} \mathbb{1}\{x^o_{k,m} = x\}}. \tag{28}$$

Specifically, since for each $m \in \mathcal{M}$ the sequence $\{x^o_{k,m}, a_{k,m}\}$ is i.i.d for $k = 1, 2, \ldots K$, by Kolmogorov's strong law of large numbers, as the number of trials $K \to \infty$, $\hat{p}_m(a|x)$ converges with probability 1 to the pmf $p_m(a|x)$. In the remainder of the paper (apart from Sec. 6), we will work with the asymptotic dataset $\mathcal{D}_M$ for IRL. In Sec. 6 we analyze the effect of finite sample size K

8. It follows from Bernstein-von Mises theorem (Le Cam, 1953) that, under mild smoothness conditions, the agent's posterior belief converges asymptotically to a normal distribution centered around the maximum likelihood estimate with covariance $\lim_{t\to\infty}(t\ I(x))^{-1}$, where I denotes the Fisher information matrix.

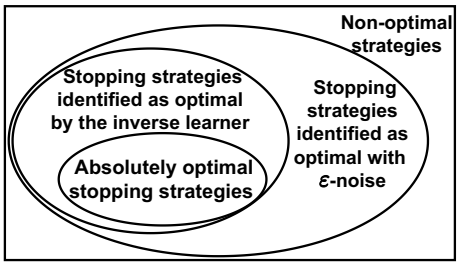

Figure 1: Given tuple $(\Xi, \mathcal{C}, \{s_m(x,a), m \in \mathcal{M}\})$, the set of stopping strategies (12) of a stopping agent identified as an optimal stopping agent by the inverse learner (Lemma 2) contains the stopping strategies of an absolutely optimal agent defined by (8), (9). Such strategies can be obtained by small perturbations of the absolute optimal strategies such that the Bayesian stopping agent's strategy in each environment still performs better than that chosen by the agent in any other environment in $\mathcal{M}$. Like Sec. 2, Sec. 3 and 4, deals with identifying such optimal strategies for the SHT and search problems. In Sec. 6, we will detect if the agent's strategies corrupted by noise (due to finite sample constraints) belong to the set of strategies identified as optimal strategies by the inverse learner.

on the inverse learner using concentration inequalities.

(A2): (A2) is necessary for the identification of an optimal stopping agent (Lemma 2) to be well-posed. Suppose (A2) does not hold. Then, for $M = 2$ and true stopping costs $s_1 = s_2$, we have $p_1(a|x) = p_2(a|x)$ in $\mathcal{D}_M$. This implies the set of feasible solutions (C_1, C_2) for the feasibility inequality (18) is the set $\{(C_1, C_2) : C_1 = C_2, C_1, C_2 \in \mathbb{R}_+\}$ and is hence, unidentifiable. [9]

2.6 Outline of proof of Theorem 3

The proof of Theorem 3 in Appendix C involves two main ideas. The first key idea is to specify a fictitious likelihood $\mathbb{P}_\mu(\tilde{y}_\pi|x)$ parametrized by the stopping strategy so that given strategy μ, observation likelihood B and prior π_0, the observation trajectory $y_{1:\tau}$ of the stopping time problem yields an identical stopping belief π_μ, i.e.,

$$\mathbb{P}(\tilde{y}_\pi|x,\mu) = \mathbb{P}\left(\{y_{1:\tau}\} : \pi_\tau = \pi|x\right).$$

A more precise statement is given in (75). In other words, a one-step Bayesian update using the likelihood $\mathbb{P}(\tilde{y}_\pi|x,\mu)$ is equivalent to the multi-step Bayesian update (3) of the state till the stopping time. This idea is shown in Fig. 2. Recall that the cumulative expected cost of the agent comprises two components, the stopping cost and cumulative continue cost. A useful property of this fictitious likelihood is that it is a sufficient statistic for the expected stopping cost $G(\cdot)$.

The second main idea is to formulate the agent's expected cumulative cost using the observed action selection policy $p(a|x)$ of the agent instead of the unobserved fictitious likelihood $p(y_{1:\tau(\mu_m)}|x)$

9. The condition $M = 1$ (or equivalently, $M = 2$ with equal stopping costs) is analogous to probing an agent with the same probe vector in classical revealed preferences (Afriat, 1967; Varian, 2012). The obtained dataset of probes and responses can be rationalized by any concave, locally non-satiated, monotone utility function thus leading to loss of identifiability of the agent's utilities.

that determines the expected stopping cost. $p_m(a|x)$ (11) is a stochastically garbled (noisy) version of $p(y_{1:\tau(\mu_m)}|x)$. We use this concept to formulate the NIAS and NIAC inequalities whose feasibility given $\mathcal{D}_M$ is necessary and sufficient for identifying an optimal stopping by a Bayesian stopping agent in multiple environments.

Showing that feasibility of the NIAS and NIAC inequalities (13), (14) is a necessary condition for the stopping strategies chosen by the Bayesian stopping agent to be optimal, (8), (12) is straightforward. The key idea in the sufficiency proof is to note that the elements of the garbling matrix that maps the fictitious observation likelihood to the action selection policy is unknown to the inverse learner. Hence, the inverse learner can arbitrarily assume $p_m(a|x)$ to be an accurate measurement of $p(y_{1:\tau(\mu_m)}|x)$. We then show that for a feasible set of viable stopping costs $\{s_m(x,a), C_m, m \in \mathcal{M}\}$ that satisfy the NIAS and NIAC inequalities, there exist a set of positive reals $\{C_m, m \in \mathcal{M}\}$ that satisfy (8), (9) with the expected cumulative continue cost incurred by the agent in the m^{th} environment set to C_m.

The NIAS and NIAC inequalities are convex in the stopping costs $s_m, m \in \mathcal{M}$. The inverse learner can solve for these convex feasibility constraints to obtain a feasible solution. Thus, we have a constructive IRL procedure for reconstructing the stopping and expected cumulative continue costs for the inverse optimal stopping time problem.

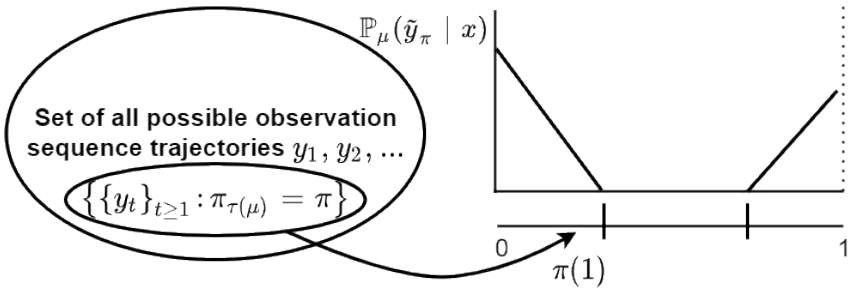

Figure 2: Schematic illustration of first main idea of proof of Theorem 3 for the case when $X = 2$. The key idea is to construct a fictitious observation likelihood $\mathbb{P}_\mu(\tilde{y}_\pi|x)$ for compact representation of the agent's expected stopping cost. The probability of generating the fictitious observation $\tilde{y}_\pi$ is equal to the probability of a sequence of observations yielding a stopping belief π for a given stopping time μ.

2.7 Summary

This section has laid the groundwork for IRL of a Bayesian stopping time agent. Specifically, we discussed the dynamics of the Bayesian stopping time agent in a single environment (2) and multiple environments (7). We then described the IRL problem that the inverse learner aims to solve. Theorem 3 gave a necessary and sufficient condition for a Bayesian stopping time agent to be identified as an optimal stopping agent when its decisions in multiple environments are observed by the inverse learner. The agent's stopping cost in each environment can be estimated by solving a convex feasibility problem. Theorem 3 forms the basis of the IRL framework in this paper. Next, we develop IRL results for 2 examples of stopping time problems, namely, sequential hypothesis testing and Bayesian search.

3. Example 1. IRL for sequential hypothesis testing (SHT)

We now discuss our first example of IRL for an optimal Bayesian stopping time problem, namely, *inverse* Sequential Hypothesis Testing (SHT). Our main result below (Theorem 6) specifies a necessary and sufficient condition for IRL in SHT. The SHT problem is a special case of the optimal Bayesian stopping problem discussed in Sec. 2.2 since the continue cost c_t (2) is a constant for all time t in the SHT problem. For our IRL task, the continue cost can be chosen as 1 WLOG.

3.1 Sequential hypothesis testing (SHT) Problem

Let $y_1, y_2, \ldots$ be a sequence of i.i.d observations. Suppose the Bayesian agent knows that the pdf of y_i is either $p(y|x=1)$ or $p(y|x=2)$. The aim of classical SHT is to decide sequentially on whether $x=1$ or $x=2$ by minimizing a combination of the continue (measurement) cost and misclassification cost. In analogy to Sec. 2.2, we now define a set of SHT environments in which a Bayesian stopping agent operates.

Definition 5 (Optimal SHT in multiple environments) *The set $\mathcal{M}$ of optimal SHT in multiple environments is a special case of optimal stopping in multiple environments Ξ_{opt} (7) with:*

- $\mathcal{X} = \{1, 2\}$, $\mathcal{Y} \subset \mathbb{R}$, $\mathcal{A} = \mathcal{X}$.
- $\mathcal{C} = \{c_t\}_{t\geq 0},\ c_t(x) = c \in \mathbb{R}^+,\ \forall x \in \mathcal{X}$ *is the constant continue cost.*
- $\{\mu_m, m \in \mathcal{M}\}$ *are the SHT stopping strategies chosen by the Bayesian agent over M SHT environments defined below.*
- $s_m(x, a)$ *is the stopping cost incurred by the agent in the m^{th} SHT environment parametrized by misclassification costs $(\bar{L}_{m,1}, \bar{L}_{m,2})$.*

$$s_m(x,a) = \begin{cases} \bar{L}_{m,1}, & \text{if } x = 1, a = 2, \\ \bar{L}_{m,2}, & \text{if } x = 2, a = 1, \\ 0, & \text{if } x = a \in \{1, 2\}. \end{cases}$$

The SHT stopping strategies in the above definition satisfy the optimality conditions in Definition 1 and can be computed using stochastic dynamic programming (Krishnamurthy, 2016). The solution for μ_m for the m^{th} SHT environment is well-known (Lovejoy, 1987) to be a stationary policy with the following threshold rule parameterized by scalars $\alpha_m, \beta_m \in (0, 1)$:

$$\mu_m(\pi) = \begin{cases} \text{choose action 2}, & \text{if } 0 \leq \pi(x=2) \leq \beta_m \\ \text{continue}, & \text{if } \beta_m < \pi(x=2) \leq \alpha_m \\ \text{choose action 1}, & \text{if } \alpha_m < \pi(x=2) \leq 1. \end{cases} \tag{29}$$

Remark: Since the SHT dynamics can be parameterized by $c, \bar{L}_1, \bar{L}_2$, we can set $c = 1$ without loss of generality since the optimal policy is unaffected. Also, the expected cumulative continue cost of the agent is simply the expected stopping time of the agent.

3.2 IRL for inverse SHT. Main assumptions

Suppose the inverse learner observes the actions of a Bayesian stopping agent in M SHT environments. In addition to assumptions (A2), we assume the following about the inverse learner performing IRL for identifying an SHT agent:

(A3) The inverse learner has the dataset

$$\mathcal{D}_M(SHT) = (\mathcal{D}_M, \{C_m, m \in \mathcal{M}\}), \tag{30}$$

where $\mathcal{D}_M$ is defined in (11), $C_m = \mathbb{E}_{\mu_m}\{\tau\}$ is the expected continue cost incurred by the Bayesian agent in the m^{th} environment.

(A4) The stopping strategies $\{\mu_m, m \in \mathcal{M}\}$ are stationary strategies characterized by the threshold structure in (29).

(A5) There exist reals $\delta_1,\ \delta_2 \in (0,1)$ such that the following conditions are satisfied:

$$(i)\ \beta_m \leq \delta_1 \leq \delta_2 \leq \alpha_m,\ \forall m \in \mathcal{M},\ (ii) \quad \delta_1/(1-\delta_1) \leq \bar{L}_{m,1}/\bar{L}_{m,2} \leq \delta_2/(1-\delta_2),$$

where α_m, β_m are the threshold values of the stationary strategy μ_m chosen by the Bayesian agent in environment m.

Remarks: (i) Assumption (A3) specifies additional information the inverse learner has for performing IRL for SHT by recording the agent decisions over $K \to \infty$ independent trials. Since the continue cost is 1, the expected cumulative continue cost is simply the expected stopping time of the agent. The inverse learner obtains an a.s. consistent estimate of the expected stopping time by computing the sample average of the K stopping times. Since the expected continue cost is simply the expected stopping time of the agent and known to the inverse learner, it is no more a feasible variable in the feasibility equations (17). This yields a smaller feasibility set for the stopping costs.
(ii) Assumption (A4) comprises partial information the inverse learner has about the stopping strategies chosen by the agent and its observation likelihood. Since the optimal stopping strategy is well-known to have a threshold structure Lovejoy (1987), the inverse learner only needs to compare the expected cost incurred from threshold policies to check for optimality and achieving IRL.
(iii) Assumption (A5) ensures the expected stopping cost of the SHT agent $G(\mu_m, s)$ (9) that depends on the unobserved strategy μ_m can be expressed in terms of the induced action selection policy $p_m(a|x)$ for any stopping cost s, i.e., $G(\mu_m, s_n) = \mathbb{E}_{p_m(a)}\{\mathbb{E}_{x\sim p_m(\cdot|a)}\{s(x,a)\}\}$.

3.3 IRL for inverse SHT. Main result

Our main result below specifies a set of linear inequalities that are necessary and sufficient for the Bayesian agent's actions observed by the inverse learner to be identified as that of an optimal SHT agent (Lemma 2). Any feasible solution constitutes a viable SHT misclassification cost for the M SHT environments in which the Bayesian agent operates.

Theorem 6 (IRL for inverse SHT) *Consider the inverse learner with dataset $\mathcal{D}_M(\mathrm{SHT})$ (30) obtained from a Bayesian agent taking actions in M SHT environments. Assume (A2) holds. Then:*
1. *Identifiability: The inverse learner can identify if the dataset $\mathcal{D}_M(\mathrm{SHT})$ is generated by an optimal SHT agent (Lemma 2).*
2. *Existence: There exists an optimal SHT agent parameterized by tuple Ξ_{opt} (7), if and only if there*

exists a feasible solution to the following convex (in stopping costs) inequalities:

$$\begin{aligned}
&\textit{Find } s_m(x,a) > 0, s_m(x,x) = 0,\ \forall x,a \in \mathcal{X},\ m \in \mathcal{M} \textit{ s.t.}\\
&\text{NIAS}: \sum_{x\in\mathcal{X}} p_m(x|a)(s_m(x,a) - s_m(x,b)) \leq 0, \forall a,b,m.\\
&\text{NIAC}^*: \left(\sum_{x,a} \pi_0(x) p_m(a|x) s_m(x,a) + C_m\right) - \left(\sum_a p_n(a) \min_b \sum_x p_n(x|a) s_m(x,b) + C_n\right) \leq 0,\\
&\qquad \forall m,n \in \mathcal{M}, m \neq n. \qquad (31)
\end{aligned}$$

(Recall that $C_m = \mathbb{E}_{\mu_m}\{\tau\}$ is known to the inverse learner, and hence is not a free variable).
3. *Reconstruction: The set-valued IRL estimates of the SHT misclassification costs $\{\bar{L}_m, m \in \mathcal{M}\}$ are defined below where $\bar{L}_m = (\bar{L}_{1,m}, \bar{L}_{2,m})$:*

$$\bar{L}_{1,m} = s_m(1,2),\ \bar{L}_{2,m} = s_m(2,1)\ \forall m \in \mathcal{M},$$

where $\{s_m(x,a), m \in \mathcal{M}\}$ is any feasible solution to the NIAS and NIAC* inequalities.■

Theorem 6 is a special instance of Theorem 3 for identifying an optimal stopping agent operating in multiple environments. The NIAC* resembles SUMCOST (18) with the only difference that C_m is the expected stopping time of the agent in environment m instead of being a feasible variable like in (18). We note that since the expected stopping time is non-convex in the agent's action selection policy $p_m(a|x)$, the inverse learner cannot use the convex reconstruction procedure of (19) to estimate the expected stopping time for any other policy.

Remarks:

1. Inverse SHT is an IRL task with partially specified costs: out of the continue and stopping costs, the continue cost incurred by the Bayesian agent is already known to the inverse learner. As a consequence, the feasibility test for identifying an optimal SHT agent imposes tighter restrictions (fewer feasible variables) compared to identifying optimal stopping in Theorem 3 and avoids degenerate feasible solutions that trivially satisfy the inequalities (31) of Theorem 6.

2. *IRL for Multi-state SHT.* Theorem 6 is independent of the number of states X. When $X > 2$, IRL for inverse SHT comprises estimating the misclassification costs $\{\bar{L}_{m,x,a},\ x \neq a, x, a \in X\}$, and is achieved by solving the feasibility inequalities (16) and (31) of Theorem 6.[10]

3. *Inverse SHT for boundedly-rational forward learner.* In Sec. 2.4, we discussed the concept of ϵ-optimality for a forward learner. Below, we briefly discuss how the NIAS and NIAC* feasibility inequalities of Theorem 6 can identify if an agent performs ϵ-optimal SHT when the inverse learner knows the agent's expected continue cost.

If NIAS and NIAC* (31) are feasible, then one cannot say if the dataset $\mathcal{D}_M$ (30) is generated from an absolutely optimal Bayesian agent (Definition 1) or an ϵ-optimal Bayesian agent (25). However, if $\mathcal{D}_M$ fails the feasibility test (31) of Theorem 6, then it is clear $\mathcal{D}_M$ results from an ϵ-optimal Bayesian agent, where a bound on ϵ can be obtained by finding the minimum relaxation needed for passing the feasibility test (31):

$$\min_{\epsilon_{\text{relax}} \geq 0} \epsilon_{\text{relax}},\ \text{ such that }\ \text{NIAS}(\mathcal{D}_M, \{s_m(x,a)\}) \leq \epsilon_{\text{relax}},\ \text{NIAS}(\mathcal{D}_M, \{s_m(x,a)\}) \leq \epsilon_{\text{relax}}. \quad (32)$$

10. Since the state and environment index suffice to denote the misclassification cost when $X = 2$, the subscript 'a' is dropped from the misclassification cost notation in Lemma 2 for notational clarity.

The ϵ-relaxation in (32) arises frequently in microeconomic theory in robustness tests to measure how far an economic agent is from satisfying economics-based rationality. Some examples of widely used robustness measures in economics literature include the Houtman index (HM-Index) (Houtman and Maks, 1985), Afriat measure (Afriat, 1972) and Varian measure (Varian et al., 1991).

3.4 Numerical example illustrating IRL for inverse SHT

We now present a toy numerical example for inverse SHT with 3 SHT environments and 3 states. The aim of this example is to illustrate the consistency property of Theorem 6. That is, that the true misclassification costs lie in the set of feasible costs computed by the inverse learner by solving the convex feasibility test of Theorem 6.

SHT environments. We consider $M = 3$ SHT environments with:

- *Prior* $\pi_0 = [0.5\ 0.5]'$.
- *Observation likelihood*: $p(y|x = 1) = \mathcal{N}(1, 2)$, $p(y|x = 2) = \mathcal{N}(-1, 2)$, where $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2.
- *Misclassification costs*:
  Environment 1: $(\bar{L}_{1,1}, \bar{L}_{1,2}) = (2, 2.5)$, Environment 2: $(\bar{L}_{2,1}, \bar{L}_{2,2}) = (4, 3)$, Environment 3: $(\bar{L}_{3,1}, \bar{L}_{3,2}) = (6, 6)$.

Inverse Learner specification. Next we consider the inverse learner. We generate $K = 10^5$ samples for the 3 SHT environments using the above parameters. Recall from Theorem 6 that the inverse learner uses the dataset $\mathcal{D}_M(\mathrm{SHT})$ to perform IRL for inverse SHT, where $\mathcal{D}_M(\mathrm{SHT})$ is defined as:

$$\mathcal{D}_M(\mathrm{SHT}) = (\pi_0, (\hat{p}_m(a|x), \sum_{k=1}^{K} \tau_k(\mu_m)/K), m \in \{1, 2, 3\}), \tag{33}$$

where $K = 10^5$, the second and third terms are the empirically calculated action selection policy and expected stopping time for SHT environments m from the 10^5 generated samples. We denote the action selection policy in (33) as $\hat{p}_m(a|x)$ and not $p_m(a|x)$ since the numerical example uses an empirical estimate.

IRL Result. The inverse learner performs IRL by using the dataset $\mathcal{D}_M(\mathrm{SHT})$ (33) to solve the linear feasibility problem in Theorem 6. The result of the feasibility test is shown in Fig. 3. The blue region is the set of feasible misclassification costs for each SHT environment. The feasible set of costs is $\{(\bar{L}_{m,1}, \bar{L}_{m,2}), m \in \{1, 2, 3\}\} \subseteq \mathbb{R}^6_+$. Fig. 3 displays the feasible misclassification costs for a single environment keeping the costs for the other two environments fixed at their true values. The need to fix costs for the other two environments for plotting the set of feasible costs is only for visualization purposes. It is not possible to plot a 6 dimensional point (vector of estimated misclassification costs for 3 SHT environments) on the 2-d plane.

The true misclassification costs for each SHT environment are highlighted by a yellow point. The key observation is that these true costs belong to the set of feasible costs (blue region) computed via Theorem 6. Thus, Theorem 6 successfully performs IRL for the SHT problem and the set of feasible misclassification costs can be reconstructed as the solution to a linear feasibility problem. Also, all points in the set of misclassification costs explain the SHT dataset equally well.

3.5 Numerical example. Regularized max-margin IRL for inverse SHT.

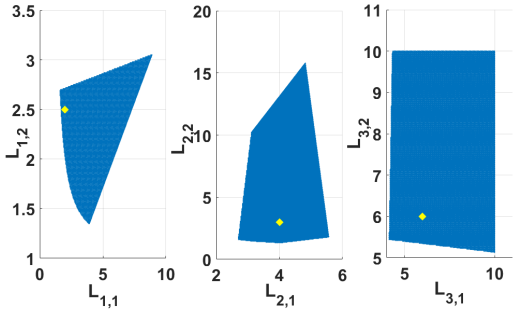

Figure 3: Inverse SHT numerical example with parameters specified in Sec. 3.4. The key observation is that the true misclassification costs (yellow points) lie in the feasible set (blue region) of costs computed via Theorem 6. This follows from the necessity proof of Theorem 6 which says if the Bayesian agent is an optimal SHT agent, then the true costs lie in the feasible set of costs that satisfy the NIAS and NIAC* inequalities. Highlighting the advantage of the set-valued estimate of our IRL algorithm, we note that all points in the blue feasible region rationalize the observed stop actions of the Bayesian agent equally well. Indeed, the feasible region shrinks with the number of environments M.

We now present a numerical example for inverse SHT involving $M = 100$ environments where we compute a point-valued IRL estimate of the SHT misclassification costs. This inference task is in contrast to the set-valued IRL flavor considered thus far in the paper. Given dataset $\mathcal{D}_M(\text{SHT})$ (33), we compute a point estimate $\bar{L}^*$ of misclassification costs that maximizes the $\mathcal{L}_2$-regularized margin of the NIAC* feasibility inequalities of Theorem 6. The point estimate $\bar{L}^*$ is inspired by max-margin IRL methods in the literature (Abbeel and Ng, 2004; Ratliff et al., 2006) and defined as:

$$\bar{L}^* = \underset{\bar{L}}{\text{argmin}} \sum_{m,n=1,m\neq n}^{M} \text{Margin}_{\mathcal{D}_M(\text{SHT})}(m,n,\bar{L}) - \lambda\|\bar{L}\|_2^2, \tag{34}$$

$$\text{Margin}_{\mathcal{D}_M(\text{SHT})}(m,n,\bar{L}) = \left(G(\hat{p}_n,\bar{L}_m) + \hat{C_n}\right) - \left(G(\hat{p}_m,\bar{L}_m) + \hat{C_m}\right), \tag{35}$$

where $G(\hat{p},\bar{L}_m)$ is the expected misclassification cost for SHT with action selection policy $\hat{p}$ and misclassification costs $\bar{L}_m$, and $\hat{C}_m = \sum_{k=1}^{K} \tau_k(\mu_m)/K)$ is the agent's expected continue cost in environment m computed empirically from K independent trials. In simple terms, (35) is the difference in expected cumulative cost between action policies $\hat{p}_m$ and $\hat{p}_n$ for a fixed misclassification cost $\bar{L}_m$. The objective function in (34) is the $\mathcal{L}_2$-norm regularized margin with which the *candidate* SHT misclassification costs pass the NIAC* convex feasibility test of (31). In (34), $\lambda > 0$ is a tunable regularization parameter and $G(\cdot)$ is the expected misclassification cost defined in (12). Setting λ to 0 yields the max-margin IRL estimate of the stopping agent's misclassification costs and lies within the feasible set of costs generated by Theorem 6. The other extreme is setting λ to ∞ which results in $\bar{L}^* = 0$.

The following numerical example illustrates regularized IRL (34) for inverse SHT.

SHT environments. We consider $M = 100$ SHT environments with:

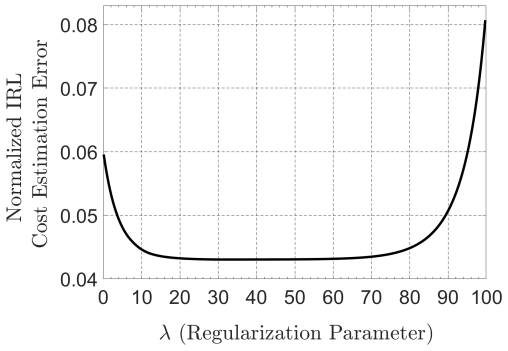

Figure 4: Inverse SHT numerical example for 100 SHT environments, with parameters specified in Sec. 3.5. The main takeaway is that regularized max-margin IRL for inverse SHT (34) can estimate the misclassification costs incurred by the stopping agent in the 100 SHT environments with up to 95% accuracy by varying the regularization parameter λ in (34).

- *Prior* $\pi_0 = [1/4 \;\; 1/4 \;\; 1/4 \;\; 1/4]'$ (The state space is now $\mathcal{X} = \{1, 2, 3, 4\}$).
- *Observation likelihood*: $p(y|x = 1) = \mathcal{N}(-2, 8),\; p(y|x = 2) = \mathcal{N}(0, 8)$, , $p(y|x = 3) = \mathcal{N}(2, 8)$ and $p(y|x = 4) = \mathcal{N}(4, 8)$.
- *Misclassification costs*: The misclassification costs $\bar{L} = \{\bar{L}_{m,x,a}\}$ in the M environments is uniformly sampled from the interval $[4, 10]^{M\times X\times(X-1)}$.

Inverse Learner Specification: The inverse learner aggregates the dataset $\mathcal{D}_M(\mathrm{SHT})$ according to the procedure described in (33) by generating $K = 10^7$ independent trials for the SHT agent in all $M = 100$ environments. Then, the inverse learner computes the regularized max-margin IRL estimate $\bar{L}^*$ by solving the optimization problem (34).

IRL Results: The inverse learner performs IRL by using the dataset $\mathcal{D}_M(\mathrm{SHT})$ to solve the optimization problem (34). Recall the dataset $\mathcal{D}_M(\mathrm{SHT})$ is generated by observing the actions of an SHT agent in multiple environments with misclassification costs $\bar{L}$. Figure 4 shows the estimation error $\|\bar{L}^* - \bar{L}\|_2/\|\bar{L}\|\|_2$ of the inverse learner's IRL estimate $\bar{L}^*$ (34) computed by the inverse learner as the regularization parameter λ in (34) is varied. The error is normalized wrt the $\mathcal{L}_2$-norm of the true misclassification costs in multiple environments incurred by the Bayesian agent whose actions comprise $\mathcal{D}_M(\mathrm{SHT})$.

The least estimation error obtained by varying λ over the interval $[0, 100]$ was observed to be 0.042. In other words, the point IRL estimate obtained by solving the optimization problem (34) can estimate the true misclassification costs of the SHT environments with up to 95% accuracy. Indeed, the estimation accuracy increases with the number of environments at the cost of greater computation resources. Second, we observed that the error starts increasing sharply from $\lambda \sim 75$. This is expected since the regularization term in (34) dominates the margin term at large values of λ.

3.6 Performance Comparison. IRL for Inverse SHT and existing IRL methods for POMDPs

In this section, we compare the IRL performance of Theorem 6 for inverse SHT against two well-known algorithms for IRL of POMDPs, namely, Max-Margin between Values (MMV) (Choi and Kim, 2011, Alg. 4) and Max-Margin between Feature Expectations (MMFE) (Choi and Kim, 2011,

Alg. 5). We compare the performance of MMV and MMFE algorithms against max-margin inverse SHT (34) with regularization parameter λ set to 0.

Recall from (30) that our inverse SHT result of Theorem 6 requires state-terminal action pairs of the SHT agent over several independent trials and the expected stopping time of the SHT agent. In comparison, MMV and MMFE do not require the expected stopping time, but instead require complete knowledge of: (a) the observation likelihood of the Bayesian agent, and (b) the beliefs of the SHT agent at every time step. Moreover, MMV and MMFE require a POMDP solver for IRL.

To compare the performance of our IRL scheme (34) against MMV and MMFE, we perform two sets of numerical experiments with different specifications of the agent's observation likelihood:
Case 1: Perfect Knowledge of SHT Model Dynamics. MMV and MMFE have perfect knowledge of the SHT agent's observation likelihood.
Case 2: Misspecified SHT Model Dynamics. MMV and MMFE have misspecified knowledge of the SHT agent's observation likelihood. For environment m the observation likelihood $p_m(y|x)$ is misspecified to be the agent's action policy $p_m(a|x)$.

EXPERIMENTAL SETUP

For our numerical experiments, we consider $M = 4$ SHT environments with:

- *Prior* $\pi_0 = [1/2 \;\; 1/2]'$ (The state space is $\mathcal{X} = \{1, 2\}$).
- *Observation likelihood*: $p(y|x = 1) = \mathcal{N}(+2, 4),\ p(y|x = 2) = \mathcal{N}(-2, 4)$,
- *Misclassification costs*: The misclassification costs $\bar{L} = \{\bar{L}_m, m \in \mathcal{M}\}$ in the M environments are uniformly sampled from the interval $[5, 25]$ for all states and actions in $\mathcal{X}$. Recall that we assume the continue cost is set to 1 WLOG.

For every environment $m = 1, 2, 3, 4$, we computed $\bar{L}_{m,\mathrm{MMV}},\ \bar{L}_{m,\mathrm{MMFE}}$ and $\bar{L}_{m,\mathrm{Margin}}$, the point-valued IRL estimate of the agent's misclassification cost from MMV, MMFE and max-margin inverse SHT (defined in (34) with regularization parameter $\lambda = 0$), respectively. For estimated misclassification cost $\bar{L}_{m,\mathrm{est}} \in \{\bar{L}_{m,\mathrm{MMV}},\ \bar{L}_{m,\mathrm{MMFE}},\ \bar{L}_{m,\mathrm{Margin}}\}$ with true cost $\bar{L}_m$ and chosen stopping strategy μ_m (Lemma 2), the normalized IRL estimation error is defined as:

$$\text{IRL Estimation Error} = \frac{|J(\mu_m, \bar{L}_m) - J(\mu_m, \bar{L}_{m,\mathrm{est}})|}{J(\mu_m, \bar{L}_m)}, \tag{36}$$

where $J(\cdot)$ is the expected cumulative cost defined in (9).

Our experimental results are displayed in Fig. 5. Our results show that our proposed IRL algorithm yields a lower IRL estimation error (36) than MMV and MMFE algorithms when model dynamics are misspecified. We observe that, on average, our max-margin IRL algorithm yields 60% lower estimation error compared to MMV and MMFE algorithms with misspecified model dynamics, and yields 27% higher estimation error compared to MMV and MMFE algorithms with accurate model dynamics.

KEY FINDINGS

Our key findings from the numerical experiments[11] can be summarized as:

- For the case of perfect knowledge of model dynamics (case 1), we observed that the MMV and MMFE algorithms of Choi and Kim (2011) perform better than max-margin IRL (34), and

11. All our numerical results are completely reproducible and can be accessed from the GitHub repository `https://github.com/KunalP117/YouTube-Commenting-Analysis`

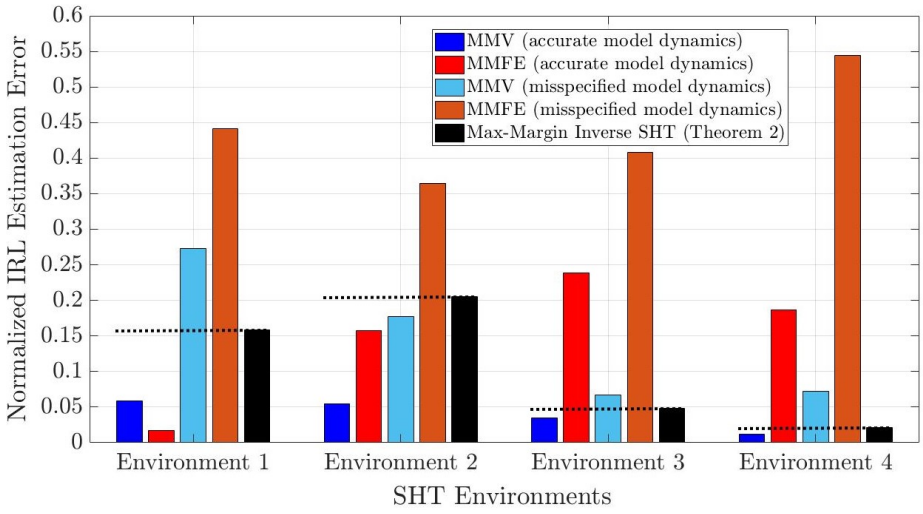

Figure 5: Inverse SHT Performance Comparison. Max-Margin NIAS-NIAC Test of Theorem 6 versus MMV and MMFE (Choi and Kim, 2011)

yield approximately $\mathbf{27}\%$ lower IRL estimation error compared to max-margin IRL. This is expected since both MMV and MMFE have access to private information the forward learner uses for decision-making and hence generates a more accurate IRL estimate.

- When the model dynamics are misspecified (case 2), our max-margin IRL algorithm outperforms both MMV and MMFE algorithms and yields approximately $\mathbf{60}\%$ lower IRL estimation error compared to MMV and MMFE.

Indeed, when no assumptions are placed on the underlying POMDP structure like in Choi and Kim (2011), achieving IRL requires perfect knowledge of the model dynamics. Hence, MMV and MMFE fail when model dynamics are misspecified.

PERSPECTIVE

Cases 1 and 2 highlight the fact that our approach is complementary to that of Choi and Kim (2011). Choi and Kim (2011) achieve IRL where the model dynamics are perfectly specified (case 1). In comparison, our IRL methods yield necessary and sufficient methods for optimal Bayesian stopping when no knowledge of model dynamics is provided to the inverse learner.

3.7 Summary

Theorem 6 specified necessary and sufficient conditions for identifying an optimal SHT agent acting in multiple environments. These conditions constitute a linear feasibility program that the inverse learner can solve to estimate SHT misclassification costs of the environments. The IRL task of solving the inverse SHT problem is more structured than the inverse optimal stopping problem in Sec. 2, since the agent's costs are partially known (expected continue cost is known) to the inverse learner. Hence, the feasible set of costs generated using Theorem 6 is smaller than that generated by Theorem 3 for the inverse SHT problem. We also proposed an IRL algorithm for point-valued estimation of the environments' misclassification costs and illustrated its performance in Sec. 3.5.

Our key finding is that this point-valued IRL algorithm reconstructs the misclassification costs with up to 95% accuracy. Recall from Sec. A.1 that an online user in multimedia platforms can be viewed as a Bayesian agent performing SHT. In the context of online multimedia platforms, the continue and stopping cost of the SHT agent can be viewed as the online user's sensing cost (attention to visual cues) and preference for viewing the online content, respectively. Hence, the numerical example in Sec. 3.5 can be viewed as an IRL methodology to reconstruct an online user's preferences for advertisements/movie thumbnails by observing his/her actions in multiple environments (webpages). We illustrate this claim in Sec. 5 with an IRL analysis on a real-world dataset. Finally, in Sec. 3.6 we compared our inverse SHT algorithm to two existing algorithms in the literature for IRL for POMDPs, namely, MMV and MMFE Choi and Kim (2011). Our key observation was that our inverse SHT algorithm outperforms MMV and MMFE in scenarios where the inverse learner has limited information about the forward learner, i.e. , the learner's model dynamics are misspecified.

4. Example 2. IRL for inverse search

In this section, we present a second example of IRL for an optimal Bayesian stopping time problem, namely, *inverse* Bayesian Search. In the search problem, a Bayesian agent sequentially searches over a set of target locations until a static (non-moving) target is found. The optimal search problem is a special case of a Bayesian multi-armed bandit problem, and also of the optimal Bayesian stopping problem discussed in Sec. 2.2 since the continue cost (2) is the cost of searching a location and the stopping cost is 0 in the Bayesian search problem. Our IRL task in this section will be to estimate the search costs.

The optimal search problem is a modification of the sequential stopping problem in Sec. 2 with the following changes:

- There is only 1 stop action but multiple continue actions, namely, which of the X locations to search at each time. We will call the continue actions as search actions, or simply, actions.
- The observation likelihood B depends both on the true state x^o and the continue action a.

Suppose an inverse learner observes the decisions of a Bayesian search agent over M search environments. The aim of the inverse search problem is to identify if the search actions of the agent are optimal and if so, estimate their search costs. Our IRL result for Bayesian search (Theorem 9 below) gives a necessary and sufficient condition for identifying an optimal search agent (formalized in Lemma 8 below) as equivalent to the existence of a feasible solution to a set of linear inequalities.

4.1 Optimal Bayesian search agent in multiple search environments

Suppose an agent searches for a target location $x \in \mathcal{X}$. When the agent chooses action $a \in \mathcal{X}$ to search location a, it obtains an observation y. Assume the agent knows the set of conditional pmfs of y, namely, $\{p(y|x^o = x), x \in \{1, 2, \dots X\}\}$. The aim of optimal search is to decide sequentially which location to search at each time to minimize the cumulative search cost until the target is found.

We define an optimal Bayesian search agent in M search environments as

$$\Xi_{opt} = (\mathcal{X}, \pi_0, \mathcal{Y}, \mathcal{A}, \boldsymbol{\alpha}, \{l_m, \mu_m, m \in \mathcal{M}\}) \tag{37}$$

where

- $\mathcal{X} = \{1, 2, \dots X\}$ is a finite set of states (target locations).
- At time 0, the true state $x^o \in \mathcal{X}$ is sampled from prior pmf π_0. This location x is not known to the agent but is known to the inverse learner (performing IRL).

- $\mathcal{Y} = \{0, 1\}$, where $y = 1$ (found) and $y = 0$ (not found) after searching a location.
- The set of actions $\mathcal{A} = \mathcal{X}$, $a \in \mathcal{A}$ is the location searched by the agent.
- The Bayesian agent in search environments m incurs instantaneous cost $l_m(a) > 0$ for searching location a.
- $\boldsymbol{\alpha} = \{\alpha(a), a \in \mathcal{A}\}$, $\alpha(a)$ is the reveal probability for location a, i.e. , the probability that the target is found when the agent searches the target location ($x = a$) in search environment $m \in \mathcal{M}$. $\boldsymbol{\alpha}$ characterizes the action dependent observation likelihood $B(y, x, a)$.

$$B(y, x, a) = p(y|x, a) = \begin{cases} \alpha(a), & y = 1, x = a \\ 1 - \alpha(a), & y = 0, x = a \\ 1, & y = 0, x \neq a. \end{cases} \tag{38}$$

  For IRL identifiability, we assume that the reveal probabilities are the same for all search environments in $\mathcal{M}$.
- $\{\mu_m, m \in \mathcal{M}\}$ are the optimal search strategies of the Bayesian agent over all environments in $\mathcal{M}$, when the agent operates sequentially on a sequence of observations $y_1, y_2, \ldots$ as discussed below in Protocol 2.

Protocol 2 *Sequential Decision-making protocol for Search:*

1. *Generate $x^o \sim \pi_0$ at time $t = 0$.*
2. *At time $t \geq 1$, agent records observation $y_t \sim B(\cdot, a_{t-1}, x^o)$.*
3. *If $y_t = 1$, then stop. Otherwise, if $y_t = 0$:*
   (i) Update belief $\pi_{t-1} \rightarrow \pi_t$ (described below).
   (ii) For search policy μ, agent takes action $a_t = \mu(\pi_t)$. (Note the first action is taken at time $t = 0$, while the first observation is at $t = 1$).
   (iii) Set $t = t + 1$ and go to Step 2.

Belief Update: *Let $\mathcal{F}_t$ denote the sigma-algebra generated by the action and observation sequence $\{a_1, y_1, \ldots a_t, y_t\}$. The agent updates its belief $\pi_t = \mathbb{P}(x^o = x|\mathcal{F}_t),\ x \in \mathcal{X}$ using Bayes formula as*

$$\pi_t = \frac{B(y_t, a_{t-1})\pi_{t-1}}{\mathbf{1}' B(y_t, a_{t-1})\pi_{t-1}}, \tag{39}$$

where $B(y, a) = \mathrm{diag}(\{B(y, x, a), x \in \mathcal{X}\})$. The belief π_t is an $X-$dimensional probability vector belonging to the $(X - 1)$ dimensional unit simplex (4).

Remark: The search agent's stopping region is simply the set of distinct vertices of the $X - 1$ dimensional unit simplex.

We define the random variable τ as the time when the agent stops (target is found).

$$\tau = \inf\ \{t > 0|\ y_t = 1\} \tag{40}$$

Clearly, the set $\{\tau = t\}$ is measurable wrt $\mathcal{F}_t$, hence, the random variable τ is adapted to the filtration $\{\mathcal{F}_t\}_{t \geq 0}$. Below, we define the optimal search strategies $\{\mu_m, m \in \mathcal{M}\}$.

Definition 7 (Search strategy optimality) *The optimal search strategy μ_m of the Bayesian agent operating according to Protocol 2 in environment $m \in \mathcal{M}$ that minimizes the agent's cumulative expected search cost is well known (Krishnamurthy, 2016) to be a stationary policy as defined below:*

$$J(\mu_m, l_m) = \min_{\mu} J(\mu, l_m) = \mathbb{E}_{\mu}\left\{\sum_{t=0}^{\tau-1} l_m(\mu(\pi_t))\right\}, \tag{41}$$

$$\mu_m(\pi) = \operatorname*{argmax}_{a\in\mathcal{A}}\left(\frac{\pi(a)\alpha}{l_m(a)}\right). \tag{42}$$

Here, $\mathbb{E}_{\mu}\{\cdot\}$ denotes expectation parametrized by μ induced by the probability measure $\{a_t, y_{t+1}\}_{t=1}^{\tau-1}$, $J(\cdot)$ denotes the expected search cost and μ belongs to the class of stationary search strategies.

Remarks. (1) Note that the minimization in (41) is over stationary search strategies. It is well known that the optimal search strategy has a threshold structure (Krishnamurthy, 2016). Since the set of all threshold strategies forms a compact set, we can replace the '$\inf$' in (9) for generic optimal stopping problems by '$\min$' in (41).
(2) Since the expected cumulative cost of an agent depends only on the search costs (for constant reveal probabilities), we can set $l_m(1) = 1,\ \forall m \in \mathcal{M}$ WLOG.

4.2 IRL for inverse search. Main result

In this subsection, we provide an inverse learner-centric view of the Bayesian stopping time problem and the main IRL result for inverse search. Suppose the inverse learner observes a search agent taking actions over M search environments where the agent performs several independent trials of Protocol 2 for Bayesian sequential search in each environment. We make the following assumptions about the inverse learner performing IRL to identify if $\mathcal{M}$ comprises an optimal search agent.

(A6) The inverse learner knows the dataset

$$\mathcal{D}_M(\text{Search}) = (\pi_0, \{g_m(a, x), m \in \mathcal{M}\}). \tag{43}$$

Here, $g_m(a, x)$ is the average number of times the agent searches location a when the target is in x in environment m:

$$g_m(a, x) = \mathbb{E}_{\mu_m}\left\{\sum_{t=1}^{\tau} \mathbb{1}\{\mu_m(\pi_t) = a\}|x\right\}. \tag{44}$$

We call $g_m(a, x)$ as the agent's *search action policy* in search environment m.

(A7) In dataset $\mathcal{D}_M(\text{Search})$, there are at least $M \geq 2$ environments with distinct search costs.

Assumption (A6) is discussed after the main result. In complete analogy to (A2), assumption (A7) is needed for identifiability of the search costs. We emphasize that the inverse learner only has the average number of times the agent searches a particular location in any environment. The inverse learner does not know the stopping time or the order in which the agent search the locations. In completely analogy to Lemma 2, Lemma 8 below specifies the inverse learner's identifiability of an optimal search agent under assumptions (A6) and (A7):

Lemma 8 (IRL identifiability of optimal Bayesian search agent) *The inverse learner identifies the tuple* Ξ_{opt} (37) *as an optimal Bayesian search agent iff* (45) *holds.*

$$J(\mu_m, l_m(a)) \leq J(\mu_n, l_m(a)),\ \forall\, m, n \in \mathcal{M},\ m \neq n. \tag{45}$$

In complete analogy with (12) in Lemma 2 for identifying an optimal stopping time agent, $J(\cdot)$ *in the above equation is the expected cumulative search cost of the agent.*

We omit the proof of Lemma 8 since it is identical to that of Lemma 2. Eq. 45 in Lemma 8 is analogous to (12) in Lemma 2. The inverse learner simply checks if the expected cumulative search cost for environment m is the smallest possible given the finite strategies $\{\mu_m, m \in \mathcal{M}\}$. We are now ready to present our main IRL result for the inverse search problem. The result specifies a set of linear inequalities that are simultaneously necessary and sufficient for a search agent's actions in multiple environments $\mathcal{M}$ to be identified as that of an optimal search agent (45).

Theorem 9 (IRL for inverse Bayesian search) *Consider the inverse learner with dataset* $\mathcal{D}_M(\text{Search})$ *(43) obtained from a search agent acting in multiple environments* $\mathcal{M}$*. Assume (A6) holds. Then:*
1. *Identifiability: The inverse learner can identify if the dataset* $\mathcal{D}_M(\text{Search})$ *is generated by an optimal search agent (Definition 8).*
2. *Existence: There exists an optimal search agent parameterized by tuple* Ξ_{opt} (37) *if and only if there exists a feasible solution to the following linear (in search costs) inequalities:*

$$\begin{aligned} &\textit{Find } l_m(a) \in \mathbb{R}_+, l_m(1) = 1 \quad \textit{s.t.} \quad \text{NIAC}^\dagger(\mathcal{D}_M(\text{Search})) \leq 0, \textit{ where} \\ &\text{NIAC}^\dagger:\ \sum_{x \in \mathcal{X}} \pi_0(x)(g_m(a,x) - g_n(a,x))\, l_m(a) < 0\ \forall m, n \in \mathcal{M},\ m \neq n. \end{aligned} \tag{46}$$

3. *Reconstruction: The set-valued IRL estimate of the agent's search costs in environments* $\mathcal{M}$ *is the set of all feasible solutions to the NIAC*† *inequalities.* ■

The proof of Theorem 9 is in Appendix D. Theorem 9 provides a set of linear inequalities whose feasibility is equivalent to identifying the optimality of a Bayesian search agent in multiple environments with different search costs. Note that Theorem 9 uses the search action policies $\{p_m(a|x), m \in \mathcal{M}\}$ to construct the expected cumulative search costs of the agent in multiple environments and verify if the inequality for identifying optimality (45) for Bayesian search holds. The key idea for the IRL result is to express the expected cost of the search agent in environment m in terms of its chosen search action policy $g_m(a, x)$ (43). Algorithms for linear feasibility such as the simplex method (Boyd and Vandenberghe, 2004) can be used to check feasibility of (46) in Theorem 9 and construct a feasible set of search costs for the optimal search agent.
Discussion of assumption (A6). To motivate (A6), suppose for each environment $m \in \mathcal{M}$, the inverse learner records the state $x_{k,m}$ and agent actions $\{a_{1:\tau_{k,m},k,m}\}$ over $k = 1, 2, \ldots K$ independent trials. Then, the variable $g_m(a, x)$ in $\mathcal{D}_M(\text{Search})$ (43) is the limit pmf of the empirical pmf $\hat{g}_m(a, x)$ as the number of trials $K \to \infty$.

$$\hat{g}_m(a, x) = \frac{\sum_{k=1}^{K} \sum_{t=1}^{\tau_{k,m}} \mathbb{1}\{x_{k,m} = x, a_{t,k,m} = a\}}{\sum_{k=1}^{K} \mathbb{1}\{x_{k,m} = x\}}. \tag{47}$$

In complete analogy to Sec. 2.5, almost sure convergence holds by Kolmogorov's strong law of large numbers. $g_m(a, x)$ is the average number of times the agent searches location a when the target is in

location x in environment m. More formally, for a fixed state x, $g_m(a, x)$ is the number of times the posterior belief of the agent visits the region in the unit simplex of pmfs where it is optimal to choose action a. In Appendix D, we discuss how the search action policy $g_m(a, x)$ can be used to express the agent's cumulative expected search cost (41) in the m^{th} environment.

Remark: Analogous to the action selection policy (28) for stopping problems with multiple stopping actions, the inverse learner uses the search action policy to identify Bayes optimality in stopping problems with multiple continue actions (and single stop action).

4.3 Numerical example illustrating IRL for inverse search

We now present a numerical example for inverse search with 3 search environments and 3 search locations. The aim of this example is to illustrate the consistency property of Theorem 9. That is, that the true search costs lie in the set of feasible costs computed by the inverse learner by solving the feasibility test of Theorem 9.

Search environments. We consider $M = 3$ search environments with:

- *Prior* $\pi_0 = [1/3\ 1/3\ 1/3]'$.
- *Search locations:* $X = A = 3$.
- *Reveal probability*: $\alpha(1) = 0.7, \alpha(2) = 0.68, \alpha(3) = 0.6$.
- *Search costs*:
  Environment 1: $l_1(1) = 1,\ l_1(2) = 3,\ l_1(3) = 4$,
  Environment 2: $l_2(1) = 1,\ l_2(2) = 1,\ l_2(3) = 2$,
  Environment 3: $l_3(1) = 1,\ l_3(2) = 0.5,\ l_3(3) = 3$.

(Recall that WLOG the search cost $l_m(1)$ can be set to 1 for all $m \in \{1, 2, 3\}$.)

Inverse Learner specification. Next we consider the inverse learner. We generate $K = 10^6$ samples for the search agent in all 3 environments using the above parameters. Recall from Theorem 9 that the inverse learner uses the dataset $\mathcal{D}_M(\text{Search})$ to perform IRL for search. Here

$$\mathcal{D}_M(\text{Search}) = (\pi_0, (\hat{g}_m(a, x), m \in \{1, 2, 3\}), \tag{48}$$

where $K = 10^6$, the second term in the dataset is the empirically calculated search action policy (47) of the agent in environment m from the 10^6 generated samples.

IRL Result. The inverse learner performs IRL by using the dataset $\mathcal{D}_M(\text{Search})$ (48) to solve the linear feasibility problem in Theorem 9. The result of the feasibility test is shown in Fig. 6. The blue region is the set of feasible search costs for each environment. The feasible set of costs is $\{(l_m(2), l_m(3), m \in \{1, 2, 3\}\} \subseteq \mathbb{R}^6_+$. For visualization purposes, Fig. 6 displays the feasible search costs for each environment in a different sub-figure. In complete analogy to Fig. 3, the feasible search costs for each environment are shown in each sub-figure by keeping the search costs of the other 2 environments fixed at their true values. The true search cost for every environment is highlighted by a yellow point. The key observation is that these true costs belong to the set of feasible costs (blue region) computed via Theorem 9. Thus, Theorem 9 successfully performs IRL for the search problem and the set of feasible search costs can be reconstructed as the solution to a linear feasibility problem.

5. Inverse Optimal Stopping for Predicting YouTube Commenting Behavior

In this section, we illustrate our IRL results for Bayesian stopping time problems on a real-world YouTube dataset. Although we use the same dataset in previous work (Hoiles et al., 2020), our IRL

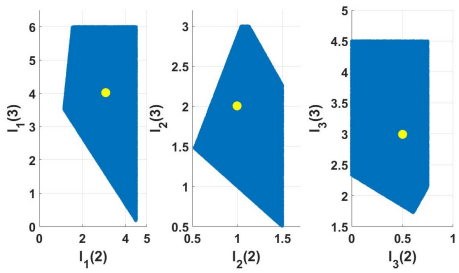

Figure 6: Numerical example for inverse search with parameters specified in Sec. 4.3. The key observation is that the true search costs (yellow points) lie in the feasible set (blue region) of costs computed via Theorem 9. This follows from the necessity proof of Theorem 9 which says if the agent is an optimal search agent, then the true costs lie in the feasible set of costs that satisfy the NIAC† inequalities.

methodology and experimental results are new. For brevity, we discuss the key differences compared to Hoiles et al. (2020) and justify our choice of Bayesian stopping for modeling user engagement on YouTube in Appendix G.

We consider a YouTube dataset comprising approximately 140000 videos across $25,000$ channels spanning 18 video categories and over 9 millions users from April 2007 to May 2015. The diversity of videos in YouTube is immense; it is intuitive to exploit this diversity for understanding how groups of YouTube users exposed to different classes of video content engage differently with YouTube. Hence, by analyzing groups of YouTube users indexed by video category, our aim is to:
(1) Identify if YouTube user engagement is consistent with Bayesian optimal stopping, and if so,
(2) Reconstruct the stopping costs of user engagement using the IRL results in this paper, and
(3) Use the reconstructed costs to predict user engagement in videos.

Our YouTube dataset does not contain any information (visual cues) about what the human user perceives from the video webpage before choosing to engage on the YouTube platform. Recall from Theorem 3 that our IRL approach does not depend on the unobserved model dynamics that generate the IRL dataset (11). This makes our IRL methodology well-suited to scenarios where the parameters of the underlying decision making process are not available in the IRL dataset. Our main conclusions from our IRL analysis of the YouTube dataset can be summarized as:

- YouTube user engagement is consistent with optimal Bayesian stopping. Based on our IRL analysis on groups of YouTube users, where each group consists of approximately 3500 viewers, the YouTube dataset (described below in (49)) satisfies the NIAS and NIAC feasibility inequalities of Theorem 3 for optimal Bayesian stopping with a high margin.
- By choosing two representative points from the feasible set of costs generated by IRL (16), (17), namely, max-margin estimate and entropy-regularized estimate defined below, we show our reconstructed IRL costs predict user engagement with high accuracy. Figure 8 illustrates the predictive performance of our IRL methodology.

5.1 YouTube Dataset and Model Parameters

Categories in YouTube (e.g. News, Gaming, Music etc.) are numbered from $1-18$ (See Fig. 7 for the full listing). The video categories have mean numbers of users ranging from 149 to 4596

for high viewcount (greater than 10000) videos and 8 to 1801 for low viewcount videos (less than 10000). Figure 7 lists each video category along with the total number of views. Note that the video categories "Unavailable" or "Removed" are videos flagged by YouTube as being suspected of violating YouTube's video policies.

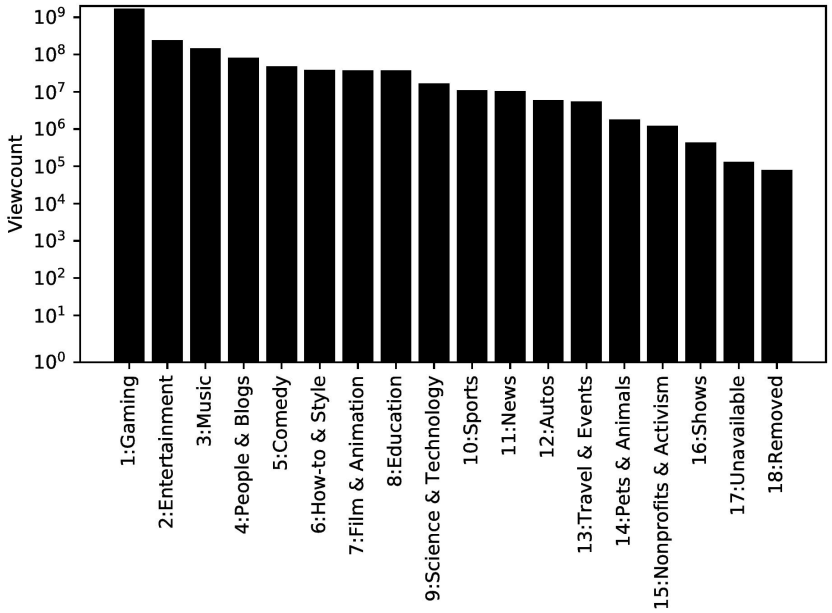

Figure 7: YouTube Dataset Overview. Viewcount summed over all videos (vertical axis) of $M = 18$ video categories. The 18 categories are listed on the horizontal axis.

The YouTube dataset contains the view counts, comment counts, likes, dislikes, thumbnail, title, and category of each video. To relate to our main IRL result of Theorem 3, we define the following:
1. *Agent*: Group of users interacting with videos in each video segment. User engagement in different video categories can be interpreted as the agent acting in multiple environments. In the rest of the section, we will use the terms 'user engagement' and 'commenting behavior' interchangeably.
2. *State (x)*: In the YouTube dataset, the state x of each video is the viewcount 1 day after the video was published. Specifically, state $x = 1$ is high viewcount (more than $10,000$ views) and $x = 2$ otherwise. In YouTube, video viewcount is the independent quantity which governs the commenting behavior since videos need to be viewed first before users can comment or rate the video.
3. *Terminal Action (a)*: In the YouTube dataset, the terminal action a is related to the overall commenting behavior[12] of the users, which is computed using the comment counts, like count, and dislike count 2 days after the video is published. The possible actions are: $a = 1$ denotes low comment count with negative sentiment, $a = 2$ denotes low comment count with neutral sentiment, $a = 3$ denotes low comment count with positive sentiment, $a = 4$ denotes high comment count with negative sentiment, $a = 5$ denotes high comment count with neutral sentiment, and $a = 6$ denotes high comment count with positive sentiment. Here negative sentiment occurs if the difference between the like count and dislike count is less than -25, neutral sentiment occurs if the difference

12. By overall commenting behavior in YouTube, we mean both the comment count and the video ratings (likes and dislikes). Another term used in the literature (Khan, 2017) is "user engagement".

lies between $-25, 25$, and positive sentiment occurs if the difference is greater than 25. A low comment count is said to occur if there are less than 100 comments, otherwise the comment count is defined to be high.

4. *Observation (y)*: The observation y for a YouTube user abstracts the visual cues a user perceives that depends on video metadata such as thumbnail, title, category etc. The observation likelihood is indicative of the attention expended by the user on a video. We note that although neither the observations y nor the observation likelihood $p(y|x)$ are contained in the YouTube dataset, our IRL algorithm abstracts away these unobserved model parameters, and still yields necessary and sufficient conditions for Bayes optimality.

4. *Environment (m)*: Environment m corresponds to each of the $M = 18$ video categories in our YouTube dataset. Fig. 7 lists each video category with the total number of views. Note that the video categories "Unavailable" or "Removed" are videos flagged by YouTube as being suspected of violating YouTube's video policies[13].

Recall from Sec. 2.3 that the inverse learner requires knowledge of the dataset $\mathcal{D}_M = (\pi_0, \boldsymbol{p})$ (11) for identifying optimal Bayesian stopping via Theorem 3. In the YouTube context, the variables $\pi_0, \boldsymbol{p} = \{p_m(a|x), m \in \mathcal{M}\}$ dataset $\mathcal{D}_M$ can be constructed as:

$$\pi_0(x) = \frac{1}{I}\sum_{i=1}^{I} \mathbb{1}\{x_i = x\}, \;\; p_m(a|x) = \frac{\sum_{i=1}^{I} \mathbb{1}\{x_i = x, a_i = a, \text{category}_i = m\}}{\sum_{i=1}^{I} \mathbb{1}\{x_i = x, \text{category}_i = m\}}, \tag{49}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function, variable i indexes the YouTube videos, $I = 140000$ is the total number of YouTube videos in the dataset, and environment $m \in \{1, 2, \ldots, 18\}$ indexes the video categories. Also, $x_i, a_i, \text{category}_i$ denote the state, action and category of the YouTube video indexed by i, where the state and action interpretations for the YouTube videos are discussed above.

5.2 YouTube Data Analysis Results

We now discuss our experimental findings from our IRL analysis on the YouTube dataset.[14] Our main task is to predict YouTube's commenting behavior, that is, the action selection policy $p_m(a|x)$ in video category m using the IRL algorithms in this paper. Our first observation is that the dataset $\mathcal{D}_M$ (49) comprising YouTube commenting behavior over $M = 18$ categories passes the convex feasibility test (16) and (17) of Theorem 3 with a high margin of 1.85×10^{-3}, where the margin is normalized by the maximum feasible cost $\max_{m,x,a} s_m(x, a)$. This shows that there exists a Bayesian stopping model that rationalizes YouTube commenting behavior.

We now illustrate how well the reconstructed costs from the feasibility test of Theorem 3 predict the commenting behavior of YouTube videos in different categories. For our prediction task, first, we randomly divided the YouTube dataset into two parts - training data (80%) and testing data (20%). Also, we consider only a subset of the 18 video categories for which the number of videos exceeds 200. This extra condition results in 9 out of 18 video categories considered for our IRL prediction analysis. For predicting commenting behavior via IRL, we first consider the training data and compute two point-valued estimates of the agent stopping costs that satisfy the NIAS and NIAC

13. Refer to `https://www.youtube.com/yt/about/policies/#community-guidelines` for details

14. All our numerical results are completely reproducible and can be accessed from the GitHub repository `https://github.com/KunalP117/YouTube-Commenting-Analysis`.

inequalities of Theorem 3, namely, max-margin IRL and entropy-regularized IRL defined below:

Max-Margin IRL :

$$\{\boldsymbol{S}_{\text{MM-IRL}},\ \epsilon^*\} \ =\ \underset{\epsilon\geq 0, \boldsymbol{S}\geq \boldsymbol{0}}{\text{argmax}}\, \epsilon,\ \text{such that}\ \text{NIAS}(\mathcal{D}_M, \boldsymbol{S}) \leq -\epsilon,\ \text{NIAC}(\mathcal{D}_M, \boldsymbol{S}) \leq -\epsilon, \tag{50}$$

Entropy-Regularized IRL :

$$\boldsymbol{S}_{\text{Ent-IRL}} \ =\ \text{Any feasible cost } \boldsymbol{S} \equiv \{s_m(x,a), x\in\mathcal{X}, a\in\mathcal{A}\}_{m=1}^M,\ \text{that satisfies} \tag{51}$$

$$(a)\ \ s_m(x,a_1) = 1,\ \forall\, x\in\mathcal{X}\ (\text{Normalization}),\ \text{and}$$

$$(b)\ \ \text{NIAS}(\mathcal{D}_M, \boldsymbol{S}) \leq 0,\ \ \text{NIAC}(\mathcal{D}_M, \boldsymbol{S}) \leq 0,\ \ \text{SUMCOST}(\mathcal{D}_M, \boldsymbol{S}, \{MI(\pi_0; p_m(a|x))\}_{m=1}^M) \leq 0.$$

In (50) and (51) above, $\boldsymbol{S} = \{s_m(x,a), m\in\mathcal{M}, x\in\mathcal{X}, a\in\mathcal{A}\}$ denotes the set of stopping costs over all environments $\mathcal{M}$, states $\mathcal{X}$ and actions $\mathcal{A}$; the NIAS, NIAC and SUMCOST feasibility inequalities are defined in (16), (17) and (18), respectively. In (51), $MI(\pi_0; p_m(a|x))$ denotes the mutual information between the agent's prior π_0 and action selection policy $p_m(a|x)$ defined as:

$$MI(\pi_0; p_m(a|x)) = \sum_{x,a} \pi_0(x)\ p_m(a|x)\ \log\left(\frac{p_m(a|x)}{\sum_x \pi_0(x)\ p_m(a|x)}\right)$$

The intuition behind (50) is clear: choose the stopping costs that pass the feasibility inequalities of Theorem 3 with the largest margin. In (51), we impose the additional constraint that the expected continue cost is the mutual information between the prior and the action selection policy. The inspiration for this information-theoretic cost stems from the seminal work of Sims (2003) who modeled human attention as a limited-capacity communication channel, and from Max-Entropy IRL (Ziebart et al., 2008) in IRL literature. Eq. 51 yields a softmax structure for the feasible stopping costs (see Appendix C.3 for a more detailed explanation); the key idea is that entropy-regularized IRL for Bayesian stopping yields a set of constant stopping costs, constant up to an affine monotone transformation.

For predicted cost $\{s_m(x,a), x\in\mathcal{X}, a\in\mathcal{A}, m\in\mathcal{M}\}$ and action selection policies $\{p_m(a|x), m\in\mathcal{M}\}$ from the training dataset, the predicted action selection policy $\hat{p}_m(a|x)$ for the test dataset is straightforwardly computed as:

$$\hat{p}_m(a|x) = \sum_{a'}\ \mathbb{1}\{a = \underset{b}{\text{argmax}} \sum_x \hat{p}_m(x|a')\ s_m(x,b)\}\ p_m(a'|x),\ \text{where} \tag{52}$$

the probability $\hat{p}_m(x|a') = \frac{\pi_{0,\text{test}}(x)\ p_m(a|x)}{\sum_x \pi_{0,\text{test}}(x)\ p_m(a|x)}$ is the predicted posterior belief of the state given action a' for the test dataset. Observe that all terms in the RHS of (52) pertain to the training dataset except for the prior $\pi_{0,\text{test}}$ that is empirically computed from the test dataset. Intuitively, (52) assumes the observation likelihood for the YouTube user in the test dataset is simply the action selection policy $p_m(a|x)$ from the training dataset. In words, the predicted action selection policy $\hat{p}_m(a|x)$ in (52) is obtained by simply summing the likelihoods of all actions $a'\in\mathcal{A}$ for which action a is optimal given posterior belief $\hat{p}(x|a')$.

Using (52), we obtained two sets of predicted action selection policies, namely, $\hat{\boldsymbol{p}}_{\text{MM-IRL}} = \{\hat{p}_m(a|x), m\in\mathcal{M}\}_{\text{MM-IRL}}$ and $\hat{\boldsymbol{p}}_{\text{Ent-IRL}} = \{\hat{p}_m(a|x), m\in\mathcal{M}\}_{\text{Ent-IRL}}$ for the test dataset, corresponding to stopping costs $\boldsymbol{S}_{\text{MM-IRL}}$ (50) and $\boldsymbol{S}_{\text{Ent-IRL}}$ (51), respectively. To comment on the prediction accuracy, we computed the chi-squared distance and total variation distance between the

true and predicted action selection policies for each video category m. [15] Figure 8 shows the IRL prediction results. We observed that for 7 out of the 9 video categories considered for IRL prediction analysis, the chi-squared and total variation distance for both sets of estimated action selection policies lie under 0.3. Hence, for 7 out of 9 video categories, our IRL algorithm successfully predicts the action selection policies in the test dataset with high accuracy. Another observation from Fig. 8 is that the max-margin IRL estimate is a more accurate predictor compared to the entropy-regularized IRL estimate and outperforms the entropy-regularized IRL in 2 out of 9 video categories.

Summary: We illustrated the predictive performance of our IRL algorithms (50), (51) on a real-world YouTube dataset. We chose two point-valued IRL estimates of stopping costs from the set of feasible costs that pass the NIAS (13) and NIAC (14) inequalities of Theorem 3, namely, max-margin IRL (50) and entropy-regularized IRL (51). We observed that both these cost estimates accurately predict YouTube commenting behavior (in terms of chi-squared and total variation distance as displayed in Fig. 8). Moreover, the max-margin IRL estimate yields a more accurate prediction compared to the entropy-regularized estimate.

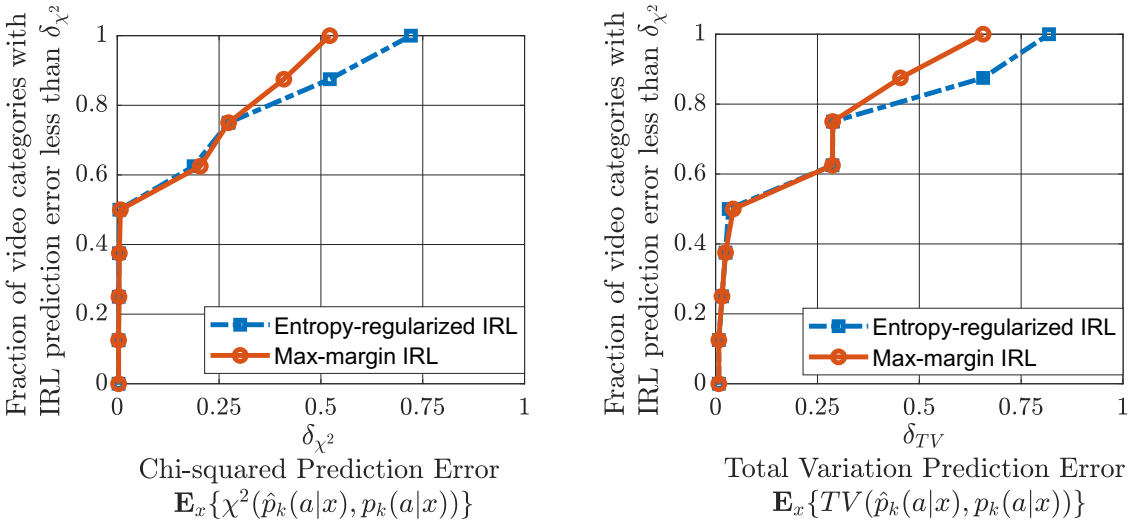

Figure 8: IRL Prediction Error for YouTube Dataset. The main takeaway is that point-valued IRL estimates that satisfy the feasibility test of Theorem 3 predict YouTube commenting behavior with high accuracy (low statistical distance between true and predicted distributions). For reconstructing the stopping costs, we choose two distinct point-valued stopping costs, namely, entropy-regularized IRL (51) and max-margin IRL (50) and performed our numerical experiments on 9 out of 18 video categories for which the video count exceeded 250. For both sets of estimated stopping costs, we observed that for 7 out of 9 YouTube video categories considered for analysis, both the chi-squared distance and total variation distance between the true and predicted action policy is less than 0.3.

15. Both chi-squared and total variation distance are normalized by definition since they take values in the interval $[0, 1]$.

6. Finite sample performance analysis of IRL decision test

Thus far, our IRL framework assumes (A1), namely, that the inverse learner has access to infinite trials of the stopping agent in M environments in order to solve the convex feasibility problem in Theorem 3. Suppose the inverse learner records only a finite number of trials and constructs its IRL dataset (7) comprising the agent's prior and empirically computed action selection policies in M environments. In this section, we address the following question: *How robust is the IRL decision test in Theorem 3 to finite sample datasets?* We now view Theorem 3 as a detector that takes in as input a noisy (empirical) dataset and outputs whether or not the observed agent is identified as an optimal stopping agent. Our aim is to provide bounds on the IRL detector's error probability in terms of the number of trials recorded by the inverse learner. We then obtain finite sample IRL results for the examples of inverse SHT and inverse search.

6.1 Finite sample statistical test for IRL

Suppose the inverse learner observes the actions of a Bayesian stopping agent in M environments. In addition to assumption (A2), we assume the following about the inverse learner for our finite sample result stated in Theorem 11 below.

(F1) The inverse learner knows the *finite dataset*

$$\widehat{\mathcal{D}}_M(\mathcal{K}) = \{\pi_0, \{\hat{p}_m(a|x), m \in \mathcal{M}\}\}, \text{ where } \mathcal{K} = \{K_{x,m}, m \in \mathcal{M}, x \in \mathcal{X}\}. \tag{53}$$

In (53), $\mathcal{K} = \{K_{x,m}, m \in \mathcal{M}, x \in \mathcal{X}\}$, $K_{x,m}$ is the number of trials recorded by the inverse learner for environment m and state x. $\hat{p}_m(a|x)$ is the empirical action selection policy of the agent in environment m computed for $K_{x,m}$ trials via (28).

(F2) The finite dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ satisfies the following inequality.

$$\varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K})), \varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K})) \geq \left(\sum_{x,m} \frac{A}{2K_{x,m}}\right)\left(\ln(2K_{x,m}/A) - \min_{x,m}\ \ln\left(2K_{x,m}/A\right)\right) \tag{54}$$

In (54), $K_m = \sum_x K_{x,m}$, $\bar{K} = K/\tau_{\max}^2$ and $\widetilde{K} = K^{-1}$. Eq. 54 imposes a lower bound on the number of samples needed for our sample complexity result of inverse optimal stopping. Eq. 54 is a sufficient condition for obtaining the constants of the sample complexity bound as the solution of a convex optimization problem; see (106) in the Appendix for more details. Variables $\varepsilon_1(\cdot), \varepsilon_2(\cdot)$ are the minimum perturbations needed for the finite dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ to satisfy and not satisfy, respectively, the NIAS and NIAC inequalities in Theorem 3, and defined formally in (57), (58) for readability.

For the reader's convenience, we discuss the assumptions (F1) and (F2) after the finite sample complexity result, Theorem 11. The feasibility test of Theorem 3 given a finite number of trials $\mathcal{K}$ can be equivalently formulated as a statistical hypothesis detection test that takes as input the finite dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ and accepts one of the two hypotheses, H_0 or H_1.

- H_0: *Null hypothesis* that the observed stopping agent is identified as an optimal agent, i.e. , the true dataset $\mathcal{D}_M$ is feasible wrt the NIAS and NIAC inequalities (13), (14) in Theorem 3.
- H_1: *Alternative hypothesis* that the observed stopping agent is *not* optimal.

Definition 10 (IRL detector for inverse optimal stopping) *Consider the inverse learner with dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$. Assume (A2) and (F1) hold. The IRL decision test $\mathrm{Test}_{\mathrm{IRL}}(\cdot)$ for inverse optimal stopping is given by:*

$$\mathrm{Test}_{\mathrm{IRL}}(\widehat{\mathcal{D}}_M(\mathcal{K})) = \begin{cases} H_0, & \text{if } \mathrm{IRL}(\widehat{\mathcal{D}}_M(\mathcal{K})) \neq \emptyset \\ H_1, & \text{if } \mathrm{IRL}(\widehat{\mathcal{D}}_M(\mathcal{K})) = \emptyset. \end{cases} \tag{55}$$

Here, $\mathrm{IRL}(\mathcal{D})$ is the set of feasible solutions to the convex NIAS and NIAC inequalities (13), (14) given dataset $\mathcal{D}$.

The statistical test defined above is a detector that accepts the null hypothesis H_0 if the finite dataset passes the feasibility test of Theorem 3 and accepts the alternative hypothesis H_1 it otherwise. Our main result stated below characterizes the performance of the feasibility test in identifying optimality given finite sample constraints, namely, provide bounds on the detector's Type-I/II error probabilities.

6.2 Main result. Finite sample analysis for IRL

Our main result below (Theorem 11) characterizes the following error probabilities of the statistical test in Definition 10:

$$\text{Type-I error prob.}: \mathbb{P}(H_0|\,\mathrm{IRL}(\widehat{\mathcal{D}}_M(\mathcal{K}) = \emptyset), \ \text{Type-II error prob.}: \mathbb{P}(H_1|\,\mathrm{IRL}(\widehat{\mathcal{D}}_M(\mathcal{K}) \neq \emptyset) \tag{56}$$

In (56), $\widehat{\mathcal{D}}_M(K)) = \emptyset$ means that the finite dataset fails the convex feasibility test for NIAC and NIAS inequalities (13), (14) and so the agent is identified as not an optimal agent. Our finite sample result in Theorem 11 below uses the dataset statistics variables $\varepsilon_1(\cdot), \varepsilon_2(\cdot), g(\cdot)$ from the finite dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ and are defined below. The quantities $\varepsilon_1(\cdot)$ and $\varepsilon_2(\cdot)$ are the minimum perturbations needed for the finite dataset to satisfy and not satisfy, respectively, the NIAS and NIAC inequalities in Theorem 3, and variable g is the constant for the error probability bounds.

NOTATION

Theorem 11 below uses the following variables:

$$\varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K})) = \min_{\{\hat{p}'_m, m\in\mathcal{M}\}} \sum_m \|\hat{p}_m - \hat{p}'_m\|_2^2 \text{ such that } \mathrm{IRL}(\{\pi_0, \{\hat{p}'_m(a|x)\}\}) \neq \emptyset. \tag{57}$$

$$\varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K})) = \min_{\{\hat{p}'_m, m\in\mathcal{M}\}} \sum_m \|\hat{p}_m - \hat{p}'_m\|_2^2 \text{ such that } \mathrm{IRL}(\{\pi_0, \{\hat{p}'_m(a|x)\}\}) = \emptyset. \tag{58}$$

$$g(\widehat{\mathcal{D}}_M(\mathcal{K})) = \left(A\sum_{x,m}\widetilde{K}_{x,m}\right)\prod_{x.m}\left(\frac{2K_{x,m}}{A}\right)^{\frac{\widetilde{K}_{x,m}}{\sum_{x,m}\widetilde{K}_{x,m}}}, \text{ where } \widetilde{K}_{x,m} = K_{x,m}^{-1} \tag{59}$$

Having defined our notation for error probability bounds, let us now state our first sample complexity result for the IRL detector (55).

Theorem 11 (Sample complexity for IRL detector) *Consider an inverse learner with finite dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ (53). The inverse learner aims to detect optimality of the stopping agent's actions using the statistical test in Definition 10. Assume (A2), (F1) and (F2) hold. Then, the Type-I and Type-II error*

probabilities (56) *of the IRL detector (Definition 10) are bounded as:*

$$\textit{Type-I error probability} \leq g(\widehat{\mathcal{D}}_M(\mathcal{K}))\ \exp\left(-\mathcal{K}_H \cdot \varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K}))\right), \tag{60}$$

$$\textit{Type-II error probability} \leq g(\widehat{\mathcal{D}}_M(\mathcal{K}))\ \exp\left(-\mathcal{K}_H \cdot \varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K}))\right). \tag{61}$$

In (59)*,* $\mathcal{K}_H = \left(\sum_{x,m} K_{x,m}^{-1}\right)^{-1}$*, and variables* $\varepsilon_1(\cdot), \varepsilon_2(\cdot)$ *and* g *are defined in* (57)*,* (58) *and* (59)*, respectively.*

The proof of Theorem 13 is in Appendix E. Below, we provide a sketch of the proof. Theorem 11 characterizes the robustness of the IRL detector in Definition 10 to finite sample constraints. It provides an upper bound on the detector's error probabilities in terms of the number of trials recorded by the inverse learner. Observe that since $\mathcal{K}_H$ is simply the unnormalized harmonic mean of $\mathcal{K}$ (68), the error rate is exponential in the *harmonic mean* of the number of trials recorded over M environments and X states.

The proof of Theorem 11 uses the two-sided Dvoretzky-Kiefer-Wolfowitz (DKW) concentration inequality (Van der Vaart, 2000; Kosorok, 2007) as the fundamental result to show that these error probabilities can be tightly bound in terms of the sample size $\mathcal{K}$ of the finite dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$. The DKW inequality provides a probabilistic bound on the deviation of the empirical cdf from the true cdf for i.i.d random variables. The i.i.d assumption holds for our detector in Definition 10 since the observed actions of the agent for a fixed state are independent and identically distributed over trials for all environments $\mathcal{M}$. To obtain our Type-I/II error bounds, we use the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality to probabilistically bound $\|p(a|x) - \hat{p}(a|x)\|_2$, the L_2-error between the empirical and true action selection policy for each environment $m \in \mathcal{M}$ and state $x \in \mathcal{X}$, followed by the union bound to bound the sum of L_2-errors due to finite sample size over all states and environments.
Discussion of Assumptions.
(F1): Given the finite dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ in (53), (F1) says that the inverse learner checks if the convex feasibility test of Theorem 3 has a feasible solution to detect an optimal stopping agent.
(F2): Abstractly, (F2) says that the inverse learner observes sufficiently many trials of the agent over all environments $\mathcal{M}$ such that the condition (54) is met.

First, for a given dataset $\mathcal{D}$, note that only one out of $\varepsilon_1(\mathcal{D}), \varepsilon_2(\mathcal{D})$ is non-zero and positive. Hence, (54) involves only the non-zero variable out of $\varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K})), \varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K}))$. Some words about the RHS of (54). $q(\mathcal{K}) - j(\mathcal{K})$ is a measure of how far is $K_{\min} = \min_{x,m} K_{x,m}$ from the remaining elements in $\mathcal{K} \backslash K_{\min}$. Since the RHS terms are of the form $\ln(z)/z$, it is easy to check that $q(\mathcal{K}) - j(\mathcal{K})$ decreases as the elements of $\mathcal{K}$ increase uniformly. As the number of samples go to infinity, the RHS in (54) tends to 0, hence the condition is almost surely satisfied for infinite samples. For finite $\mathcal{K}$, checking if (54) holds requires the inverse learner to solve an optimization problem (for the LHS) and perform MX multiplication operations and MX addition operations to compute the RHS of (54). As a practical estimate, for the inverse SHT task in Sec. 3.5 for 100 SHT environments, we observed that the inequality in (54) is satisfied if the samples exceeded $\sim 10^3$ for each environment.

6.3 Example 1. Finite sample effects for IRL in inverse SHT

We next turn to a finite sample analysis of IRL for inverse sequential hypothesis testing (SHT). Recall from Theorem 6 that identifying optimality of SHT is equivalent to feasibility of the linear

inequalities NIAS and NIAC*. The inverse learner's SHT dataset comprises both the agent action selection policies and the expected stopping times to perform IRL compared to only the action selection policies for inverse optimal stopping. Hence, in addition to the DKW inequality, our main result, Theorem 13 also uses the Hoeffding's inequality (Boucheron et al., 2013) to account for the finite sample effect[16] on the computation of the expected stopping time.

Assumptions and Detection Test. Suppose the inverse learner observes the actions of the Bayesian stopping agent over M SHT environments. We assume the following about the inverse learner for our finite sample result stated below for the inverse SHT problem.

(F3) The inverse learner uses the *finite SHT dataset*

$$\widehat{\mathcal{D}}_M(\mathcal{K}) = \{\pi_0, \{\hat{p}_m(a|x), \hat{C}_m, m \in \mathcal{M}\}\} \tag{62}$$

to detect if the stopping agent is an optimal SHT agent or not. The variable $\mathcal{K}$ defined in (53) is the number of trials recorded by the inverse learner, $\hat{C}_m$ is the sample average of the agent's stopping time in the m^{th} environment. $\hat{p}_m(a|x)$ is the agent's empirical action selection policy computed for $K_{x,m}$ trials via (28) in the m^{th} environment.

(F4) The inverse learner knows $\tau_{\max} = \inf\ \{t > 0 \mid \mathbb{P}(\tau \leq t) = 1, \forall m \in \mathcal{M}\}$, an upper bound on the stopping time of the SHT strategies chosen by the agent in all environments $\mathcal{M}$.

(F5) The finite dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ satisfies the following inequality.

$$\begin{aligned}
&\varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K})), \varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K})) \geq q(\mathcal{K}) - j(\mathcal{K}), \text{ where} \\
&q(\mathcal{K}) = \sum_{x,m} \frac{\ln(2K_{x,m}/A)}{2K_{x,m}/A} + \sum_m \frac{\ln(2\bar{K}_m)}{2\bar{K}_m}, \\
&j(\mathcal{K}) = \min_m \left(\min_x \ln\left(\frac{2K_{x,m}}{A}\right), \ln\left(\frac{\bar{K}_m}{\tau_{\max}^2}\right)\right) \left(A \sum_{x,m} \frac{K_{x,m}^{-1}}{2} + \sum_m \frac{K_m^{-1}}{2} \right)
\end{aligned} \tag{63}$$

In (63), $K_m = \sum_x K_{x,m}$, $\bar{K} = K/\tau_{\max}^2$ and $\widetilde{K} = K^{-1}$. Analogous to (54) in assumption (F2) for finite sample complexity of IRL for optimal stopping, (63) imposes a lower bound on the number of samples needed for our sample complexity result of inverse SHT. Eq. 63 is a sufficient condition for obtaining the constants of the sample complexity bound as the solution of a convex optimization problem. $\varepsilon_1(\cdot), \varepsilon_2(\cdot)$ are the minimum perturbations needed for the finite dataset to satisfy and not satisfy, respectively, the linear NIAS and NIAC* inequalities in Theorem 6, and defined formally in 65). The quantities $q(\cdot), j(\cdot)$ are decreasing functions of the sample size $\mathcal{K}$. For the reader's convenience, we discuss the assumptions (F3)-(F5) after the finite sample complexity result, Theorem 13. Analogous to Definition 10, the statistical detection test for the inverse SHT problem is defined below. It takes in as input a finite (noisy) dataset and outputs one of the two hypotheses- H_0 (agent is an optimal SHT agent) or H_1 (agent is not an optimal SHT agent).

Definition 12 (IRL decision test for inverse SHT) *Consider the inverse learner with dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ (62). Assume (A2) (A4), (A5) and (F3) hold. The IRL detector $\text{Test}_{\text{IRL}}(\cdot)$ for the inverse SHT problem*

16. Hoeffding's inequality applies to bounded r.v.s , and is true for SHT since the stopping time τ is finite almost surely.

is given by:

$$\text{Test}_{\text{IRL}}(\widehat{\mathcal{D}}_M(\mathcal{K})) = \begin{cases} H_0, & \textit{if } \text{IRL}_{\text{SHT}}(\widehat{\mathcal{D}}_M(\mathcal{K})) \neq \emptyset \\ H_1, & \textit{if } \text{IRL}_{\text{SHT}}(\widehat{\mathcal{D}}_M(\mathcal{K})) = \emptyset. \end{cases} \tag{64}$$

Here, $\text{IRL}_{\text{SHT}}(\mathcal{D})$ *is the set of feasible solutions to the linear NIAS and NIAC** *inequalities (16), (31) in Theorem 6 given dataset* $\mathcal{D}$*.*

Main Result. Finite Sample analysis for inverse SHT

We now present our finite sample result for IRL of the inverse SHT problem. It provides bounds for the Type-I/II error probabilities of the IRL detector (64) in terms of the sample size of $\widehat{\mathcal{D}}_M(\mathcal{K})$ (62).

Notation. Theorem 13 below uses the following variables:

$$\begin{aligned}
&\varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K})): \min_{\{\hat{p}'_m, \hat{C}'_m\}} \sum_m \|\hat{p}_m - \hat{p}'_m\|_2^2 + (\hat{C}_m - \hat{C}'_m)^2, \ \text{IRL}_{\text{SHT}}(\{\pi_0, \{\hat{p}'_m, C'_m\}\}) \neq \emptyset \\
&\varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K})): \min_{\{\hat{p}'_m, \hat{C}'_m\}} \sum_m \|\hat{p}_m - \hat{p}'_m\|_2^2 + (\hat{C}_m - \hat{C}'_m)^2, \ \text{IRL}_{\text{SHT}}(\{\pi_0, \{\hat{p}'_m, C'_m\}\}) = \emptyset \\
&i(\widehat{\mathcal{D}}_M(\mathcal{K})) = \mathcal{K}_H(\text{SHT}) \cdot h(\widehat{\mathcal{D}}_M(\mathcal{K})), \text{ where } \mathcal{K}_H(\text{SHT}) = A \sum_{x,m} K_{x,m}^{-1} + \tau_{\max}^2 \sum_m K_m^{-1}, \text{ and} \\
&h(\widehat{\mathcal{D}}_M(\mathcal{K})) = \prod_m \left(\left(2\bar{K}_m\right)^{\bar{K}_m^{-1}} \prod_x \left(\frac{2\bar{K}_{x,m}}{A} \right)^{A\widetilde{K}_{x,m}} \right)^{\mathcal{K}_H^{-1}(\text{SHT})}.
\end{aligned} \tag{65}$$

In (65), $K_m = \sum_x K_{x,m}$, $\bar{K} = K/\tau_{\max}^2$ and $\widetilde{K} = K^{-1}$. Analogous to the finite sample result for inverse optimal stopping, $\varepsilon_1(\cdot)$, $\varepsilon_2(\cdot)$ defined above are the minimum perturbations needed for the finite SHT dataset to satisfy and not satisfy, respectively, the NIAS and NIAC* inequalities in Theorem 6. Compared to the minimum perturbations defined in (57) and (58) for inverse optimal stopping, the key distinction is that $\varepsilon_1(\cdot)$ and $\varepsilon_2(\cdot)$ in (65) also involve perturbations in the expected continue cost of the agent. The variable i in (65) is the error constant for the finite sample error bounds for inverse SHT; variable $\mathcal{K}_H(\text{SHT})$ can be interpreted as a weighted harmonic mean of the recorded trials $\mathcal{K}$ (62).

Theorem 13 (Sample complexity for inverse SHT) *Consider an inverse learner with dataset* $\widehat{\mathcal{D}}_M(\mathcal{K})$ *(53) detecting if the agent acting in multiple environments* $\mathcal{M}$ *is an optimal SHT agent using the statistical test in Definition 12. Assume (F3)-(F5) hold. Then, the Type-I and Type-II error probabilities of the IRL detector (Definition 12) are bounded as:*

$$\textit{Type-I error probability} \leq i(\widehat{\mathcal{D}}_M(\mathcal{K})) \exp\left(-2\, \mathcal{K}_H(\text{SHT}) \cdot \varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K}))\right), \tag{66}$$

$$\textit{Type-II error probability} \leq i(\widehat{\mathcal{D}}_M(\mathcal{K})) \exp\left(-2\, \mathcal{K}_H(\text{SHT}) \cdot \varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K}))\right). \tag{67}$$

The proof of Theorem 13 is in Appendix E. Theorem 13 characterizes the robustness of the linear feasibility test in Theorem 6 to finite sample constraints. Compared to Theorem 11, the finite sample result of Theorem 15 requires accounting for the empirical estimate of the agent's expected continue cost. Hence, in addition to the DKW inequality, the proof of Theorem 13 uses Hoeffding's inequality to bound the empirical estimation error for the expected continue cost.

Discussion of Assumptions.
(F3): (F3) specifies the inverse learner's dataset for the inverse SHT problem computed from a finite number of trials.
(F4) says the inverse learner knows the upper bound of the agent's stopping times over all environments. This assumption is crucial for our main result since the Hoeffding's inequality (for bounding the finite sample effect of the expected stopping time) requires this knowledge.
(F5): The condition (63) in (F5) is analogous to assumption (F2) for the finite sample result for IRL of optimal stopping. (F5) admits a close form expression for the error bounds in Theorem 13. Abstractly, (F5) says that the number of samples recorded by the inverse learner is sufficiently large so that the condition (63) is satisfied.

6.4 Example 2. Finite sample effects for IRL in inverse search

We now analyze the finite sample effect of IRL for inverse search. Recall from Theorem 9 that optimal search is equivalent to feasibility of the linear NIAC† inequalities. Our main result below, namely, Theorem 15, characterizes the robustness of the feasibility test (wrt the NIAC† inequality) for detecting optimal search under finite sample constraints. It turns out that Theorem 15 is a special case of Theorem 11, our finite sample complexity result for IRL of inverse optimal stopping.

Main assumptions and detection test. Suppose the inverse learner observes the actions of a Bayesian stopping agent. We assume the following about the inverse learner:

(F6) Instead of (A6), the inverse learner uses the *finite dataset*

$$\widehat{\mathcal{D}}_M(\mathcal{K}) = \{\pi_0, \{\hat{g}_m(a,x), m \in \mathcal{M}\}\} \tag{68}$$

to detect if the agent performs optimal search or not. $\hat{g}_m(a,x)$ is the empirical search action policy of the agent defined in (47) and $\mathcal{K} = \{K_{x,m}, x \in \mathcal{X}, m \in \mathcal{M}\}$ denotes the number of trials recorded by the inverse learner in state x, where m indexes the search environment.

(F7) The prior belief of the targets π_0 is a uniform prior, i.e., $\pi_0(x) = 1/X$. Also, the reveal probability $\alpha(a)$ is the same for all actions $a \in \mathcal{A}$, i.e., $\alpha(a) = \alpha$. Although the variable α is unknown to the inverse learner, it satisfies the following inequality.

$$\alpha \geq \max\left\{\alpha^*, 1 - \frac{\min_{a\in\mathcal{A}} l_m(a)}{\max_{a\in\mathcal{A}} l_m(a)},\ \forall m \in \mathcal{M}\right\} \tag{69}$$

For the reader's convenience, we discuss the assumptions (F6) and (F7) after the finite sample complexity result, Theorem 15. We now define the statistical detection test for the inverse search problem. It takes in as input the finite (noisy) dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ (68) and detects one of the two hypotheses- H_0 (agent performs optimal search) or H_1 (agent does *not* perform optimal search).

Definition 14 (IRL decision test for inverse search) *Consider the inverse learner with dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ (68). Assume (A7), (F6) holds. The IRL detector $\mathrm{Test}_{\mathrm{IRL}}(\cdot)$ for the inverse search problem is given by:*

$$\mathrm{Test}_{\mathrm{IRL}}(\widehat{\mathcal{D}}_M(\mathcal{K})) = \begin{cases} H_0, & \text{if } \mathrm{IRL}_{\mathrm{Search}}(\widehat{\mathcal{D}}_M(\mathcal{K})) \neq \emptyset \\ H_1, & \text{if } \mathrm{IRL}_{\mathrm{Search}}(\widehat{\mathcal{D}}_M(\mathcal{K})) = \emptyset. \end{cases} \tag{70}$$

Here, $\mathrm{IRL}_{\mathrm{Search}}(\mathcal{D})$ is the set of feasible solutions to the linear NIAC† inequalities (46) in Theorem 9 given dataset $\mathcal{D}$.

Main Result. Finite Sample Result for Inverse Search.
We now present Theorem 15, our finite sample result for IRL of the inverse search problem. Theorem 15 provides bounds for the Type-I/II and posterior Type-I/II error probabilities of the IRL detector in Definition 14 in terms of the sample size of the finite search dataset and uses the following variables.

$$\varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K})) = \min_{\{\hat{g}'_m, m\in\mathcal{M}\}} \sum_m \|\hat{g}_m - \hat{g}'_m\|_2^2, \ \mathrm{IRL}_{\mathrm{Search}}(\{\pi_0, \hat{g}'_m(a,x), m\in\mathcal{M}\}) \neq \emptyset.$$
$$\varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K})) = \min_{\{\hat{g}'_m, m\in\mathcal{M}\}} \sum_m \|\hat{g}_m - \hat{g}'_m\|_2^2, \ \mathrm{IRL}_{\mathrm{Search}}(\{\pi_0, \hat{g}'_m(a,x), m\in\mathcal{M}\}) = \emptyset.$$

The variables $\varepsilon_1(\cdot),\ \varepsilon_2(\cdot)$ are the minimum perturbations needed for the finite search dataset to satisfy and not satisfy, respectively, the linear NIAC† inequalities (46) in Theorem 9.

Theorem 15 (Sample complexity for inverse search) *Consider an inverse learner with dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ (53) detecting if a Bayesian stopping agent is performing optimal search by using the statistical test in Definition 14. Assume (F6) and (F7) hold. Then, the Type-I and Type-II error probabilities for the IRL detector (Definition 14) are bounded as:*

$$\textit{Type-I error probability} \leq \frac{(1-\alpha^*)A}{\varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K}))(\alpha^*)^2}\left(\sum_{x,m} K_{x,m}^{-1/2}\right)^2, \tag{71}$$

$$\textit{Type-II error probability} \leq \frac{(1-\alpha^*)A}{\varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K}))(\alpha^*)^2}\left(\sum_{x,m} K_{x,m}^{-1/2}\right)^2. \tag{72}$$

The proof of Theorem 15 is in Appendix D. Theorem 15 characterizes the robustness of the linear feasibility test in Theorem 9 to finite sample constraints. It upper bounds the probability of incorrectly detecting the Bayesian agent as an optimal search agent or not an optimal search agent, in terms of the number of trials recorded by the inverse learner.

Discussion of assumptions.
(F6): Assumption (F6) specifies the inverse learner's dataset for the inverse search problem computed from a finite number of trials.
(F7): (F7) says that the agent has the same reveal probability for all locations for all environments and the inverse learner knows this reveal probability is greater than a certain value. This assumption can be viewed as an analogy of having the same instantaneous continue cost for the agent solving the SHT problem. The condition (69) results in the optimal search strategy of the agent to be periodic for all environments - the agent searches each location exactly once in a particular order (depends on the agent's search costs) and repeats this cycle till the target is located. This allows the search action policy $g_m(a,x)$ to be written in terms of the conditional pdf of the stopping time $\mathbb{P}_{\mu_m}(\tau|x)$ (see Appendix F). By analyzing the finite sample effects of the stopping time due to the added structure, Theorem 15 results. Note that Theorem 15 does not require the inverse learner to have information about the true stopping time of the agent, but only the empirical search action policy.

Summary

For finite sample observations, this section presented an IRL detector for optimality of a Bayesian stopping agent (Definition 10) and provided error bounds of the detector (Theorems 11). We also

presented finite sample IRL detectors for optimality in SHT and Bayesian search (Definitions 12, 14) and obtained error bounds of the detector in terms of the sample size (Theorem 13, 15). The key idea behind the sample complexity results is the construction of a probabilistic bound on the minimum perturbation needed to satisfy and not satisfy, respectively, the feasibility inequalities for optimality to bound the Type-I and Type-II error probabilities of the IRL detector, respectively.

7. Discussion and extensions

This paper has proposed Bayesian revealed preference methods for inverse reinforcement learning (IRL) in partially observed environments. Specifically, we considered IRL for a Bayesian agent performing multi-horizon sequential stopping. The results in this paper achieve IRL under the following restrictions on the inverse learner: (1) The inverse learner does not know if the decision maker is an optimal Bayesian stopping agent (2) The inverse learner does not know the agent's observation likelihood and (3) IRL for noisy datasets. Our IRL algorithms first identify if the agent is behaving in an optimal manner, and if so, estimate their stopping costs. The inverse learner can at best identify optimality of an agent's strategy wrt to its strategies chosen in other environments, a notion intuitively explained in the introduction and defined formally in Lemma 2. To illustrate our IRL approach, we considered two examples of sequential stopping problems, namely, sequential hypothesis testing (SHT) and Bayesian search and provided algorithms to estimate their misclassification/search costs.

Our main results were:

1. Specifying necessary and sufficient conditions for the decisions taken by a Bayesian agent in multiple environments to be identified as optimal sequential stopping and generating set-valued estimates of their stop costs (Theorem 3). To the best of our knowledge, our IRL results for Bayesian stopping time problems when the inverse learner has no knowledge of the agent's dynamics is novel.
2. Constructing convex feasibility based IRL algorithms for set-valued estimation of misclassification for an SHT agent (Theorem 6) and search costs for a search agent (Theorem 9) when decisions from infinite trials of the agent are available in multiple SHT and search environments, respectively. These results are special cases of Theorem 3 due to additional structure of the SHT and search problem compared to generic sequential stopping problems.
3. Proposing IRL detection tests for detecting optimality of sequential stopping, SHT and search when only a finite number of agent decisions are observed.
4. Providing sample complexity bounds on the Type-I/II and posterior Type-I/II error probabilities of the above detection tests under finite sample constraints (Theorems 11, 13 and 15).

EXTENSIONS

This paper identifies optimal stopping behavior in a Bayesian agent by observing their actions without external interference. A natural extension is to consider the controlled IRL setting where the inverse learner is an *active* entity that can influence/control the actions of the agent. This leads to the question: *How to influence the agent's actions so as to better identify optimal stopping behavior and estimate the agent costs more efficiently?*

Another question is: *How to formulate conditions under which the set-valued cost estimates for an agent in finitely many environments tends to a point-valued estimate as the number of environments tend to infinity?* In classical revealed preference theory, the papers Reny (2015);

Mas-Colell (1978) characterize properties of utility functions that rationalize infinite datasets. It is worthwhile generalizing these results to a Bayesian IRL setting.

Recent advances in deep IRL (Wulfmeier et al., 2015b; You et al., 2019) use deep neural networks as function approximators for the underlying feature space. Our current research aims to extend the results in this paper to deep-IRL for inverse optimal stopping where the inverse learner does not know the underlying state space and relies on neural networks for feature space approximation.

The IRL methodology of the paper assumes the analyst has no knowledge of the agent's observation likelihood. If the inverse learner knows *a priori* that the agent must choose its observation likelihood from a finite set, the inverse learner cannot rely on NIAS and NIAC in Theorem 3 for checking optimal Bayesian stopping. Instead, one must adapt adaptive search techniques for identifying the optimal observation likelihood that is (a) consistent with the inverse learner's dataset and (b) optimizes the agent's objective. If the agent's observation likelihood is known to be multi-variate Gaussian, then one can use the tree search approach that has seen success in applications such as adversarial tracking (Lan and Schwager, 2013) and motion planning (Bry and Roy, 2011). Extending the IRL results in this paper to tree-based adaptive search techniques is a subject of current research.

Finally, it is worthwhile exploring IRL for stopping time problems using the iterative update approach of Abbeel and Ng (2004) and the MCMC based sampling approach of Ziebart et al. (2008).

Acknowledgments

This research was funded in part the U. S. Army Research Office under grant W911NF-21-1-0093, National Science Foundation under grant CCF-2112457, and Air Force Office of Scientific Research under grant FA9550-22-1-0016

Appendix A. Context and Perspective. IRL for Bayesian stopping problems

A.1 Literature and Applications. IRL in Bayesian stopping problems

IRL methods have been successfully applied to areas like robotics (Kretzschmar et al., 2016), user engagement in multimedia social networks such as YouTube (Hoiles et al., 2020), autonomous navigation by Abbeel and Ng (2004); Ziebart et al. (2008) and Sharifzadeh et al. (2016) and inverse cognitive radar (Krishnamurthy, 2020; Krishnamurthy et al., 2020). Below we discuss several real-world examples where an analyst aggregates data from a Bayesian stopping time agent, and has no knowledge of the agent's observation likelihood for solving the IRL problem.

- *Consumer Insights and Ad Design Research:* Online multimedia are sequential Bayesian decision makers (Ratcliff and Smith, 2004; Krajbich et al., 2010); they accumulate evidence sequentially from audio-visual cues on the screen and then take an action (for example, playing a video, clicking on an ad etc.). In advertisement design, an analyst observes how an online user (stopping time agent) reacts to a pop-up advertisement in multiple environments, where the environment is characterized by the current web-page, content and position of the ad etc. In consumer research for online movie platforms, the analyst observes whether an online user clicks on a movie thumbnail or not in multiple scenarios, where the scenario depends on factors like user's past history of movies watched and neighboring movie thumbnails. The decision process of the online user (forward learner) in both these examples can be embedded into an SHT framework, where the sensing cost is the cost of attention to visual cues and the stopping (terminal) cost measures the online user's preferences for viewing the advertisement/movie. By characterizing the content reactivity of online users in different multimedia platforms, IRL for stopping time problems is useful for targeted ad-design and content recommendation.
- *Electronic counter-countermeasures in electronic Warfare:* Sequential Bayesian jamming models are extensively used in Electronic Counter Measures (ECM) for mitigating radar systems; see Arik and Akan (2015); Song et al. (2016) and Arasaratnam et al. (2006) for details. The proposed IRL algorithms can be used for Electronic Counter Counter Measures (ECCM) by the radar system to reverse engineer the adversarial ECM algorithms and avoid performance mitigation, hence extending the paper Krishnamurthy et al. (2020) to the Bayesian case. For instance, suppose an adversarial radar uses Bayesian search to identify valuable targets like in Bourgault et al. (2003). Using IRL for inverse search, a radar analyst can use the estimated search costs of the adversarial radar for effectively designing the targets to avoid being easily detected. Self et al. (2019); Xue et al. (2021) develop inverse optimal control (IOC) based IRL algorithms for reconstructing adversary intent for tracking control. Our work complements Xue et al. (2021) since it allows one to still achieve IRL without knowledge of model dynamics, as is common to assume in the literature.
- *Interpretable ML for Smart Healthcare:* Recently, sequential Bayesian models for assisting medical diagnoses have been aggressively used in smart healthcare like in Nishio et al. (2018); Oniśko and Druzdzel (2013); Exarchos et al. (2013); Jack Lee and Chu (2012) and Thakor et al. (1994). These trained models are usually only accessible in an abstracted black-box format in an executable software application. *Our IRL algorithm provides an interpretable Bayesian decision model for these assistive algorithms.* Interpretability in AI-enabled healthcare (Ahmad et al., 2018) facilitates informed decisions for the debugging and improvement of assistive diagnoses.

A.2 Related works in IRL

We now summarize the key IRL works in the literature and compare them to our paper.

(a) IRL in fully observed environments: Traditional IRL (Ng et al., 2000; Abbeel and Ng, 2004) aims to estimate an unknown deterministic reward function of an agent by observing the optimal actions of the agent in a Markov decision process (MDP) setting. The key assumption is the existence of an optimal policy. Our convex feasibility approach for IRL in stopping time problems can be viewed as a generalization of the feasibility inequalities in Ng et al. (2000, Theorem 3). Ng et al. (2000, Theorem 3) compute a feasible set of rewards that ensure the agent's policy outperforms all other policies. Since the set of policies for an MDP is finite, Ng et al. (2000, Theorem 3) comprises a finite set of linear inequalities. In comparison, the set of policies for a partially observed MDP (POMDP) is infinite. From the feasible set of rewards, Ng et al. (2000); Ratliff et al. (2006) choose the max-margin reward, i.e. , the reward that maximizes the regularized sum of differences between the performance of the observed policy and all other policies. In Sec. 3.4, we compute a regularized max-margin estimate of costs for inverse SHT and plot the reconstruction error. Abbeel and Ng (2004) achieve IRL by devising iterative algorithms for estimating the agent's reward function. Abstractly, the key idea is to terminate the iterative process once the value function of the rewards converges to an ϵ interval.

Ziebart et al. (2008) use the principle of maximum entropy for achieving IRL of an optimal agent, wherein the agent's policy is subject to a Shannon mutual information regularization. This regularization facilitates expressing the optimal policy in closed form; the optimal policy turns out to be softmax in terms of the Q-function of the MDP. Jeon et al. (2020) extend Ziebart et al. (2008) to a more general regularization setup that also admits a closed form solution to the optimal policy in terms of strongly convex functions for regularization, for examples, the Tsallis entropy (Lee et al., 2020) that generalizes Shannon entropy. Solving the IRL task with zero dynamics knowledge has also been explored in the literature. Herman et al. (2016) append the IRL task with simultaneous learning of model dynamics, specifically, the agent's transition kernel. The key idea in the approach is to maximize the log-likelihood of sampled trajectories wrt the appended parameter space that parametrizes the agent's rewards and transition kernel. Our problem setting differs from Herman et al. (2016) in that we operate in the non-parametric partially observed setting regime where the observation likelihood of the agent is unknown and not necessarily parametrizable. Indeed, our results can be specialized to parameter families of observation likelihood known to the analyst, and is a subject of current research.

Levine and Koltun (2012) generalize IRL to continuous space processes and circumvent the problem of finding the optimal policy for candidate reward functions. Recently Fu et al. (2017); Wulfmeier et al. (2015a); Sharifzadeh et al. (2016) and Finn et al. (2016) used deep neural networks for IRL to estimate agent rewards parametrized by complicated non-linear functions. Ramachandran and Amir (2007) achieve IRL when the agent's rewards are sampled from a prior distribution and the demonstrator's trajectories update the posterior belief of the reward. Building on the seminal work of Rust (1994), Kim et al. (2021); Cao et al. (2021) study identifiability of parameters for structure MDPs in IRL. In analogy, in this paper, we provide identifiability conditions for a subset of POMDPs, namely, Bayesian stopping problems.

(b) IRL in partially observed environments: The influential works of Choi and Kim (2009, 2011) and Makino and Takeuchi (2012) are the first works on IRL in a POMDP setting. They extend traditional IRL (Ng et al., 2000; Abbeel and Ng, 2004) to an infinite state space (space of posterior

beliefs of the agent). Makino and Takeuchi (2012) extend Bayesian IRL (Ramachandran and Amir, 2007) for MDPs to the POMDP setting. In analogy to Bayesian IRL, the aim is to compute the posterior distribution of reward functions given an observation dataset. The assumption of a softmax action policy suffices to compute the likelihood of the observation dataset given a reward function, and hence, bypasses the need to compare the agent's performance with respect to other candidate policies.

Since our work is closely related to Choi and Kim (2009, 2011), we briefly review their approach. In Choi and Kim (2009, 2011), the inverse learner first checks if the agent chooses the optimal action given a particular posterior belief, for *finitely many beliefs* aggregated from the observed trajectories of belief-action pairs. This is analogous to our NIAS condition (16) in Theorem 6, where we check if the agent's terminal action is optimal given its terminal belief. Next, the inverse learner check if the agent's policy is optimal with respect to a *finite set of policies* that deviate from the observed policy by a single step. This resembles our NIAC condition (31) in Theorem 3 where we check for optimality of the Bayesian decision maker's actions in multiple environments.

As Choi and Kim (2009, 2011) mention, this approach to checking for optimality only gives rise to a necessary condition, and not a necessary and sufficient condition like in Ng et al. (2000); Abbeel and Ng (2004), where the number of policies are finite. In other words, without prior information about the nature of the Q-function given a policy, it is impossible to check for global optimality, that is, find a reward function that outperforms *all* other policies (infinitely many policies).

Our Bayesian revealed preference based approach is *complementary* to Choi and Kim (2011). While Choi and Kim (2011) develop IRL methods for POMDPs with no assumption on problem structure, we consider a subset of POMDPs, namely, Bayesian stopping time problems. Due to the structure of stopping time problems, we show that our IRL algorithms *do not* require knowledge of the observation likelihood of the decision maker, nor require solving a POMDP. Indeed, IRL for generic POMDPs is non-identifiable if the inverse learner does not know the model dynamics, nor can solve a POMDP. To test for optimality, our IRL algorithms rely on the decision maker's strategies from multiple environments, where every environment differs in the terminal cost. Decision strategy in multiple environments can be viewed as a surrogate for performance wrt different policies. To summarize, our work builds on the seminal work of Choi and Kim (2011) with the key discerning features of our IRL methodology being: (1) Unobservability of agent dynamics, (2) No assumptions on decision optimality, and (3) IRL generalization for empirical (noisy) datasets with performance guarantees via finite sample complexity.

(c) Inverse Rational Control (IRC). IRC (Kwon et al., 2020) is a closely related field to IRL in partially observed environments. IRC models sub-optimality in decision makers as a misspecified reward function and aims to estimate this reward. The IRC task comprises two sub-tasks:
First, the inverse learner constructs a map from a continuous space of reward functions parameterized by θ to the reward's optimal policy.
Second, based on a finite observation dataset $\mathcal{D}$, the underlying hyperparameter θ is estimated as the maximum likelihood estimate $\operatorname{argmax}_\theta \mathbb{P}(\mathcal{D}|\theta)$.

In comparison, our approach bypasses the first sub-task in IRC by checking the feasibility of a finite set of convex inequalities. Given the information available to the inverse learner, these inequalities are both necessary and sufficient conditions for identifying optimality of a decision maker's decisions in multiple environments. Indeed, increasing the number of environments in which the decision

maker's actions are observed decreases the size of the feasible set of rationalizing rewards, and hence increases the precision of our set-valued IRL cost estimate.[17]

(d) Revealed Preference. The key formalism used in this paper to achieve IRL is *Bayesian revealed preferences* studied in microeconomics by Martin (2014); Caplin and Martin (2015); Caplin and Dean (2015); Caplin et al. (2019). Non-parametric estimation of cost functions given a finite length time series of decisions and budget constraints is the central theme in the area of classical (non-Bayesian) revealed preferences in microeconomics, starting with Afriat (1967); Samuelson (1938) where necessary and sufficient conditions for constrained utility maximization are given; see also Varian (1982, 2012); Woodford (2012) and more recently in machine learning (Lopes et al., 2009).

(e) Examples of Bayesian stopping time problems. After constructing an IRL framework for general stopping time problems, this paper discusses two important examples, namely, inverse sequential hypothesis testing and inverse Bayesian search. Below we briefly motivate these examples.
Example 1. Inverse Sequential Hypothesis Testing (SHT). Sequential hypothesis testing (SHT) (Poor, 1993; Ross, 2014) is widely studied in detection theory. The inverse SHT problem of estimating misclassification costs by observing the decisions of an SHT agent has not been addressed. Estimating SHT misclassification costs is useful in adversarial inference problems. For example, by observing the actions of an adversary, an inverse learner can estimate the adversary's utility and predict its future decisions.
Example 2. Inverse Bayesian Search. In Bayesian search, each agent sequentially searches locations until a stationary (non-moving) target is found. Bayesian search (Ross, 2014) is used in vehicular tracking (Wong et al., 2005), image processing (Pele and Werman, 2008) and cognitive radars (Goodman et al., 2007). IRL for Bayesian search requires the inverse learner to estimate the search costs by observing the search actions taken by a Bayesian search agent in multiple environments with different search costs.

Bayesian search is a special case of the Bayesian multi-armed bandit problem (Gittins, 1989; Bubeck et al., 2012). *A promising extension of our IRL approach would be to solve inverse Bayesian bandit problems, namely, estimate the Gittins indices of the bandit arms.* Regarding the literature in inverse bandits, Chan et al. (2019) propose a real-time assistive procedure for a human performing a bandit task based on the history of actions taken by the human. Noothigattu et al. (2021) solve the inverse bandit problem by assuming the inverse learner knows the variance of the stochastic reward; in comparison out setup assumes no knowledge of the rewards.

Appendix B. Proof of Lemma 2

Proof. Suppose $\mathcal{D}_M$ is generated by a Bayesian agent performing optimal stopping (Definition 1) in M environments. By definition, the following conditions hold:

$$\mu_m(\pi,\tau) = \underset{a\in\mathcal{A}}{\operatorname{argmin}}\, \pi' \bar{s}_{m,a},\ J(\mu_m, s_m) = \inf_{\mu\in\boldsymbol{\mu}} J(\mu, s_m), \tag{73}$$

where $J(\cdot)$ (10) is the expected cumulative cost comprising the expected stopping cost $G(\cdot)$ and expected cumulative continue cost $C(\cdot)$ Since the set of chosen strategies $\boldsymbol{\mu}_{\mathcal{M}} \subset \boldsymbol{\mu}$, the set of *all*

17. Revealed preference micro-economics (Mas-Colell, 1978; Reny, 2015) have studied the consistency of the set-valued approach to eliciting agent rewards. Mas-Colell (1978) specifies conditions under which the feasible set of utility functions reconstructed from a dataset of agent actions converges to a point for infinite datasets. Reny (2015) constructs a quasi-concave utility function that rationalizes an infinite dataset.

admissible policies, the feasibility of (73) implies the following conditions hold:

$$\begin{aligned} \mu_m(\pi, \tau) &= \operatorname*{argmin}_{a \in \mathcal{A}} \pi' \bar{s}_{m,a}, \\ J(\mu_m, s_m) &= \min_{\mu \in \boldsymbol{\mu}_{\mathcal{M}}} J(\mu, s_m). \end{aligned} \tag{74}$$

Since $\boldsymbol{\mu}_{\mathcal{M}}$ is finite, the 'inf' in (73) can be replaced with 'min' in (74). The second condition in (74) is simply a reformulation of (12). Hence, the 'if' statement of Lemma 2 is proved.

We now prove the 'only if' direction. Suppose the inverse learner has access to $\mathcal{D}_M$ aggregated from a Bayesian agent's actions in M environments. Specifically, the inverse learner only knows the agent's incurred expected costs finitely many policies $\mu_m \in \boldsymbol{\mu}_{\mathcal{M}}$. Alternatively, the sole knowledge of $\mathcal{D}_M$ implies that the inverse learner *does not know* the agent's expected stopping cost, nor the expected cumulative continue cost if the agent chooses any policy $\mu \in \boldsymbol{\mu} \backslash \boldsymbol{\mu}_{\mathcal{M}}$. Condition (8) is independent of the agent's policy, and only depends on the agent's stopping belief. However, (10) requires the inverse learner to compare the expected cost of the agent's strategy μ_m in environment m against infinitely many strategies $\mu \in \boldsymbol{\mu}$. Due to inverse learner's limited knowledge, the best the inverse learner can do to check if (10) holds is to check the feasibility of (74). ■

Appendix C. Proof of Theorem 3

We first introduce an observation likelihood α_m over a fictitious observation space $\mathcal{Y}_\pi$ with generic element $\tilde{y}_\pi$ for stopping strategy $\mu_m, m \in \mathcal{M}$:

$$\alpha_m(\tilde{y}_\pi | x) = \sum_{\bar{y}: \pi_\tau = \pi} \left(\prod_{t=1}^{\tau} B(y_t, x) \right) \tag{75}$$

Here $\bar{y}$ denotes a sequence of observations $y_1, y_2, \ldots$ and τ is the random stopping time for strategy μ_m defined in (6). $\alpha_m(\tilde{y}_\pi | x)$ is the likelihood of all observation sequences $\bar{y}$ such that given true state $x^o = x$ and stopping strategy μ_m, the agent's belief state at the stopping time τ is π. Equivalently, $\alpha_m(\tilde{y}_\pi | x)$ is the conditional probability density of the agent's stopping belief for strategy μ_m. By definition, the mapping from $\tilde{y}_\pi$ to stopping belief π is one-to-one. Hence, $|\mathcal{Y}_\pi| = |\Delta(\mathcal{X})|$, where $\Delta(\mathcal{X})$ denotes the $X - 1$ dimensional unit simplex of pmfs.

Next, we re-formulate the expected stopping cost $G(\mu_m, s_m)$ defined in (10) for stopping cost s_m in terms of the fictitious observation likelihood defined in (75).

$$G(\mu_m, s_m) = \mathbb{E}_{\mu_m} \left\{ \pi'_\tau \bar{s}_{m, a_\tau} \right\} = \int_{\mathcal{Y}_\pi} \underbrace{\left(\sum_{x \in \mathcal{X}} \alpha_m(\tilde{y}_\pi | x) \pi_0(x) \right)}_{\text{Marginal distribution of } \tilde{y}_\pi} \min_{a \in \mathcal{A}} \pi' \bar{s}_{m,a} \, d\tilde{y}_\pi \tag{76}$$

In the above equation, the summation within the parentheses is the unconditional probability density of the stopping belief π given stopping strategy μ_m. Also, as described above, $\alpha_m(\tilde{y}_\pi | x)$ is the likelihood of all observation sequences $\bar{y}$ such that given true state $x^o = x$ and stopping strategy μ_m, the agent's belief state at the stopping time τ is π. We are now ready to prove necessity and sufficiency of the NIAS, NIAC inequalities (13), (14) in Theorem 3 for identifying an optimal stopping agent (Lemma 2).

C.1 Necessity of NIAS, NIAC inequalities

Recall from Theorem 3 that the analyst knows the agent's action selection policy in multiple environments. The action selection policy $p_m(a|x)$ in $\mathcal{D}_M$ is a stochastically garbled version of $\alpha_m(\tilde{y}_\pi|x)$ defined in (75) and $p_m(x|a)$ is a stochastic garbling of the agent's stopping belief π when the stop action is a. The action selection policy can be rewritten as follows

$$p_m(a|x) = \int_{\mathcal{Y}_\pi} p_m(a|\tilde{y}_\pi)\, \alpha_m(\tilde{y}_\pi|x)\, d\tilde{y}_\pi \tag{77}$$

$$\implies p_m(x|a) = \int_{\mathcal{Y}_\pi} \frac{p_m(a|\tilde{y}_\pi)\alpha_m(\tilde{y}_\pi|x)\pi_0(x)}{p_m(a)}\, d\tilde{y}_\pi = \int_{\mathcal{Y}_\pi} p_m(\tilde{y}_\pi|a)\, \pi(x)\, d\tilde{y}_\pi, \tag{78}$$

where π is the agent's stopping belief and $p_m(\tilde{y}_\pi|a)$ is the probability density of the fictitious observations $\tilde{y}_\pi$ conditioned on the stop action a.

C.1.1 NIAS

Let action a be the optimal stop action (8) for stopping belief π of the m^{th} agent in $\mathcal{A}$. Then,

$$\sum_{x\in\mathcal{X}} \pi(x)(s_m(x,a) - s_m(x,b)) \leq 0,\ \forall a, b \in \mathcal{A} \tag{79}$$

$$\implies \int_{\mathcal{Y}_\pi} \sum_{x\in\mathcal{X}} \pi(x)(s_m(x,a) - s_m(x,b))\, p_m(\tilde{y}_\pi|a) d\tilde{y}_\pi \leq 0 \tag{80}$$

$$\implies \sum_{x\in\mathcal{X}} \left(\int_{\mathcal{Y}_\pi} p_m(\tilde{y}_\pi|a)\, \pi(x)\, d\tilde{y}_\pi \right) (s_m(x,a) - s_m(x,b)) \leq 0 \tag{81}$$

$$\implies \boxed{\sum_{x\in\mathcal{X}} p_m(x|a)(s_m(x,a) - s_m(x,b)) \leq 0.} \tag{82}$$

Eq. 79 says that the expected stop cost given belief π is minimum for stop action a. Here, the expectation is taken over the finite state set $\mathcal{X}$. The LHS of (80) is the expected value of the LHS of (79) taken over the space of fictitious observations $\mathcal{Y}_\pi$ wrt the probability density $p_m(\tilde{y}_\pi|a)$. Since $|s_m(x,a) - s_n(x,a)|$ is bounded, the integral on the LHS of (80) is finite. Hence, by Fubini's theorem, we can change the order of summation to get (81). The first term in the integral of (81) is equal to $p_m(x|a)$ from (78) which results in the final NIAS inequality (82).

C.1.2 NIAC

Define $\tilde{G}_{m,n}$ as expected stopping cost when the fictitious observation likelihood is $p_m(a|x)$ and stopping cost is $s_n(x,a)$. Then:

$$\tilde{G}_{n,m} = \sum_{a\in\mathcal{A}} \left(\sum_{x\in\mathcal{X}} p_n(a|x)\pi_0(x) \right) \min_{b\in\mathcal{A}} \sum_{x\in\mathcal{X}} p_m(x|a) s_m(x,b). \tag{83}$$

It follows from Blackwell dominance (Blackwell, 1953) that:

$$\tilde{G}_{n,m} \geq G(\mu_n, s_m) \text{ for all } m,\ n, \tag{84}$$

since the kernel $p_m(y_{1:\tau(\mu_m)}|x)$ Blackwell dominates the action selection policy $p_m(a|x)$. A key observation is that, in (84), equality holds for $m = n$ and is straightforward to show using Jensen's inequality. To summarize, we have the following inequality:

$$G_{m,m} = \tilde{G}_{m,m},\ G_{n,m} \leq \tilde{G}_{n,m} \tag{85}$$

For any set of environment indices $\{m_1, m_2, \ldots m_I\} \subset \mathcal{M},\ (m_{I+1} = m_1)$, we have the following inequality from (12) in Lemma 2:

$$\sum_{i=1}^{I} G(\mu_{m_{i+1}}, s_{m_i}) - G(\mu_{m_i}, s_{m_i}) \geq \sum_{i=1}^{I} C(\mu_{m_i}) - C(\mu_{m_{i+1}}) = 0$$

Combining the inequalities (84) with the above inequality, we get the NIAC inequality:

$$\sum_{i=1}^{I} \tilde{G}_{m_{i+1},m_i} - \tilde{G}_{m_i,m_i} \geq 0 \tag{86}$$

C.2 Sufficiency of NIAS, NIAC inequalities for Bayes optimal stopping

The inverse learner only has access to the agent's prior π_0 over the state space $\mathcal{X}$ and action selection policy $p_m(a|x)$ induced by the agent's policy in environment m. For a finite set of fictitious observations, the sufficiency proof assumes that there exists a one-to-one correspondence between the fictitious observation $\tilde{y}_\pi$ to the terminal action a. If the observation space $\mathcal{Y}$ is continuous-valued, the sufficiency proof assumes there exist disjoint subsets $\mathcal{Y}_\pi(a) \subset \mathcal{Y}_\pi$ and pdfs f_a with support $\mathcal{Y}_\pi(a)$ such that the fictitious observation likelihood $\alpha_m(\tilde{y}_\pi|x)$ can be expressed as:

$$\alpha_m(\tilde{y}_\pi|x) = f_a(\tilde{y}_\pi)\, p_m(a|x),\ \forall x \in \mathcal{X}, \tilde{y}_\pi \in \mathcal{Y}_\pi(a),\ a \in \mathcal{A}. \tag{87}$$

It follows straightforwardly from the fictitious observation likelihood expression in (87) that, for all $\tilde{y}_\pi \in \mathcal{Y}_\pi(a)$, the belief is simply $p(\cdot|a)$:

$$p_m(x|\tilde{y}_\pi) \underset{\text{(from (87))}}{=} \frac{f_a(\tilde{y}_\pi)\, p_m(a|x)\pi_0(x)}{\sum_{x'} f_a(\tilde{y}_\pi)\, p_m(a|x')\pi_0(x')} = \frac{p_m(a|x)\pi_0(x)}{\sum_{x'} p_m(a|x')\pi_0(x')} = p_m(x|a) \tag{88}$$

The important but subtle consequence of this assumption is that the expected stopping cost (12) $G_{m,n}$ is *equal* to the surrogate cost $\tilde{G}_{m,n}$ (83) for all $m, n \in \{1, 2, \ldots, M\}$.

C.2.1 NIAS

Suppose NIAS inequality holds, that is, for all $m \in \mathcal{M}$,

$$a = \underset{b \in \mathcal{A}}{\operatorname{argmin}} \sum_{x \in \mathcal{X}} p_m(x|a) s_m(x, b),\ \forall a \in \mathcal{A}. \tag{89}$$

Since the set $\{p_m(x|a), a \in \mathcal{A}\}$ constitutes the set of all stopping beliefs when $\alpha_m(\tilde{y}_\pi|x) = p_m(a|x)$, the following condition holds from (89).

$$\mu_m(\pi, \tau) = \underset{a \in \mathcal{A}}{\operatorname{argmin}}\, \pi' \bar{s}_{m,a}. \tag{90}$$

Eq. 90 is precisely (8), which says the agent chooses the stop action that minimizes its stopping cost given its stopping belief. Hence, it only remains to show that (12) in Lemma 2 holds to complete our sufficiency proof.

C.2.2 NIAC

Assuming the NIAS condition (82) holds, we use the concept of KKT multipliers from duality theory (Boyd and Vandenberghe, 2004, Sec. 5.5) to show that NIAC (86) is sufficient for (12) in Lemma 2 for optimal Bayesian stopping to hold. To do so, we use Lemma 16 below for linear assignment problems to show the existence of scalars C_m that satisfy (12); the feasibility inequality of interest is stated in (93) below. We now state Lemma 16 which can be viewed as a variational form of the NIAC inequality:

Lemma 16 *Suppose NIAC* (86) *holds. Then:*
(a) The solution of the following linear assignment problem is the identity map, that is, the optimal assignment map $x^*_{m,n}$ is given by $x^*_{m,n} = 1$ if $m = n$, and 0 otherwise:

$$\text{minimize}_{x_{m,n}} \sum_{m,n=1}^{M} x_{m,n}\,\tilde{G}_{m,n},\ \textit{subject to:} \tag{91}$$
$$\sum_n x_{m,n} \geq 1,\ \sum_m x_{m,n} \geq 1,\ x_{m,n} \geq 0\ \ \forall\, m,n \in \{1,2,\ldots,M\}.$$

(b) The KKT multipliers corresponding to the solution of the above assignment problem solve the feasibility condition of (12) *in Lemma 2.*

Proof.
(a) Let $x^*_{m,n}$ denote the optimal solution to the optimization problem (91). Indeed, since (91) is an LP, $x^*_{m,n} \in \{0,1\}$. We can prove by contradiction that if NIAC (86) holds, then the optimal assignment variables $x^*_{m,n}$ is Kronecker delta, that is, $x^*_{m,n} = 1$ if $m = n$ and 0:
Choose any arbitrary feasible $x_{m,n} \in \{0,1\}$. Consider the sequence of indices $I \equiv \{1, h_x(1),\ h_x \circ h_x(1),\ \ldots,\ (h_x\circ)^{M-2}h(1)\}$, where $h_x(m) = m'$ is the unique (due to assignment constraints in (91)) index $m' \in \mathcal{M}$ for which $x_{m,m'} = 1$ and '$\circ$' denotes the function composition operator. From invoking NIAC (86) on the index sequence I, we observe that $\sum_{m,n=1}^{M} x_{m,n}\tilde{G}_{m,n} \leq \sum_m \tilde{G}_{m,m} = \sum_{m,n=1}^{M} x^*_{m,n}\tilde{G}_{m,n}$, where $x^*_{m,n} = 1$ if $m = n$ and 0 otherwise. Hence, the identity map solves the assignment problem (91).

(b) We now write down the Karush-Kuhn-Tucker (KKT) conditions (Boyd and Vandenberghe, 2004, pg. 121) for the assignment problem (91) at the optimal solution $\{x^*_{m,n}\}_{m,n=1}^{M}$ that are first-order necessary conditions for optimality:

$$\text{There exist scalars } \lambda_{1,m},\ \lambda_{2,m},\ \lambda_{3,m,n} \geq 0,\ m,n \in \{1,2,\ldots M\},\ \text{such that:}$$
$$\text{(i) For } n = m:\ \tilde{G}_{m,m} = \lambda_{1,m} + \lambda_{2,m},\ \text{(ii) For } n \neq m:\ \tilde{G}_{n,m} = \lambda_{1,m} + \lambda_{2,n} + \lambda_{3,n,m}. \tag{92}$$

The scalars $\lambda_{1,m}$ and $\lambda_{2,n}$ in (92) correspond to KKT multipliers associated with the inequality constraints $(-\sum_n\ x_{m,n}) \leq -1$ and $(-\sum_m\ x_{m,n}) \leq -1$ in (91), respectively. We note that both sets of inequalities are active at $\{x^*_{m,n}\}_{m,n=1}^{M}$, the solution of (91). The scalar $\lambda_{m,n}$ is the KKT multiplier associated with the inequality constraint $(-x_{m,n}) \leq 0$, where the inequality is active only for $x^*_{m,n},\ m \neq n$. For any pair of environments $m,n \in \{1,2,\ldots,M\}$, the following inequalities

result due to the KKT conditions in (92):

$$\begin{aligned}
&\tilde{G}_{m,m} - \lambda_{2,m} = \tilde{G}_{n,m} - \lambda_{2,n} - \lambda_{3,n,m} \leq \tilde{G}_{n,m} - \lambda_{2,n} \quad (\text{since } \lambda_{3,n,m} \geq 0) \\
\Leftrightarrow\ &\tilde{G}_{m,m} + (\max_{m'} \lambda_{2,m'} - \lambda_{2,m}) \leq \tilde{G}_{n,m} + (\max_{m'} \lambda_{2,m'} - \lambda_{2,n}) \\
\Leftrightarrow\ &\tilde{G}_{m,m} + C_m \leq \tilde{G}_{n,m} + C_n \\
&(\text{by replacing } (\max_{m'} \lambda_{2,m'} - \lambda_{2,m}) \text{ with the variable } C_m \text{ for all } m \in \mathcal{M})
\end{aligned} \tag{93}$$

We now reconstruct an estimate of the agent's expected continue cost $\widehat{C}$ below and show (12) holds for Bayes optimal stopping. With $\boldsymbol{p}_\mu = p_\mu(a|x)$ denoting the action selection policy induced by a stopping strategy μ, consider the following reconstructed estimate of the agent's expected continue cost $\widehat{C}(\mu)$ in terms of the feasible variables $\{C_m\}_{m=1}^M$ (93):

$$\begin{aligned}
\widehat{C}(\mu) &= \max_{m=1,2,\ldots,M} \left\{ C_m + G_{m,m} - \tilde{G}(\mu, s_m) \right\}, \text{ where} \\
\tilde{G}(\mu, s_m) &= \sum_{a \in \mathcal{A}} \left(\sum_{x \in \mathcal{X}} p_\mu(a|x) \pi_0(x) \right) \min_{b \in \mathcal{A}} \sum_{x \in \mathcal{X}} p_\mu(x|a) s_m(x, b)
\end{aligned} \tag{94}$$

In (94), $\tilde{G}(\mu, s_m)$ denotes the expected stopping cost of the Bayesian agent with strategy μ and stopping costs $s_m(x, a)$ *assuming a one-to-one map between the set of observations* $y_{1:\tau(\mu)}$ *to action* a.[18] The variable $p_\mu(x|a)$ is the posterior belief of the state computed using Bayes rule as:

$$p_\mu(x|a) = \frac{\pi_0(x) p_\mu(a|x)}{\sum_{x'} \pi_0(x') p_\mu(a|x')}$$

Indeed, if the mapping from the fictitious observation set $\mathcal{Y}_\pi$ to the action set $\mathcal{A}$ is assumed to be one-to-one, the expected stopping cost can be expressed in terms of the action selection policy $\boldsymbol{p}_\mu$ induced by the stopping strategy μ. From (93), it is straightforward to show that $C(\mu_m) = C_m$. Hence, replacing C_m in (93) with $\widehat{C}(\mu_m)$ yields the following inequalities:

$$\begin{aligned}
&\tilde{G}_{m,m} + \widehat{C}(\mu_m) \leq \tilde{G}_{n,m} + \widehat{C}(\mu_n) \\
\Leftrightarrow\ &\text{A cumulative running cost can be reconstructed (94) such that condition (12) in Lemma 2} \\
&\text{holds with expected stopping costs } \tilde{G}_{n,m},\ m, n \in \{1, 2, \ldots, M\}. \blacksquare
\end{aligned} \tag{95}$$

In words, for a feasible set of stopping costs $\{s_m\}_{m=1}^M$ such that NIAS and NIAC hold, the Bayesian agent's (unobserved) strategies satisfy optimal Bayesian stopping (12). Moreover, for every feasible set of costs $\{s_m\}_{m=1}^M$, the term $(\max_{m'} \lambda_{2,m'} - \lambda_{2,m})$ denotes the expected continue cost incurred by the Bayesian agent due to choosing strategy μ_m, and $\tilde{G}_{m,n}$ denotes the agent's incurred expected stopping cost in environment n if it chooses strategy μ_m.

18. While it may seem counter-intuitive to assume a one-to-one mapping from the fictitious observation space to action space, one can show for convex costs like entropic costs (Shannon-Gibbs, Rényi and Tsallis) that the *optimal* mapping is one-to-one. The key idea is to show that having a many-to-one map with the same expected stopping cost is sub-optimal in that the agent incurs a strictly larger expected continue cost; see Matějka and McKay (2015, Lemma 1) for a more detailed explanation.

C.3 Remarks

1. *IRL for inverse SHT.* For the inverse SHT problem discussed in Sec. 3, the inverse learner knows C_m, the expected cumulative continue cost for the agent in environment m. Hence, the inverse learner can identify optimal SHT simply by checking if the NIAS inequality (82) and the following inequality is feasible:

$$\tilde{G}_{m,m} - \tilde{G}_{n,m} \leq C_n - C_m, \ \forall m, n \in \mathcal{M}, \tag{96}$$

where $\tilde{G}(\cdot)$ is the expected stop cost defined in (83) and C_m is the expected continue cost of the agent in environment m now known to the inverse learner. Due to (A5), $\tilde{G}_{m,\cdot} = G_{m,\cdot}$. Hence, (96) is equivalent to (12) in Lemma 2. We term the inequality in (96) as NIAC$*$ and use it in Theorem 6 for IRL for inverse SHT.

2. *Different observation likelihoods for different environments.* Theorem 3 is a purely *data-centric* approach for IRL that makes no assumptions on the agent's observation likelihood. If the NIAS and NIAC inequalities have a feasible solution, then the inverse learner's dataset $\mathcal{D}_M$ can be rationalized by a Bayesian agent that acts optimally (in the sense of Lemma 2) and has a fixed observation likelihood over all M environments. It may very well be the case that the Bayesian agent has a different observation likelihood in different environments, but Theorem 3 is opaque to this condition.

If the inverse learner knows *a priori* that the Bayesian agent uses a different observation likelihood for different environments, we need stronger conditions to achieve IRL. A sufficient condition for identifying optimal Bayesian stopping with distinct observation likelihoods in different environments is to assume the expected cumulative continue cost of the agent is independent of the observation likelihood. One example that satisfies this assumption is the entropic continue cost:

$$c_t = \lambda \left(H(\pi_t) - \mathbb{E}\{H(\pi_{t+1}) | \pi_t\} \right), \ t \geq 0, \ \lambda > 0 \tag{97}$$

where $H(p) = -\sum_i p_i \log(p_i)$ is the entropy of pmf p. The above choice of continue cost has two advantages:

(i) The expected *cumulative* continue cost for agent m is simply $H(\pi_0) - \mathbb{E}_a\{H(p_m(a|x))\}$, and is independent of the observation likelihood; see Matějka and McKay (2015, Lemma 1) for a discussion on how conditioned on state x, the optimal mapping from the space of fictitious observations $\tilde{y}_\pi$ (75) to the space of actions $\mathcal{A}$ is one-to-one due to the convexity of the entropic cost.

(ii) The inverse learner can test for 'absolute optimality' (8), (9) of the Bayesian agent's decisions and does not require observing the agent's behavior in multiple environments. Using the method of Lagrange multipliers, it is straightforward to show that for environment m, the following relation holds between the agent's stopping costs and its observed decisions for optimal Bayesian stopping:

$$p_m(a|x) = \frac{p_m(a) \ \exp(-s_m(x,a)/\lambda)}{\sum_{b \in \mathcal{A}} p_m(b) \ \exp(-s_m(x,b)/\lambda)}, \ \forall \, a, x, m, \tag{98}$$

where $\lambda > 0$ is a feasible variable that parametrizes the continue cost (97), and $p_m(a)$ is the marginal distribution of the action a in environment m. IRL is achieved by checking for the feasibility condition of Caplin et al. (2019, Eq. 3, Proposition 1) and solving the above set of equations (98) for $s_m(x,a)$; observe that there is no assumption of a fixed observation likelihood for the Bayesian agent across environments and the IRL estimate returns an ordinal estimate of the agent's stopping costs.

Appendix D. Proof of Theorem 9

We will show (46) is equivalent to the condition for identifying search optimality (45) in two steps. For a fixed stationary search strategy $\mu : \pi \to a$ and search cost $\{l(a), a \in \mathcal{A}\}$, we first express the expected cumulative search cost in terms of the search action policy $g(x, a)$ (43) and the prior π_0.

$$J(\mu, l) = \mathbb{E}_\mu \left\{ \sum_{t=1}^{\tau} l(\mu(\pi_t)) \right\} = \mathbb{E}_\mu \left\{ \sum_{a\in\mathcal{A}} l(a) \left(\sum_{t=1}^{\tau} \mathbb{1}\{\mu(\pi_t) = a\} \right) \right\}$$
$$= \sum_{a\in\mathcal{A}} l(a) \sum_{x\in\mathcal{X}} \pi_0(x) \mathbb{E}_\mu \left\{ \sum_{t=1}^{\tau} \mathbb{1}\{\mu(\pi_t) = a\} | x \right\} = \sum_{x\in\mathcal{X}, a\in\mathcal{A}} \pi_0(x) g(x, a) l(a).$$

Now, consider the set of search strategies $\{\mu_m, m \in \mathcal{M}\}$.

$$(45) \equiv \mu_m \in \underset{\{\mu_n, n\in\mathcal{M}\}}{\operatorname{argmin}} J(\mu_n, l_m) \iff J(\mu_m, l_m) - J(\mu_n, l_m) \le 0, \ m, n \in \mathcal{M}.$$
$$\iff \sum_{x\in\mathcal{X}, a\in\mathcal{A}} \pi_0(x)(g_m(a, x) - g_n(a, x))\, l_m(a) \le 0 \equiv (46).$$

■

Appendix E. Proof of Theorem 13

We divide the proof of Theorem 11 into 4 steps:
Step 1. Using Dvoretzky-Kiefer-Wolfowitz (DKW) inequality (Kosorok, 2007) to bound the deviation of the empirically computed action selection policy $\hat{p}_m(a|x)$ from $p_m(a|x)$.
The DKW inequality (Van der Vaart, 2000) provides a finite sample characterization of the asymptotic result of Glivenko-Cantelli theorem by quantifying the convergence rate of the empirical cdf to the true cdf. Let $F_m(a|x)$ and $\hat{F}_m(a|x)$ denote the cdfs of $p_m(a|x)$ and $\hat{p}_m(a|x)$, respectively. From the two-sided DKW inequality, the following inequalities result:

$$1 - 2\exp\left(-2K_{x,m}\varepsilon^2\right) \le \mathbb{P}\left(\max_{a\in\mathcal{A}} |\hat{F}_m(a|x) - F_m(a|x)| < \varepsilon \right) \le \mathbb{P}\left(\max_{a\in\mathcal{A}} |\hat{F}_m(a|x) - F_m(a|x)| \le \varepsilon \right)$$
$$\le \mathbb{P}\left(|p_m(a|x) - \hat{p}_m(a|x)| \le \varepsilon, \forall a \right) \le \mathbb{P}\left(\sum_{a\in\mathcal{A}} |p_m(a|x) - \hat{p}_m(a|x)|^2 \le A\varepsilon^2 \right).$$

For a fixed state x and environment m, let $\varepsilon_{x,m}$ bound the error $|\hat{p}_m(a|x) - p_m(a|x)|, \ \forall a \in \mathcal{A}$. With $\varepsilon^2_{\max} = A(\sum_{x,m} \varepsilon^2_{x,m})$, we have the following probabilistic bound on the L_2-error between the true and empirical action selection policies, summed over all states, actions and environments:

$$\boxed{\mathbb{P}\left(\sum_{a,x,m} |p_m(a|x) - \hat{p}_m(a|x)|^2 \le \varepsilon^2_{\max} \right) \ge \prod_{x,m} 1 - 2\exp\left(-2K_{x,m}\varepsilon^2_{x,m}\right)}. \tag{99}$$

Step 2. Using Hoeffding's Inequality (Boucheron et al., 2013) to bound the deviation of the sample average of the SHT stopping times $\hat{C}_m$ from the true value C_m.

The inverse learner knows the agent's stopping time $\tau \in [1, \tau_{\max}]$ for all M environments. Analogous to (99), for a fixed environment m, let η_m bound the error $|\hat{C}_m - C_m|$. With $\eta^2_{\max} = \sum_m \eta^2_m$, we have the following probabilistic bound on the L_2-error between the true and empirical expected stopping times of the agent, summed over all M environments via the two-sided Hoeffding's inequality:

$$\begin{aligned}
&\mathbb{P}(|\hat{C}_m - C_m| \leq \eta_m) \geq 1 - 2\exp\left(-2K_m\eta^2_m/\tau^2_{\max}\right) \\
\implies\ &\mathbb{P}(\sum_{x,m} |\hat{C}_m - C_m|^2 \leq \eta^2_{\max}) \geq \mathbb{P}(|\hat{C}_m - C_m| \leq \eta_m,\ \forall m \in \mathcal{M}) \\
\implies\ &\boxed{\mathbb{P}(\sum_{x,m} |\hat{C}_m - C_m|^2 \leq \eta^2_{\max}) \geq \prod_m 1 - 2\exp\left(-\frac{2K_m\eta^2_m}{\tau^2_{\max}}\right)}, \text{ where } K_m = \sum_x K_{x,m}
\end{aligned} \tag{100}$$

Step 3. Using the union bound on error bounds from steps 1 and 2 to bound the cumulative deviation of empirically computed action selection policies and expected stopping times.
Our aim is to construct a tight bound on the probability of the event $E_{pert}(\delta_{\max})$, where $E_{pert}(\delta_{\max})$ is defined as:

$$E_{pert}(\delta_{\max}) \equiv \{\ \{\hat{p}_m(a|x), \hat{C}_m\}\ \big|\ \sum_{x,m} |p_m(a|x) - \hat{p}_m(a|x)|^2 + \sum_x |C_m - \hat{C}_m|^2 \leq \delta^2_{\max}\}, \tag{101}$$

We note that $\mathbb{P}(E_{pert}(\delta_{\max}))$ bounds the Type-I and Type-II IRL error probabilities for suitable choices of $\delta_{\max}$. The Type-I error probability is bounded by $1 - \mathbb{P}(E_{pert})$ when $\delta^2_{\max}$ in (101) is set to $\varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K}))$. Also, the Type-II error probability is bounded by $1 - \mathbb{P}(E_{pert})$ when $\delta^2_{\max}$ in (101) is set to $\varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K}))$. Recall from (57), (58) that $\varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K}))$ and $\varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K}))$ correspond to the minimum L_2-perturbation needed for the *finite* IRL dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ (53) to pass and fail, respectively, the NIAS and NIAC conditions of Theorem 3 for inverse optimal Bayesian stopping.

For a fixed error tuple $\{\varepsilon_{x,m}, \eta_m\}$, consider the surrogate event $E(\{\varepsilon_{x,m}, \eta_m\})$ defined as:

$$E(\{\varepsilon_{x,m}, \eta_m\}) = \{\ \{\hat{p}_m(a|x), \hat{C}_m\}\ \big|\ \hat{p}_m(a|x) - p_m(a|x)| \leq \varepsilon_{x,m}, |\hat{C}_m - C_m| \leq \eta_m,\ \forall\, a, x, m\}. \tag{102}$$

Clearly, $E(\{\varepsilon_{x,m}, \eta_m\}) \subseteq E_{pert}(\delta_{\max})$ if $\delta^2_{\max}$ in (101) is equal to $A\sum_{x,m}\varepsilon^2_{x,m} + \sum_m \eta^2_{x,m}$. Combining the error bounds in (99) and (100) via a union bound to bound $\mathbb{P}(E(\{\varepsilon_{x,m}, \eta_m\}))$ yields the following inequality:

$$\begin{aligned}
&\mathbb{P}(E_{pert}(\delta_{\max})) \geq \mathbb{P}(E(\{\varepsilon_{x,m}, \eta_m\})) \\
\geq\ &\prod_{x,m} 1 - 2\exp\left(-2K_{x,m}\varepsilon^2_{x,m}\right) \prod_m 1 - 2\exp\left(-2K_m\eta^2_m/\tau^2_{\max}\right) &(103)\\
\geq\ &1 - \sum_{x,m} 2\exp\left(-2K_{x,m}\varepsilon^2_{x,m}\right) - \sum_m 2\exp\left(-2K_m\eta^2_m/\tau^2_{\max}\right) &(104)\\
\implies\ &\boxed{\mathbb{P}(E_{pert}(\delta_{\max})) \geq 1 - \sum_{x,m} 2\exp\left(-2K_{x,m}\varepsilon^2_{x,m}\right) - \sum_m 2\exp\left(-2K_m\eta^2_m/\tau^2_{\max}\right)}, &(105)
\end{aligned}$$

where $\delta^2_{\max} = A\sum_{x,m}\varepsilon^2_{x,m} + \sum_m \eta^2_m$. The inequality in (103) is simply a union bound on the error bounds in (99) and (100). The inequality in (104) holds due to Assumption (F5) that says the

analyst observes the agent's stopping action over sufficiently many trials. If the expected stopping time is known accurately, that is, $\hat{C}_m = C_m$ for all $m \in \mathcal{M}$, Assumption (F5) specializes to Assumption (F2).

Step 4. Obtaining a tight bound on the error probability computer in step 3.

Eq. 105 provides a probabilistic bound on the perturbation of the empirical dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ from the asymptotic dataset $\mathcal{D}_M$ in terms of the sample size $\mathcal{K} = \{K_{x,m}, x \in \mathcal{X}, m \in \mathcal{M}\}$, for a fixed error sequence $\{\varepsilon_{x,m}, \eta_m\}$. Our final step is to maximize the RHS in (105) subject to the constraint $A(\sum_{x,m} \varepsilon_{x,m}^2) + \sum_m \eta_m^2 = \delta_{\max}^2$ to obtain the tightest bound on $\mathbb{P}(E_{pert}(\delta_{\max}))$ (105). Equivalently, our aim is to minimize the following objective function:

$$\boxed{\min_{\{\varepsilon_{x,m}^2, \eta_m^2\} \geq 0} \sum_{x,m} 2\exp\left(-2K_{x,m}\varepsilon_{x,m}^2\right) + \sum_m 2\exp\left(-2K_m\eta_m^2/\tau_{\max}^2\right), \text{ s.t. } A\sum_{x,m}\varepsilon_{x,m}^2 + \sum_m \eta_m^2 = \delta_{\max}^2}. \tag{106}$$

We observe that the terms $\exp\left(-2K_{x,m}\varepsilon_{x,m}^2\right)$ and $\exp\left(-2K_m\eta_m^2/\tau_{\max}^2\right)$ are convex in $\varepsilon_{x,m}^2$ and η_m^2, respectively. Also, Assumption (F4) ensures the values of the terms $\varepsilon_{x,m}^2, \eta_m^2$ are bounded away from 0 at the local optimum of (106) computed via the method of Lagrange multipliers. Assumption (F5) ensures the $\varepsilon_{x,m}^2, \eta_m^2$ in (107) satisfy the Slater's condition for regularity. Since the equality constraint is linear, and the objective function is convex in the feasible variables $\varepsilon_{x,m}^2$ and η_m^2, (106) constitutes a convex optimization problem whose solution can be computed using the method of Lagrange multipliers (since Assumption (F4) ensures inactive inequality constraints at the optimal solution). Finally, the solution of the optimization problem (106) can be expressed as:

$$\boxed{\varepsilon_{x,m}^2 = (A\widetilde{K}_{x,m}/2)\left(\ln(\lambda) + \ln(2K_{x,m}/A)\right),\ \eta_{x,m}^2 = (\widetilde{\bar{K}}_m/2)(\ln(\lambda) + \ln(2\bar{K}_m))}, \text{ where}$$

$$\ln(\lambda) = \frac{\delta_{\max}^2 - A\sum_{x,m}\ln((2K_{x,m}/A)^{A\widetilde{K}_{x,m}/2}) - \sum_m \ln(2\bar{K}_m^{\widetilde{\bar{K}}_m/2})}{(A\sum_{x,m}\widetilde{K}_{x,m} + \sum_{m\in\mathcal{M}}\widetilde{\bar{K}}_m)/2}\ . \tag{107}$$

In the above equations, $\bar{K}_{(\cdot)} = K_{(\cdot)}/\tau_{\max}^2$, $\widetilde{K} = K^{-1}$. Subtracting the objective function of (106) evaluated at the optimal values of $\varepsilon_{x,m}^2$ and η_m^2 (107) from 1, and setting $\delta_{\max}^2$ to $\varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K}))$ and $\varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K}))$, respectively, yield lower bounds for Type-I and Type-II error probabilities of the IRL detector, respectively. ■

Remark. The proof of Theorem 11 for finite sample complexity of Theorem 3 is identical to the above proof structure (except that there is no step 2) and hence, omitted for brevity.

Appendix F. Proof of Theorem 15

To prove Theorem 15, we first state and prove an auxiliary result, namely, Proposition 17, below. Theorem 15 is a special case of Proposition 17 as discussed below.

Proposition 17 *Given dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ and (F7), the deviation of the finite sample search action policy $\hat{g}_m(a,x)$ and the true search action policy $g_m(a,x)$ can be bounded in terms of the number of samples $\mathcal{K} = \{K_{x,m}\}$ as follows.*

$$\mathbb{P}\left(\sum_{a,x,m} |g_m(a,x) - \hat{g}_m(a,x)|^2 \leq \epsilon\right) \geq 1 - \sum_{x,m}\frac{u(\widehat{\mathcal{D}}_M(\mathcal{K}))}{\epsilon K_{x,m}^{1/2}}, \tag{108}$$

where $u(\widehat{\mathcal{D}}_M(\mathcal{K})) = \frac{(1-\alpha^*)}{(\alpha^*)^2} X \sum_{x,m} K_{x,m}^{-1/2}$ *and* $\hat{g}_m(a,x)$ *is the sample average of the number of times the agent searches location* a *given state* x *in environment* m*.*

Proof. Assumption (F7) implies that for any environment $m \in \mathcal{M}$, given prior π_0, the agent's optimal search sequence $a_0, a_1, a_2, \ldots$ is periodic, i.e. , $a_t = a_{t+A}$. In other words, in any interval of A time steps, the agent searches each location exactly once in a particular (unknown to the inverse learner) order.

Consider the agent's search policy $g_m(a,x)$ defined in (43). Below we express $g_m(a,x)$ in terms of the pdf of the stopping time of the search process.

$$g_m(a,x) = \mathbb{E}_{\mu_m}\left\{\sum_{t=1}^{\tau} \mathbb{1}\{\mu_m(\pi_t) = a\} | x^o = x\right\} = \sum_{t=1}^{\infty} \mathbb{P}_{\mu_m}\left(\tau = t | x^o = x\right)\left(\lfloor t/X \rfloor + r(x,a)\right).$$

$$= \sum_{t=1}^{\infty} \lfloor t/X \rfloor \mathbb{P}_{\mu_m}\left(\tau = t | x^o = x\right) + r(x,a) = \mathbb{E}_{\mu_m}\{\lfloor \tau/X \rfloor \, | x\} + r_m(x,a) = \frac{1}{\alpha} + r_m(x,a).$$

Here, α denotes the reveal probability of the agent and $\lfloor \cdot \rfloor$ denotes the floor function. $r(x,a) = 1$ if agent searches location a prior to location x in one search cycle from time $t = 0 \to X-1$ in environment m, and 0 otherwise. The final equality follows from the fact that conditioned on the true state $x^o = x$, the random variable $\lfloor \tau/X \rfloor$ follows a geometric distribution with parameter α (unknown) due to (F7).

Consider now the quantity $|\hat{g}_m(a,x) - g_m(a,x)|$. Define $\hat{\mathbb{E}}_{\mu_m}\{\tau/X\} = \frac{\sum_{k=1}^{K_{x,m}} \lfloor \tau_{x,m,k}/X \rfloor}{K_{x,m}}$, the sample average of the normalized stopping time $\lfloor \tau/X \rfloor$ computed from $\widehat{\mathcal{D}}_M(\mathcal{K})$. Then,

$$|\hat{g}_m(a,x) - g_m(a,x)| = |\mathbb{E}_{\mu_m}\{\lfloor \tau/X \rfloor \, | x\} + r_m(x,a) - \hat{\mathbb{E}}_{\mu_m}\{\lfloor \tau/X \rfloor \, | x\} - r_m(x,a)|$$

$$= \left| \frac{1}{\alpha} - \hat{\mathbb{E}}_{\mu_m}\{\lfloor \tau/X \rfloor \, | x\} \right| \text{ (equal for all } a \text{ for a fixed } x).$$

$\hat{\mathbb{E}}_{\mu}\{\lfloor \tau/X \rfloor | x\}$ is an unbiased estimator of $\mathbb{E}_{\mu}\{\lfloor \tau/X \rfloor | x\}$ with variance $(1-\alpha)/K_{x,m}\alpha^2$. Using Chebyshev's inequality for random variables with finite variance to bound $|\hat{g}_m(a,x) - g_m(a,x)|$ for fixed a, x, m, the following inequality results

$$\mathbb{P}\left(|\hat{g}_m(a,x) - g_m(a,x)| \leq \varepsilon\right) \geq 1 - \frac{(1-\alpha)}{K_{x,m}(\alpha\varepsilon)^2} \tag{109}$$

For any set of positive reals $\{\varepsilon_{x,m}, x \in \mathcal{X}, m \in \mathcal{M}\}$ s.t. $|g_m(a,x) - \hat{g}_m(a,x)| \leq \varepsilon_{x,m}$ and $(A\sum_{x,m} \varepsilon_{x,m}^2) \leq \varepsilon_{\max}^2$, we have

$$\mathbb{P}(\sum_{x,m} |\hat{g}_m(a,x) - g_m(a,x)|^2 \leq \varepsilon_{\max}^2) \geq \prod_{x,m} 1 - \frac{(1-\alpha)}{K_{x,m}(\alpha\varepsilon_{x,m})^2}$$

$$\geq 1 - \sum_{x,m} \frac{(1-\alpha)}{K_{x,m}(\alpha\varepsilon_{x,m})^2} \geq 1 - \sum_{x,m} \frac{(1-\alpha^*)}{K_{x,m}(\alpha^*\varepsilon_{x,m})^2}. \tag{110}$$

Since $\frac{(1-\alpha)}{K_{x,m}(\alpha\varepsilon_{x,m})^2}$ is decreasing in $\varepsilon_{x,m}$, the tightest lower bound is achieved for the above inequality when $\sum_{x,m} \varepsilon_{x,m}^2 = \varepsilon_{\max}^2$ and is the solution to the following constrained optimization problem.

$$\min_{\{\varepsilon_{x,m}, x\in\mathcal{X}, m\in\mathcal{M}\}} \sum_{x,m} \frac{(1-\alpha^*)}{K_{x,m}(\alpha^*\varepsilon_{x,m})^2} \text{ s.t. } A\sum_{x,m} \varepsilon_{x,m}^2 = \varepsilon_{\max}^2. \tag{111}$$

Moreover, since the objective function in (111) is convex in $\varepsilon_{x,m}^2$ and constraint is affine in $\varepsilon_{x,m}^2$, the method of Lagrange multipliers (Boyd and Vandenberghe, 2004) yields necessary and sufficient conditions for an optimal solution to the above optimization problem if the solution obtained is positive for all $x \in \mathcal{X}, m \in \mathcal{M}$. The optimal value of $\varepsilon_{x,m}^2 = \frac{\varepsilon_{\max}^2 K_{x,m}^{-1/2}}{A \sum_{x,m} K_{x,m}^{-1/2}} > 0$ and thus minimizes the objective function in (111). Plugging this value in (110) and setting $\varepsilon_{\max}^2 = \epsilon$ yields the bound in the RHS of (108) and completes the proof for Proposition 17.

To obtain the error bounds (71), note that setting $\varepsilon = \varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K}))$ in (108) and subtracting the objective function from 1 bounds from below the Type-I error probability (see Sec. 6.2 for a detailed explanation). Similarly, setting $\varepsilon = \varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K}))$ in (108) and subtracting the objective function from 1 bounds from below the Type-II error probability of the IRL detector which completes the proof. ■

Appendix G. Context. IRL For Predicting YouTube Commenting Behavior

Our previous work (Hoiles et al., 2020) analyzes YouTube user engagement from a behavioral economics viewpoint. Although we use the same dataset for our numerical experiments in this paper, we emphasize that the IRL approach in this paper is new and differs from Hoiles et al. (2020) as:
(1) In Hoiles et al. (2020), we check if YouTube engagement is consistent with rationally inattentive utility maximization behavior Caplin and Martin (2015), a *static* decision model studied widely in behavioral and information economics. In comparison, our aim here is to test if the YouTube dataset satisfies Bayes optimal stopping, a *dynamic* decision model.
(2) The inference algorithms in Hoiles et al. (2020) considers *pairs* of video categories to reconstruct the underlying utility function of the YouTube user. In this paper, our IRL approach considers *all* 18 *YouTube video categories (described in Sec. 5.1 below) simultaneously in the feasibility test* for reconstructing the underlying stopping costs of the YouTube user, and hence, fully exploits the diversity in engagement behavior.
(3) In Hoiles et al. (2020), we perform a naive prediction analysis of YouTube user engagement using a *maximum a posteriori* (MAP) approach. In this paper, we predict the distribution of user engagement behavior via two representative point estimates of the recovered stopping costs and show the statistical similarity of the predicted distribution to the true engagement distribution.

YouTube user engagement and Bayesian stopping.
YouTube is a social multimedia platform where human users interact with video content on YouTube channels by posting comments and rating videos. Empirical studies (Khan (2017); Kwon and Gruzd (2017); Alhabash et al. (2015); Aprem and Krishnamurthy (2017)) show that the comments and ratings from users are influenced by the thumbnail, title, category, and perceived popularity of each video. Models for human decision making in the context of online multimedia platforms have been studied extensively in the literature. Two widely-used classes of models that motivate us to understand YouTube user engagement from the lens of Bayesian stopping are 'parallel constraint satisfaction models' and 'evidence accumulation models'. *Parallel constraint satisfaction models* (Glöckner and Betsch (2008); McClelland and Rumelhart (1989)) assume that information is screened sequentially to highlight salient alternatives and final choice is made when the decision maker reaches sufficient internal coherence. *Evidence accumulation models* (Krajbich et al. (2010); Ratcliff and Smith (2004)) model consumers' attention by drift-diffusion models that accumulate evidence based on whether they are fixating their gaze on either the product or its price. The decision is taken when any of the alternatives' evidence threshold level is achieved.

Both classes of models described above have one aspect in common - the decision maker makes a final choice *after* sequentially accumulating information, and naturally fits our Bayesian stopping time framework. In terms of YouTube webpage parameters, we hypothesize the YouTube user is a Bayesian agent that sequentially consumes webpage cues such as thumbnail, title and perceived popularity and incurs a cost of attention, followed by engaging on the YouTube platform and incurring a terminal cost. Our IRL aim in this section is to identify using the YouTube dataset, if YouTube users engage 'optimally' in a Bayesian stopping sense.

References

P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

S. N. Afriat. The construction of utility functions from expenditure data. *International economic review*, 8(1):67–77, 1967.

S. N. Afriat. Efficiency estimation of production functions. *International economic review*, pages 568–598, 1972.

S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013.

M. A. Ahmad, C. Eckert, and A. Teredesai. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pages 559–560, 2018.

S. Alhabash, J.-h. Baek, C. Cunningham, and A. Hagerstrom. To comment or not to comment?: How virality, arousal level, and commenting behavior on YouTube videos affect civic behavioral intentions. *Computers in human behavior*, 51:520–531, 2015.

A. Aprem and V. Krishnamurthy. Utility change point detection in online social media: A revealed preference framework. *IEEE Transactions on Signal Processing*, 65(7), April 2017.

I. Arasaratnam, S. Haykin, T. Kirubarajan, and F. A. Dilkes. Tracking the mode of operation of multi-function radars. In *2006 IEEE Conference on Radar*, pages 6–pp. IEEE, 2006.

M. Arik and O. B. Akan. Enabling cognition on electronic countermeasure systems against next-generation radars. In *MILCOM 2015-2015 IEEE Military Communications Conference*, pages 1103–1108. IEEE, 2015.

D. P. Bertsekas. Dynamic programming and optimal control 4th edition, volume ii. *Athena Scientific*, 2015.

D. Blackwell. Equivalent comparisons of experiments. *The annals of mathematical statistics*, pages 265–272, 1953.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

F. Bourgault, T. Furukawa, and H. F. Durrant-Whyte. Optimal search for a lost target in a Bayesian world. In *Field and service robotics*, pages 209–222. Springer, 2003.

S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

D. Brown, W. Goo, P. Nagarajan, and S. Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International conference on machine learning*, pages 783–792. PMLR, 2019.

A. Bry and N. Roy. Rapidly-exploring random belief trees for motion planning under uncertainty. In *2011 IEEE International Conference on Robotics and Automation*, pages 723–730. IEEE, 2011.

S. Bubeck, N. Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

H. Cao, S. Cohen, and L. Szpruch. Identifiability in inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12362–12373, 2021.

A. Caplin and M. Dean. Revealed preference, rational inattention, and costly information acquisition. *The American Economic Review*, 105(7):2183–2203, 2015.

A. Caplin and D. Martin. A testable theory of imperfect perception. *The Economic Journal*, 125 (582):184–202, 2015.

A. Caplin, M. Dean, and J. Leahy. Rational inattention, optimal consideration sets, and stochastic choice. *The Review of Economic Studies*, 86(3):1061–1094, 2019.

C. P. Chamley. *Rational herds: Economic models of social learning*. Cambridge University Press, 2004.

L. Chan, D. Hadfield-Menell, S. Srinivasa, and A. Dragan. The assistive multi-armed bandit. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 354–363. IEEE, 2019.

J. Choi and K.-E. Kim. Inverse reinforcement learning in partially observable environments. In *Twenty-First International Joint Conference on Artificial Intelligence*, 2009.

J. Choi and K.-E. Kim. Inverse reinforcement learning in partially observable environments. *Journal of Machine Learning Research*, 12:691–730, 2011.

K. P. Exarchos, T. P. Exarchos, C. V. Bourantas, M. I. Papafaklis, K. K. Naka, L. K. Michalis, O. Parodi, and D. I. Fotiadis. Prediction of coronary atherosclerosis progression using dynamic Bayesian networks. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3889–3892. IEEE, 2013.

C. Finn, S. Levine, and P. Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *International conference on machine learning*, pages 49–58, 2016.

X. Fontaine, Q. Berthet, and V. Perchet. Regularized contextual bandits. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2144–2153. PMLR, 2019.

J. Fu, K. Luo, and S. Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

J. C. Gittins. *Multi–armed Bandit Allocation Indices*. Wiley, 1989.

A. Glöckner and T. Betsch. Modeling option and strategy choices with connectionist networks: Towards an integrative model of automatic and deliberate decision making. *MPI Collective Goods Preprint*, 2(2008), 2008.

M. Gombolay, R. Jensen, J. Stigile, S.-H. Son, and J. Shah. Apprenticeship scheduling: Learning to schedule from human experts. AAAI Press/international joint conferences on artificial intelligence, 2016.

N. A. Goodman, P. R. Venkata, and M. A. Neifeld. Adaptive waveform design and sequential hypothesis testing for target recognition with active sensors. *IEEE Journal of Selected Topics in Signal Processing*, 1(1):105–113, 2007.

M. Herman, T. Gindele, J. Wagner, F. Schmitt, and W. Burgard. Inverse reinforcement learning with simultaneous estimation of rewards and dynamics. In *Artificial intelligence and statistics*, pages 102–110. PMLR, 2016.

W. Hoiles, V. Krishnamurthy, and K. Pattanayak. Rationally Inattentive Inverse Reinforcement Learning Explains YouTube commenting behavior. *The Journal of Machine Learning Research*, 21(170):1–39, 2020.

J. Hong, B. Kveton, M. Zaheer, and M. Ghavamzadeh. Hierarchical Bayesian bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 7724–7741. PMLR, 2022.

M. Houtman and J. Maks. Determining all maximal data subsets consistent with revealed preference. *Kwantitatieve methoden*, 19(1):89–104, 1985.

J. Jack Lee and C. T. Chu. Bayesian clinical trials in action. *Statistics in medicine*, 31(25):2955–2972, 2012.

W. Jeon, C.-Y. Su, P. Barde, T. Doan, D. Nowrouzezahrai, and J. Pineau. Regularized inverse reinforcement learning. *arXiv preprint arXiv:2010.03691*, 2020.

M. L. Khan. Social media engagement: What motivates user participation and consumption on YouTube? *Computers in Human Behavior*, 66:236–247, 2017.

K. Kim, S. Garg, K. Shiragur, and S. Ermon. Reward identification in inverse reinforcement learning. In *International Conference on Machine Learning*, pages 5496–5505. PMLR, 2021.

M. R. Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media, 2007.

I. Krajbich, C. Armel, and A. Rangel. Visual fixations and the computation and comparison of value in simple choice. *Nature neuroscience*, 13(10):1292–1298, 2010.

H. Kretzschmar, M. Spies, C. Sprunk, and W. Burgard. Socially compliant mobile robot navigation via inverse reinforcement learning. *The International Journal of Robotics Research*, 35(11): 1289–1307, 2016.

V. Krishnamurthy. *Partially observed Markov decision processes*. Cambridge University Press, 2016.

V. Krishnamurthy. Adversarial radar inference. from inverse tracking to inverse reinforcement learning of cognitive radar. *arXiv preprint arXiv:2002.10910*, 2020.

V. Krishnamurthy and B. Wahlberg. Partially observed markov decision process multiarmed bandits—structural results. *Mathematics of Operations Research*, 34(2):287–302, 2009.

V. Krishnamurthy, D. Angley, R. Evans, and B. Moran. Identifying cognitive radars-inverse reinforcement learning using revealed preferences. *IEEE Transactions on Signal Processing*, 68: 4529–4542, 2020.

H. K. Kwon and A. Gruzd. Is offensive commenting contagious online? examining public vs interpersonal swearing in response to Donald Trump's YouTube campaign videos. *Internet Research*, 27(4):991–1010, 2017.

M. Kwon, S. Daptardar, P. Schrater, and X. Pitkow. Inverse rational control with partially observable continuous nonlinear dynamics. *arXiv preprint arXiv:2009.12576*, 2020.

X. Lan and M. Schwager. Planning periodic persistent monitoring trajectories for sensing robots in Gaussian random fields. In *2013 IEEE International Conference on Robotics and Automation*, pages 2415–2420. IEEE, 2013.

L. Le Cam. On some asymptotic properties of maximum likelihood estimates and related bayes' estimates. *Univ. Calif. Publ. in Statist.*, 1:277–330, 1953.

G. Lee, M. Luo, F. Zambetta, and X. Li. Learning a super mario controller from examples of human play. In *2014 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE, 2014.

K. Lee, S. Kim, S. Lim, S. Choi, M. Hong, J. I. Kim, Y.-L. Park, and S. Oh. Generalized Tsallis entropy reinforcement learning and its application to soft mobile robots. In *Robotics: Science and Systems*, volume 16, pages 1–10, 2020.

S. Levine and V. Koltun. Continuous inverse optimal control with locally optimal examples. *arXiv preprint arXiv:1206.4617*, 2012.

M. Lopes, F. Melo, L. Montesano, and J. Santos-Victor. Active learning for reward estimation in inverse reinforcement learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 31–46. Springer, 2009.

W. S. Lovejoy. On the convexity of policy regions in partially observed systems. *Operations Research*, 35(4):619–621, 1987.

T. Makino and J. Takeuchi. Apprenticeship learning for model parameters of partially observable environments. *arXiv preprint arXiv:1206.6484*, 2012.

D. Martin. Bayesian revealed preferences. *Available at SSRN 2393035*, 2014.

A. Mas-Colell. On revealed preference analysis. *The Review of Economic Studies*, 45(1):121–131, 1978.

F. Matějka and A. McKay. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–98, 2015.

J. L. McClelland and D. E. Rumelhart. *Explorations in parallel distributed processing: A handbook of models, programs, and exercises*. MIT press, 1989.

A. Y. Ng, S. J. Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.

M. Nishio, M. Nishizawa, O. Sugiyama, R. Kojima, M. Yakami, T. Kuroda, and K. Togashi. Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization. *PloS one*, 13(4):e0195875, 2018.

R. Noothigattu, T. Yan, and A. D. Procaccia. Inverse reinforcement learning from like-minded teachers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35-10, pages 9197–9204, 2021.

A. Oniśko and M. J. Druzdzel. Impact of precision of Bayesian network parameters on accuracy of medical diagnostic systems. *Artificial intelligence in medicine*, 57(3):197–206, 2013.

C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of markov decision processes. *Mathematics of operations research*, 12(3):441–450, 1987.

O. Pele and M. Werman. Robust real-time pattern matching using Bayesian sequential hypothesis testing. *IEEE transactions on pattern analysis and machine intelligence*, 30(8):1427–1443, 2008.

H. V. Poor. *An Introduction to Signal Detection and Estimation*. Springer-Verlag, New York, 2 edition, 1993.

D. Ramachandran and E. Amir. Bayesian inverse reinforcement learning. In *IJCAI*, volume 7, pages 2586–2591, 2007.

R. Ratcliff and P. L. Smith. A comparison of sequential sampling models for two-choice reaction time. *Psychological review*, 111(2):333, 2004.

N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich. Maximum margin planning. In *Proceedings of the 23rd international conference on Machine learning*, pages 729–736, 2006.

P. J. Reny. A characterization of rationalizable consumer behavior. *Econometrica*, 83(1):175–192, 2015.

P. Rolland, L. Viano, N. Schürhoff, B. Nikolov, and V. Cevher. Identifiability and generalizability from multiple experts in inverse reinforcement learning. *arXiv preprint arXiv:2209.10974*, 2022.

S. M. Ross. *Introduction to stochastic dynamic programming*. Academic press, 2014.

J. Rust. Structural estimation of markov decision processes. *Handbook of econometrics*, 4:3081–3143, 1994.

P. A. Samuelson. A note on the pure theory of consumer's behaviour. *Economica*, 5(17):61–71, 1938.

R. Self, M. Harlan, and R. Kamalapurkar. Online inverse reinforcement learning for nonlinear systems. In *2019 IEEE conference on control technology and applications (CCTA)*, pages 296–301. IEEE, 2019.

S. Sharifzadeh, I. Chiotellis, R. Triebel, and D. Cremers. Learning to drive using inverse reinforcement learning and deep q-networks. *arXiv preprint arXiv:1612.03653*, 2016.

C. Sims. Implications of rational inattention. *Journal of monetary Economics*, 50(3):665–690, 2003.

H. Song, M. Xiao, J. Xiao, Y. Liang, and Z. Yang. A POMDP approach for scheduling the usage of airborne electronic countermeasures in air operations. *Aerospace Science and Technology*, 48: 86–93, 2016.

N. V. Thakor, A. Natarajan, and G. F. Tomaselli. Multiway sequential hypothesis testing for tachyarrhythmia discrimination. *IEEE Transactions on Biomedical Engineering*, 41(5):480–487, 1994.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

H. R. Varian. The nonparametric approach to demand analysis. *Econometrica: Journal of the Econometric Society*, pages 945–973, 1982.

H. R. Varian. Revealed preference and its applications. *The Economic Journal*, 122(560):332–338, 2012.

H. R. Varian et al. *Goodness-of-fit for revealed preference tests*. Department of Economics, University of Michigan Ann Arbor, 1991.

E.-M. Wong, F. Bourgault, and T. Furukawa. Multi-vehicle Bayesian search for multiple lost targets. In *Proceedings of the 2005 ieee international conference on robotics and automation*, pages 3169–3174. IEEE, 2005.

M. Woodford. Inattentive valuation and reference-dependent choice. *Unpublished Manuscript, Columbia University*, 2012.

M. Wulfmeier, P. Ondruska, and I. Posner. Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888*, 2015a.

M. Wulfmeier, P. Ondruska, and I. Posner. Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888*, 2015b.

W. Xue, P. Kolaric, J. Fan, B. Lian, T. Chai, and F. L. Lewis. Inverse reinforcement learning in tracking control based on inverse optimal control. *IEEE Transactions on Cybernetics*, 2021.

C. You, J. Lu, D. Filev, and P. Tsiotras. Advanced planning for autonomous vehicles using reinforcement learning and deep inverse reinforcement learning. *Robotics and Autonomous Systems*, 114: 1–18, 2019.

B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy Inverse Reinforcement Learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.