Inverse Active Sensing: Modeling and Understanding Timely Decision-Making

Daniel Jarrett [1] **Mihaela van der Schaar** [1 2]

Abstract

Evidence-based decision-making entails collecting (costly) observations about an underlying phenomenon of interest, and subsequently committing to an (informed) decision on the basis of accumulated evidence. In this setting, *active sensing* is the goal-oriented problem of efficiently selecting which acquisitions to make, and when and what decision to settle on. As its complement, *inverse active sensing* seeks to uncover an agent's preferences and strategy given their observable decision-making behavior. In this paper, we develop an expressive, unified framework for the general setting of evidence-based decision-making under endogenous, context-dependent time pressure—which requires negotiating (subjective) tradeoffs between accuracy, speediness, and cost of information. Using this language, we demonstrate how it enables *modeling* intuitive notions of surprise, suspense, and optimality in decision strategies (the forward problem). Finally, we illustrate how this formulation enables *understanding* decision-making behavior by quantifying preferences implicit in observed decision strategies (the inverse problem).

1. Introduction

Modeling decision-making processes is a central concern in computational and behavioral science, with important applications to medicine (Li et al., 2015), economics (Clithero, 2018), and cognition (Drugowitsch et al., 2014). In evidence-based decision-making, the agent first *collects* a series of observations about an underlying phenomenon of interest, then subsequently *commits* to an informed decision based on the accumulated evidence. As popular examples, consider the problems of hypothesis testing, medical diagnostics, and employee hiring: In each case, the decision-maker first conducts *acquisitions* for information (i.e. hypothesis tests, diagnostic procedures, and candidate interviews), the results on which the final *decision* is then based (i.e. the selected hypothesis, the declared disease, and the hiring decision).

In this context, *active sensing* is the goal-directed task of selecting which acquisitions to make, when to stop gathering information, and what decision to ultimately settle on. Active sensing strategies have been studied for such applications as multi-hypothesis testing (Naghshvar et al., 2013), sensory inference (Ahmad & Yu, 2013), and visual search (Butko & Movellan, 2010). These are typically formulated simply as sequential identification problems with an infinite horizon—that is, of minimizing inaccuracies against a unit sampling cost. However, for the general task of evidence-based decision-making, two critical shortcomings bear emphasis—a lack of *expressivity*, and a need for *specification*.

First, any sufficiently realistic decision model must account for the presence, endogeneity, and context-dependence of *time pressure*. While deadlines are studied in Frazier et al. (2008) and Dayanik & Yu (2013), they are external variables, and their settings are passive (i.e. sampling from a single exogenous supply of information). This is unrealistic: While lengthy aptitude tests may be more discriminative for recruiting purposes, their grueling nature may also cause more candidates drop out of the pipeline entirely. Similarly, the probability of an adverse medical event that aborts the diagnostic process depends on both the nature of the test chosen (i.e. *endogenous*) and the underlying disease itself (i.e. *context-dependent*). What we desire is a more expressive framework capable of modeling such tradeoffs.

Second, even the simplest decision models suffer from a need for specification. At a minimum, they require explicit knowledge of the relative penalties of decision inaccuracies and costs of acquisition. The need for complete specification

Problem Setting	Decisions	Acquisitions	Outcomes
Hypothesis Testing	Hypotheses	Hyp. Tests	Observations
Medical Diagnosis	Diseases	Diag. Tests	Results
Cognitive Science	Responses	Perceptions	Evidence
Sensory Inference	Targets	Fixations	Sensations
Marketing & Sales	Demographic	Outreaches	Engagements
Recruiting & Hiring	Hire or Fire	Interviews	Assessments

Table 1. Applications and terminology in timely decision-making.

[1]Department of Mathematics, University of Cambridge, UK. [2]Department of Electrical Engineering, UCLA, USA. Correspondence to: Daniel Jarrett <daniel.jarrett@maths.cam.ac.uk>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

Table 2. Comparison of models for timely decision-making. Our general framework accounts for the endogeneity (due to λ) and context-dependence (due to θ) of time pressure, as well as differential costs of acquisition, deadline penalties, and preferences. [1] Ahmad & Yu (2013), [2] Chernoff (1959), [3] Naghshvar et al. (2013), [4] Alaa & van der Schaar (2016), [5] Dayanik & Yu (2013), [6] Frazier et al. (2008). While the second row of models incorporates (external) deadlines, they do not consider *active* sensing—the first only considers sampling from a single stream, and the latter two only consider a *passive* supply of information, whence the problem readily reduces to optimal stopping.

Framework	Accuracy of Decision	Breach of Deadline	Cost of Acquisition	Time Pressure
Ahmad,[1] Chernoff,[2] Naghshvhar,[3] etc.	$\sum_{\theta'} \eta_{\mathrm{a},\theta'} \mathbb{1}_{\{\theta=\theta', \theta\neq\hat{\theta}\}}$	-	$\eta_{\mathrm{c}} \tau$	$\mathbb{P}\{\delta = t\} = 0$
Alaa,[4] Dayanik,[5] Frazier,[6] etc.	$\sum_{\theta'} \eta_{\mathrm{a},\theta'} \mathbb{1}_{\{\theta=\theta', \theta\neq\hat{\theta}, \tau<\delta\}}$	$\eta_{\mathrm{b}} \mathbb{1}_{\{\tau=\delta\}}$	$\eta_{\mathrm{c}} \tau$	e.g.[5] $\mathbb{P}\{\delta = t\} = p(1-p)^t$
(Ours)	$\sum_{\theta'} \eta_{\mathrm{a},\theta'} \mathbb{1}_{\{\theta=\theta', \theta\neq\hat{\theta}, \tau<\delta\}}$	$\sum_{\theta'} \eta_{\mathrm{b},\theta'} \mathbb{1}_{\{\theta=\theta', \tau=\delta\}}$	$\sum_{t=0}^{\tau-1} \eta_{\mathrm{c},\lambda_t} c_{\lambda_t}$	$\mathbb{P}\{\delta = t\} = p_{\theta,\lambda_t} \prod_{t'=0}^{t-1} (1 - p_{\theta,\lambda_{t'}})$

of (subjective) *preferences* severely dampens the practical utility of any analysis. For instance, we expect doctors to care much more about correctly diagnosing a lethal disease than another condition that presents with similar symptoms. Do they actually? By how much? Similarly, do recruiters care more about identifying the best candidates, or simply avoiding the worst at all costs? What we desire is a way to perform *inverse active sensing*—that is, to uncover preferences that effectively underlie observed decision behavior.

Contributions. We tackle both challenges simultaneously. In this paper, we first develop an expressive, unified framework for decision-making under endogenous and context-dependent time pressure. In this formulation, a decision-maker is required to negotiate the *subjective* tradeoff between accuracy, speediness, and the cost of information. Second, using this language, we demonstrate how it enables modeling intuitive notions of surprise, suspense, and optimality in decision strategies (the *forward* problem). Finally, we illustrate how this formulation enables understanding decision-making behavior by quantifying preferences implicit in observed decision strategies (the *inverse* problem).

Implications. Decision-making behavior is heterogeneous, and different agents are driven by different priorities. The implications are clear: An expressive forward model allows *prescribing* (optimal) decision-making in the presence of subjective preferences, while an inverse procedure allows *describing* (observed) decision-making in terms of preferences implicit among agents and institutions. In medicine, some populations may be subject to less rigorous diagnostic scrutiny than others (McKinlay et al., 2007), and test prescriptions often skewed by financial incentive (Song et al., 2010). The potential for detecting biases and quantifying hidden priorities in decision systems offers a first step towards a more methodical understanding of clinical practice.

2. Timely Decision-Making

First, we formulate the problem of timely decision-making (Section 2.1), and derive the Bayesian recognition model for an agent (Section 2.2). Next, we characterize optimal active sensing strategies (Section 3.1), on the basis of which we formalize and describe a solution for inverse active sensing (Section 3.2). Figure 1 provides a high-level block-diagram. **NOTE**: As a guide for anchoring our subsequent developments, see Figure 2 (on page 7) for a map of our key results.

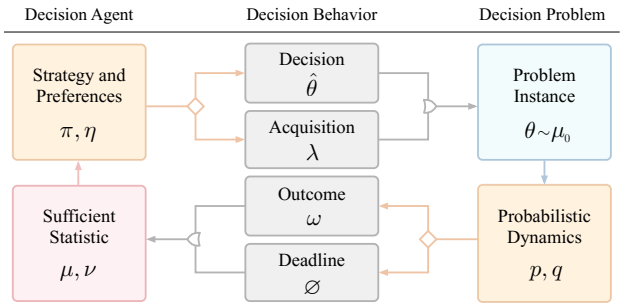


Figure 1. Active sensing and inverse active sensing. Decision problems are characterized by their probabilistic dynamics (right), and each instance is drawn from a known distribution (blue). Decision agents are characterized by their strategy and preferences (left), and maintain an internal representation according to a Bayesian recognition model (red). When an agent is presented with a problem, *active sensing* produces observable behavior (center). Conversely, given an agent's observed behavior with respect to a problem, *inverse active sensing* seeks to recover their preferences and strategy.

2.1. Decision Problem

Let Θ give the space of decisions (e.g. possible diagnoses), Λ the space of acquisitions (e.g. medical tests), and Ω the space of outcomes (e.g. diagnostic results) of acquisitions. We consider the setting where these spaces are finite, but note that our analysis easily extends to continuous outcomes, or distinct spaces of outcomes per test. Briefly, the goal of an agent is to commit to a decision $\hat{\theta} \in \Theta$ (e.g. issue an official diagnosis) at some decision time $\sigma \in \mathbb{N}$ before a probabilistic deadline $\delta \in \mathbb{N}$ (e.g. complication of the underlying disease). The active sensing challenge is in adaptively, sequentially choosing *which* acquisitions to perform, *when* to stop gathering information, and *what* decision to settle on.

The outcome $\omega \in \Omega$ of each acquisition $\lambda \in \Lambda$ is a random variable distributed according to the (stationary) generating distribution $q_{\theta,\lambda}(\omega) \doteq \mathbb{P}\{\omega|\theta;\lambda\}$, where $\theta \in \Theta$ is the unknown latent multinoulli variable representing the correct

decision (e.g. the true underlying disease). We assume that $\{q_{\theta,\lambda}\}_{\theta\in\Theta,\lambda\in\Lambda}$ are known (e.g. the known power of each medical test). Each λ_t conducted at time $t \in \mathbb{N}$ yields a corresponding outcome ω_{t+1} at the next step, and the outcomes are conditionally independent over time. For brevity, let $\lambda_{0:t}$ denote $\{\lambda_0, ..., \lambda_t\}$, and analogously $\omega_{1:t}$ for $\{\omega_1, ..., \omega_t\}$.

If the probabilistic deadline δ interrupts the trial, further interaction is void. The deadline is a random variable distributed as $\mathbb{P}\{\delta = t\} = p_{\theta,\lambda_t}\prod_{t'=0}^{t-1}(1 - p_{\theta,\lambda_{t'}})$ where constants $p_{\theta,\lambda} \in (0, 1)$ are specific to the acquisition λ (i.e. endogenous) and latent variable θ (i.e. context-dependent). This is in contrast to typical sequential identification models with either no deadline or external deadlines (see Table 2). We assume that $\{p_{\theta,\lambda}\}_{\theta\in\Theta,\lambda\in\Lambda}$ are known (e.g. the known risks of complication for various diseases and procedures). The presence of time pressure is important—in general decision-makers are not given an infinite window of opportunity to ponder their options; moreover, more informative acquisitions are often riskier (e.g. more invasive).

Episodes and Risks. Each *problem instance* is drawn as $\theta \sim \mu_0$ for some prior μ_0. In the most general case, statistics μ_0 may be stratified by subpopulation (e.g. as a function of patient demographics), or even as a learned mapping from covariates. Here we simply take it that μ_0 is known (e.g. from medical experience or literature), and defer further aspects of modeling for later work. A *decision episode* is characterized by the tuple $(\lambda_{0:\tau-1}, \tau, \hat{\theta})$, where $\tau = \min\{\delta, \sigma\}$ denotes the *stopping time* for the episode; note that $\hat{\theta} = \varnothing$ should the deadline occur before a decision is registered.

Each episode is generated by a *decision strategy* π, which produces—possibly stochastically—for each time t either a (continuing) acquisition $\lambda_t \in \Lambda$ or (terminating) decision $\hat{\theta} \in \Theta$. Let c_λ denote the immediate fixed cost of performing λ (e.g. the monetary expense of ordering a test), and let coefficient vectors $\eta_{\mathrm{a}} \in \mathbb{R}_+^{|\Theta|}$, $\eta_{\mathrm{b}} \in \mathbb{R}_+^{|\Theta|}$, and $\eta_{\mathrm{c}} \in \mathbb{R}_+^{|\Lambda|}$ respectively denote preference weights assigned to the importance of deciding *accurately* (i.e. on the correct decision), *speedily* (i.e. before the deadline), and *efficiently* (i.e. with minimal cost). Then the loss function is given as follows:

$$\begin{aligned}&\ell(\lambda_{0:\tau-1}, \tau, \hat{\theta}; \eta) \qquad (1)\\ &\doteq \textstyle\sum_{\theta'\in\Theta}\eta_{\mathrm{a},\theta'}\mathbb{1}_{\{\theta=\theta',\theta\neq\hat{\theta},\tau<\delta\}} && \triangleleft \text{ Accuracy of Decision}\\ &+ \textstyle\sum_{\theta'\in\Theta}\eta_{\mathrm{b},\theta'}\mathbb{1}_{\{\theta=\theta',\tau=\delta\}} && \triangleleft \text{ Breach of Deadline}\\ &+ \textstyle\sum_{t=0}^{\tau-1}\eta_{\mathrm{c},\lambda_t}c_{\lambda_t} && \triangleleft \text{ Cost of Acquisition}\end{aligned}$$

where we explicitly indicate dependence on $\eta \doteq (\eta_{\mathrm{a}}, \eta_{\mathrm{b}}, \eta_{\mathrm{c}})$. Then risk associated with executing strategy π is given by:

$$L(\pi; \eta) \doteq \mathbb{E}_{p,q}[\ell(\lambda_{0:\tau-1}, \tau, \hat{\theta}; \eta)|\mu_0, \pi] \qquad (2)$$

where the expectation is taken with respect to problem dynamics p and q. Importantly, this gives the most flexible framework: In addition to incorporating differential costs of acquisition $c \in \mathbb{R}_+^{|\Lambda|}$, the subjective penalties now depend on both θ and λ (see Table 2). For instance, failing to correctly diagnose a relatively innocuous condition may incur less damage than a lethal disease (viz. η_{a}); likewise, scaring away a good candidate with tough interviews may entail a larger sacrifice than losing a bad one to attrition (viz. η_{b}).

2.2. Beliefs and Information

In order to model and understand decision strategies, we first describe an agent's Bayesian recognition model for representing information. Note that this is not an assumption: We are *not* effectively assuming that real-world decision-makers indeed perform exact Bayesian inference—they most likely do not; instead, we are simply deriving a compact (internal) representation for use in describing their (external) behavior.

To this end, we highlight two implications of the endogeneity and context-dependence of time pressure. First, since acquisitions are no longer independent from time-to-event (i.e. survival), we cannot use the standard Bayes update as in prior work (Proposition 1). Second, each step now conveys two pieces of information: one from the acquired outcomes, and another from the process survival itself (Proposition 2).

Proposition 1 (Sufficient Statistic) Let $\nu_t \doteq \mathbb{1}_{\{\delta>t\}}$ denote the *survival* process, with initial value $\nu_0 = 1$. Then the *posterior* process $\mu_t \in \Delta(\Theta)$ is given by the following:

$$\begin{aligned}\mu_t = (1-\nu_{t-1})\mu_{t-1} + ((1-\nu_t)\bar{M}(\lambda_{t-1}, \mu_{t-1})\\ + \nu_t M(\lambda_{t-1}, \mu_{t-1}, \omega_t))\nu_{t-1}\end{aligned} \qquad (3)$$

where the *continual* update $M : \Lambda\times\Delta(\Theta)\times\Omega \to \Delta(\Theta)$ returns a distribution assigning to element θ the probability:

$$\frac{(1 - p_{\theta,\lambda_{t-1}})q_{\theta,\lambda_{t-1}}(\omega_t)\mu_{t-1}(\theta)}{\sum_{\theta'\in\Theta}(1 - p_{\theta',\lambda_{t-1}})q_{\theta',\lambda_{t-1}}(\omega_t)\mu_{t-1}(\theta')} \qquad (4)$$

and where the *terminal* update $\bar{M} : \Lambda \times \Delta(\Theta) \to \Delta(\Theta)$ returns a distribution assigning to element θ the probability:

$$p_{\theta,\lambda_{t-1}}\mu_{t-1}(\theta)/\textstyle\sum_{\theta'\in\Theta}p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta') \qquad (5)$$

Moreover, the sequence $(\mu_t, \nu_t)_{t=0}^{\infty}$ is a *controlled Markov process*, where the control inputs are the acquisitions λ_t.

Proof. Appendix C. □

This allows us to formally define decision strategies π as maps from μ_t, ν_t into $\Delta(\Lambda \cup \Theta)$. However, the dynamics of acquisition and survival are *entangled* here. The following result separately identifies the two sources of information:

Proposition 2 (Active and Passive Information) The information gleaned from (costly) acquisitions and (costless) observations of survival can be uniquely decomposed as:

$$\mu_t = \tilde{\mu}_t + \alpha_t + \beta_t \qquad (6)$$

where $\tilde{\mu}_t$ is a *martingale* that captures information obtained from the (actively) acquired results, the (continual) compensator $\alpha_t = A(\mu_{t-1}, \lambda_{t-1}, \nu_{t-1}, \nu_t)$ (passively) incorporates the bias from the ongoing process *survival* (where $\alpha_0 = 0$):

$$\alpha_t(\theta) = \alpha_{t-1}(\theta) - \mu_{t-1}(\theta)\nu_{t-1}\nu_t \cdot (p_{\theta,\lambda_{t-1}} - \bar{p}_{\mu_t,\lambda_{t-1}})/(1 - \bar{p}_{\mu_t,\lambda_{t-1}}) \quad (7)$$

and $\beta_t = B(\mu_{t-1}, \lambda_{t-1}, \nu_{t-1}, \nu_t)$ is the (terminal) compensator that analogously incorporates the bias from process *stoppage* (where $\beta_0 = 0$)—if the deadline were breached:

$$\beta_t(\theta) = \beta_{t-1}(\theta) + \mu_{t-1}(\theta)\nu_{t-1}(1 - \nu_t) \cdot (p_{\theta,\lambda_{t-1}} - \bar{p}_{\mu_t,\lambda_{t-1}})/\bar{p}_{\mu_t,\lambda_{t-1}} \quad (8)$$

where for brevity we denote the weighted average posterior probability of failure $\bar{p}_{\mu_t,\lambda_{t-1}} \doteq \sum_{\theta' \in \Theta} p_{\theta',\lambda_{t-1}} \mu_{t-1}(\theta')$.

Proof. Appendix C. □

This is intuitive: Before the deadline, the posterior process for any $\theta \in \Theta$ behaves like a *supermartingale* whenever the corresponding deadline risk $p_{\theta,\lambda_{t-1}}$ is greater than the average $\bar{p}_{\mu_t,\lambda_{t-1}}$, and behaves like a *submartingale* where it is less risky. Equality holds only if the deadline is exogenous, whence we recover the classic sequential identification setting where the posterior process is (always) a martingale.

3. Strategies and Preferences

Having formalized the decision problem and recognition model, we are ready for the forward and inverse problems. First, we consider optimal active sensing strategies (translating preferences *into* behavior). Our results then enable inverse active sensing (inferring preferences *from* behavior).

3.1. Optimal Active Sensing

Two distinguishing characteristics of our timely decision framework is that it requires *active* strategies, and that decisions are made under *time pressure*. This is in contrast to the (passive) settings in Dayanik & Yu (2013) and Frazier et al. (2008) with only a single choice of acquisition, where the decision problem readily reduces to optimal stopping. This is also in contrast to the (infinite) decision horizons in Ahmad & Yu (2013) and Naghshvar & Javidi (2011), where optimal strategies are free from considerations of survival.

First, we characterize the optimal value function, and show that it is unique and computable (Proposition 3). We then describe the optimal choice between continuing and terminating (Proposition 4). Finally, we interpret the risk-benefit tradeoff underlying the optimal acquisition (Proposition 5).

To begin, observe that at each time t, we wish to minimize the *to-go* component of risk, motivating the value function:

$$V^\pi(\mu_t, \nu_t; \eta) \doteq \mathbb{E}_{p,q}[\ell(\lambda_{0:\tau-1}, \tau, \hat{\theta}; \eta) | \lambda_{0:t-1}, \mu_t, \nu_t, \pi] - \sum_{t'=0}^{t-1} \eta_{\mathrm{c},\lambda_{t'}} c_{\lambda_{t'}} \quad (9)$$

for strategy π and preferences η. Now, it is tempting to immediately identify the optimal value function $V^*(\mu_t, \nu_t; \eta)$ with the fixed point of a dynamic programming operator. However, similar to even the passive case of Dayanik & Yu (2013), active sensing is *not* a discounted (nor fixed-horizon) problem; further, the stopping time itself is an endogenous (choice) random variable. Consequently, such an operator is not necessarily contractive (hence the optimal value is not necessarily unique or computable). Fortunately, we can leverage the (almost surely) finite decision deadline, and the following result assures us that these properties still hold:

Proposition 3 (Optimal Value) The optimal value function $V^*(\mu_t, \nu_t; \eta)$ is a fixed point of the operator $\mathbb{B}$ defined over the space of functions $V \in \mathbb{R}_+^{\Delta(\Theta)\times\{0,1\}}$ as follows:

$$(\mathbb{B}V)(\mu_t, \nu_t; \eta) = \min\{\inf_{\hat{\theta}' \in \Theta} \bar{Q}_{\hat{\theta}'}(\mu_t, \nu_t; \eta), \inf_{\lambda_t' \in \Lambda} Q_{\lambda_t'}(\mu_t, \nu_t; \eta)\} \quad (10)$$

where the (continual) Q-factors for *acquisitions* quantify the risk-to-go upon performing acquisition λ_t, given by:

$$Q_{\lambda_t}(\mu_t, \nu_t; \eta) = (1 - \nu_t)V(\mu_t, 0; \eta) + \eta_{\mathrm{c},\lambda_t} c_{\lambda_t} + \nu_t \mathbb{E}_{p,q}[V(\mu_{t+1}, \nu_{t+1}; \eta) | \lambda_t, \mu_t, \nu_t = 1] \quad (11)$$

and the (terminal) Q-factors for *decisions* quantify the risk upon settling on the final choice of decision $\hat{\theta}$, given by:

$$\bar{Q}_{\hat{\theta}}(\mu_t, \nu_t; \eta) = (1 - \nu_t)\sum_{\theta' \in \Theta} \eta_{\mathrm{b},\theta'} \mu_t(\theta') + \nu_t \sum_{\theta' \in \Theta, \theta \neq \hat{\theta}} \eta_{\mathrm{a},\theta'} \mu_t(\theta') \quad (12)$$

Moreover, the operator $\mathbb{B}$ is *contractive*, and the optimal value function is therefore the *unique* fixed point admitted.

Proof. Appendix C. □

As a result, we have that $V^*(\mu_t, \nu_t; \eta)$ is (uniquely) identifiable, and is (iteratively) computable via successive approximations. Now, the natural question becomes *when* to keep collecting information, versus stopping and committing to a decision. The following gives a geometric characterization:

Proposition 4 (Continuation and Termination) Denote by $m_\theta \in \Delta(\Theta)$ each vertex in the simplex, and let the optimal *aggregate* Q-factor for continuation be given by:

$$Q^*(\mu_t, \nu_t; \eta) \doteq \inf_{\lambda_t' \in \Lambda} Q^*_{\lambda_t'}(\mu_t, \nu_t; \eta) \quad (13)$$

and likewise $\bar{Q}(\mu_t, \nu_t; \eta) \doteq \inf_{\hat{\theta}' \in \Theta} \bar{Q}_{\hat{\theta}'}(\mu_t, \nu_t; \eta)$. Then Q^* is a *concave* function with respect to μ_t, and moreover takes on values strictly greater than $\bar{Q}$ at every vertex m_θ:

$$\forall m_\theta : Q^*(m_\theta, \nu_t; \eta) > \bar{Q}(m_\theta, \nu_t; \eta) \quad (14)$$

Hence the *termination set* $\mathcal{T}$ is the (disjoint) union of $|\Theta|$ *convex* regions delimited by the intersection of Q^* and $\bar{Q}$:

$$\mathcal{T}(\eta) = \{\mu_t : Q^*(\mu_t, \nu_t; \eta) \geq \bar{Q}(\mu_t, \nu_t; \eta)\} \quad (15)$$

and contains each of the simplex vertices. Finally, the (possibly null) *continuation set* is its complement $\Delta(\Theta) \setminus \mathcal{T}$.

Proof. Appendix C. □

In the *passive* setting (i.e. with a single acquisition choice), we are done. For *active* sensing, the key question concerns *which* acquisition to perform. Intuitively, we expect this choice to be made with respect to some notion of maximal information gain. Simultaneously, we also expect this to be

balanced against the deadline risk associated with different acquisitions. Again, this tradeoff is critical: An aggressive, extensive email survey may be maximally informative for marketing outreach, but may also be most likely to cause recipients to unsubscribe from the campaign entirely; also recall our earlier examples on candidate tests and interviews.

Negotiating Tradeoffs. The tradeoff between *surprise* (i.e. information gain) and *suspense* (i.e. riskiness of actions) has been explored in prior work on Bayesian reasoning in economics (Ely et al., 2015) and in binary decisions with samples from a single continuous stream (Alaa & van der Schaar, 2016). We first formalize these notions appropriately in the context of active sensing for timely decisions.

Two distinctions warrant attention: First, the informativeness of an acquisition must be *timely* (i.e. arriving *before* the deadline). Now, the classical definition for the informativeness of an acquisition simply measures the difference between the prior and expected posterior values for some appropriate measure of risk or uncertainty (DeGroot et al., 1962). While this notion of surprise readily applies to the infinite-horizon setting (Naghshvar et al., 2013), here we only care about informativeness while the process is still alive: Realizing the correct decision *after* the opportunity has closed carries no value for the original decision problem.

Second, here the riskiness of different choices of acquisitions is *subjective* (i.e. weighted by an agent's preferences). Now, the notion of suspense is previously simply taken with respect to the the posterior survival probability (Alaa & van der Schaar, 2016). In the presence of preferences, we now care about the *preference-weighted* survival function.

Formally, define Markov operator $\mathbb{M}_{\lambda_t}$ for any appropriate measure of risk or uncertainty $U : \Delta(\Theta) \times \{0,1\} \to \mathbb{R}_+$:

$$\begin{aligned}(\mathbb{M}_{\lambda_t} U)(\mu_t, \nu_t) = (1-\nu_t)U(\mu_t, \nu_t) + \\ \nu_t \mathbb{E}_{p,q}[U(M(\lambda_t, \mu_t, \omega_{t+1}), \nu_{t+1}) | \lambda_t, \mu_t, \nu_{t+1} = 1]\end{aligned} \tag{16}$$

capturing the expected posterior value of U should the deadline not intercede, and simply returns the prior if the process were already dead. Then we naturally have the following:

(*Timely*) **Surprise**. With respect to $U : \Delta(\Theta) \times \{0,1\} \to \mathbb{R}_+$, the informativeness of any choice of acquisition λ_t in reference to the statistic (μ_t, ν_t) is given by the following:

$$\begin{aligned}I_t(\lambda_t) \doteq U(\mu_t, \nu_t) \\ - (1 - \bar{p}_{\mu_t, \lambda_{t-1}})(\mathbb{M}_{\lambda_t} U)(\mu_t, \nu_t)\end{aligned} \tag{17}$$

(*Subjective*) **Suspense**. With respect the importance parameters η, the preference-weighted posterior probability of survival until after acquisition λ_t is given by the following:

$$S_t(\lambda_t) \doteq 1 - \frac{\sum_{\theta' \in \Theta} \eta_{\mathrm{b},\theta'} p_{\theta',\lambda_t} \mu_t(\theta')}{\sum_{\theta' \in \Theta} \eta_{\mathrm{b},\theta'} \mu_t(\theta')} \tag{18}$$

These generalized notions of surprise and suspense simply inherit existing definitions while accounting for the presence of time pressure and subjective preferences. In particular, setting $p_{\theta,\lambda} = 0$ recovers the original (infinite-horizon) definition of information gain (Naghshvar et al., 2013), whence the optimal strategy simply greedily maximizes information. Likewise, setting $\eta_{\mathrm{b},\theta} = 1$ recovers the original (preference-free) definition of suspense (Alaa & van der Schaar, 2016).

We can now formally characterize the risk-benefit tradeoff, highlighting the expressivity of our framework. It turns out that the optimal acquisition λ_t^* (i.e per the optimal strategy) naturally strikes a balance between surprise and suspense:

Proposition 5 (Surprise and Suspense) When $\mu_t \notin \mathcal{T}(\eta)$, the optimal acquisition directly trades off surprise and suspense (in addition to the immediate cost of acquisition):

$$\lambda_t^* = \arg\sup_{\lambda_t \in \Lambda} h(I_t(\lambda_t), S_t(\lambda_t)) - \eta_{\mathrm{c},\lambda_t} c_{\lambda_t} \tag{19}$$

where h is increasing in $I_t(\lambda_t)$ and $S_t(\lambda_t)$, and the uncertainty function for the information gain is taken as $U = V^*$.

Proof. Appendix C. □

Depending on an individual's preferences, this tradeoff automatically expresses a range of behaviors from "surprise-optimal" to "suspense-optimal". Note that this is absent in passive settings such as Frazier et al. (2008) and Dayanik & Yu (2013)—i.e. no choice of surprise; it is also absent in settings with no deadline risk, such as Ahmad & Yu (2013) and Naghshvar et al. (2013)—i.e. no element of suspense.

Finally, for completeness we also state the optimal decision $\hat{\theta}^*$ when $\mu_t \in \mathcal{T}(\eta)$. Immediately from Equation 12 for $\bar{Q}$:

$$\hat{\theta}^* = \arg\sup_{\hat{\theta} \in \Theta} \eta_{\mathrm{a},\hat{\theta}} \mu_t(\hat{\theta}) \tag{20}$$

Together, this (decision) rule and (acquisition) rule of Proposition 5 fully identify the optimal active sensing strategy.

3.2. Inverse Active Sensing

In the opposite direction, the inverse active sensing ("IAS") problem translates (observed) behavior back to (unobserved) preferences. We should be precise with semantics: Neither are we *assuming* that real-world agents indeed act a certain way, nor are we *prescribing* that they act optimally. Instead, our objective is thoroughly *descriptive*: Based on an agent's behavior, what do they appear (ceteris paribus) to effectively prioritize? For instance, what diseases are more important to diagnose correctly, and which tests are being over-prescribed? While actual behavior may not be induced (explicitly) by conscious optimization with respect to preferences (e.g. bounded rationality), we wish to understand them in terms of what is—in effect—prioritized (implicitly).

Inverse Optimization. We approach IAS from an inverse optimization ("IO") perspective with multiple observations. Broadly, IO deals with finding an objective function to best explain a set of observations (Bärmann et al., 2017; Dong et al., 2018), and is applicable to a wide variety of underlying problems. Specifically, an objective is determined by a (fixed) *parameter* and a (variable) *signal*. Different sig-

Table 3. Examples of inverse optimization problems with respect to various underlying settings. A generic formulation is shown for each example category. [1] Bärmann et al. (2017), [2] Dempe & Lohse (2006), [3] Dong et al. (2018), [4] Keshavarz et al. (2011), [5] Heuberger (2004), [6] Krumke et al. (1998), [7] Yang & Zhang (1999), [8] Ahmadian et al. (2018), [9] Abbeel & Ng (2004), [10] Ziebart et al. (2008). Notation: dimensions m, n, k indicate arbitrary, problem-dependent dimensions; G, D respectively denote graphs and digraphs with vertices V and edges E; $\mathcal{S}_G$ is the set of spanning trees of G and $\mathcal{P}_D$ is the set of paths in D; ϕ are known basis features, and w their respective weights.

Framework	Problem Class	Objective Function	Signal	Parameter	Response
Bärmann,[1] Dempe,[2] etc.	Inverse Linear Optimization	$b_1^\top x, Ax = b_2, x \geq 0$	$A \in \mathbb{R}^{m\times n}$	$b \in \mathbb{R}^{m+n}$	$x \in \mathbb{R}^n$
Dong,[3] Keshavarz,[4] etc.	Inverse Convex Optimization	$f(a,b,x), g(a,b,x) \leq 0$	$a \in \mathbb{R}^m$	$b \in \mathbb{R}^k$	$x \in \mathbb{R}^n$
Heuberger,[5] Krumke,[6] etc.	Inverse Minimum Spanning Tree	$c^\top A_\psi, \psi \in \mathcal{S}_G$	$G = \langle V, E\rangle$	$c \in \mathbb{R}^E$	$\psi \subseteq E$
Yang,[7] Ahmadian,[8] etc.	Inverse Integral Shortest Paths	$c^\top A_\xi, \xi \in \mathcal{P}_D$	$D; s, t \in V$	$c \in \mathbb{R}^E$	$\xi_{(s,t)} \subseteq E$
Abbeel,[9] Ziebart,[10] etc.	Inverse Reinforcement Learning	$\sum_{t=0}^\infty \mathbb{E}[\gamma^t w^\top \phi(S_t)]$	$S_{0:T}$	$w \in \mathbb{R}^{\lvert\Phi\rvert}$	$A_{0:T}$
(Ours)	Inverse Active Sensing	$\mathbb{E}[\ell(\lambda_{0:\tau-1}, \tau, \hat{\theta}; \eta)]$	$\langle \omega_{1:\tau-1}, \mu_0\rangle$	$\eta, \rho \in \mathcal{H}\times\mathbb{R}$	$\langle \lambda_{0:\tau-1}, \tau, \hat{\theta}\rangle$

nals induce different instances of the optimization problem, hence different *responses* (i.e. solutions) from an optimizing agent. Given a collection of (observed) signal-response pairs, we seek to infer the (unobserved) parameter. Table 3 sets out classic examples of this paradigm—along with IAS.

Three considerations drive our approach:

1. **Uncertainty**. Almost always, more than one configuration will accord with observed behavior. In other settings where the focus is typically on the would-be *performance* of acting under the recovered objective, this is hardly an issue. In contrast, our focus is on *understanding* drivers of behavior, so extracting a single configuration is of limited utility. Ideally, we wish to work with a distribution.
2. **Bounded Rationality**. Instead of conditioning purely on Bayes-optimal strategies, we may additionally consider other well-studied objectives in behavioral literature—such as greedy look-ahead (Najemnik & Geisler, 2005) or information-maximizing (Butko & Movellan, 2010) strategies. We may even wish to compare how well different classes of strategies describe observed behavior.
3. **Imperfect Response**. Even if we concede a known class of strategies, observed responses are often imperfect due to compliance, measurement noise, implementation error, and model uncertainty (Aswani et al., 2018; Esfahani et al., 2018). We would like to account (probabilistically) for the fact that the observed response to a signal may deviate from the perfect (i.e. objective-maximizing) choice.

Posterior Inference. We consider a *Bayesian* approach to IAS; this accommodates [1]. Compacting notation, first let $\tilde{\lambda}_t \in \Lambda \cup \Omega$ denote either λ_t (prior to stopping) or $\hat{\theta} \in \Omega$. Likewise, let $\tilde{\omega}_t \in \Omega \cup \{\varnothing\}$ indicate ω_t or $\omega_\tau = \varnothing$. Then

$$\mathcal{D} \doteq \{(\tilde{\lambda}_{n,t}, \tilde{\omega}_{n,t+1})_{t=0}^{\tau_n - 1}\}_{n=1}^N \tag{21}$$

denotes a collection of acquisitions and outcomes, with decision episodes indexed $n \in \{1, ..., N\}$. Let π be drawn from some prior $\mathbb{P}\{\pi\}$ over the space of strategies $\mathcal{P}$. Then

$$\mathbb{P}_{p,q}\{\pi|\mathcal{D}\} = \frac{\mathbb{P}_{p,q}\{\mathcal{D}|\pi\}\mathbb{P}\{\pi\}}{\int_{\mathcal{P}} \mathbb{P}_{p,q}\{\mathcal{D}|\pi\} d\mathbb{P}\{\pi\}} \tag{22}$$

is the posterior over strategies—given the observed decision behavior. Next, we specify what each strategy π consists in.

Behavioral Strategies. Most of the time, what we aim to do is inverse *optimal* active sensing (cf. inverse *optimization*) —that is, to locate preferences most consistent with an agent making Bayes-optimal acquisitions. But what if we want to accommodate a different criterion? Humans may act myopically (e.g. greedy lookahead), pursue approximate goals (e.g. infomax), or simply follow rulebooks (e.g. attribute-wise scoring schemes common in hiring and medical settings). For this, we need the notion of *generalized* Q-factors.

Let a strategy be characterized by its set of (not necessarily Bayes-optimal) factors $\{Q_\lambda^\kappa : \Delta(\Theta)\times\{0,1\}\to\mathbb{R}_+\}_{\lambda\in\Lambda}$, where κ identifies the sensing criterion; this accommodates [2]. Denote with $\tilde{Q}_\lambda^\kappa$ either Q_λ^κ or $\bar{Q}_{\hat{\theta}}$; so $\tilde{Q}_{\tilde{\lambda}_t}^\kappa(\mu_t, \nu_t; \eta)$ encodes the *desirability* of $\tilde{\lambda}_t$ under μ_t, ν_t given η. For instance, the greedy look-ahead criterion ("GL") simply corresponds to:

$$Q_{\lambda_t}^{\mathrm{GL}}(\mu_t, \nu_t; \eta) \doteq \eta_{\mathrm{c},\lambda_t} c_{\lambda_t} + g(\eta_{\mathrm{d}}) + \mathbb{E}_{p,q}[\bar{Q}(\mu_{t+1}, \nu_{t+1}; \eta)|\lambda_t, \mu_t, \nu_t] \tag{23}$$

where g is some function of decision-threshold parameters $\eta_{\mathrm{d}} \in \mathbb{R}_+^{|\Theta|}$, and here $\eta \doteq (\eta_{\mathrm{b}}, \eta_{\mathrm{c}}, \eta_{\mathrm{d}})$. As an arbitrary functional of η, we can likewise encode any criteria of choice, such as infomax Q_λ^{IM} by way of (weighted) entropy, and (of course) the Bayes-optimal case Q_λ^*. Given some $\tilde{Q}^\kappa \doteq \{\tilde{Q}_{\tilde{\lambda}}^\kappa\}_{\tilde{\lambda}\in\Lambda\cup\Theta}$, we consider the *Boltzmann* behavioral strategy with inverse temperature ρ; this accommodates [3]:

$$\pi_\rho^\kappa(\tilde{\lambda}_t|\mu_t, \nu_t; \eta) \doteq \frac{\exp(-\rho\tilde{Q}_{\tilde{\lambda}_t}^\kappa(\mu_t, \nu_t; \eta))}{\sum_{\tilde{\lambda}'_t\in\Lambda\cup\Theta}\exp(-\rho\tilde{Q}_{\tilde{\lambda}'_t}^\kappa(\mu_t, \nu_t; \eta))} \tag{24}$$

Formally, then, a strategy π is specified by (κ, η, ρ). If we restrict our attention to known classes $\kappa \in \{\mathrm{GL}, \mathrm{IM}, *, ...\}$, then the prior $\mathbb{P}\{\pi\}$ is equivalently captured by $\mathbb{P}\{\kappa, \eta, \rho\} = \mathbb{P}\{\kappa\}\mathbb{P}\{\eta|\kappa\}\mathbb{P}\{\rho\}$. (The conditioning on κ accommodates a single global space of preference weights for different κ).

Now, $\mathcal{D}$ depends on the dynamics of both the decision problem and recognition model. However, the latter involves no uncertainty (we impose a Bayesian recognition model), and the former simply drops out when evaluating the posterior:

Proposition 6 (Strategy Posterior) The posterior $\mathbb{P}\{\pi|\mathcal{D}\}$ over $\mathcal{P}$ (Equation 22) satisfies the following proportionality:

$$\mathbb{P}\{\pi^{\kappa}_{\rho}(\ldots;\eta)|\mathcal{D}\} \propto \mathbb{P}\{\kappa\}\mathbb{P}\{\eta|\kappa\}\mathbb{P}\{\rho\} \cdot \prod_{n=1}^{N}\prod_{t=0}^{\tau_n-1}\pi^{\kappa}_{\rho}(\tilde{\lambda}_{n,t}|\mu_{n,t},\nu_{n,t};\eta) \tag{25}$$

where $\mu_{n,t}$ is recursively computed via update M, $\nu_{n,t}=1$ prior to stopping, and $\pi^{\kappa}_{\rho}(\ldots;\eta)$ is defined as in Equation 24.

Proof. Appendix C. □

Note that setting $\mathbb{P}\{\kappa\}$ as the Dirac delta centered on the Bayes-optimal criterion recovers the case of inverse *optimal* active sensing; this is the formulation presented in Table 3.

Maximum A Posteriori. Seeking the MAP estimate effectively reduces IAS to a (posterior) optimization problem. Let $\mathcal{H}=\mathbb{R}^d$ denote the d-dimensional (global) space of η, and $\mathcal{K}$ the space of κ. Then the MAP estimate is given as follows, where $\mathrm{LSE}[\cdot]$ denotes logsumexp and $\mathcal{P} \doteq \mathcal{K}\times\mathcal{H}\times\mathbb{R}$:

$$\begin{aligned}\operatorname*{argmax}_{(\kappa,\eta,\rho)\in\mathcal{P}} \Big\{ &\log\mathbb{P}\{k\} - \log\mathbb{P}\{\eta|\kappa\} - \log\mathbb{P}\{\rho\} \\ &- \textstyle\sum_{n=1}^{N}\sum_{t=0}^{\tau_n-1}\big(\rho\tilde{Q}^{\kappa}_{\tilde{\lambda}_{n,t}}(\mu_{n,t},\nu_{n,t};\eta) \\ &+ \mathrm{LSE}_{\tilde{\lambda}'_{n,t}\in\Lambda\cup\Theta}[-\rho\tilde{Q}^{\kappa}_{\tilde{\lambda}_{n,t}}(\mu_{n,t},\nu_{n,t};\eta)]\big)\Big\}\end{aligned} \tag{26}$$

Given some (finite) set of known κ's, we can simply compute the MAP for each (over $\mathcal{H}\times\mathbb{R}$), then compare over $\mathcal{K}$ using $\mathbb{P}\{\kappa\}$. This can be done via standard gradient methods or numerical optimization. Using Bayes-optimal strategies, it is easy to show differentiability with respect to η and ρ:

Proposition 7 (Differentiable Posterior) Assuming differentiable priors $\mathbb{P}\{\eta|*\}, \mathbb{P}\{\rho\}$, the posterior $\mathbb{P}\{\eta,\rho|*,\mathcal{D}\}$ for optimal strategies is differentiable (almost everywhere).

Proof. Appendix C. □

Sampling from Posterior. Instead of a point estimate, we can generate samples from the posterior for each κ. MCMC sampling is common in inverse problem settings (Ye et al., 2019; Bardsley & Fox, 2012; Ramachandran & Amir, 2007). We perform geometric random walks over coordinates of a lattice in $\mathcal{H}\times\mathbb{R}$ (Frieze et al., 1994; Applegate et al., 1990) to yield samples from posterior $\mathbb{P}\{\eta,\rho|\kappa,\mathcal{D}\}$ (Algorithm 1). Consider a discrete subset $\mathcal{L}$ of $\mathbb{R}^{d+1}$ comprising coordinates that are integer multiples of a chosen resolution. The algorithm simply tries to move to one of its neighbors $\mathcal{N}$ at each step, with acceptance ratios determined by posteriors.

Algorithm 1 Posterior Sampler for IAS

Input: Decision behavior $\mathcal{D}$ and priors $\mathbb{P}\{\eta|\kappa\}, \mathbb{P}\{\rho\}$
Randomly select $(\eta,\rho)_0 \in \mathcal{L}$
$\tilde{Q}_0 \leftarrow \texttt{ActiveSensing}(\kappa,\eta_0)$
for $i = 1, \ldots$ **do**
  Randomly select $(\eta,\rho)' \in \mathcal{N}((\eta,\rho)_{i-1})$ ▷ neighbor
  $\tilde{Q}' \leftarrow \texttt{ActiveSensing}(\kappa,\eta')$
  $R \leftarrow \mathbb{P}\{(\eta,\rho)'|\kappa,\mathcal{D}\}/\mathbb{P}\{(\eta,\rho)_{i-1}|\kappa,\mathcal{D}\}$
  w.p. $\min\{1,R\}$ **do** $(\eta,\rho)_i \leftarrow (\eta,\rho)'$ ▷ accept
  otherwise $(\eta,\rho)_i \leftarrow (\eta,\rho)_{i-1}$ ▷ reject
Output: Estimate of posterior $\hat{\mathbb{P}}\{\eta,\rho|\kappa,\mathcal{D}\}$

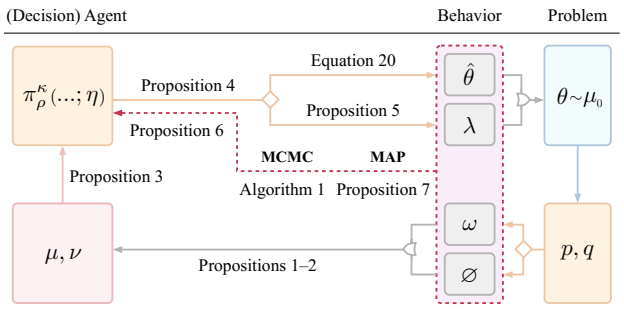


Figure 2. Active sensing and inverse active sensing (redux). Forward direction shown in clockwise solid, and inverse in purple dashed. Using a Bayesian recognition model (Propositions 1–2), the optimal value is computable (Proposition 3) via dynamic programming to obtain the optimal strategy (Propositions 4–5, Equation 20). In IAS, the strategy can be inferred (Proposition 6) via MAP estimation (Proposition 7) or via MCMC sampling (Algorithm 1).

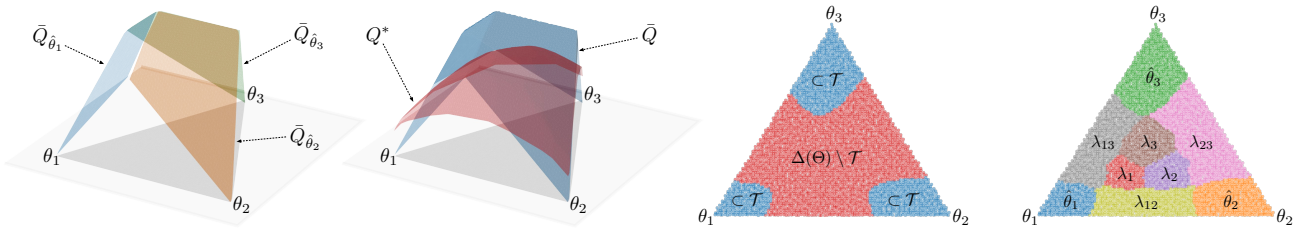
In a nutshell, we summarize the entire framework by harking back to Figure 1. We now have all the tools for IAS: Figure 2 shows a map of our key developments in both the forward (clockwise solid) and inverse (purple dashed) directions.

4. Illustrative Examples

We show archetypical examples that exercise our framework through numerical simulation. Examples 1–2 give intuition for optimal active sensing, and 3–5 exemplify potential applications of IAS. Due to space limitation, commentary is necessarily brief; Appendix A gives more context and detail.

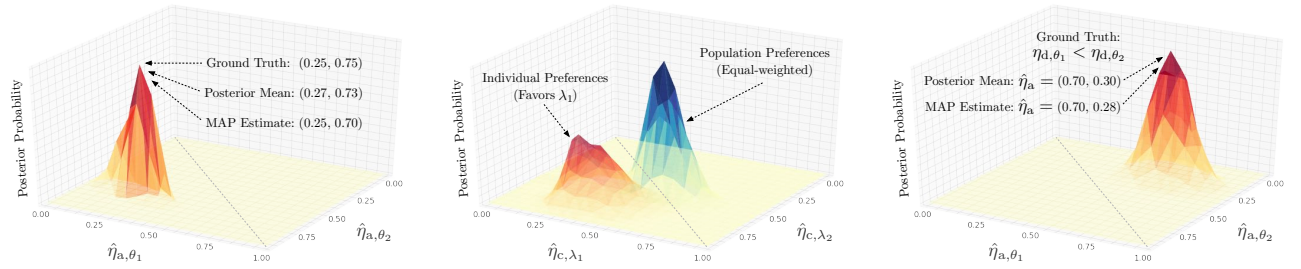
Example 1 (Ternary Hypothesis) We first give *geometric* intuition for the forward problem—exercising Propositions 4–5 and Equation 20. Consider the ternary hypothesis space $\Theta = \{\theta_1,\theta_2,\theta_3\}$, where the decision-maker is equipped with unary tests $\lambda_1,\lambda_2,\lambda_3$ (each probabilistically confirming or denying an individual hypothesis) and binary tests $\lambda_{12},\lambda_{23},\lambda_{13}$ (each probabilistically distinguishing between pairs of hypotheses). Figures 3(a)–(d) depict the output of (optimal) active sensing (by dynamic programming, cf. Proposition 3) in the posterior simplex, showing the relationships among Q-factors and their intersections, acquisitions and decisions, as well as continuation and termination sets.

Example 2 (Decision Tree) We illustrate *belief trajectories* for a common class of decision problems. Consider a medical diagnosis setting where the disease space $\Theta = \{\theta_1,\theta_2,\theta_3,\theta_4\}$ is arranged in a hierarchy, and where the diagnostic agent has access to a (top-level) test λ_0 that (probabilistically) distinguishes between the groups $\{\theta_1,\theta_2\}$ vs. $\{\theta_3,\theta_4\}$ (and are otherwise uninformative), and (level-2) tests λ_{12} and λ_{34} that respectively (probabilistically) distinguish between θ_1 vs. θ_2, and θ_3 vs. θ_4 (and are otherwise uninformative). Figures 3(e)–(g) visualize episodes for different cost-sensitivity preferences η_c, as well as verifying our intuition that the optimal strategy navigates *down* the tree.



(a) Q-factors for Decisions (b) Intersection of Q^* and $\bar{Q}$ (c) Continue vs. Terminate (d) Acquisitions vs. Decisions

Figure 3. Optimal active sensing. Posterior simplex (before deadline) for Example 1. (a) Aggregate Q-factor for decisions, cf. Equation 20. (b) Intersection of aggregate Q-factors for acquisitions and decisions, cf. Proposition 4. (c) Continuation and termination sets; the latter comprises convex regions around vertices, where $Q^* \geq \bar{Q}$; the former is its complement. (d) Complete strategy map; continuation and termination sets are partitioned into individual acquisition and decision regions, cf. Proposition 5, Equation 20 (*continued on page 9*).



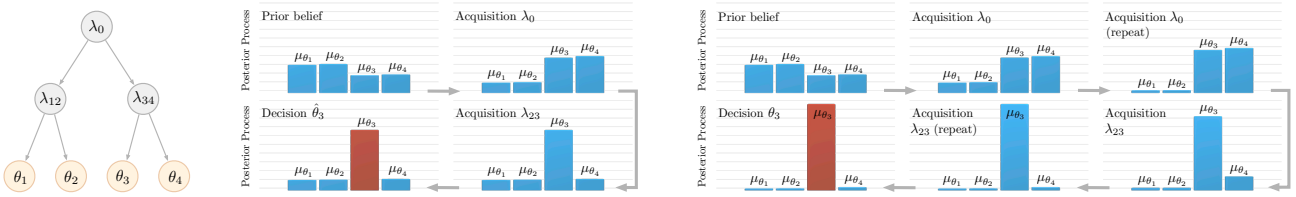
(a) Recovering Distribution of Preferences (b) Comparing Distributions of Preferences (c) Understanding Effective Preferences

Figure 4. Inverse active sensing. (Relevant dimensions of) posteriors $\mathbb{P}\{\eta, \rho | *, \mathcal{D}\}$ for Examples 3–5 (cf. Proposition 6). Each distribution is generated as 1000 MCMC samples, cf. Algorithm 1; MAP estimates are per Equation 26. $N = 300$ episodes are simulated for optimal softmax agents in (a) and the (biased) "individual" agent in (b), and $N = 1000$ for the (unbiased) "population" agent in (b); a greedy lookahead softmax agent ($N = 300$) is used in (c). Uniform priors $\mathbb{P}\{\eta | *\}$ and $\mathbb{P}\{\rho\}$ are employed in all instances (*continued on page 9*).

Example 3 (Differential Importance) We show an archetypical application of IAS in recovering preferences from behavior—exercising Propositions 6–7 and Algorithm 1. Consider a *preoperative testing* problem, where the aim is to confirm (θ_1) or deny (θ_2) the absence of comorbidities that may complicate surgery. Given the downside risk, we certainly hope to verify that Type I errors are taken more seriously than Type II (i.e. accuracy weights $\eta_{a,\theta_2} > \eta_{a,\theta_1}$). Suppose that (unbeknownst to us) this is true in practice: we simulate a random collection $\mathcal{D}$ of decision episodes for a Bayes-optimal softmax decision agent driven by $\eta_a = (0.25, 0.75)$. Figure 4(a) depicts the output of our MAP and MCMC solutions, showing (relevant dimensions of) recovered estimates for optimal softmax strategies, along with the true weights.

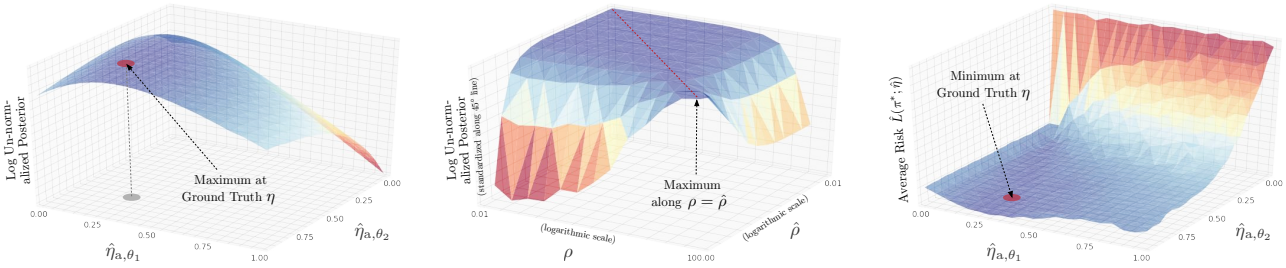
Example 4 (Differential Treatment) We highlight the applicability of IAS in comparing preferences *across* different agents or populations. Consider the problem of detecting the phenomenon of *prescription bias* with respect to two different diagnostic tests (λ_1, λ_2) for the same disease. Absent bias, by definition it must be the case that all $\eta_{c,\lambda}$ take on identical values. Suppose that (unbeknownst to us) one hospital secretly favors λ_1 (ceteris paribus) more than λ_2, unlike the rest of the hospital network; we simulate episodes for each accordingly. Figure 4(b) shows (relevant dimensions of) the output of Algorithm 1 and Equation 26—for both the institution in question and the population (of other institutions). The former's bias in favor of λ_1 (i.e. with cost-sensitivity weights $\eta_{c,\lambda_1} < \eta_{c,\lambda_2}$) is evident, in contrast with the (apparently) unbiased behavior of the latter ($\eta_{c,\lambda_1} \approx \eta_{c,\lambda_2}$).

Example 5 (Effective Preferences) Finally, we give an example that emphasizes the *interpretative* nature of IAS. Consider an agent who (unbeknownst to us) behaves myopically (cf. greedy lookahead), with a higher decision-threshold parameter for one hypothesis (i.e. $\eta_{d,\theta_1} < \eta_{d,\theta_2}$). Now, if we were just interested in the generic question of what best describes their behavior, we would simply run IAS across the entire space $\mathcal{P}$—including all classes κ of interest. In this case, for instance, we would—unsurprisingly—recover some configuration (GL, η, ρ) as MAP estimate. But suppose we are actually interested in *interpreting* their behavior *as if* they were optimal with respect to some preferences (which we wish to identify). We are now asking the question: No matter what your internal decision-making processes are, what are you *effectively* prioritizing? For instance, if the medical consensus is that λ_1 is more important to catch than λ_2, then recovering the *effective* values of η_a (via $\kappa = *$) would give an immediate assessment of this. Figure 4(c) shows (relevant dimensions of) our IAS output on the (greedily) simulated decision episodes, where we find (ceteris paribus) that thresholds $\eta_{d,\theta_1} < \eta_{d,\theta_2}$ *effectively* translate into accuracy weights $\eta_{a,\theta_1} > \eta_{a,\theta_2}$, which (in this case) accords—at least ordinally—with the medical consensus.



(e) Decision Tree (f) Typical Belief Trajectory (g) Typical Belief Trajectory with Lowered η_c

Figure 3. Optimal active sensing (continued from page 8). Example 2 considers a medical diagnosis setting with diseases $\theta_1, \theta_2, \theta_3, \theta_4$ arranged in a hierarchy (e)—each λ (probabilistically) distinguishes between its child elements. Intuitively, we expect that the optimal strategy navigate *down* the decision-tree, sequentially going from high-level tests to low-level tests before declaring specific diagnoses. Figure (f) shows a typical *belief trajectory* for the optimal strategy computed; observe from its decision-behavior that it indeed successively narrows down the space of hypotheses through the tree. Figure (g) additionally shows the effect of uniformly decreasing the cost-sensitivity parameter η_c—as expected, the optimal strategy can now afford to "double-check" each test result before committing down each branch.



(d) Preference Weights and Posteriors (e) Inverse Temperature and Posteriors (f) Preference Weights and Average Risk

Figure 4. Inverse active sensing (continued from page 8). For additional visual intuition, (d) computes the (log un-normalized) posterior in relation to (relevant dimensions of) the space of preferences η for Example 3; we observe (as expected) that the posterior is maximized at values that coincide with the ground truth. Similarly, (e) shows the (log un-normalized) posterior in relation to the inverse temperature ρ in the context of Example 3; here we explicitly simulate a range of ground-truth values and observe (as expected) that the posterior is maximal along the 45° line (to make this clearer, we standardize all probabilities along this line). Finally, in (f) we observe (as expected) that the (Bayes-optimal) strategy induced by the true parameters is in fact the strategy that achieves the lowest average (ground-truth) risk.

Discussion. In this work, we developed a unified theoretical framework for evidence-based decision-making under time pressure, illustrating how it enables modeling intuitive tradeoffs in decision strategies, and understanding behavior by quantifying preferences implicit in observed strategies.

In modeling the forward problem, our formulation of active sensing inherits several assumptions from prior literature (Ahmad & Yu, 2013; Alaa & van der Schaar, 2016; Chernoff, 1959; Dayanik & Yu, 2013; Frazier et al., 2008; Naghshvar et al., 2013)—that the spaces of decisions, acquisitions, and outcomes are given; that the distributions of deadlines, outcomes, and problem instances are known or must be appropriately estimated beforehand; and that the outcomes of acquisitions are conditionally independent over time—which is not always true depending on the type of acquisition in question, and is a shortcoming of this approach that requires extra care in practical applications. On the other hand, our framework departs from prior work by way of expressivity and specification—in accounting for the presence, endogeneity, and context-dependence of time pressure; in accommodating differential costs of acquisition and penalties for inaccuracies and deadline breaches; and in modeling preference weights directly via behavioral data without prior specification. Where need be, note that it is possible for future work to generalize the present approach to continuous outcome spaces, or disjoint outcome spaces Ω_λ per acquisition; to incorporate learnable mappings from instance-specific features $\mathbf{x}$ to priors $\mu_0(\mathbf{x})$; and to allow unknown environment parameters to be jointly estimated (although complexity may be of concern in high dimensions).

In approaching the inverse problem, we inherit a data-driven formulation of inverse optimization—i.e. where solutions to multiple problem instances are observed (Aswani et al., 2018; Bärmann et al., 2017; Dong et al., 2018; Esfahani et al., 2018). Further, our Bayesian method is similar to prior approaches in inverse problems settings (Bardsley & Fox, 2012; Ramachandran & Amir, 2007; Ye et al., 2019), and the posterior sampler bears resemblance to a Bayesian solution to inverse optimal control (Abbeel & Ng, 2004), which analogously adapts geometric random walks to the underlying parameter space. The complexity of Algorithm 1 is therefore identical; it has been shown that such a Markov chain is rapidly-mixing (i.e. sampling terminates in a polynomially-bounded number of steps) under some assumptions (Ramachandran & Amir, 2007; Applegate et al., 1990). For a more detailed survey of related literature, see Appendix B.

Acknowledgments

This work was supported by Alzheimer's Research UK (ARUK), the US Office of Naval Research (ONR), and the National Science Foundation (NSF): grant numbers 1407712, 1462245, 1524417, 1533983, 1722516. We thank all reviewers for their generous comments and suggestions.

References

Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *ICML*, 2004.

Ahmad, S. and Yu, A. J. Active sensing as bayes-optimal sequential decision making. *UAI*, 2013.

Ahmadian, S., Bhaskar, U., Sanità, L., and Swamy, C. Algorithms for inverse optimization problems. In *ESA*, 2018.

Ahuja, K., Zame, W., and van der Schaar, M. Dpscreen: Dynamic personalized screening. In *NIPS*, 2017.

Ahuja, R. K. and Orlin, J. B. Inverse optimization. *Operations Research*, 49(5):771–783, 2001.

Alaa, A. M. and van der Schaar, M. Balancing suspense and surprise: Timely decision making with endogenous information acquisition. In *NIPS*, pp. 2910–2918, 2016.

Applegate, D., Kannan, R., et al. Sampling and integration of near log-concave functions. In *ACM symposium on Theory of computing*, pp. 156–163, 1990.

Aswani, A., Shen, Z.-J., and Siddiq, A. Inverse optimization with noisy data. *Operations Research*, 66(3), 2018.

Atia, G. K. and Veeravalli, V. V. Controlled sensing for sequential multihypothesis testing. In *IEEE International Symposium on Information Theory*. IEEE, 2012.

Augenblick, N. and Rabin, M. Belief movement, uncertainty reduction, and rational updating. *Mimeo*, 2018.

Bardsley, J. M. and Fox, C. An mcmc method for uncertainty quantification in nonnegativity constrained inverse problems. *Inverse Problems in Science and Engineering*, 20(4):477–498, 2012.

Bärmann, A., Pokutta, S., and Schneider, O. Emulating the expert: inverse optimization through online learning. In *Proc. 34th ICML*, pp. 400–410. JMLR. org, 2017.

Bertsekas, D. P., Bertsekas, D. P., Bertsekas, D. P., and Bertsekas, D. P. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

Bertsimas, D., Gupta, V., and Paschalidis, I. C. Data-driven estimation in equilibrium using inverse optimization. *Mathematical Programming*, 153(2):595–633, 2015.

Blahut, R. Hypothesis testing and information theory. *IEEE Transactions on Information Theory*, 20(4), 1974.

Bock, M., Fritsch, G., and Hepner, D. L. Preoperative laboratory testing. *Anesthesiology clinics*, 34(1), 2016.

Butko, N. J. and Movellan, J. R. Infomax control of eye movements. *IEEE Transactions on Autonomous Mental Development*, 2(2):91–107, 2010.

Castro, R. and Nowak, R. Active sensing and learning. *Foundations and Applications of Sensor Mgmt*, 2009.

Chernoff, H. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770, 1959.

Clithero, J. A. Response times in economics: Looking through the lens of sequential sampling models. *Journal of Economic Psychology*, 2018.

Dayanik, S. and Yu, A. J. Reward-rate maximization in sequential identification under a stochastic deadline. *SIAM Journal on Control and Optimization*, 51(4), 2013.

DeGroot, M. H. et al. Uncertainty, information, and sequential experiments. *Annals of Mathematical Stat.*, 1962.

Dempe, S. and Lohse, S. Inverse linear programming. In *Recent Advances in Optimization*. Springer, 2006.

Dong, C., Chen, Y., and Zeng, B. Generalized inverse optimization through online learning. In *NIPS*, 2018.

Drugowitsch, J., Moreno-Bote, R., and Pouget, A. Relation between belief and performance in perceptual decision making. *PloS one*, 9(5):e96511, 2014.

Ely, J., Frankel, A., and Kamenica, E. Suspense and surprise. *Journal of Political Economy*, 123(1):215–260, 2015.

Esfahani, P. M., Shafieezadeh-Abadeh, S., et al. Data-driven inverse optimization with imperfect information. *Mathematical Programming*, 167(1):191–234, 2018.

Frazier, Peter, A. J. Y. et al. Sequential hypothesis testing under stochastic deadlines. In *Advances in neural information processing systems*, pp. 465–472, 2008.

Freedman, A. N., Seminara, D., Gail, M. H., et al. Cancer risk prediction models. *National Cancer Institute*, 2005.

Frieze, A., Kannan, R., and Polson, N. Sampling from log-concave distributions. *Annals of App. Prob.*, 1994.

Gail, M. H. Personalized estimates of breast cancer risk in clinical practice and public health. *Statistics in medicine*, 30(10):1090–1104, 2011.

Guo, S., Sanner, S., and Bonilla, E. V. Gaussian process preference elicitation. In *NeurIPS*, pp. 262–270, 2010.

Hayashi, M. Discrimination of two channels by adaptive methods and its application to quantum system. *IEEE Transactions on Information Theory*, 55(8), 2009.

Heuberger, C. Inverse combinatorial optimization. *Journal of combinatorial optimization*, 8(3):329–361, 2004.

Iyengar, G. and Kang, W. Inverse conic programming with applications. *Operations Research Letters*, 33(3), 2005.

Janisch, J. et al. Classification with costly features using deep reinforcement learning. In *AAAI*, volume 33, 2019.

Jarrett, D. and van der Schaar, M. Target-embedding autoencoders for supervised representation learning. *International Conference on Learning Representations*, 2020.

Jarrett, D., Yoon, J., and van der Schaar, M. Dynamic prediction in clinical survival analysis using temporal convolutional networks. *IEEE Journal of Biomedical and Health Informatics*, 2019.

Keshavarz, A., Wang, Y., and Boyd, S. Imputing a convex objective function. In *2011 IEEE International Symposium on Intelligent Control*, pp. 613–619. IEEE, 2011.

Krumke, S. O., Marathe, M. V., et al. Approximation algorithms for certain network improvement problems. *Journal of Combinatorial Optimization*, 2(3):257–288, 1998.

Li, A., Jin, S., Zhang, L., and Jia, Y. A sequential decision-theoretic model for medical diagnostic system. *Technology and Health Care*, 23(s1):S37–S42, 2015.

Lorden, G. Nearly-optimal sequential tests for finitely many parameter values. *The Annals of Statistics*, 1977.

Martin, S. K. and Cifu, A. S. Routine preoperative laboratory tests for elective surgery. *Jama*, 318(6), 2017.

McKinlay, J. B., Link, C. L., et al. Sources of variation in physician adherence with clinical guidelines. *Journal of general internal medicine*, 22(3):289–296, 2007.

Naghshvar, M. and Javidi, T. Information utility in active sequential hypothesis testing. In *Allerton Conference on Communication, Control, and Computing*. IEEE, 2011.

Naghshvar, M., Javidi, T., et al. Active sequential hypothesis testing. *The Annals of Statistics*, 41(6):2703–2738, 2013.

Najemnik, J. and Geisler, W. S. Optimal eye movement strategies in visual search. *Nature*, 434(7031):387, 2005.

National Center for Complementary & Integrative Health, U. S. Clinical practice guidelines, nccih.nih.gov/health/providers/clinicalpractice.htm. 2017.

National Guideline Centre, U. K. Preoperative tests (update): routine preoperative tests for elective surgery. 2016.

Nitinawarat, S., Atia, G. K., and Veeravalli, V. V. Controlled sensing for multihypothesis testing. *IEEE Transactions on Automatic Control*, 58(10):2451–2464, 2013.

Petousis, P., Han, S. X., Hsu, W., and Bui, A. A. Generating reward functions using irl towards individualized cancer screening. In *International Workshop on Artificial Intelligence in Health*, pp. 213–227. Springer, 2018.

Polyanskiy, Y. and Verdu, S. Binary hypothesis testing with feedback. In *Information Theory and Applications*, 2011.

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

Schaefer, A. J. Inverse integer programming. *Optimization Letters*, 3(4):483–489, 2009.

Schroeder, C. E., Wilson, D. A., Radman, T., Scharfman, H., and Lakatos, P. Dynamics of active sensing and perceptual selection. *Current opinion in neurobiology*, 20(2):172–176, 2010.

Smedley, N. F., Chau, N., Petruse, A., and Hsu, W. A platform for generating and validating breast risk models from clinical data. *age*, 14(12):13, 2011.

Song, Y., Skinner, J., Bynum, J., Sutherland, J., Wennberg, J. E., and Fisher, E. S. Regional variations in diagnostic practices. *New England Journal of Medicine*, (1), 2010.

Tammemägi, M. C., Katki, H. A., Hocking, W. G., et al. Selection criteria for lung-cancer screening. *New England Journal of Medicine*, 368(8):728–736, 2013.

Tuncel, E. On error exponents in hypothesis testing. *IEEE Transactions on Information Theory*, 51(8), 2005.

Vendrov, I., Lu, T., Huang, Q., and Boutilier, C. Gradient-based optimization for bayesian preference elicitation. *AAAI Conference on Artificial Intelligence*, 2020.

Wald, A., Wolfowitz, J., et al. Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, 19(3):326–339, 1948.

Wray, K. H. and Zilberstein, S. A pomdp formulation of proactive learning. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Yang, C. and Zhang, J. Two general methods for inverse optimization problems. *Applied Mathematics Letters*, 12 (2):69–72, 1999.

Yang, S. C.-H., Wolpert, D. M., and Lengyel, M. Theoretical perspectives on active sensing. *Current opinion in behavioral sciences*, 11:100–108, 2018.

Ye, N., Roosta-Khorasani, F., and Cui, T. Optimization methods for inverse problems. In *Matrix Annals 2017*, pp. 121–140. Springer, 2019.

Yoon, J., Zame, W. R., and van der Schaar, M. Deep sensing: Active sensing using multi-directional recurrent neural networks. In *ICLR*, 2018.

Yu, S., Krishnapuram, B., Rosales, R., and Rao, R. B. Active sensing. In *AISTATS*, pp. 639–646, 2009.

Ziebart, B. D., Maas, A., Bagnell, J. A., and Dey, A. K. Maximum entropy inverse reinforcement learning. 2008.

A. Notes on Simulations

Context for Example 1. Propositions 4–5 and Equation 20 give a theoretical characterization of optimal active sensing. The aim of this example is to visualize the geometry of the forward problem in the simplex, illustrating these various results through a non-trivial example. In addition to the main points to note in the captions to Figures 3(a)–(d), in this example we set $\eta_{\text{a},\theta_1}<\eta_{\text{a},\theta_2}<\eta_{\text{a},\theta_3}$ and likewise $\eta_{\text{b},\theta_1}<\eta_{\text{b},\theta_2}<\eta_{\text{b},\theta_3}$, where $\eta_{\text{a},\theta}<\eta_{\text{b},\theta}$ for all θ (as is often the case—for medical diagnosis, for instance—failing to make a decision before the deadline is at least as bad as making the incorrect decision); observe that this preference ordering among the hypotheses is reflected in the termination regions in Figure 3(c): The optimal strategy most readily commits to θ_3 since it is the most important to catch, whereas it can afford to be surer of θ_1 before committing to it. Finally, note that Proposition 5 operates implicitly behind Figure 3(d): In this example, we set $p_{\text{unary}}, q_{\text{unary}}$ and $p_{\text{binary}}, q_{\text{binary}}$ such that the former are more powerful but more risky, and the latter are less powerful but less risky, which induces a surprise-suspense tradeoff; note that increasing the power (or decreasing the risk) of unary tests would naturally expand the (inner) acquisition regions or $\lambda_1, \lambda_2, \lambda_3$ relative to $\lambda_{12}, \lambda_{23}, \lambda_{13}$, and vice versa in the opposite direction. (Moreover, the tradeoff in Equation 20 is similarly (but trivially) implicit in Figure 3(a): The peak of the Q-factor for decisions gravitates away from vertices with higher $\eta_{\text{a},\theta}$).

Context for Example 2. While Example 1 illustrates properties of the optimal Q-factors, Example 2 and Figure 3(e)–(g) visualizes the optimal strategy *in action* (i.e. showing typical belief trajectories) through an intuitive example from medical diagnosis. Consider the diagnostic problem with diseases $\theta_1, \theta_2, \theta_3, \theta_4$ arranged in a hierarchy as in Figure 3(e) such that each test λ probabilistically distinguishes between its child elements, which can be specific diseases, groups of diseases, or even disease stages as in progressive cognitive impairment (Jarrett et al., 2019); for real-world analogies see for instance National Guideline Centre (2016); National Center for Complementary & Integrative Health (2017). We naturally expect that the optimal strategy navigate *down* the decision-tree, starting first from high-level tests, then onto low-level tests, before finally declaring specific diagnoses of diseases. Panel (f) shows a typical *belief trajectory* for the optimal strategy; observe from its decision behavior that it indeed successively narrows down the space of hypotheses through the tree. Panel (g) additionally shows the effect of uniformly decreasing the cost-sensitivity parameter η_{c}: as expected, the optimal strategy now affords to "double-check" test results before committing to a branch.

Context for Example 3. Unlike the previous two (which serve to illustrate our results for the forward problem), this gives an archetypical example exercising the *full* framework for IAS that we have been building towards. In this case, we specifically use the problem of preoperative testing as a concrete setting, but more broadly we are simply demonstrating the central capability of IAS—that is, in understanding preferences from behavior: Given the decision-behavior of an agent acting according to unknown preferences, can we recover their preferences? To do so, here we perform inverse optimal active sensing on a simulated agent that in fact behaves as $\kappa = *$ (i.e. the model matches the behavior); in Example 5, we highlight the interpretive nature of IAS through a more general example (where there is a mismatch). First, we simulate a collection $\mathcal{D}$ of 300 decision episodes for a Bayes-optimal softmax agent with access to a single preoperative test for surgery-complicating comorbidities. The agent is driven by $\eta_{\text{a}} = (0.25, 0.75)$; that is, Type I errors are taken more seriously than Type II errors—but this is (of course) unknown from the IAS point of view, and the pretext is that we wish to estimate η_{a} from $\mathcal{D}$. Complete IAS (cf. Proposition 6) would yield an estimate for the full tuple (κ, η, ρ); in Figure 4(a) we show dimensions of the result for $\kappa = *$ relevant to this example. The MAP estimate is computed as Equation 26, and the posterior as Algorithm 1. For additional visual intuition, Figures 4(d)–(f) depict the (log un-normalized) posterior probabilities in relation to values of η and ρ in this example, and also verify numerically—through 10,000 random episodes—that the (Bayes-optimal) strategy induced by the true parameter values is in fact the strategy with the lowest average (ground-truth) risk.

Context for Example 4. Clearly IAS allows analyzing preference weights *within* a decision-agent (i.e. differential importances)—that is our objective from the beginning. However, we are often also interested in comparing preference weights *across* agents and/or populations. In the case of healthcare, for instance, current diagnostic guidelines are largely based only on consensus (Martin & Cifu, 2017), with remarkable physician-, provider-, and population-level variability in clinical practice even among routine procedures (Song et al., 2010), which may incur significant harms and costs (Bock et al., 2016). This example illustrates the potential use of IAS in assessing such differences in behavior. As a concrete setting, consider the phenomenon of *prescription bias* w.r.t. two different diagnostic tests (λ_1, λ_2) for the same disease. Using our timely decision-making framework, prescription bias is naturally defined, simulated, and detected as inequalities between $\eta_{\text{c},\lambda}$ for different λ. Ceteris paribus, we simulate the presence of bias in an "individual" institution of of interest (via 300 trajectories driven by cost-sensitivity weights $\eta_{\text{c},\lambda_1}<\eta_{\text{c},\lambda_2}$); similarly, we simulate the absence of bias in the broader "population" (via 1000 trajectories driven by $\eta_{\text{c},\lambda_1}\approx\eta_{\text{c},\lambda_2}$). (Two runs of) IAS would yield estimates (κ, η, ρ) each for the individual and population parameters; in Figure 4(b) we show relevant dimensions of the results for $\kappa = *$, where we observe the apparent deviation of the individual's preferences from that of the population.

Context for Example 5. While Examples 3–4 show the result of IAS with $\kappa = *$ on an agent that behaves as $\kappa = *$, here we emphasize the *interpretive* nature of IAS for understanding decision-making behavior through a more general example—where there is a mismatch. Of course, the (obvious) caveat here—as in any parameter estimation problem—is that the mismatch cannot be too large. Clearly a complete mismatch would yield nonsensical results in IAS: consider a strategy that simply selects acquisitions and decisions uniformly at random. In practice, however, while there may be a range of (active sensing) decision-making behaviors in the world, we generally expect that they be (somewhat imperfect) approximations to the optimal strategy. For instance, the acquisition behavior induced by the greedy generalized Q-factor (Equation 23) can be seen as a one-step approximation to Q^*_λ where (apart from the soft decision-threshold) V^* is simply replaced by $\bar{Q}$. Figure 4(c) shows what happens when we interpret behavior (unbeknownst to us) generated as $\kappa = \text{GL}$, in terms of the *effective* preferences under $\kappa = *$—namely, that (ceteris paribus) greedy look-ahead behavior driven by $\eta_{\text{d},\theta_1} < \eta_{\text{d},\theta_2}$ is roughly equivalent to $\eta_{\text{a},\theta_1} > \eta_{\text{a},\theta_2}$. This (perhaps obvious) point is worth belaboring—that is, while decision agents may not necessarily be optimal in practice, this has little bearing on the fact that inverse optimal active sensing can still be able to provide a common yardstick by which different decision behaviors can be quantified and compared.

Computation. For all examples, agents are simulated with inverse temperature $\rho = 10$. The precise setting is unimportant, and we observe that similar results obtain for an order of magnitude larger or smaller; however, note that very large values result in more deterministic behavior, which may not be realistic ($\rho = \infty$ gives fully-deterministic strategies), and very small values result in more random behavior, which may result in difficulties in parameter estimation ($\rho = 0$ gives strategies that are completely random). For MCMC, we choose the lattice given by the union of $\mathcal{G}_\eta \cap [0,1]^d \in \mathcal{H}$ and $\mathcal{G}_\rho \in \mathbb{R}$, where $\mathcal{G}_\eta \doteq \{x : x_j \text{ is an integer multiple of } r\}$ with $r = 0.05$ being our choice of resolution for the elements of η (and j being the index into elements of x), and where $\mathcal{G}_\rho \doteq \{0.01, 0.03, ..., 30, 100\}$ is the set of roughly (logarithmically) uniformly-spaced values for ρ. Note that restricting the values of η to $[0,1]$ by itself involves no loss of expressivity, since different values of ρ are equivalent to a scaling of the Q-factors, which (by linearity of expectations) is equivalent to a scaling of all elements of η. What does have an effect on expressivity is the choice of resolution r; now, our goal is to understand the *relative* magnitudes of preference weights underlying decision behaviors, and setting $r = 0.05$ with the $[0,1]$ bounds means that we can already represent relative importance weights taking on values up to a maximum of 20 times each another. (In practice, if IAS still returns estimates with elements at opposing boundaries of the lattice, this may indicate that we need to further increase the resolution—e.g. by setting $r = 0.01$, which would allow representing relative importance weights up to 100 times one another). For each inverse example, the posterior distributions (using uniform priors) are generated as 1000 samples; with 300 initial "burn-in" samples discarded.

Modeling Priors. We briefly mention here a point for (more applied) future work. In this paper we focus on developing a theoretical framework and demonstrating archetypical examples for modeling and understanding timely decision-making behavior. Therefore we do not concern ourselves with the (separate but related) problem of obtaining or modeling the priors μ_0 themselves. Recall from Section 2.1 that we simply take it that μ_0 for a given problem instance is available from an agent's experience, medical literature, etc. (Again, however, bear in mind the interpretive nature of IAS: we are *not* effectively assuming that decision-makers themselves possess such exact and common knowledge). In our numerical examples, we simulate episodes for $\mathcal{D}$ with μ_0 uniformly randomly scattered throughout the simplex. In practical applications with real-world input data, we probably wish to model μ_0 based on additional input (clearly, having a single constant prior may not provide nearly enough variation for meaningful estimation of preference weights). Any such model necessarily depends on the specific context; however, while we defer this topic to future work with a more applied focus, we note that in many cases existing domain-specific models (such as those in medicine) can be more or less adapted for this purpose. See Petousis et al. (2018) for an example where such models are deployed for modeling initial beliefs also in an inverse setting (although with a very different approach, detailed in the next section). In the context of medical diagnosis, for instance, one can consider a rich literature of feature-based models (Freedman et al., 2005), including the widely used and validated Tammemägi and Gail risk models (Tammemägi et al., 2013; Gail, 2011; Smedley et al., 2011) for lung cancer and breast cancer, which can consider a variety of baseline features such as age, race, body mass, smoking status, family history, and previous biopsies in generating accurate priors for use.

B. Related Work

In this paper, we develop an expressive theoretical framework for evidence-based decision-making under time pressure, and illustrate how it enables modeling and understanding decision behavior via optimal and inverse active sensing. As such, it lends itself to contextualization within broader notions of both the forward and inverse problem settings. While relevant works have been noted throughout the manuscript, here we provide a more detailed overview.

Active Sensing. In the broadest sense, active sensing refers to the general process of directing one's attention towards ex-

Table 4. Comparison with related work in sequential analysis. Viewed from the perspective of sequential analysis, our decision problem can be framed as one of active multiple-hypothesis testing via adaptive and sequential sensing in the presence stochastic, endogenous, and context-dependent time pressure. An exemplary work is shown for each category. Importantly, we focus on the significance of *subjective preferences*, and develop a most general framework accommodating both *forward* (i.e. modeling) & *inverse* (i.e. understanding) problems.

Literature	Acquisition	Decision	Strategy	Evidence	Costs	Horizon	Deadline	Problem
Wald et al. (1948)	Passive	Binary	-	Sequential	Fixed	No	-	Forward
Blahut (1974)	Passive	Binary	-	Batch	Fixed	No	-	Forward
Bertsekas et al. (1995)	Passive	Binary	-	Sequential	Fixed	Fixed	External	Forward
Frazier et al. (2008)	Passive	Binary	-	Sequential	Fixed	Stochastic	External	Forward
Lorden (1977)	Passive	Multiple	-	Sequential	Fixed	No	-	Forward
Tuncel (2005)	Passive	Multiple	-	Batch	Fixed	No	-	Forward
Dayanik & Yu (2013)	Passive	Multiple	-	Sequential	Fixed	Stochastic	External	Forward
Polyanskiy & Verdu (2011)	Active	Binary	Fixed	Sequential	Fixed	No	-	Forward
Hayashi (2009)	Active	Binary	Adaptive	Batch	Fixed	No	-	Forward
Naghshvar & Javidi (2011)	Active	Binary	Adaptive	Sequential	Fixed	No	-	Forward
Nitinawarat et al. (2013)	Active	Multiple	Fixed	Batch	Fixed	No	-	Forward
Atia & Veeravalli (2012)	Active	Multiple	Adaptive	Batch	Fixed	No	-	Forward
Naghshvar et al. (2013)	Active	Multiple	Adaptive	Sequential	Fixed	No	-	Forward
(Ours)	Active	Multiple	Adaptive	Sequential	Differential	Stochastic	Endogenous	Forward+Inverse

tracting *task-relevant* information through interaction with the world (Yang et al., 2018). This broad notion of intentional information gathering has been applied in various settings such as multi-view learning (Yu et al., 2009), sensory processing (Schroeder et al., 2010), personalized screening (Ahuja et al., 2017), time-series prediction (Yoon et al., 2018), and black-box classification (Janisch et al., 2019). While most applications focus on crafting function approximators to optimize performance on the downstream task, our focus is instead in developing an expressive framework for modeling and understanding the decision process itself.

Timely Decision-Making. In particular, we study active sensing for the general problem of timely decision-making—that is, the goal-directed task of selecting which acquisitions to make, when to stop gathering information, and what decision to ultimately settle on. As such, it is related to the sequential identification problem in statistics (Naghshvar et al., 2013), neuroscience (Ahmad & Yu, 2013), and economics (Augenblick & Rabin, 2018)—where a hypothesis is selected following observations of relevant evidence. Starting with the seminal work on binary hypothesis testing (Wald et al., 1948), a variety of studies have aimed to characterize a range of heuristic and/or optimal strategies, with such extensions as deadline pressure (Frazier et al., 2008), incorporating active choice (Castro & Nowak, 2009), and comparisons of behavioral strategies (Ahmad & Yu, 2013). We emphasize the goal-directed nature of active sensing in general (and our timely decision-making setting): this is in contrast to pure exploration and surveillance problems, which do not involve a specific task (the decision problem).

Generalized Setting. Several key distinctions warrant special attention (see Table 4). We consider the most flexible decision-making setting: (1) acquisitions are *active*—i.e. involving choices among multiple competing sensory options; (2) strategies are *adaptive*—i.e. admitting context-dependent choices determined on the fly; and (3) samples are *sequential*—i.e. requiring a variable number of observations per the endogenous choice of stopping and issuing a decision. These distinctions are critical—for instance, if sampling were passive (e.g. single stream of observations), then the task readily reduces to the well-studied problem of optimal stopping (Frazier et al., 2008; Dayanik & Yu, 2013). Further, as motivated throughout, we additionally account for (4) *differential* costs of acquisition and the presence of (5) stochastic, *endogenous*, and *context-dependent* time pressure. Perhaps most importantly, we accommodate modeling and understanding (6) *subjective* preferences in decision behavior, and uniquely focus on *both* forward and inverse problems in our active sensing framework. Table 4 sets out a comparison with related work in sequential analysis in general, and Table 2 specifically as pertains timely decision-making. In this view, our work develops a most generalized framework to analyze both optimal and inverse problems.

Inverse Active Sensing. For the inverse direction, we approach the problem from an *inverse optimization* perspective. In general, IO turns optimization problems on their heads: Given (one or more) solutions to some problem, the goal is to infer (parameters of) the objective function (Ahuja & Orlin, 2001). IO has been applied to a broad range of underlying problems, including inverse linear (Dempe & Lohse, 2006) and integer (Schaefer, 2009) programming, inverse convex optimization (Keshavarz et al., 2011), inverse conic programming (Iyengar & Kang, 2005), and any manner of inverse combinatorial optimization problems (Heuberger, 2004). Table 3 shows inverse (optimal) active sensing alongside example formulations for some classic IO applications.

Multiple Observations. In particular, inverse active sensing can be interpreted as a form of *data-driven* IO with multiple

Table 5. Summary comparison of IAS *and* IRL. Although the two classes of IO problems share superficial resemblance from the perspective of inverse learning from multiple observations, they have vastly different goals and multiple crucial distinctions. In particular, while learning medical diagnosis behavior can be alternatively cast in IRL as a generic *apprenticeship* problem, our proposed IAS framework is much better suited for *modeling* and *understanding* the decision process itself in timely decision-making settings. [1] Petousis et al. (2018).

Approach	Markov Process	Stopping Time	Behavior Parameters	Modeling Acquisitions	Modeling Decisions	Time Pressure	Parameters Interpretable	Downstream Goal	Accuracy of Decision
IRL (Petousis)[1]	States with Transitions	Fixed	Per-State Rewards	Yes	No	No	No	Apprentice-ship	Objective, Imposed
IAS **(Ours)**	Posterior & Survival	Stochastic, Endogenous	Risk-based Preferences	Yes	Yes	Yes	Yes	Understan-ding	Subjective, Learned

observations (of solutions). Methods for data-driven IO are increasingly relevant with the exponentially growing availability of electronic patient data (Jarrett & van der Schaar, 2020), and have been studied as pertains to imperfect information (Esfahani et al., 2018) and noisy observations (Aswani et al., 2018), as well as using online learning (Bärmann et al., 2017; Dong et al., 2018). Now, a popular application of this paradigm is inverse reinforcement learning ("IRL"), which deals with inferring the reward function for a reinforcement learning agent (Abbeel & Ng, 2004; Ziebart et al., 2008). Although IRL may appear to bear resemblance to IAS, they have vastly different goals and a number of crucial distinctions. These are best highlighted by direct comparison with Petousis et al. (2018), which applies IRL for apprenticeship of expert cancer screening behavior (see Table 5). In the first instance, (1) the typical goal of IRL lies in *apprenticeship*; to that end, the central concern is in replicating some notion of ("true") performance, using (potentially black-box) reward functions as an intermediary to parameterize behavior. In contrast, in IAS the goal lies in *modeling* and *understanding* the decision process itself (in timely decision-making settings); to that end, the central concern is in recovering a (transparent) description of an agent's (subjective) preferences. This distinction becomes apparent in a number of aspects that render IRL unsuitable for our purposes. An immediate difference lies in (2) the nature of the Markov process in question: Recall that our formulation tracks a posterior process (cf. Proposition 1) over the hypothesis space, with survival itself is informative (cf. Proposition 2). Applying the IRL formulation instead as in Petousis et al. (2018), the "state space" is taken to be the space of hypotheses; the Markov process tracks where the agent him-/herself is located within the hypothesis space, and the "transitions" model the agent probabilistically moving between hypotheses over time. Now, (3) this abstraction is inherently opaque: What does it mean for the agent to "be" somewhere, and what how do the transition probabilities inform our understanding of what an agent prioritizes? This is fine simply as a mathematical intermediary to parameterize behavior, but is by no means interpretable as a vehicle for understanding behavior (see also point 5). In contrast, IAS purely focuses on the specific task of estimating preferences for understanding. Moreover, (4) these transition parameters must be concomitantly learned, which adds an (unnecessary) layer of approximation. Equally importantly, (5) in the IRL formulation (as is typical), the observed behavior is parameterized (and learned) in terms of per-state (and action) rewards, which—in timely decision-making—are *not* amenable to interpretation: What does it mean to reward the agent for being "in" a given (intermediary) hypothesis at each point in time? Again, this is fine purely as mathematical means to parameterize data (e.g. in their apprenticeship setting), but makes less sense for our purposes of understanding. Instead, we directly parameterize behaviors as importance weights assigned to inherently interpretable elements of the loss function (Equation 1). On a more technical note—but perhaps even more significantly: (6) in our framework, not only is the stopping time itself is an endogenous variable, it is modeled as a conscious choice (cf. Proposition 4); this is critical, since the ultimate decision itself is in some sense the whole point. In contrast, the IRL formulation (as is typical) employs fixed horizons, and does not accommodate modeling the conscious choice of stopping. In fact, to assess apprenticeship, the "accuracy" of their learned behavior is quantified via the post-hoc choice of equating some acquisitions to "positive" diagnoses (and others to "negative" diagnoses); accuracies (e.g. Type I and II errors) are therefore *objective* and *imposed* for evaluation. In contrast, we seek to model the entire decision process endogenously (not just acquisition behavior) via *subjective* preferences over accuracies, deadlines, and costs—which are *learned*. Last but not least is the technical distinction that (7) the contractive property of the operator $\mathbb{B}$ is not readily guaranteed in our setting (cf. Proposition 3); this is in contrast with typical reinforcement learning (and IRL) settings with fixed or infinite discounted horizons. Table 5 summarizes main distinctions between the problem classes.

Bayesian Approach. In terms of the objective, typical IO settings are chiefly concerned with notions of identifiability and optimality—that is, in recovering either some notion of a "true" parameter, or in prescribing behavior that performs "as well as" (or better than) observed solutions per the "true" parameter (this obviously includes inverse reinforcement learning). Instead, the focus of IAS is on describing and understanding observed decision behavior; thus we embrace non-identifiability—after all, we seek the *range* of strategies

and preferences that can interpret or best explain behavior (there is no single right answer). In this sense, we are more aligned with Bayesian approaches to inverse problem settings (Ye et al., 2019; Bardsley & Fox, 2012; Ramachandran & Amir, 2007), which avoid confronting the convexity assumptions of duality-based approaches (Bertsimas et al., 2015; Keshavarz et al., 2011), nor the intractability of non-convex solutions (Aswani et al., 2018; Esfahani et al., 2018).

Preference Elicitation. Finally, for completeness we note that preference elicitation is a well-studied problem in computational and social science: A range of works have approached the problem of (interactive) preference elicitation using gaussian processes (Guo et al., 2010), Markov decision processes (Wray & Zilberstein, 2016), and differentiable networks (Vendrov et al., 2020). However, these lines of work are very differet in that what is being modeled (and optimized) is the process of *explicitly* reaching out and querying user preferences efficiently—that is, the active preference elicitation task itself constitutes the forward problem. In contrast, our focus is on *implicitly* understanding strategies and preferences from observed decision behavior.

Relationship with POMDPs. Throughout this work, we have taken a "bottom-up" approach in contextualizing our developments—that is, by taking the basic case of sequential identification and "generalizing" from there, which highlights structural results specific to the timely decision-making problem. As its complement, it is equally possible to take an opposite "top-down" approach—that is, by taking the generic POMDP formalism and "specializing" from there. In particular, the timely decision-making problem can be formulated as a POMDP with $|\Theta|$ decision states plus an additional "terminal" state, with transitions from each of the former into the latter, and self-loops for all states; stepwise decomposing Equation 1 yields a "reward". For instance, for the decision tree from Example 2, the POMDP would consist of the state space $\mathcal{S} = \{\theta_0, \theta_1, \theta_2, \theta_3, \theta_4\}$ where θ_0 is absorbing, action space $\mathcal{A} = \{\lambda_0, \lambda_{12}, \lambda_{34}, \theta_1, \theta_2, \theta_3, \theta_4\}$, emission kernels that correspond to generating distributions $\{q_{\theta,\lambda}\}_{\theta\in\Theta,\lambda\in\Lambda}$, and transition kernels to $\{p_{\theta,\lambda}\}_{\theta\in\Theta,\lambda\in\Lambda}$.

In light of this correspondence to POMDPs, note that Proposition 1 follows by construction, providing an alternative proof. Note, however, that Propositions 2–5 are structural results specific to active sensing for timely decision-making; in particular, we note—analogously to the passive case of Dayanik & Yu (2013)—that Proposition 2 is not free due to the fact that this is neither a fixed-horizon nor discounted problem; likewise, concavity of Q is similar to—but not the same as—the classic PWLC result. That said, the fact that the (forward) active sensing problem can be re-cast as a POMDP does mean that we can use generic algorithms to accomplish the inner-loop `ActiveSensing` sub-procedure in Algorithm 1 (bar minor technicalities in translation, such as the fact that applying off-the-shelf POMDP solvers requires the use of some nominal discount rate $\gamma<1$ to guarantee convergence). In our simulations, we verify using implementations from http://pomdp.org/code/index.html and http://github.com/AdaCompNUS/sarsop for our examples that all results are virtually identical for any solver of choice, such as PBVI and SARSOP (with γ nominally set to 0.99).

In the inverse direction, as noted above IAS (with optimal κ) is likewise related to inverse optimal control; by casting the forward problem generically as a POMDP, solving the inverse optimal active sensing problem in our framework can be interpreted by analogy to a model-based, Bayesian solution to inverse reinforcement learning, but with partially-observable states instead, and a reward function parameterized by stepwise decomposing Equation 1; though beyond the scope of this work, it is conceivable to derive "max-margin", "max-likelihood", etc. versions of IAS (with optimal κ) in addition to the MAP and MCMC versions presented here. Finally, note that non-Bayes-optimal strategies can alternatively be modeled by defining rewards as sums of hand-crafted features, or by using "belief-dependent" POMDPs. In the former case, however, this may require more prior knowledge than we have access to, and—more importantly—may not result in an interpretable functional form amenable to comparing preferences across decision agents (a key mission objective of ours); in the latter, note that approximating the forward solution to belief-dependent POMDPs in general requires that rewards be convex in μ—which may be difficult to satisfy or verify in practice.

C. Proofs

Proposition 1 (Sufficient Statistic) Let $\nu_t \doteq \mathbb{1}_{\{\delta>t\}}$ denote the *survival* process, with initial value $\nu_0 = 1$. Then the *posterior* process $\mu_t \in \Delta(\Theta)$ is given by the following:

$$\begin{aligned}\mu_t = (1-\nu_{t-1})\mu_{t-1} + ((1-\nu_t)\bar{M}(\lambda_{t-1},\mu_{t-1}) \\ + \nu_t M(\lambda_{t-1},\mu_{t-1},\omega_t))\nu_{t-1}\end{aligned} \tag{27}$$

where the *continual* update $M : \Lambda\times\Delta(\Theta)\times\Omega \to \Delta(\Theta)$ returns a distribution assigning to element θ the probability:

$$\frac{(1-p_{\theta,\lambda_{t-1}})q_{\theta,\lambda_{t-1}}(\omega_t)\mu_{t-1}(\theta)}{\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda_{t-1}})q_{\theta',\lambda_{t-1}}(\omega_t)\mu_{t-1}(\theta')} \tag{28}$$

and where the *terminal* update $\bar{M} : \Lambda\times\Delta(\Theta) \to \Delta(\Theta)$ returns a distribution assigning to element θ the probability:

$$p_{\theta,\lambda_{t-1}}\mu_{t-1}(\theta)/\textstyle\sum_{\theta'\in\Theta}p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta') \tag{29}$$

Moreover, the sequence $(\mu_t,\nu_t)_{t=0}^{\infty}$ is a *controlled Markov process*, where the control inputs are the acquisitions λ_t.

Proof. For $\bar{M}$, we want that θ be assigned the probability:

$$\mathbb{P}_{p,q}\{\theta|\lambda_{t-1},\mu_{t-1},\nu_{t-1}=1,\nu_t=0\} \tag{30}$$

$$= \frac{\mathbb{P}_{p,q}\{\theta,\nu_t=0|\lambda_{t-1},\mu_{t-1},\nu_{t-1}=1\}}{\mathbb{P}_{p,q}\{\nu_t=0|\lambda_{t-1},\mu_{t-1},\nu_{t-1}=1\}} \tag{31}$$

$$= \frac{\mathbb{P}_p\{\nu_t = 0|\theta, \lambda_{t-1}, \nu_{t-1} = 1\}\mu_{t-1}(\theta)}{\sum_{\theta' \in \Theta} \mathbb{P}_p\{\nu_t = 0|\theta, \lambda_{t-1}, \nu_{t-1} = 1\}\mu_{t-1}(\theta)} \quad (32)$$

$$= \frac{p_{\theta,\lambda_{t-1}}\mu_{t-1}(\theta)}{\sum_{\theta' \in \Theta} p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta')} \quad (33)$$

For M, we want that θ be assigned the probability:

$$\mathbb{P}_{p,q}\{\theta|\lambda_{t-1}, \mu_{t-1}, \nu_t = 1, \omega_t\} \quad (34)$$

$$= \frac{\mathbb{P}_{p,q}\{\theta, \nu_t = 1, \omega_t|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1} = 1\}}{\mathbb{P}_{p,q}\{\nu_t = 1, \omega_t|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1} = 1\}} \quad (35)$$

$$= \mathbb{P}_p\{\theta, \nu_t = 1|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1} = 1\} \cdot \mathbb{P}_q\{\omega_t|\theta, \lambda_{t-1}, \nu_t = 1\} / \sum_{\theta' \in \Theta}(\mathbb{P}_p\{\theta', \nu_t = 1|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1} = 1\}\mathbb{P}_q\{\omega_t|\theta', \lambda_{t-1}, \nu_t = 1\}) \quad (36)$$

$$= \mathbb{P}_p\{\nu_t = 1|\theta, \lambda_{t-1}, \nu_{t-1} = 1\} \cdot \mathbb{P}_q\{\omega_t|\theta, \lambda_{t-1}, \nu_t = 1\}\mu_{t-1}(\theta) / \sum_{\theta' \in \Theta}\mathbb{P}_p\{\nu_t = 1|\theta, \lambda_{t-1}, \nu_{t-1} = 1\}\mathbb{P}_q\{\omega_t|\theta', \lambda_{t-1}, \nu_t = 1\}\mu_{t-1}(\theta) \quad (37)$$

$$= \frac{(1 - p_{\theta,\lambda_{t-1}})q_{\theta,\lambda_{t-1}}(\omega_t)\mu_{t-1}(\theta)}{\sum_{\theta' \in \Theta}(1 - p_{\theta',\lambda_{t-1}})q_{\theta',\lambda_{t-1}}(\omega_t)\mu_{t-1}(\theta')} \quad (38)$$

where we used $\mathbb{P}\{\theta|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1} = 1\} = \mu_{t-1}(\theta)$. To show this is a controlled Markov process, first note that:

$$\mathbb{P}_{p,q}\{\mu_t|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1}, \nu_t\} \quad (39)$$

$$= (1 - \nu_{t-1})\mathbb{P}_p\{\mu_t|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1} = 0, \nu_t = 0\} + ((1 - \nu_t)\mathbb{P}_p\{\mu_t|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1} = 1, \nu_t = 0\} + \nu_t\mathbb{P}_{p,q}\{\mu_t|\lambda_{t-1}, \mu_{t-1}, \nu_t = 1\})\nu_{t-1} \quad (40)$$

$$= (1 - \nu_{t-1})\mathbb{1}_{\{\mu_t = \mu_{t-1}\}} + ((1 - \nu_t)\mathbb{1}_{\{\mu_t = \bar{M}(\lambda_{t-1}, \mu_{t-1})\}} + \nu_t \frac{\mathbb{P}_{p,q}\{\mu_t, \nu_t = 1|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1} = 1\}}{\mathbb{P}_p\{\nu_t = 1|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1} = 1\}})\nu_{t-1} \quad (41)$$

$$= (1 - \nu_{t-1})\mathbb{1}_{\{\mu_t = \mu_{t-1}\}} + ((1 - \nu_t)\mathbb{1}_{\{\mu_t = \bar{M}(\lambda_{t-1}, \mu_{t-1})\}} + \nu_t \sum_{\omega'_t \in \Omega}(\mathbb{1}_{\{\mu_t = M(\lambda_{t-1}, \mu_{t-1}, \omega'_t)\}} \cdot \frac{\sum_{\theta' \in \Theta}(1 - p_{\theta',\lambda_{t-1}})q_{\theta',\lambda_{t-1}}(\omega_t)\mu_{t-1}(\theta')}{1 - \sum_{\theta' \in \Theta} p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta')}))\nu_{t-1} \quad (42)$$

Then the joint probability of the tuple is given by:

$$\mathbb{P}_{p,q}\{\mu_t, \nu_t|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1}\} \quad (43)$$

$$= \mathbb{P}_{p,q}\{\mu_t|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1}, \nu_t\} \cdot \mathbb{P}_p\{\nu_t|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1}\} \quad (44)$$

$$= (1 - \nu_{t-1})\mathbb{1}_{\{\mu_t = \mu_{t-1}\}} + ((1 - \nu_t) \cdot \mathbb{1}_{\{\mu_t = \bar{M}(\lambda_{t-1}, \mu_{t-1})\}} \sum_{\theta' \in \Theta} p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta') + \nu_t \sum_{\omega'_t \in \Omega}(\mathbb{1}_{\{\mu_t = M(\lambda_{t-1}, \mu_{t-1}, \omega'_t)\}} \cdot \sum_{\theta' \in \Theta}(1 - p_{\theta',\lambda_{t-1}})q_{\theta',\lambda_{t-1}}(\omega_t)\mu_{t-1}(\theta')))\nu_{t-1} \quad (45)$$

and for any $f : \Delta(\Theta) \times \{0, 1\} \to \mathbb{R}_+$ we have:

$$\mathbb{E}_{p,q}[f(\mu_t, \nu_t)|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1}] \quad (46)$$

$$= \mathbb{E}_{p,q}[(1 - \nu_{t-1})f(\mu_{t-1}, 0) + ((1 - \nu_t)f(\bar{M}(\lambda_{t-1}, \mu_{t-1}), 0) + \nu_t \cdot f(M(\lambda_{t-1}, \mu_{t-1}, \omega_t), 1))\nu_{t-1}|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1}] \quad (47)$$

$$= (1 - \nu_{t-1})f(\mu_{t-1}, 0) + (f(\bar{M}(\lambda_{t-1}, \mu_{t-1}), 0) \cdot \sum_{\theta' \in \Theta} p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta') + \sum_{\omega'_t \in \Omega}(f(M(\lambda_{t-1}, \mu_{t-1}, \omega_t), 1) \cdot \sum_{\theta' \in \Theta}(1 - p_{\theta',\lambda_{t-1}})q_{\theta',\lambda_{t-1}}(\omega_t)\mu_{t-1}(\theta')))\nu_{t-1} \quad (48)$$

where we used the fact that $\mathbb{P}_p\{\nu_t = 1|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1} = 1\} = 1 - \sum_{\theta' \in \Theta} p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta')$, that $\mathbb{P}_p\{\nu_t = 0|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1} = 1\} = \sum_{\theta' \in \Theta} p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta')$. Likewise, it is also trivial to see that $\mathbb{P}_p\{\nu_t = 0|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1} = 0\} = 1$, as well as $\mathbb{P}_p\{\nu_t = 1|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1} = 0\} = 0$.

Proposition 2 (Active and Passive Information) The information gleaned from (costly) acquisitions and (costless) observations of survival can be uniquely decomposed as:

$$\mu_t = \tilde{\mu}_t + \alpha_t + \beta_t \quad (49)$$

where $\tilde{\mu}_t$ is a *martingale* that captures information obtained from the (actively) acquired results, the (continual) compensator $\alpha_t = A(\mu_{t-1}, \lambda_{t-1}, \nu_{t-1}, \nu_t)$ (passively) incorporates the bias from the ongoing process *survival* (where $\alpha_0 = 0$):

$$\alpha_t(\theta) = \alpha_{t-1}(\theta) - \mu_{t-1}(\theta)\nu_{t-1}\nu_t \cdot (p_{\theta,\lambda_{t-1}} - \bar{p}_{\mu_t,\lambda_{t-1}})/(1 - \bar{p}_{\mu_t,\lambda_{t-1}}) \quad (50)$$

and $\beta_t = B(\mu_{t-1}, \lambda_{t-1}, \nu_{t-1}, \nu_t)$ is the (terminal) compensator that analogously incorporates the bias from process *stoppage* (where $\beta_0 = 0$)—if the deadline were breached:

$$\beta_t(\theta) = \beta_{t-1}(\theta) + \mu_{t-1}(\theta)\nu_{t-1}(1 - \nu_t) \cdot (p_{\theta,\lambda_{t-1}} - \bar{p}_{\mu_t,\lambda_{t-1}})/\bar{p}_{\mu_t,\lambda_{t-1}} \quad (51)$$

where for brevity we denote the weighted average posterior probability of failure $\bar{p}_{\mu_t,\lambda_{t-1}} \doteq \sum_{\theta' \in \Theta} p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta')$.

Proof. First, writing out the expectation:

$$\mathbb{E}_{p,q}[\mu_t|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1}, \nu_t] \quad (52)$$

$$= \mathbb{E}_{p,q}[(1 - \nu_{t-1})\mu_{t-1} + ((1 - \nu_t)\bar{M}(\lambda_{t-1}, \mu_{t-1}) + \nu_t M(\lambda_{t-1}, \mu_{t-1}, \omega_t))\nu_{t-1}|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1}, \nu_t] \quad (53)$$

$$= (1 - \nu_{t-1})\mu_{t-1} + ((1 - \nu_t)\bar{M}(\lambda_{t-1}, \mu_{t-1}) + \nu_t \sum_{\omega'_t \in \Omega}(M(\lambda_{t-1}, \mu_{t-1}, \omega'_t) \cdot \frac{\sum_{\theta' \in \Theta}(1 - p_{\theta',\lambda_{t-1}})q_{\theta',\lambda_{t-1}}(\omega_t)\mu_{t-1}(\theta')}{1 - \sum_{\theta' \in \Theta} p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta')}))\nu_{t-1} \quad (54)$$

Then for element θ, this is equal to:

$$(1 - \nu_{t-1})\mu_{t-1} + ((1 - \nu_t)\frac{p_{\theta,\lambda_{t-1}}\mu_{t-1}(\theta)}{\sum_{\theta' \in \Theta} p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta')} + \nu_t \sum_{\omega'_t \in \Omega} \frac{(1 - p_{\theta,\lambda_{t-1}})q_{\theta,\lambda_{t-1}}(\omega_t)\mu_{t-1}(\theta)}{1 - \sum_{\theta' \in \Theta} p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta')})\nu_{t-1} \quad (55)$$

$$= \mu_{t-1} + ((1 - \nu_t)\frac{p_{\theta,\lambda_{t-1}}\mu_{t-1}(\theta)}{\sum_{\theta' \in \Theta} p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta')}$$

$$+\nu_t\frac{(1-p_{\theta,\lambda_{t-1}})\mu_{t-1}(\theta)}{1-\sum_{\theta'\in\Theta}p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta')}-\mu_{t-1})\nu_{t-1} \quad (56)$$

$$=\mu_{t-1}+((1-\nu_t)\cdot\frac{p_{\theta,\lambda_{t-1}}-\sum_{\theta'\in\Theta}p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta')}{\sum_{\theta'\in\Theta}p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta')}\mu_{t-1}(\theta)-\nu_t\frac{p_{\theta,\lambda_{t-1}}-\sum_{\theta'\in\Theta}p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta')}{1-\sum_{\theta'\in\Theta}p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta')}\mu_{t-1}(\theta))\nu_{t-1} \quad (57)$$

Therefore it is straightforward to define the functions $\alpha_t = A(\mu_{t-1},\lambda_{t-1},\nu_{t-1},\nu_t)$, $\beta_t = B(\mu_{t-1},\lambda_{t-1},\nu_{t-1},\nu_t)$, as well as $\tilde{\mu}_t = \mu_t - \alpha_t - \beta_t$, where $\alpha_0 = \beta_0 = 0$ and:

$$\alpha_t(\theta)=\alpha_{t-1}(\theta)-\mu_{t-1}(\theta)\cdot\frac{p_{\theta,\lambda_{t-1}}-\sum_{\theta'\in\Theta}p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta')}{1-\sum_{\theta'\in\Theta}p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta')}\nu_{t-1}\nu_t \quad (58)$$

$$\beta_t(\theta)=\beta_{t-1}(\theta)+\mu_{t-1}(\theta)\cdot\frac{p_{\theta,\lambda_{t-1}}-\sum_{\theta'\in\Theta}p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta')}{\sum_{\theta'\in\Theta}p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta')}(1-\nu_t)\nu_{t-1} \quad (59)$$

Finally, for $\tilde{\mu}_t$ observe that:

$$\alpha_t+\beta_t=\sum_{t'=1}^{t}(\mathbb{E}_{p,q}[\mu_{t'}-\mu_{t'-1}|\lambda_{t'-1},\mu_{t'-1},\nu_{t'-1},\nu_{t'}]) \quad (60)$$

therefore the difference between two steps is:

$$\tilde{\mu}_t-\tilde{\mu}_{t-1}=\mu_t-\mu_{t-1}-\mathbb{E}_{p,q}[\mu_t-\mu_{t-1}|\lambda_{t-1},\mu_{t-1},\nu_{t-1},\nu_t] \quad (61)$$

hence—taking expectations—we can write:

$$\mathbb{E}_{p,q}[\tilde{\mu}_t-\tilde{\mu}_{t-1}|\lambda_{t-1},\mu_{t-1},\nu_{t-1}]=0 \quad (62)$$

$$\Rightarrow\ \mathbb{E}_{p,q}[\tilde{\mu}_t|\lambda_{t-1},\mu_{t-1},\nu_{t-1}]=\tilde{\mu}_{t-1} \quad (63)$$

Proposition 3 (Optimal Value) The optimal value function $V^*(\mu_t,\nu_t;\eta)$ is a fixed point of the operator $\mathbb{B}$ defined over the space of functions $V\in\mathbb{R}_+^{\Delta(\Theta)\times\{0,1\}}$ as follows:

$$(\mathbb{B}V)(\mu_t,\nu_t;\eta)=\min\{\inf_{\hat{\theta}'\in\Theta}\bar{Q}_{\hat{\theta}'}(\mu_t,\nu_t;\eta),\inf_{\lambda'_t\in\Lambda}Q_{\lambda'_t}(\mu_t,\nu_t;\eta)\} \quad (64)$$

where the (continual) Q-factors for *acquisitions* quantify the risk-to-go upon performing acquisition λ_t, given by:

$$Q_{\lambda_t}(\mu_t,\nu_t;\eta)=(1-\nu_t)V(\mu_t,0;\eta)+\eta_{\mathrm{c},\lambda_t}c_{\lambda_t}+\nu_t\mathbb{E}_{p,q}[V(\mu_{t+1},\nu_{t+1};\eta)|\lambda_t,\mu_t,\nu_t=1] \quad (65)$$

and the (terminal) Q-factors for *decisions* quantify the risk upon settling on the final choice of decision $\hat{\theta}$, given by:

$$\bar{Q}_{\hat{\theta}}(\mu_t,\nu_t;\eta)=(1-\nu_t)\textstyle\sum_{\theta'\in\Theta}\eta_{\mathrm{b},\theta'}\mu_t(\theta')+\nu_t\sum_{\theta'\in\Theta,\theta\neq\hat{\theta}}\eta_{\mathrm{a},\theta'}\mu_t(\theta') \quad (66)$$

Moreover, the operator $\mathbb{B}$ is *contractive*, and the optimal value function is therefore the *unique* fixed point admitted.

Proof. Each of the Q-factors for decisions is given by:

$$\bar{Q}_{\hat{\theta}}(\mu_t,\nu_t;\eta) \quad (67)$$

$$\doteq\mathbb{E}_{p,q}[\ell(\lambda_{0:\tau-1},\tau,\hat{\theta};\eta)|\lambda_{0:t-1},\tau=t,\hat{\theta},\mu_t,\nu_t]-\textstyle\sum_{t'=0}^{t-1}\eta_{\mathrm{c},\lambda_{t'}}c_{\lambda_{t'}} \quad (68)$$

$$=\mathbb{E}_{p,q}\big[\textstyle\sum_{\theta'\in\Theta}\eta_{\mathrm{a},\theta'}\mathbb{1}_{\{\theta=\theta',\theta\neq\hat{\theta},\tau<\delta\}}+\sum_{\theta'\in\Theta}\eta_{\mathrm{b},\theta'}\mathbb{1}_{\{\theta=\theta',\tau=\delta\}}+\sum_{t'=0}^{\tau-1}\eta_{\mathrm{c},\lambda_{t'}}c_{\lambda_{t'}}|\lambda_{0:t-1},\tau=t,\hat{\theta},\mu_t,\nu_t\big]-\textstyle\sum_{t'=0}^{t-1}\eta_{\mathrm{c},\lambda_{t'}}c_{\lambda_{t'}} \quad (69)$$

$$=\mathbb{E}_{p,q}\big[\textstyle\sum_{\theta'\in\Theta}\eta_{\mathrm{a},\theta'}\mathbb{1}_{\{\theta=\theta',\theta\neq\hat{\theta},t<\delta\}}+\sum_{\theta'\in\Theta}\eta_{\mathrm{b},\theta'}\mathbb{1}_{\{\theta=\theta',t=\delta\}}|\hat{\theta},\mu_t,\nu_t\big] \quad (70)$$

$$=\nu_t\textstyle\sum_{\theta'\in\Theta,\theta\neq\hat{\theta}}\eta_{\mathrm{a},\theta'}\mu_t(\theta')+(1-\nu_t)\sum_{\theta'\in\Theta}\eta_{\mathrm{b},\theta'}\mu_t(\theta') \quad (71)$$

For acquisitions, first observe that:

$$Q_{\lambda_t}(\mu_t,\nu_t;\eta) \quad (72)$$

$$\doteq\mathbb{E}_{p,q}[V(\mu_{t+1},\nu_{t+1};\eta)|\lambda_t,\mu_t,\nu_t]+\eta_{\mathrm{c},\lambda_t}c_{\lambda_t} \quad (73)$$

$$=(1-\nu_t)\mathbb{E}_{p,q}[V(\mu_{t+1},\nu_{t+1};\eta)|\lambda_t,\mu_t,\nu_t=0]+\mathbb{E}_{p,q}[V(\mu_{t+1},\nu_{t+1};\eta)|\lambda_t,\mu_t,\nu_t=1]\nu_t+\eta_{\mathrm{c},\lambda_t}c_{\lambda_t} \quad (74)$$

$$=(1-\nu_t)V(\mu_t,0;\eta)+\eta_{\mathrm{c},\lambda_t}c_{\lambda_t}+(\mathbb{E}_{p,q}[V((1-\nu_{t+1})\bar{M}(\lambda_t,\mu_t)+\nu_{t+1}M(\lambda_t,\mu_t,\omega_{t+1}),1;\eta)|\lambda_t,\mu_t,\nu_t=1])\nu_t \quad (75)$$

For the expectation term:

$$\mathbb{E}_{p,q}[V((1-\nu_{t+1})\bar{M}(\lambda_t,\mu_t)+\nu_{t+1}M(\lambda_t,\mu_t,\omega_{t+1}),1;\eta)|\lambda_t,\mu_t,\nu_t=1] \quad (76)$$

$$=\textstyle\sum_{\omega'_{t+1}\in\Omega}(\mathbb{P}_{p,q}\{\nu_{t+1}=1,\omega_{t+1}|\lambda_t,\mu_t,\nu_t=1\}\cdot V(M(\lambda_t,\mu_t,\omega_{t+1}),1;\eta))+\mathbb{P}_p\{\nu_{t+1}=0|\lambda_t,\mu_t,\nu_t=1\}V(\bar{M}(\lambda_t,\mu_t),0;\eta) \quad (77)$$

$$=\textstyle\sum_{\omega'_{t+1}\in\Omega}(V(M(\lambda_t,\mu_t,\omega_{t+1}),1;\eta)\cdot\sum_{\theta'\in\Theta}(\mathbb{P}_p\{\theta',\nu_{t+1}=1|\lambda_t,\mu_t,\nu_t=1\}\cdot\mathbb{P}_q\{\omega_{t+1}|\theta',\lambda_t,\nu_{t+1}=1\}))+\mathbb{P}_p\{\nu_{t+1}=0|\lambda_t,\mu_t,\nu_t=1\}V(\bar{M}(\lambda_t,\mu_t),0;\eta) \quad (78)$$

$$=\textstyle\sum_{\omega'_{t+1}\in\Omega}(V(M(\lambda_t,\mu_t,\omega_{t+1}),1;\eta)\cdot\sum_{\theta'\in\Theta}(\mathbb{P}_p\{\nu_{t+1}=1|\theta,\lambda_t,\nu_t=1\}\cdot\mathbb{P}_q\{\omega_{t+1}|\lambda_t,\theta',\nu_{t+1}=1\}\mu_t(\theta)))+V(\bar{M}(\lambda_t,\mu_t),0;\eta)\sum_{\theta'\in\Theta}p_{\theta',\lambda_t}\mu_t(\theta') \quad (79)$$

$$=\textstyle\sum_{\omega'_{t+1}\in\Omega}(V(M(\lambda_t,\mu_t,\omega_{t+1}),1;\eta)\cdot\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda_t})q_{\theta',\lambda_t}(\omega_{t+1})\mu_t(\theta'))+V(\bar{M}(\lambda_t,\mu_t),0;\eta)\sum_{\theta'\in\Theta}p_{\theta',\lambda_t}\mu_t(\theta') \quad (80)$$

Therefore each Q-factor for acquisition is given by:

$$Q_{\lambda_t}(\mu_t,\nu_t;\eta)=(1-\nu_t)V(\mu_t,0;\eta)+\eta_{\mathrm{c},\lambda_t}c_{\lambda_t}+\big(V(\bar{M}(\lambda_t,\mu_t),0;\eta)\textstyle\sum_{\theta'\in\Theta}p_{\theta',\lambda_t}\mu_t(\theta')+\sum_{\omega'_{t+1}\in\Omega}(V(M(\lambda_t,\mu_t,\omega'_{t+1}),1;\eta)\cdot\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda_t})q_{\theta',\lambda_t}(\omega'_{t+1})\mu_t(\theta'))\big)\nu_t \quad (81)$$

For the contractive property, we want that $\|\mathbb{B}V^i - \mathbb{B}V^j\| \leq \gamma\|V^i - V^j\|$ for some $\gamma < 1$, but where we do *not* have the benefit of an explicit discount factor γ for this purpose. For notational brevity, in the following we omit the functional dependence of value functions and Q-factors on η:

$$
\begin{aligned}
&|(\mathbb{B}V^i)(\mu_t,\nu_t) - (\mathbb{B}V^j)(\mu_t,\nu_t)| && (82)\\
=&|\min\{\inf_{\hat\theta'\in\Theta}\bar Q^i_{\hat\theta'}(\mu_t,\nu_t), \inf_{\lambda'_t\in\Lambda} Q^i_{\lambda'_t}(\mu_t,\nu_t)\}\\
&-\min\{\inf_{\hat\theta'\in\Theta}\bar Q^j_{\hat\theta'}(\mu_t,\nu_t), \inf_{\lambda'_t\in\Lambda} Q^j_{\lambda'_t}(\mu_t,\nu_t)\}| && (83)\\
=&|\min\{\inf_{\hat\theta\in\Theta}(\nu_t\textstyle\sum_{\theta'\in\Theta,\theta\neq\hat\theta}\eta_{\mathrm{a},\theta'}\mu_t(\theta') + (1-\nu_t)\\
&\cdot\textstyle\sum_{\theta'\in\Theta}\eta_{\mathrm{b},\theta'}\mu_t(\theta')), \inf_{\lambda'_t\in\Lambda}((1-\nu_t)V^i(\mu_t,0)\\
&+\big(\textstyle\sum_{\omega'_{t+1}\in\Omega}(V^i(M(\lambda'_t,\mu_t,\omega_{t+1}),1)\\
&\cdot\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda'_t})q_{\theta',\lambda'_t}(\omega_{t+1})\mu_t(\theta'))\\
&+V^i(\bar M(\lambda'_t,\mu_t),0)\textstyle\sum_{\theta'\in\Theta}p_{\theta',\lambda'_t}\mu_t(\theta')\big)\nu_t\\
&+\eta_{\mathrm{c},\lambda'_t}c_{\lambda'_t})\}\\
&-\min\{\inf_{\hat\theta\in\Theta}(\nu_t\textstyle\sum_{\theta'\in\Theta,\theta\neq\hat\theta}\eta_{\mathrm{a},\theta'}\mu_t(\theta') + (1-\nu_t)\\
&\cdot\textstyle\sum_{\theta'\in\Theta}\eta_{\mathrm{b},\theta'}\mu_t(\theta')), \inf_{\lambda'_t\in\Lambda}((1-\nu_t)V^j(\mu_t,0)\\
&+\big(\textstyle\sum_{\omega'_{t+1}\in\Omega}(V^j(M(\lambda'_t,\mu_t,\omega_{t+1}),1)\\
&\cdot\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda'_t})q_{\theta',\lambda'_t}(\omega_{t+1})\mu_t(\theta'))\\
&+V^j(\bar M(\lambda'_t,\mu_t),0)\textstyle\sum_{\theta'\in\Theta}p_{\theta',\lambda'_t}\mu_t(\theta')\big)\nu_t\\
&+\eta_{\mathrm{c},\lambda'_t}c_{\lambda'_t})\}| && (84)\\
\leq&|\inf_{\lambda'_t\in\Lambda}((1-\nu_t)V^i(\mu_t,0) + \eta_{\mathrm{c},\lambda'_t}c_{\lambda'_t}\\
&+\big(\textstyle\sum_{\omega'_{t+1}\in\Omega}(V^i(M(\lambda'_t,\mu_t,\omega_{t+1}),1)\\
&\cdot\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda'_t})q_{\theta',\lambda'_t}(\omega_{t+1})\mu_t(\theta'))\\
&+V^i(\bar M(\lambda'_t,\mu_t),0)\textstyle\sum_{\theta'\in\Theta}p_{\theta',\lambda'_t}\mu_t(\theta')\big)\nu_t)\\
&-\inf_{\lambda'_t\in\Lambda}((1-\nu_t)V^j(\mu_t,0) + \eta_{\mathrm{c},\lambda'_t}c_{\lambda'_t}\\
&+\big(\textstyle\sum_{\omega'_{t+1}\in\Omega}(V^j(M(\lambda'_t,\mu_t,\omega_{t+1}),1)\\
&\cdot\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda'_t})q_{\theta',\lambda'_t}(\omega_{t+1})\mu_t(\theta'))\\
&+V^j(\bar M(\lambda'_t,\mu_t),0)\textstyle\sum_{\theta'\in\Theta}p_{\theta',\lambda'_t}\mu_t(\theta')\big)\nu_t)| && (85)\\
=&|(1-\nu_t)V^i(\mu_t,0) + \eta_{\mathrm{c},\lambda^*_t}c_{\lambda^*_t}\\
&+\big(\textstyle\sum_{\omega'_{t+1}\in\Omega}(V^i(M(\lambda^*_t,\mu_t,\omega_{t+1}),1)\\
&\cdot\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda^*_t})q_{\theta',\lambda^*_t}(\omega_{t+1})\mu_t(\theta'))\\
&+V^i(\bar M(\lambda^*_t,\mu_t),0)\textstyle\sum_{\theta'\in\Theta}p_{\theta',\lambda^*_t}\mu_t(\theta')\big)\nu_t\\
&-\inf_{\lambda'_t\in\Lambda}((1-\nu_t)V^j(\mu_t,0) + \eta_{\mathrm{c},\lambda'_t}c_{\lambda'_t}\\
&+\big(\textstyle\sum_{\omega'_{t+1}\in\Omega}(V^j(M(\lambda'_t,\mu_t,\omega_{t+1}),1)\\
&\cdot\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda'_t})q_{\theta',\lambda'_t}(\omega_{t+1})\mu_t(\theta'))\\
&+V^j(\bar M(\lambda'_t,\mu_t),0)\textstyle\sum_{\theta'\in\Theta}p_{\theta',\lambda'_t}\mu_t(\theta')\big)\nu_t)| && (86)\\
\leq&|(1-\nu_t)V^i(\mu_t,0) + \eta_{\mathrm{c},\lambda^*_t}c_{\lambda^*_t}\\
&+\big(\textstyle\sum_{\omega'_{t+1}\in\Omega}(V^i(M(\lambda^*_t,\mu_t,\omega_{t+1}),1)\\
&\cdot\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda^*_t})q_{\theta',\lambda^*_t}(\omega_{t+1})\mu_t(\theta'))\\
&+V^i(\bar M(\lambda^*_t,\mu_t),0)\textstyle\sum_{\theta'\in\Theta}p_{\theta',\lambda^*_t}\mu_t(\theta')\big)\nu_t\\
&-(1-\nu_t)V^j(\mu_t,0) - \eta_{\mathrm{c},\lambda^*_t}c_{\lambda^*_t}\\
&-\big(\textstyle\sum_{\omega'_{t+1}\in\Omega}(V^j(M(\lambda^*_t,\mu_t,\omega_{t+1}),1)\\
&\cdot\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda^*_t})q_{\theta',\lambda^*_t}(\omega_{t+1})\mu_t(\theta'))\\
&-V^j(\bar M(\lambda^*_t,\mu_t),0)\textstyle\sum_{\theta'\in\Theta}p_{\theta',\lambda^*_t}\mu_t(\theta')\big)\nu_t| && (87)\\
=&|\textstyle\sum_{\omega'_{t+1}\in\Omega}((V^i(M(\lambda^*_t,\mu_t,\omega_{t+1}),1)\\
&-V^j(M(\lambda^*_t,\mu_t,\omega_{t+1}),1))\\
&\cdot\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda^*_t})q_{\theta',\lambda^*_t}(\omega_{t+1})\mu_t(\theta')\nu_t)| && (88)\\
\leq&|(1-\inf_{\theta'\in\Theta,\lambda'\in\Lambda}p_{\theta',\lambda'})\\
&\cdot\textstyle\sum_{\omega'_{t+1}\in\Omega}((V^i(M(\lambda^*_t,\mu_t,\omega_{t+1}),1)\\
&-V^j(M(\lambda^*_t,\mu_t,\omega_{t+1}),1))\\
&\cdot\textstyle\sum_{\theta'\in\Theta}q_{\theta',\lambda^*_t}(\omega_{t+1})\mu_t(\theta')\nu_t)| && (89)\\
\leq&(1-\inf_{\theta'\in\Theta,\lambda'\in\Lambda}p_{\theta',\lambda'})\\
&\cdot\sup_{\mu'_{t+1}\in\Delta(\Theta)}|V^i(\mu'_{t+1},1) - V^j(\mu'_{t+1},1)| && (90)\\
\leq&\gamma\|V^i - V^j\| && (91)
\end{aligned}
$$

where in the fourth equality $\lambda^*_t \doteq \arg\inf_{\lambda'_t\in\Lambda} Q^k_{\lambda'_t}(\mu_t,\nu_t)$ in which $k \doteq \arg\inf_{k'\in\{i,j\}}\inf_{\lambda'_t\in\Lambda} Q^{k'}_{\lambda'_t}(\mu_t,\nu_t)$, and in the last step $\gamma \doteq 1 - \inf_{\theta'\in\Theta,\lambda'\in\Lambda}p_{\theta',\lambda'} < 1$, and we also used the fact that $V(\mu_t,0) = \bar Q_{\hat\theta}(\mu_t,0) = \sum_{\theta'\in\Theta}\eta_{\mathrm{b},\theta'}\mu_t(\theta')$.

For the uniqueness property, consider two such fixed points V^* and V'^*. But $\|V^* - V'^*\| = \|\mathbb{B}V^* - \mathbb{B}V'^*\| \leq \gamma\|V^* - V'^*\|$, therefore it must the case that $\|V^* - V'^*\| = 0$.

Proposition 4 (Continuation and Termination) Denote by $m_\theta \in \Delta(\Theta)$ each vertex in the simplex, and let the optimal *aggregate* Q-factor for continuation be given by:

$$Q^*(\mu_t,\nu_t;\eta) \doteq \inf_{\lambda'_t\in\Lambda} Q^*_{\lambda'_t}(\mu_t,\nu_t;\eta) \tag{92}$$

and likewise $\bar Q(\mu_t,\nu_t;\eta) \doteq \inf_{\hat\theta'\in\Theta}\bar Q_{\hat\theta'}(\mu_t,\nu_t;\eta)$. Then Q^* is a *concave* function with respect to μ_t, and moreover takes on values strictly greater than $\bar Q$ at every vertex m_θ:

$$\forall m_\theta : Q^*(m_\theta,\nu_t;\eta) > \bar Q(m_\theta,\nu_t;\eta) \tag{93}$$

Hence the *termination set* $\mathcal{T}$ is the (disjoint) union of $|\Theta|$ *convex* regions delimited by the intersection of Q^* and $\bar Q$:

$$\mathcal{T}(\eta) = \{\mu_t : Q^*(\mu_t,\nu_t;\eta) \geq \bar Q(\mu_t,\nu_t;\eta)\} \tag{94}$$

and contains each of the simplex vertices. Finally, the (possibly null) *continuation set* is its complement $\Delta(\Theta)\setminus\mathcal{T}$.

Proof. We first show that V^* is concave. Since V^* is the limit of successive approximations by application of $\mathbb{B}$, we simply want to show if V is concave that $\mathbb{B}V$ is then concave. Suppose V is concave. Since $\mathbb{B}V$ is the minimum between $\inf_{\hat\theta'\in\Theta}\bar Q_{\hat\theta'}(\mu_t,\nu_t;\eta)$ and $\inf_{\lambda'_t\in\Lambda} Q_{\lambda'_t}(\mu_t,\nu_t;\eta)$ and the former clearly concave, it remains to show that each Q_{λ_t} in the latter is concave. This is obvious for $\nu_t = 0$ since $V(\mu_t,0) = \bar Q_{\hat\theta}(\mu_t,0)$ is concave. For $\nu_t = 1$, we want that $\sum_{\omega'_{t+1}\in\Omega}(V(M(\lambda_t,\mu_t,\omega_{t+1}),1;\eta)\sum_{\theta'\in\Theta}(1-$

$p_{\theta',\lambda_t})q_{\theta',\lambda_t}(\omega_{t+1})\mu_t(\theta'))$ be concave. Let $\upsilon \in (0,1)$. We similarly omit functional dependence on η for brevity:

$$\begin{aligned}
&\upsilon\textstyle\sum_{\omega'_{t+1}\in\Omega}(V(M(\lambda_t,\mu_t,\omega_{t+1}),1)\\
&\cdot\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda_t})q_{\theta',\lambda_t}(\omega_{t+1})\mu_t(\theta'))\\
&+(1-\upsilon)\textstyle\sum_{\omega'_{t+1}\in\Omega}(V(M(\lambda_t,\mu'_t,\omega_{t+1}),1)\\
&\cdot\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda_t})q_{\theta',\lambda_t}(\omega_{t+1})\mu'_t(\theta')) &(95)\\
=&\textstyle\sum_{\omega'_{t+1}\in\Omega}((\upsilon V(M(\lambda_t,\mu_t,\omega_{t+1}),1)\\
&\cdot\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda_t})q_{\theta',\lambda_t}(\omega_{t+1})\mu_t(\theta')\\
&/(\upsilon\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda_t})q_{\theta',\lambda_t}(\omega_{t+1})\mu_t(\theta')\\
&+(1-\upsilon)\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda_t})q_{\theta',\lambda_t}(\omega_{t+1})\mu'_t(\theta'))\\
&+(1-\upsilon)V(M(\lambda_t,\mu'_t,\omega_{t+1}),1)\\
&\cdot\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda_t})q_{\theta',\lambda_t}(\omega_{t+1})\mu'_t(\theta')\\
&/(\upsilon\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda_t})q_{\theta',\lambda_t}(\omega_{t+1})\mu_t(\theta')\\
&+(1-\upsilon)\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda_t})q_{\theta',\lambda_t}(\omega_{t+1})\mu'_t(\theta')))\\
&\cdot(\upsilon\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda_t})q_{\theta',\lambda_t}(\omega_{t+1})\mu_t(\theta')\\
&+(1-\upsilon)\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda_t})q_{\theta',\lambda_t}(\omega_{t+1})\mu'_t(\theta'))) &(96)\\
\leq&\textstyle\sum_{\omega'_{t+1}\in\Omega}(V((\upsilon M(\lambda_t,\mu_t,\omega_{t+1})\\
&\cdot\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda_t})q_{\theta',\lambda_t}(\omega_{t+1})\mu_t(\theta')\\
&+(1-\upsilon)M(\lambda_t,\mu'_t,\omega_{t+1})\\
&\cdot\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda_t})q_{\theta',\lambda_t}(\omega_{t+1})\mu'_t(\theta'))\\
&/(\upsilon\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda_t})q_{\theta',\lambda_t}(\omega_{t+1})\mu_t(\theta')\\
&+(1-\upsilon)\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda_t})q_{\theta',\lambda_t}(\omega_{t+1})\mu'_t(\theta')),1)\\
&\cdot(\upsilon\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda_t})q_{\theta',\lambda_t}(\omega_{t+1})\mu_t(\theta')\\
&+(1-\upsilon)\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda_t})q_{\theta',\lambda_t}(\omega_{t+1})\mu'_t(\theta'))) &(97)\\
=&\textstyle\sum_{\omega'_{t+1}\in\Omega}(V(M(\upsilon\mu_t+(1-\upsilon)\mu'_t))\\
&\cdot\textstyle\sum_{\theta'\in\Theta}((1-p_{\theta',\lambda_t})q_{\theta',\lambda_t}(\omega_{t+1})\\
&\cdot(\upsilon\mu_t(\theta')+(1-\upsilon)\mu'_t(\theta')))) &(98)
\end{aligned}$$

Now, V^* is simply the limit of successive approximations by application of $\mathbb{B}$, so by induction V^* is concave. Finally, each $Q^*_{\lambda_t}$ and therefore Q^* is concave since V^* is concave.

For the inequality, note if $\nu_t = 1$ then $\bar{Q}_{\hat{\theta}}$ at each vertex is simply zero for any choice of $\hat{\theta}$. But clearly $Q^*_{\lambda_t}$ is at least $\eta_{\mathrm{c},\lambda_t}c_{\lambda_t}$ for any choice of λ_t, so it must be true that $Q^* > \bar{Q}$. Finally, consider the intersection (if any) of Q^* and $\bar{Q}_{\hat{\theta}}$ when $\nu_t = 1$, for any $\hat{\theta}$. Let $\mu_t, \mu'_t \in \Delta(\Theta)$ be two points for which $\hat{\theta} = \arg\inf_{\hat{\theta}'\in\Theta}\bar{Q}_{\hat{\theta}'}(\cdot,\nu_t;\eta)$. Since the former is concave and the latter is affine, we can write:

$$\begin{aligned}
&\bar{Q}_{\hat{\theta}}(\upsilon\mu_t+(1-\upsilon)\mu_t,1;\eta) &(99)\\
=&\upsilon\bar{Q}_{\hat{\theta}}(\mu_t,1;\eta)+(1-\upsilon)\bar{Q}_{\hat{\theta}}(\mu_t,1;\eta) &(100)\\
=&\upsilon V^*(\mu_t,1;\eta)+(1-\upsilon)V^*(\mu_t,1;\eta) &(101)\\
\leq&V^*(\upsilon\mu_t+(1-\upsilon)\mu_t,1;\eta) &(102)\\
\leq&\bar{Q}(\upsilon\mu_t+(1-\upsilon)\mu_t,1;\eta) &(103)\\
\leq&\bar{Q}_{\hat{\theta}}(\upsilon\mu_t+(1-\upsilon)\mu_t,1;\eta) &(104)
\end{aligned}$$

for $\upsilon \in (0,1)$, hence the set $\bar{Q}_{\hat{\theta}} < Q^*$ is convex. Finally, the overall termination set $\mathcal{T}(\eta)$ is the union of $|\Theta|$ such regions. For completeness, consider the other (trivial) case where $\nu_t = 0$; clearly $Q^*_{\lambda_t} = \bar{Q} + \eta_{\mathrm{c},\lambda_t}c_{\lambda_t}$, so convexity is automatic and there is no intersection (i.e. $\mathcal{T}(\eta)$ is empty).

Proposition 5 (Surprise and Suspense) When $\mu_t \notin \mathcal{T}(\eta)$, the optimal acquisition directly trades off surprise and suspense (in addition to the immediate cost of acquisition):

$$\lambda^*_t = \arg\sup_{\lambda_t\in\Lambda} h(I_t(\lambda_t), S_t(\lambda_t)) - \eta_{\mathrm{c},\lambda_t}c_{\lambda_t} \quad (105)$$

where h is increasing in $I_t(\lambda_t)$ and $S_t(\lambda_t)$, and the uncertainty function for the information gain is taken as $U = V^*$.

Proof. Each optimal Q-factor for acquisitions is given by:

$$\begin{aligned}
&Q^*_{\lambda_t}(\mu_t,\nu_t;\eta) &(106)\\
=&(1-\nu_t)V^*(\mu_t,0;\eta)+\eta_{\mathrm{c},\lambda_t}c_{\lambda_t}\\
&+(\mathbb{E}_{p,q}[V^*((1-\nu_{t+1})\bar{M}(\lambda_t,\mu_t)\\
&+\nu_{t+1}M(\lambda_t,\mu_t,\omega_{t+1}),1;\eta)|\lambda_t,\mu_t,\nu_t=1])\nu_t &(107)\\
=&(1-\nu_t)V^*(\mu_t,0;\eta)+\eta_{\mathrm{c},\lambda_t}c_{\lambda_t}\\
&+(V^*(\bar{M}(\lambda_t,\mu_t),0;\eta)\textstyle\sum_{\theta'\in\Theta}p_{\theta',\lambda_t}\mu_t(\theta')\\
&+\textstyle\sum_{\omega'_{t+1}\in\Omega}(V^*(M(\lambda_t,\mu_t,\omega'_{t+1}),1;\eta)\\
&\cdot\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda_t})q_{\theta',\lambda_t}(\omega'_{t+1})\mu_t(\theta')))\nu_t &(108)
\end{aligned}$$

Note that the expectation term can also be expressed:

$$\begin{aligned}
&\mathbb{E}_{p,q}[V^*((1-\nu_{t+1})\bar{M}(\lambda_t,\mu_t)\\
&+\nu_{t+1}M(\lambda_t,\mu_t,\omega_{t+1}),1;\eta)|\lambda_t,\mu_t,\nu_t=1] &(109)\\
=&\mathbb{P}_p\{\nu_{t+1}=1|\lambda_t,\mu_t,\nu_t=1\}\\
&\cdot\mathbb{E}_{p,q}[V^*(M(\lambda_t,\mu_t,\omega_{t+1}),1;\eta)|\lambda_t,\mu_t,\\
&\nu_{t+1}=1]+\mathbb{P}_p\{\nu_{t+1}=0|\lambda_t,\mu_t,\nu_t=1\}\\
&\cdot V^*(\bar{M}(\lambda_t,\mu_t),0;\eta) &(110)
\end{aligned}$$

So we can rewrite:

$$\begin{aligned}
&\textstyle\sum_{\omega'_{t+1}\in\Omega}\mathbb{P}_{p,q}\{\nu_{t+1}=1,\omega_{t+1}|\lambda_t,\mu_t,\nu_t=1\}\\
&\cdot V^*(M(\lambda_t,\mu_t,\omega_{t+1}),1;\eta) &(111)\\
=&\textstyle\sum_{\omega'_{t+1}\in\Omega}(V^*(M(\lambda_t,\mu_t,\omega_{t+1}),1;\eta)\\
&\cdot\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda_t})q_{\theta',\lambda_t}(\omega_{t+1})\mu_t(\theta')) &(112)\\
=&\mathbb{P}_p\{\nu_{t+1}=1|\lambda_t,\mu_t,\nu_t=1\}\\
&\cdot\mathbb{E}_{p,q}[V^*(M(\lambda_t,\mu_t,\omega_{t+1}),1;\eta)|\lambda_t,\mu_t,\\
&\nu_{t+1}=1] &(113)
\end{aligned}$$

Hence each Q-factor for acquisitions can be expressed:

$$\begin{aligned}
&Q^*_{\lambda_t}(\mu_t,\nu_t;\eta) &(114)\\
=&(1-\nu_t)V^*(\mu_t,0;\eta)+\eta_{\mathrm{c},\lambda_t}c_{\lambda_t}\\
&+(V^*(\bar{M}(\lambda_t,\mu_t),0;\eta)\textstyle\sum_{\theta'\in\Theta}p_{\theta',\lambda_t}\mu_t(\theta')\\
&+\textstyle\sum_{\omega'_{t+1}\in\Omega}(V^*(M(\lambda_t,\mu_t,\omega'_{t+1}),1;\eta)\\
&\cdot\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda_t})q_{\theta',\lambda_t}(\omega'_{t+1})\mu_t(\theta')))\nu_t &(115)\\
=&(1-\nu_t)V^*(\mu_t,0;\eta)+\eta_{\mathrm{c},\lambda_t}c_{\lambda_t}
\end{aligned}$$

$$
\begin{aligned}
&+ (V^*(\bar{M}(\lambda_t, \mu_t), 0; \eta) \textstyle\sum_{\theta' \in \Theta} p_{\theta', \lambda_t} \mu_t(\theta') \\
&+ \mathbb{P}_p\{\nu_{t+1} = 1 | \lambda_t, \mu_t, \nu_t = 1\} \\
&\cdot \mathbb{E}_{p,q}[V^*(M(\lambda_t, \mu_t, \omega_{t+1}), 1; \eta) | \lambda_t, \mu_t, \\
&\quad \nu_{t+1} = 1]) \nu_t && (116) \\
=&(1 - \nu_t) V^*(\mu_t, 0; \eta) + \eta_{\mathrm{c}, \lambda_t} c_{\lambda_t} \\
&+ (\frac{\sum_{\theta' \in \Theta} \eta_{\mathrm{b}, \theta'} p_{\theta', \lambda_t} \mu_t(\theta')}{\sum_{\theta' \in \Theta} p_{\theta', \lambda_{t-1}} \mu_{t-1}(\theta')} \textstyle\sum_{\theta' \in \Theta} p_{\theta', \lambda_t} \mu_t(\theta') \\
&- (V^*(\mu_t, 1; \eta) - \mathbb{P}_p\{\nu_{t+1} = 1 | \lambda_t, \mu_t, \nu_t = 1\} \\
&\cdot \mathbb{E}_{p,q}[V^*(M(\lambda_t, \mu_t, \omega_{t+1}), 1; \eta) | \lambda_t, \mu_t, \\
&\quad \nu_{t+1} = 1]) + V^*(\mu_t, 1; \eta)) \nu_t && (117) \\
=&(1 - \nu_t) V^*(\mu_t, 0; \eta) - I_t(\lambda_t) + \eta_{\mathrm{c}, \lambda_t} c_{\lambda_t} \\
&+ (\frac{\sum_{\theta' \in \Theta} \eta_{\mathrm{b}, \theta'} p_{\theta', \lambda_t} \mu_t(\theta')}{\sum_{\theta' \in \Theta} \eta_{\mathrm{b}, \theta'} \mu_t(\theta')} \textstyle\sum_{\theta' \in \Theta} \eta_{\mathrm{b}, \theta'} \mu_t(\theta') \\
&+ V^*(\mu_t, 1; \eta)) \nu_t && (118) \\
=&(1 - \nu_t) V^*(\mu_t, 0; \eta) - I_t(\lambda_t) + \eta_{\mathrm{c}, \lambda_t} c_{\lambda_t} \\
&+ ((1 - S_t(\lambda_t)) \textstyle\sum_{\theta' \in \Theta} \eta_{\mathrm{b}, \theta'} \mu_t(\theta') \\
&+ V^*(\mu_t, 1; \eta)) \nu_t && (119)
\end{aligned}
$$

Consider $\nu_t = 1$, and suppose $\mu_t \in \mathcal{T}(\eta)$. Then:

$$
\begin{aligned}
Q^*_{\lambda_t} &= V^*(\mu_t, 1; \eta) - I_t(\lambda_t) \\
&\quad - (S_t(\lambda_t) - 1) \textstyle\sum_{\theta' \in \Theta} \eta_{\mathrm{b}, \theta'} \mu_t(\theta') + \eta_{\mathrm{c}, \lambda_t} c_{\lambda_t} && (120) \\
&= -h(I_t(\lambda_t), S_t(\lambda_t)) + \eta_{\mathrm{c}, \lambda_t} c_{\lambda_t} && (121)
\end{aligned}
$$

for some h increasing in $I_t(\lambda_t)$ and $S_t(\lambda_t)$, since other terms do not depend on the choice of λ_t. Hence minimizing $Q^*_{\lambda_t}$ is equivalent to maximizing $h(I_t(\lambda_t), S_t(\lambda_t)) - \eta_{\mathrm{c}, \lambda_t} c_{\lambda_t}$. For completeness, consider also $\nu_t = 0$. But clearly $\mathcal{T}(\eta)$ is empty since $Q^*_{\lambda_t} = \bar{Q} + \eta_{\mathrm{c}, \lambda_t} c_{\lambda_t}$, therefore $\mu_t \notin \mathcal{T}(\eta)$ and there is no acquisition hence no tradeoff.

Proposition 6 (Strategy Posterior) The posterior $\mathbb{P}\{\pi | \mathcal{D}\}$ over $\mathcal{P}$ (Equation 22) satisfies the following proportionality:

$$
\begin{aligned}
\mathbb{P}\{\pi^\kappa_\rho(...; \eta) | \mathcal{D}\} &\propto \mathbb{P}\{\kappa\} \mathbb{P}\{\eta | \kappa\} \mathbb{P}\{\rho\} \\
&\quad \cdot \textstyle\prod_{n=1}^N \prod_{t=0}^{\tau_n - 1} \pi^\kappa_\rho(\tilde{\lambda}_{n,t} | \mu_{n,t}, \nu_{n,t}; \eta) && (122)
\end{aligned}
$$

where $\mu_{n,t}$ is recursively computed via update M, $\nu_{n,t}{=}1$ prior to stopping, and $\pi^\kappa_\rho(...; \eta)$ is defined as in Equation 24.

Proof. First, the likelihood term is given by:

$$
\begin{aligned}
&\mathbb{P}_{p,q}\{\mathcal{D} | \pi^\kappa_\rho(...; \eta)\} && (123) \\
=&\mathbb{P}_{p,q}\{(\tilde{\lambda}_{n,0:\tau-1}, \tilde{\omega}_{n,1:\tau})_{n=1}^N | \kappa, \eta, \rho\} && (124) \\
=&\textstyle\int_{\Delta(\Theta)} \prod_{n=1}^N \prod_{t=0}^{\tau-1} (\mathbb{P}\{\tilde{\lambda}_{n,t} | \mu_{n,t}, \nu_{n,t}, \kappa, \eta, \rho\} \\
&\cdot \mathbb{P}_{p,q}\{\nu_{n,t+1}, \tilde{\omega}_{n,t+1} | \tilde{\lambda}_{n,t}, \mu_{n,t}, \nu_{n,t}\}) \\
&\quad d\mathbb{P}\{\mu_{n,t+1} | \tilde{\lambda}_{n,t}, \mu_{n,t}, \tilde{\omega}_{n,t+1}\} && (125) \\
=&\textstyle\prod_{n=1}^N \prod_{t=0}^{\tau-1} (\mathbb{P}\{\tilde{\lambda}_{n,t} | \mu_{n,t} = \\
&\quad M(\lambda_{n,t-1}, \mu_{n,t-1}, \tilde{\omega}_{n,t}), \nu_{n,t}, \kappa, \eta, \rho\} \\
&\cdot \mathbb{P}_{p,q}\{\nu_{n,t+1}, \tilde{\omega}_{n,t+1} | \tilde{\lambda}_{n,t}, \mu_{n,t}, \nu_{n,t}\}) && (126) \\
=&\textstyle\prod_{n=1}^N \prod_{t=0}^{\tau-1} \pi^\kappa_\rho(\tilde{\lambda}_{n,t} | M(\lambda_{n,t-1}, \mu_{n,t-1}, \tilde{\omega}_{n,t}), \\
&\quad \nu_{n,t}; \eta) \mathbb{P}_{p,q}\{\nu_{n,t+1}, \tilde{\omega}_{n,t+1} | \tilde{\lambda}_{n,t}, \mu_{n,t}, \nu_{n,t}\} && (127)
\end{aligned}
$$

where for third equality recall the Bayesian recognition model (which involves no uncertainty), and the fourth equality is just our definition of a strategy. So the posterior is:

$$
\begin{aligned}
&\mathbb{P}_{p,q}\{\pi^\kappa_\rho(...; \eta) | \mathcal{D}\} && (128) \\
=&\frac{1}{Z} \mathbb{P}\{\kappa\} \mathbb{P}\{\eta | \kappa\} \mathbb{P}\{\rho\} \textstyle\prod_{n=1}^N \prod_{t=0}^{\tau-1} (\\
&\pi^\kappa_\rho(\tilde{\lambda}_{n,t} | M(\lambda_{n,t-1}, \mu_{n,t-1}, \tilde{\omega}_{n,t}), \nu_{n,t}; \eta) \\
&\cdot \mathbb{P}_{p,q}\{\nu_{n,t+1}, \tilde{\omega}_{n,t+1} | \tilde{\lambda}_{n,t}, \mu_{n,t}, \nu_{n,t}\}) && (129)
\end{aligned}
$$

where the normalizing constant is given by:

$$
\begin{aligned}
Z =&\textstyle\int_{\mathcal{K}} \int_{\mathcal{H}} \int_{\mathbb{R}} \prod_{n=1}^N \prod_{t=0}^{\tau-1} (\\
&\pi^\kappa_\rho(\tilde{\lambda}_{n,t} | M(\lambda_{n,t-1}, \mu_{n,t-1}, \tilde{\omega}_{n,t}), \nu_{n,t}; \eta) \\
&\cdot \mathbb{P}_{p,q}\{\nu_{n,t+1}, \tilde{\omega}_{n,t+1} | \tilde{\lambda}_{n,t}, \mu_{n,t}, \nu_{n,t}\}) \\
&d\mathbb{P}\{\rho\} d\mathbb{P}\{\eta | \kappa\} d\mathbb{P}\{\kappa\} && (130)
\end{aligned}
$$

Note that the dynamics term does not depend on κ, η, or ρ and cancels out from the numerator and denominator, so:

$$
\begin{aligned}
&\mathbb{P}\{\pi^\kappa_\rho(...; \eta) | \mathcal{D}\} && (131) \\
=&\frac{1}{Z'} \mathbb{P}\{\kappa\} \mathbb{P}\{\eta | \kappa\} \mathbb{P}\{\rho\} \textstyle\prod_{n=1}^N \prod_{t=0}^{\tau_n - 1} (\\
&\pi^\kappa_\rho(\tilde{\lambda}_{n,t} | M(\lambda_{n,t-1}, \mu_{n,t-1}, \tilde{\omega}_{n,t}), \nu_{n,t}; \eta)) && (132)
\end{aligned}
$$

Proposition 7 (Differentiable Posterior) Assuming differentiable priors $\mathbb{P}\{\eta | *\}, \mathbb{P}\{\rho\}$, the posterior $\mathbb{P}\{\eta, \rho | *, \mathcal{D}\}$ for optimal strategies is differentiable (almost everywhere).

Proof. First, we show each $\tilde{Q}^*_{\tilde{\lambda}_{n,t}}(\mu_{n,t}, \nu_{n,t}; \eta)$ is concave in η, for which it is sufficient to show each $V^*(\mu_{n,t}, \nu_{n,t}; \eta)$ is concave. Let π be the Bayes-optimal strategy corresponding to the point $\upsilon\eta + (1 - \upsilon)\eta'$ for $\upsilon \in (0, 1)$. Then:

$$
\begin{aligned}
&V^*(\mu_{n,t}, \nu_{n,t}; \upsilon\eta + (1 - \upsilon)\eta') && (133) \\
=&V^\pi(\mu_{n,t}, \nu_{n,t}; \upsilon\eta + (1 - \upsilon)\eta') && (134) \\
=&\upsilon V^\pi(\mu_{n,t}, \nu_{n,t}; \eta) + (1 - \upsilon) V^\pi(\mu_{n,t}, \nu_{n,t}; \eta') && (135) \\
\geq&\upsilon V^*(\mu_{n,t}, \nu_{n,t}; \eta) + (1 - \upsilon) V^*(\mu_{n,t}, \nu_{n,t}; \eta') && (136)
\end{aligned}
$$

where the second equality follows from linearity of expectations, and the inequality from the fact that any the optimal strategy for η and η' respectively is by definition at least as good as any other strategy π (which in this case is only known to be optimal for some other point $\upsilon\eta + (1 - \upsilon)\eta'$). But for any function $f : \mathbb{R}^d \to \mathbb{R}$ for some finite d that is concave, the set of points of non-differentiability is at most countable. Therefore $\tilde{Q}^*_{\tilde{\lambda}_{n,t}}(\mu_{n,t}, \nu_{n,t}; \eta)$ is differentiable (almost everywhere). Now, the likelihood is a differentiable in ρ and in each $\tilde{Q}^*_{\tilde{\lambda}_{n,t}}(\mu_{n,t}, \nu_{n,t}; \eta)$, so the posterior is differentiable (almost everywhere) in η and ρ as long as the priors $\mathbb{P}\{\eta | *\}$ and $\mathbb{P}\{\rho\}$ themselves are differentiable.