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	Apprenticeship	Objective, Imposed
IAS **(Ours)**	Posterior & Survival	Stochastic, Endogenous	Risk-based Preferences	Yes	Yes	Yes	Yes	Understanding	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 $M:\Lambda\times\Delta(\Theta)\times\Omega\rightarrow\Delta(\Theta)$ 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 $\bar{M}:\Lambda\times\Delta(\Theta)\rightarrow\Delta(\Theta)$ 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:

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

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

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

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

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

$$\begin{aligned}
&\mathbb{P}_{p,q}\{\theta|\lambda_{t-1}, \mu_{t-1}, \nu_t = 1, \omega_t\} && (34)\\
=&\frac{\mathbb{P}_{p,q}\{\theta, \nu_t = 1, \omega_t|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1} = 1\}}{\mathbb{P}_{p,q}\{\nu_t = 1, \omega_t|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1} = 1\}} && (35)\\
=&\mathbb{P}_p\{\theta, \nu_t = 1|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1} = 1\} \cdot \mathbb{P}_q\{\omega_t|\theta,\\
&\lambda_{t-1}, \nu_t = 1\}/\textstyle\sum_{\theta'\in\Theta}(\mathbb{P}_p\{\theta', \nu_t = 1|\lambda_{t-1},\\
&\mu_{t-1}, \nu_{t-1} = 1\}\mathbb{P}_q\{\omega_t|\theta', \lambda_{t-1}, \nu_t = 1\}) && (36)\\
=&\mathbb{P}_p\{\nu_t = 1|\theta, \lambda_{t-1}, \nu_{t-1} = 1\} \cdot \mathbb{P}_q\{\omega_t|\theta,\\
&\lambda_{t-1}, \nu_t = 1\}\mu_{t-1}(\theta)/\textstyle\sum_{\theta'\in\Theta}\mathbb{P}_p\{\nu_t = 1|\theta,\\
&\lambda_{t-1}, \nu_{t-1} = 1\}\mathbb{P}_q\{\omega_t|\theta', \lambda_{t-1}, \nu_t = 1\}\mu_{t-1}(\theta) && (37)\\
=&\frac{(1 - p_{\theta,\lambda_{t-1}})q_{\theta,\lambda_{t-1}}(\omega_t)\mu_{t-1}(\theta)}{\sum_{\theta'\in\Theta}(1 - p_{\theta',\lambda_{t-1}})q_{\theta',\lambda_{t-1}}(\omega_t)\mu_{t-1}(\theta')} && (38)
\end{aligned}$$

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:

$$\begin{aligned}
&\mathbb{P}_{p,q}\{\mu_t|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1}, \nu_t\} && (39)\\
=&(1 - \nu_{t-1})\mathbb{P}_p\{\mu_t|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1} = 0, \nu_t = 0\}\\
&+ ((1 - \nu_t)\mathbb{P}_p\{\mu_t|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1} = 1, \nu_t = 0\}\\
&+ \nu_t\mathbb{P}_{p,q}\{\mu_t|\lambda_{t-1}, \mu_{t-1}, \nu_t = 1\})\nu_{t-1} && (40)\\
=&(1 - \nu_{t-1})\mathbb{1}_{\{\mu_t=\mu_{t-1}\}}\\
&+ ((1 - \nu_t)\mathbb{1}_{\{\mu_t=\bar{M}(\lambda_{t-1},\mu_{t-1})\}}\\
&+ \nu_t\frac{\mathbb{P}_{p,q}\{\mu_t, \nu_t = 1|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1} = 1\}}{\mathbb{P}_p\{\nu_t = 1|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1} = 1\}})\nu_{t-1} && (41)\\
=&(1 - \nu_{t-1})\mathbb{1}_{\{\mu_t=\mu_{t-1}\}}\\
&+ ((1 - \nu_t)\mathbb{1}_{\{\mu_t=\bar{M}(\lambda_{t-1},\mu_{t-1})\}}\\
&+ \nu_t\textstyle\sum_{\omega'_t\in\Omega}(\mathbb{1}_{\{\mu_t=M(\lambda_{t-1},\mu_{t-1},\omega'_t)\}}\\
&\cdot \frac{\sum_{\theta'\in\Theta}(1 - p_{\theta',\lambda_{t-1}})q_{\theta',\lambda_{t-1}}(\omega_t)\mu_{t-1}(\theta')}{1 - \sum_{\theta'\in\Theta}p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta')}))\nu_{t-1} && (42)
\end{aligned}$$

Then the joint probability of the tuple is given by:

$$\begin{aligned}
&\mathbb{P}_{p,q}\{\mu_t, \nu_t|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1}\} && (43)\\
=&\mathbb{P}_{p,q}\{\mu_t|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1}, \nu_t\}\\
&\cdot \mathbb{P}_p\{\nu_t|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1}\} && (44)\\
=&(1 - \nu_{t-1})\mathbb{1}_{\{\mu_t=\mu_{t-1}\}} + ((1 - \nu_t)\\
&\cdot \mathbb{1}_{\{\mu_t=\bar{M}(\lambda_{t-1},\mu_{t-1})\}}\textstyle\sum_{\theta'\in\Theta}p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta')\\
&+ \nu_t\textstyle\sum_{\omega'_t\in\Omega}(\mathbb{1}_{\{\mu_t=M(\lambda_{t-1},\mu_{t-1},\omega'_t)\}}\cdot\\
&\textstyle\sum_{\theta'\in\Theta}(1 - p_{\theta',\lambda_{t-1}})q_{\theta',\lambda_{t-1}}(\omega_t)\mu_{t-1}(\theta')))\nu_{t-1} && (45)
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}_{p,q}[f(\mu_t, \nu_t)|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1}] && (46)\\
=&\mathbb{E}_{p,q}[(1 - \nu_{t-1})f(\mu_{t-1}, 0)\\
&+ ((1 - \nu_t)f(\bar{M}(\lambda_{t-1}, \mu_{t-1}), 0) + \nu_t\cdot\\
&f(M(\lambda_{t-1}, \mu_{t-1}, \omega_t), 1))\nu_{t-1}|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1}] && (47)\\
=&(1 - \nu_{t-1})f(\mu_{t-1}, 0) + (f(\bar{M}(\lambda_{t-1}, \mu_{t-1}), 0)\\
&\cdot \textstyle\sum_{\theta'\in\Theta}p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta')\\
&+ \textstyle\sum_{\omega'_t\in\Omega}(f(M(\lambda_{t-1}, \mu_{t-1}, \omega_t), 1)\cdot\\
&\textstyle\sum_{\theta'\in\Theta}(1 - p_{\theta',\lambda_{t-1}})q_{\theta',\lambda_{t-1}}(\omega_t)\mu_{t-1}(\theta')))\nu_{t-1} && (48)
\end{aligned}$$

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:

$$\mu_t = \tilde{\mu}_t + \alpha_t + \beta_t \tag{49}$$

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 $\alpha_0 = 0$):

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

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:

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

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:

$$\begin{aligned}
&\mathbb{E}_{p,q}[\mu_t|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1}, \nu_t] && (52)\\
=&\mathbb{E}_{p,q}[(1 - \nu_{t-1})\mu_{t-1} + ((1 - \nu_t)\bar{M}(\lambda_{t-1}, \mu_{t-1})\\
&+\nu_t M(\lambda_{t-1}, \mu_{t-1}, \omega_t))\nu_{t-1}|\lambda_{t-1}, \mu_{t-1}, \nu_{t-1}, \nu_t] && (53)\\
=&(1 - \nu_{t-1})\mu_{t-1} + ((1 - \nu_t)\bar{M}(\lambda_{t-1}, \mu_{t-1})\\
&+ \nu_t\textstyle\sum_{\omega'_t\in\Omega}(M(\lambda_{t-1}, \mu_{t-1}, \omega'_t)\\
&\cdot\frac{\sum_{\theta'\in\Theta}(1 - p_{\theta',\lambda_{t-1}})q_{\theta',\lambda_{t-1}}(\omega_t)\mu_{t-1}(\theta')}{1 - \sum_{\theta'\in\Theta}p_{\theta',\lambda_{t-1}}\mu_{t-1}(\theta')}))\nu_{t-1} && (54)
\end{aligned}$$

Then for element θ, this is equal to:

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

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

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

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:

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

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

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

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

therefore the difference between two steps is:

$$\begin{aligned}\tilde{\mu}_t-\tilde{\mu}_{t-1}&=\mu_t-\mu_{t-1}\\&-\mathbb{E}_{p,q}[\mu_t-\mu_{t-1}|\lambda_{t-1},\mu_{t-1},\nu_{t-1},\nu_t]\end{aligned} \tag{61}$$

hence—taking expectations—we can write:

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

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

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:

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

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

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

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

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

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:

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

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

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

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

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

For acquisitions, first observe that:

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

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

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

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

For the expectation term:

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

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

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

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

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

Therefore each Q-factor for acquisition is given by:

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

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:

$$|(\mathbb{B}V^i)(\mu_t,\nu_t) - (\mathbb{B}V^j)(\mu_t,\nu_t)| \tag{82}$$

$$\begin{aligned}
=&|\min\{\inf_{\hat{\theta}'\in\Theta}\bar{Q}^i_{\hat{\theta}'}(\mu_t,\nu_t),\inf_{\lambda'_t\in\Lambda}Q^i_{\lambda'_t}(\mu_t,\nu_t)\}\\
&-\min\{\inf_{\hat{\theta}'\in\Theta}\bar{Q}^j_{\hat{\theta}'}(\mu_t,\nu_t),\inf_{\lambda'_t\in\Lambda}Q^j_{\lambda'_t}(\mu_t,\nu_t)\}|
\end{aligned}\tag{83}$$

$$\begin{aligned}
=&|\min\{\inf_{\hat{\theta}\in\Theta}(\nu_t{\textstyle\sum}_{\theta'\in\Theta,\theta\neq\hat{\theta}}\eta_{\mathrm{a},\theta'}\mu_t(\theta') + (1-\nu_t)\\
&\cdot{\textstyle\sum}_{\theta'\in\Theta}\eta_{\mathrm{b},\theta'}\mu_t(\theta')),\inf_{\lambda'_t\in\Lambda}((1-\nu_t)V^i(\mu_t,0)\\
&+\big({\textstyle\sum}_{\omega'_{t+1}\in\Omega}(V^i(M(\lambda'_t,\mu_t,\omega_{t+1}),1)\\
&\cdot{\textstyle\sum}_{\theta'\in\Theta}(1-p_{\theta',\lambda'_t})q_{\theta',\lambda'_t}(\omega_{t+1})\mu_t(\theta'))\\
&+V^i(\bar{M}(\lambda'_t,\mu_t),0){\textstyle\sum}_{\theta'\in\Theta}p_{\theta',\lambda'_t}\mu_t(\theta')\big)\nu_t\\
&+\eta_{\mathrm{c},\lambda'_t}c_{\lambda'_t})\}\\
&-\min\{\inf_{\hat{\theta}\in\Theta}(\nu_t{\textstyle\sum}_{\theta'\in\Theta,\theta\neq\hat{\theta}}\eta_{\mathrm{a},\theta'}\mu_t(\theta') + (1-\nu_t)\\
&\cdot{\textstyle\sum}_{\theta'\in\Theta}\eta_{\mathrm{b},\theta'}\mu_t(\theta')),\inf_{\lambda'_t\in\Lambda}((1-\nu_t)V^j(\mu_t,0)\\
&+\big({\textstyle\sum}_{\omega'_{t+1}\in\Omega}(V^j(M(\lambda'_t,\mu_t,\omega_{t+1}),1)\\
&\cdot{\textstyle\sum}_{\theta'\in\Theta}(1-p_{\theta',\lambda'_t})q_{\theta',\lambda'_t}(\omega_{t+1})\mu_t(\theta'))\\
&+V^j(\bar{M}(\lambda'_t,\mu_t),0){\textstyle\sum}_{\theta'\in\Theta}p_{\theta',\lambda'_t}\mu_t(\theta')\big)\nu_t\\
&+\eta_{\mathrm{c},\lambda'_t}c_{\lambda'_t})\}|
\end{aligned}\tag{84}$$

$$\begin{aligned}
\leq&|\inf_{\lambda'_t\in\Lambda}((1-\nu_t)V^i(\mu_t,0)+\eta_{\mathrm{c},\lambda'_t}c_{\lambda'_t}\\
&+\big({\textstyle\sum}_{\omega'_{t+1}\in\Omega}(V^i(M(\lambda'_t,\mu_t,\omega_{t+1}),1)\\
&\cdot{\textstyle\sum}_{\theta'\in\Theta}(1-p_{\theta',\lambda'_t})q_{\theta',\lambda'_t}(\omega_{t+1})\mu_t(\theta'))\\
&+V^i(\bar{M}(\lambda'_t,\mu_t),0){\textstyle\sum}_{\theta'\in\Theta}p_{\theta',\lambda'_t}\mu_t(\theta')\big)\nu_t)\\
&-\inf_{\lambda'_t\in\Lambda}((1-\nu_t)V^j(\mu_t,0)+\eta_{\mathrm{c},\lambda'_t}c_{\lambda'_t}\\
&+\big({\textstyle\sum}_{\omega'_{t+1}\in\Omega}(V^j(M(\lambda'_t,\mu_t,\omega_{t+1}),1)\\
&\cdot{\textstyle\sum}_{\theta'\in\Theta}(1-p_{\theta',\lambda'_t})q_{\theta',\lambda'_t}(\omega_{t+1})\mu_t(\theta'))\\
&+V^j(\bar{M}(\lambda'_t,\mu_t),0){\textstyle\sum}_{\theta'\in\Theta}p_{\theta',\lambda'_t}\mu_t(\theta')\big)\nu_t)|
\end{aligned}\tag{85}$$

$$\begin{aligned}
=&|(1-\nu_t)V^i(\mu_t,0)+\eta_{\mathrm{c},\lambda^*_t}c_{\lambda^*_t}\\
&+\big({\textstyle\sum}_{\omega'_{t+1}\in\Omega}(V^i(M(\lambda^*_t,\mu_t,\omega_{t+1}),1)\\
&\cdot{\textstyle\sum}_{\theta'\in\Theta}(1-p_{\theta',\lambda^*_t})q_{\theta',\lambda^*_t}(\omega_{t+1})\mu_t(\theta'))\\
&+V^i(\bar{M}(\lambda^*_t,\mu_t),0){\textstyle\sum}_{\theta'\in\Theta}p_{\theta',\lambda^*_t}\mu_t(\theta')\big)\nu_t\\
&-\inf_{\lambda'_t\in\Lambda}((1-\nu_t)V^j(\mu_t,0)+\eta_{\mathrm{c},\lambda'_t}c_{\lambda'_t}\\
&+\big({\textstyle\sum}_{\omega'_{t+1}\in\Omega}(V^j(M(\lambda'_t,\mu_t,\omega_{t+1}),1)\\
&\cdot{\textstyle\sum}_{\theta'\in\Theta}(1-p_{\theta',\lambda'_t})q_{\theta',\lambda'_t}(\omega_{t+1})\mu_t(\theta'))\\
&+V^j(\bar{M}(\lambda'_t,\mu_t),0){\textstyle\sum}_{\theta'\in\Theta}p_{\theta',\lambda'_t}\mu_t(\theta')\big)\nu_t)|
\end{aligned}\tag{86}$$

$$\begin{aligned}
\leq&|(1-\nu_t)V^i(\mu_t,0)+\eta_{\mathrm{c},\lambda^*_t}c_{\lambda^*_t}\\
&+\big({\textstyle\sum}_{\omega'_{t+1}\in\Omega}(V^i(M(\lambda^*_t,\mu_t,\omega_{t+1}),1)\\
&\cdot{\textstyle\sum}_{\theta'\in\Theta}(1-p_{\theta',\lambda^*_t})q_{\theta',\lambda^*_t}(\omega_{t+1})\mu_t(\theta'))\\
&+V^i(\bar{M}(\lambda^*_t,\mu_t),0){\textstyle\sum}_{\theta'\in\Theta}p_{\theta',\lambda^*_t}\mu_t(\theta')\big)\nu_t\\
&-(1-\nu_t)V^j(\mu_t,0)-\eta_{\mathrm{c},\lambda^*_t}c_{\lambda^*_t}\\
&-\big({\textstyle\sum}_{\omega'_{t+1}\in\Omega}(V^j(M(\lambda^*_t,\mu_t,\omega_{t+1}),1)\\
&\cdot{\textstyle\sum}_{\theta'\in\Theta}(1-p_{\theta',\lambda^*_t})q_{\theta',\lambda^*_t}(\omega_{t+1})\mu_t(\theta'))\\
&-V^j(\bar{M}(\lambda^*_t,\mu_t),0){\textstyle\sum}_{\theta'\in\Theta}p_{\theta',\lambda^*_t}\mu_t(\theta')\big)\nu_t|
\end{aligned}\tag{87}$$

$$\begin{aligned}
=&|{\textstyle\sum}_{\omega'_{t+1}\in\Omega}((V^i(M(\lambda^*_t,\mu_t,\omega_{t+1}),1)\\
&-V^j(M(\lambda^*_t,\mu_t,\omega_{t+1}),1))\\
&\cdot{\textstyle\sum}_{\theta'\in\Theta}(1-p_{\theta',\lambda^*_t})q_{\theta',\lambda^*_t}(\omega_{t+1})\mu_t(\theta')\nu_t)|
\end{aligned}\tag{88}$$

$$\begin{aligned}
\leq&|(1-\inf_{\theta'\in\Theta,\lambda'\in\Lambda}p_{\theta',\lambda'})\\
&\cdot{\textstyle\sum}_{\omega'_{t+1}\in\Omega}((V^i(M(\lambda^*_t,\mu_t,\omega_{t+1}),1)\\
&-V^j(M(\lambda^*_t,\mu_t,\omega_{t+1}),1))\\
&\cdot{\textstyle\sum}_{\theta'\in\Theta}q_{\theta',\lambda^*_t}(\omega_{t+1})\mu_t(\theta')\nu_t)|
\end{aligned}\tag{89}$$

$$\begin{aligned}
\leq&(1-\inf_{\theta'\in\Theta,\lambda'\in\Lambda}p_{\theta',\lambda'})\\
&\cdot\sup_{\mu'_{t+1}\in\Delta(\Theta)}|V^i(\mu'_{t+1},1)-V^j(\mu'_{t+1},1)|
\end{aligned}\tag{90}$$

$$\leq\gamma\|V^i-V^j\| \tag{91}$$

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 $V(\mu_t,0)=\bar{Q}_{\hat{\theta}}(\mu_t,0)={\textstyle\sum}_{\theta'\in\Theta}\eta_{\mathrm{b},\theta'}\mu_t(\theta')$.

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 ${\textstyle\sum}_{\omega'_{t+1}\in\Omega}(V(M(\lambda_t,\mu_t,\omega_{t+1}),1;\eta){\textstyle\sum}_{\theta'\in\Theta}(1-$

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

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

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

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

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

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 $\eta_{\text{c},\lambda_t}c_{\lambda_t}$ 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:

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

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 $Q^*_{\lambda_t} = \bar{Q} + \eta_{\text{c},\lambda_t}c_{\lambda_t}$, 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):

$$\lambda^*_t = \arg\sup\nolimits_{\lambda_t\in\Lambda} h(I_t(\lambda_t), S_t(\lambda_t)) - \eta_{\text{c},\lambda_t}c_{\lambda_t}\tag{105}$$

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:

$$Q^*_{\lambda_t}(\mu_t,\nu_t;\eta)\tag{106}$$

$$\begin{aligned}
&=(1-\nu_t)V^*(\mu_t,0;\eta)+\eta_{\text{c},\lambda_t}c_{\lambda_t}\\
&\quad+(\mathbb{E}_{p,q}[V^*((1-\nu_{t+1})\bar{M}(\lambda_t,\mu_t)\\
&\quad+\nu_{t+1}M(\lambda_t,\mu_t,\omega_{t+1}),1;\eta)|\lambda_t,\mu_t,\nu_t=1])\nu_t
\end{aligned}\tag{107}$$

$$\begin{aligned}
&=(1-\nu_t)V^*(\mu_t,0;\eta)+\eta_{\text{c},\lambda_t}c_{\lambda_t}\\
&\quad+(V^*(\bar{M}(\lambda_t,\mu_t),0;\eta)\textstyle\sum_{\theta'\in\Theta}p_{\theta',\lambda_t}\mu_t(\theta')\\
&\quad+\textstyle\sum_{\omega'_{t+1}\in\Omega}(V^*(M(\lambda_t,\mu_t,\omega'_{t+1}),1;\eta)\\
&\quad\cdot\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda_t})q_{\theta',\lambda_t}(\omega'_{t+1})\mu_t(\theta')))\nu_t
\end{aligned}\tag{108}$$

Note that the expectation term can also be expressed:

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

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

So we can rewrite:

$$\begin{aligned}
&\textstyle\sum_{\omega'_{t+1}\in\Omega}\mathbb{P}_{p,q}\{\nu_{t+1}=1,\omega_{t+1}|\lambda_t,\mu_t,\nu_t=1\}\\
&\quad\cdot V^*(M(\lambda_t,\mu_t,\omega_{t+1}),1;\eta)
\end{aligned}\tag{111}$$

$$\begin{aligned}
&=\textstyle\sum_{\omega'_{t+1}\in\Omega}(V^*(M(\lambda_t,\mu_t,\omega_{t+1}),1;\eta)\\
&\quad\cdot\textstyle\sum_{\theta'\in\Theta}(1-p_{\theta',\lambda_t})q_{\theta',\lambda_t}(\omega_{t+1})\mu_t(\theta'))
\end{aligned}\tag{112}$$

$$\begin{aligned}
&=\mathbb{P}_p\{\nu_{t+1}=1|\lambda_t,\mu_t,\nu_t=1\}\\
&\quad\cdot\mathbb{E}_{p,q}[V^*(M(\lambda_t,\mu_t,\omega_{t+1}),1;\eta)|\lambda_t,\mu_t,\\
&\quad\nu_{t+1}=1]
\end{aligned}\tag{113}$$

Hence each Q-factor for acquisitions can be expressed:

$$Q^*_{\lambda_t}(\mu_t,\nu_t;\eta)\tag{114}$$

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

$$=(1-\nu_t)V^*(\mu_t,0;\eta)+\eta_{\text{c},\lambda_t}c_{\lambda_t}$$

$$\begin{aligned}&+(V^*(\bar{M}(\lambda_t,\mu_t),0;\eta)\textstyle\sum_{\theta'\in\Theta}p_{\theta',\lambda_t}\mu_t(\theta')\\&+\mathbb{P}_p\{\nu_{t+1}=1|\lambda_t,\mu_t,\nu_t=1\}\\&\cdot\mathbb{E}_{p,q}[V^*(M(\lambda_t,\mu_t,\omega_{t+1}),1;\eta)|\lambda_t,\mu_t,\\&\nu_{t+1}=1])\nu_t\end{aligned}\tag{116}$$

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

$$\begin{aligned}=&(1-\nu_t)V^*(\mu_t,0;\eta)-I_t(\lambda_t)+\eta_{\mathrm{c},\lambda_t}c_{\lambda_t}\\&+(\frac{\sum_{\theta'\in\Theta}\eta_{\mathrm{b},\theta'}p_{\theta',\lambda_t}\mu_t(\theta')}{\sum_{\theta'\in\Theta}\eta_{\mathrm{b},\theta'}\mu_t(\theta')}\textstyle\sum_{\theta'\in\Theta}\eta_{\mathrm{b},\theta'}\mu_t(\theta')\\&+V^*(\mu_t,1;\eta))\nu_t\end{aligned}\tag{118}$$

$$\begin{aligned}=&(1-\nu_t)V^*(\mu_t,0;\eta)-I_t(\lambda_t)+\eta_{\mathrm{c},\lambda_t}c_{\lambda_t}\\&+((1-S_t(\lambda_t))\textstyle\sum_{\theta'\in\Theta}\eta_{\mathrm{b},\theta'}\mu_t(\theta')\\&+V^*(\mu_t,1;\eta))\nu_t\end{aligned}\tag{119}$$

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

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

$$\quad=-h(I_t(\lambda_t),S_t(\lambda_t))+\eta_{\mathrm{c},\lambda_t}c_{\lambda_t}\tag{121}$$

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:

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

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:

$$\mathbb{P}_{p,q}\{\mathcal{D}|\pi^\kappa_\rho(...;\eta)\}\tag{123}$$

$$=\mathbb{P}_{p,q}\{(\tilde{\lambda}_{n,0:\tau-1},\tilde{\omega}_{n,1:\tau})_{n=1}^N|\kappa,\eta,\rho\}\tag{124}$$

$$\begin{aligned}=&\textstyle\int_{\Delta(\Theta)}\prod_{n=1}^N\prod_{t=0}^{\tau-1}(\mathbb{P}\{\tilde{\lambda}_{n,t}|\mu_{n,t},\nu_{n,t},\kappa,\eta,\rho\}\\&\cdot\mathbb{P}_{p,q}\{\nu_{n,t+1},\tilde{\omega}_{n,t+1}|\tilde{\lambda}_{n,t},\mu_{n,t},\nu_{n,t}\})\\&d\mathbb{P}\{\mu_{n,t+1}|\tilde{\lambda}_{n,t},\mu_{n,t},\tilde{\omega}_{n,t+1}\}\end{aligned}\tag{125}$$

$$\begin{aligned}=&\textstyle\prod_{n=1}^N\prod_{t=0}^{\tau-1}(\mathbb{P}\{\tilde{\lambda}_{n,t}|\mu_{n,t}=\\&M(\lambda_{n,t-1},\mu_{n,t-1},\tilde{\omega}_{n,t}),\nu_{n,t},\kappa,\eta,\rho\}\\&\cdot\mathbb{P}_{p,q}\{\nu_{n,t+1},\tilde{\omega}_{n,t+1}|\tilde{\lambda}_{n,t},\mu_{n,t},\nu_{n,t}\})\end{aligned}\tag{126}$$

$$\begin{aligned}=&\textstyle\prod_{n=1}^N\prod_{t=0}^{\tau-1}\pi^\kappa_\rho(\tilde{\lambda}_{n,t}|M(\lambda_{n,t-1},\mu_{n,t-1},\tilde{\omega}_{n,t}),\\&\nu_{n,t};\eta)\mathbb{P}_{p,q}\{\nu_{n,t+1},\tilde{\omega}_{n,t+1}|\tilde{\lambda}_{n,t},\mu_{n,t},\nu_{n,t}\}\end{aligned}\tag{127}$$

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:

$$\mathbb{P}_{p,q}\{\pi^\kappa_\rho(...;\eta)|\mathcal{D}\}\tag{128}$$

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

where the normalizing constant is given by:

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

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

$$\mathbb{P}\{\pi^\kappa_\rho(...;\eta)|\mathcal{D}\}\tag{131}$$

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

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:

$$V^*(\mu_{n,t},\nu_{n,t};\upsilon\eta+(1-\upsilon)\eta')\tag{133}$$

$$=V^\pi(\mu_{n,t},\nu_{n,t};\upsilon\eta+(1-\upsilon)\eta')\tag{134}$$

$$=\upsilon V^\pi(\mu_{n,t},\nu_{n,t};\eta)+(1-\upsilon)V^\pi(\mu_{n,t},\nu_{n,t};\eta')\tag{135}$$

$$\geq\upsilon V^*(\mu_{n,t},\nu_{n,t};\eta)+(1-\upsilon)V^*(\mu_{n,t},\nu_{n,t};\eta')\tag{136}$$

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 $f:\mathbb{R}^d\rightarrow\mathbb{R}$ 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.