Rational Inattention when Decisions Take Time*

Benjamin Hébert †
Stanford University

Michael Woodford ‡
Columbia University

April 21, 2022

Abstract

Decisions take time, and the time taken to reach a decision is likely to be informative about the cost of more precise judgments. We formalize this insight in the context of a dynamic model of optimal evidence accumulation. We provide conditions under which the resulting belief dynamics resemble either diffusion processes or processes with large jumps. We then consider a limit under which the state-contingent choice probabilities predicted by our model are identical to those predicted by a static rational inattention model, providing a micro-foundation for such models. In the diffusion case, our model provides a normative foundation for a variant of the drift-diffusion model from mathematical psychology.

*The authors would like to thank Mark Dean, Sebastian Di Tella, Mira Frick, Xavier Gabaix, Matthew Gentzkow, Mike Harrison, Emir Kamenica, Divya Kirti, Jacob Leshno, Stephen Morris, José Scheinkman, Ilya Segal, Ran Shorrer, Joel Sobel, Miguel Villas-Boas, Ming Yang, and participants at the Cowles Theory conference, 16th SAET Conference, Barcelona GSE Summer Conference on Stochastic Choice, Stanford GSB research lunch, 2018 ASSA meetings, UC Berkeley theory seminar, and UC San Diego theory seminar for helpful discussions on this topic, Tianhao Liu and Oliver Xie for excellent research assistance, and the NSF for research support. We would particularly like to thank Philipp Strack and Doron Ravid for discussing an earlier version of the paper, and Simon Kelly for sharing data from Kelly et al. [2021]. We would also like to thank the editor (Pietro Ortoleva), associate editor, and six anonymous referees for constructive feedback that helped improve the paper. Portions of this paper circulated previously as the working papers "Rational Inattention with Sequential Information Sampling," "Rational Inattention in Continuous Time," and "Information Costs and Sequential Information Sampling," and appeared in Benjamin Hébert's Ph.D. dissertation at Harvard University. All remaining errors are our own.

†Hébert: Stanford University. Email: bhebert@stanford.edu.

‡Woodford: Columbia University. Email: mw2230@columbia.edu.

1 Introduction

It is common in economic modeling to assume that, when presented with a choice set, a decision maker (DM) will choose the option that is ranked highest according to a coherent preference ordering. However, observed choices in experimental settings often appear to be random, and while this could reflect random variation in preferences, it is often more sensible to view choice as imprecise. Models of rational inattention (such as Matêjka et al. [2015]) formalize this idea by assuming that the DM chooses her action based on a signal that provides only an imperfect indication of the true state. The information structure that generates this signal is optimal, in the sense of allowing the best possible joint distribution of states and actions, net of a cost of information. In the terminology of Caplin and Dean [2015], models of rational inattention make predictions about patterns of state-dependent stochastic choice. These predictions will depend in part on the nature of the information cost, and several recent papers have attempted to recover information costs from observed behavior in laboratory experiments (Caplin and Dean [2015], Dean and Neligh [2019]).

However, in both laboratory experiments and real-world economic settings, decisions take time, and the time required to make a decision is likely to be informative about the nature of information costs.[1] In this paper, we develop a framework to study rational inattention problems in which decisions take time, providing a means of connecting decision times to information costs and state-dependent stochastic choice.

There is an extensive literature in mathematical psychology that focuses on these issues. Variants of the drift-diffusion model (DDM, Ratcliff [1985], Ratcliff and Rouder [1998], Wagenmakers et al. [2007]) also make predictions about stopping times and state-dependent stochastic choice.[2] In particular, these models are designed to match the empirical observation that hasty decisions are likely to be of lower quality.[3] However, these models are not based on optimizing behavior, and this raises a question as to the extent to which they can be regarded as structural; it is unclear how the parameters of the DDM

[1] On the usefulness more generally of data on response times for drawing inferences about the nature of the random error involved in choices, see Alós-Ferrer et al. [2021].

[2] DDM models were originally developed to explain imprecise perceptual classifications. See Woodford [2020] for a more general discussion of the usefulness of the analogy between perceptual classification errors and imprecision in economic decisions.

[3] The existence of a speed-accuracy trade-off is well-documented in perceptual classification experiments (e.g., Schouten and Bekker [1967]). Variants of the DDM that have been fit to stochastic choice data include Busemeyer and Townsend [1993] and more recently Krajbich et al. [2014] and Clithero [2018]; see Fehr and Rangel [2011] for a review of other early work. Shadlen and Shohamy [2016] provide a neural-process interpretation of sequential-sampling models of choice.

model should be expected to change when incentives or the costs of delay change, and this limits the use of the model for making counter-factual predictions. The framework we develop includes as a special case variants of the DDM model, while at the same time making predictions about state-dependent stochastic choice that, in some cases, match those of a static rational inattention (RI) model. Consequently, our framework is able to both speak to the relationship between stopping times and state-dependent stochastic choice (unlike standard RI models) and make counter-factual predictions (unlike standard DDM models).

We propose a class of rational inattention models in which the DM's imprecise perception of the decision problem evolves over time, and an optimization problem determines a joint probability distribution over stopping times and choices. We give conditions under which the dynamics of the belief state prior to stopping will be a pure diffusion (as assumed in the DDM), or alternatively will be a pure jump process (as in the models of Che and Mierendorff [2019] and Zhong [2019]). The diffusion case requires a limit assumption: that discounting effects are negligible relative to the opportunity cost of time. Away from this case, beliefs will follow a pure jump process, as shown by Zhong [2019]. For this more general case, we give conditions under which beliefs will evolve as a series of jumps, approximating a diffusion in the aforementioned limit, and conditions under which the DM will act immediately after the first jump.

Our model is particularly tractable under this limit assumption. In this case, beliefs follow a Markov process and move in a space whose dimensionality is one less than the number of actions (e.g. a line in the case of a binary decision problem, as assumed in the DDM). Our results therefore contribute to the literature on DDM-style models by presenting a model with many features of the DDM, but that — because it is developed as an optimizing model — makes predictions about how decision boundaries and choice probabilities should change in response to changes in incentives.

We also characterize the boundaries of the stopping regions and the predicted ex ante probabilities of different actions in this limiting case, as functions of model parameters including the opportunity cost of time. The key to this characterization is a demonstration that the resulting state-dependent stochastic choice probabilities of our continuous-time model are equivalent to those of a static RI model. Thus in addition to providing foundations for interest in DDM-like models of the decision process, our paper provides novel foundations for interest in static RI problems. For example, we provide conditions under which the predictions of our model will be equivalent to those of a static RI model with the mutual-information cost function proposed by Sims [2010] — and thus equivalent to the model of

stochastic choice analyzed by Matêjka et al. [2015] — but the foundations that we provide for this model do not rely on an analogy with rate-distortion theory in communications engineering (the original motivation for the proposal of Sims).

More generally, as noted above, we show that any cost function for a static RI model in the uniformly posterior-separable family studied by Caplin et al. [Forthcoming] can be justified by the process of sequential evidence accumulation that we describe. This includes the neighborhood-based cost functions discussed in Hébert and Woodford [2021], that lead to predictions that differ from those of the mutual-information cost function in ways that arguably better resemble the behavior observed in experiments such as those of Dean and Neligh [2019]. Our result provides both a justification for using such cost functions in static RI problems, and an answer (not given by static RI theory alone) to the question of how the cost function should change as the opportunity cost of time changes.

The connection that we establish between the choice probabilities implied by a dynamic model of optimal evidence accumulation and those implied by an equivalent static RI model holds both in the case that the belief dynamics in the dynamic model are described by a pure diffusion process and in the case that they are described by a jump process; thus we also show that with regard to these particular predictions, these two types of dynamic models are equivalent. However, the predictions of the two types of model differ with regard to the distribution of decision times, so that it is possible in principle to use empirical evidence to determine which better describes actual decision making.

The key to our analysis is a continuous-time model of optimal evidence accumulation, in which beliefs are martingales (as implied by Bayes' rule). The evolution of beliefs in our model is limited only by a constraint on the rate of information arrival, specified in terms of a posterior-separable cost function. This flexibility is consistent with the spirit of the literature on rational inattention, but with some noteworthy differences. Much of the previous literature considers a static problem, in which a decision is made after a single noisy signal is obtained by the DM. This allows the set of possible signals to be identified with the set of possible decisions, which is no longer true in our dynamic setting.

Steiner et al. [2017] also discuss a dynamic model of rational inattention. In their model, because of the assumed information cost, it is never optimal to acquire information other than what is required for the current action. As a result, in each period of their discrete-time model, the set of possible signals can again be identified with the possible actions at that time. We instead consider situations in which evidence is accumulated over time before any action is taken, as in the DDM; this requires us to model the stochastic

evolution of a belief state that is not simply an element of the set of possible actions.[4] Our central concerns are to study the conditions under which the resulting continuous-time model of optimal information sampling gives rise to belief dynamics and stochastic choices similar to those implied by a DDM-like model, and to study how variations in the opportunity cost of time or the payoffs of actions should affect stochastic choice.

A number of prior papers have endogenized aspects of a DDM-like process. Moscarini and Smith [2001] consider both the optimal intensity of information sampling per unit of time and the optimal stopping problem, when the only possible kind of information is given by the sample path of a Brownian motion with a drift that depends on the unknown state, as assumed in the DDM.[5] Fudenberg et al. [2018] consider a variant of this problem with a continuum of possible states, and an exogenously fixed sampling intensity.[6] Woodford [2014] takes as given the kind of stopping rule posited by the DDM, but allows a very flexible choice of the information sampling process, as in theories of rational inattention. Our approach differs from these earlier efforts in seeking to endogenize *both* the nature of the information that is sampled at each stage of the evidence accumulation process and the stopping rule that determines how much evidence is collected before a decision is made.[7]

Section 2 introduces our continuous-time evidence-accumulation problem, and presents some preliminary results. In section 3, we define two special conditions that information costs may satisfy: a "preference for gradual learning" or a "preference for discrete learning." These properties represent the conditions under which we can show that the optimal belief dynamics will evolve either as a sequence of bounded jumps (a diffusion in the limit case) or a single jump. In section 4 we demonstrate that the state-dependent choice probabilities predicted by our model in the limit case are equivalent to those predicted by a static rational inattention model with a uniformly posterior-separable cost function. In section 5

[4] Our model differs from the one analyzed by Steiner et al. [2017] in several respects. First, as just noted, we study a setting in which the DM takes an action only once, and chooses when to stop and take an action. Second, we consider a much more general class of information costs, as opposed to assuming the mutual information cost. And third, we assume that the DM has a motive to smooth her information gathering over time, rather than learn all of the relevant information at a single point in time.

[5] Moscarini and Smith [2001] allow the instantaneous variance of the observation process to be freely chosen (subject to a cost), but this is equivalent to changing how much of the sample path of a given Brownian motion can be observed by the DM within a given amount of clock time.

[6] See also Tajima et al. [2016] for analysis of a related class of models, and Tajima et al. [2019] for an extension to the case of more than two alternatives.

[7] Both Morris and Strack [2019] and Zhong [2019] adopt our approach, and obtain special cases of the relationship between static and dynamic models of optimal information choice that we present below. Che and Mierendorff [2019] and Zhong [2019] both differ from our treatment in not considering conditions under which beliefs will evolve as a diffusion process.

we discuss how the diffusion and jump cases can nonetheless be distinguished using data on response times. Section 6 concludes.

2 Dynamic Models of Rational Inattention

Let X be a finite set of possible states of nature. The state of nature is determined ex-ante, does not change over time, but is not known to the DM. Let $q_t \in \mathscr{P}(X)$ denote the DM's beliefs at time $t \in [0,\infty)$, where $\mathscr{P}(X)$ is the probability simplex defined on X. We will represent q_t as vector in $\mathbb{R}_+^{|X|}$ whose elements sum to one, each of which corresponds to the likelihood of a particular element of X, and use the notation $q_{t,x}$ to denote the likelihood under the DM's beliefs at time t over the true state being $x \in X$.

At each time t, the DM can either stop and choose an action from a finite set A, or continue to acquire information. Let τ denote the time at which the DM stops and makes a decision, with $\tau = 0$ corresponding to making a decision without acquiring any information. The DM receives utility $u_{a,x}$ if she takes action a and the true state of the world is x, and pays a flow cost of delay per unit time, $\kappa \geq 0$, until an action is taken. Let $\hat{u}(q_\tau)$ be the payoff (not including the cost of delay) of taking an optimal action under beliefs q_τ:

$$\hat{u}(q_\tau) = \max_{a \in A} \sum_{x \in X} q_{\tau,x} u_{a,x}.$$

We assume $u_{a,x}$ is strictly positive, and discuss the implications of this assumption below.

If the DM does not stop and act, she can gather information. We adopt the rational inattention approach to information acquisition and assume that the DM can choose any process for beliefs satisfying "Bayes-consistency," subject to a further constraint (specified below) on the rate of information acquisition. In a single-period model, Bayes-consistency requires that the expectation of the posterior beliefs be equal to the prior beliefs. The continuous-time analog of this requirement is that beliefs must be a martingale.

Let the DM's initial beliefs be $\bar{q}_0 \in \mathscr{P}(X)$, and let $\Omega = \mathbb{D}(\mathbb{R}^{|X|+1})$ (i.e. the space of possible paths of a càdlàg $\mathbb{R}^{|X|+1}$-valued process).[8] We allow the DM to choose any filtered probability space on Ω, $(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_{t \in \mathbb{R}_+}, P)$, and stochastic process, $q : \Omega \times \mathbb{R}_+ \to \mathscr{P}(X)$, such that q_t is a càdlàg $\{\mathscr{F}_t\}$-martingale and $q_0 = \bar{q}_0$, subject to an additional constraint specified below in Equation (3).

[8] Using this space as opposed to $\mathbb{D}(\mathscr{P}(X))$ allows us to consider stopping strategies τ that are not measurable in beliefs.

Example. A Markovian diffusion: The DM could choose

$$dq_t = Diag(q_{t^-})\sigma(q_{t^-}) \cdot dB_t \tag{1}$$

where $Diag(q_{t^-})$ is a diagonal matrix with q_{t^-} on the diagonal, σ is an $|X| \times (|X|-1)$ matrix-valued function and B_t is an $(|X|-1)$-dimensional Brownian motion. To ensure that q_t remains in the simplex, we must have $q^T \cdot \sigma(q) = \vec{0}$ for all $q \in \mathscr{P}(X)$.

Example. K Markovian jump processes: The DM could choose, for some integer $K > 0$,

$$dq_t = -\sum_{k=1}^{K} \psi_k(q_{t^-}) z_k(q_{t^-}) dt + \sum_{k=1}^{K} z_k(q_{t^-}) dJ_t^k, \tag{2}$$

where each J_t^k is an independent Poisson process with intensity $\psi_k(q_{t^-})$. To ensure that beliefs remain in the simplex and satisfy Bayes-consistency, the functions z_k must be such that, for all $q \in \mathscr{P}(X)$, $q + z_k(q)$ is also in the simplex and absolutely continuous with respect to q.

In both of these examples, we assume sufficient regularity to ensure the existence and uniqueness of a solution to the SDE.[9] These two examples could also be combined, to generate a jump-diffusion process. The quantities σ_t and $z_{k,t}$ could also be allowed to vary with time in a more complex way, rather than having to be functions of the current belief q_{t^-} as specified above.[10]

We assume that DM is subject to a constraint on how fast her beliefs can evolve, specified in terms of a "posterior-separable" constraint on the rate of information flow (as in the static rational inattention problems considered by Caplin et al. [Forthcoming]). Posterior-separable constraints are defined in terms of a divergence, $D : \mathscr{P}(X) \times \mathscr{P}(X) \to \mathbb{R}_+$, which is defined for all $(q', q) \in \mathscr{P}(X) \times \mathscr{P}(X)$ such that $q' \ll q$.[11] By the definition of a divergence, $D(q'||q)$ is zero if and only if $q' = q$, and strictly positive otherwise. We extend D to $R_+^{|X|} \times R_+^{|X|}$ by assuming the function to be homogenous of degree one in each of

[9] Because $\mathscr{P}(X)$ is a compact subset of $\mathbb{R}^{|X|}$, Lipschitz continuity of the σ, ψ, and z functions is sufficient; see e.g. Pham [1998].

[10] That is, we do not assume the belief process is Markovian, and whether or not such policies are optimal is not the focus of our analysis. Our results will in some cases show that Markovian optimal policies exist.

[11] We assume here that D is finite for q', q on the boundary of the simplex, provided that $q' \ll q$, as is true for example of the widely-used Kullback-Leibler divergence. But our results could readily be extended to cover the case in which D is infinite for such values.

its arguments. We also assume that it is strongly convex in its first argument and twice continuously-differentiable in both arguments.[12]

We require the DM's belief process to satisfy

$$\limsup_{h\downarrow 0} \frac{1}{h} E^P[D(q_{t+h}||q_{t^-})|\mathscr{F}_{t^-}] \leq \chi, \tag{3}$$

where $\chi > 0$ is a finite constant.[13] This constraint can be understood as the continuous-time analog of requiring that $E_{t^-}[D(q_{t+h}||q_{t^-})] \leq \chi h$ in a discrete time model with time interval h. Note also that in what follows, we will use the notation $E_t[\cdot]$ to indicate $E^P[\cdot|\mathscr{F}_t]$.

The divergence D that defines this constraint is central to our analysis. Specifically, we investigate the relationship between properties of this divergence and properties of optimal beliefs processes chosen by the DM. One of the core questions we consider is whether it is optimal for the DM to choose a beliefs process that diffuses (as in our first example above) or jumps (as in our second example above). When considering the constraint in the context of a diffusion, only the local properties of the divergence D are relevant. These local properties are summarized by the Hessian matrix that characterizes $D(q'||q)$ up to second order when q' is close to q. Define $\bar{k}(q)$ as the $|X| \times |X|$ matrix-valued function defined on the interior of the simplex by

$$\bar{k}_{x,x'}(q) = \frac{\partial^2 D(q'||q)}{\partial q'_x \partial q'_{x'}}|_{q'=q}, \tag{4}$$

and extended to the boundary by continuity. As the following example illustrates, this matrix-valued function characterizes the constraint (3) for diffusion processes.[14]

Example. A Markovian diffusion: in the context of the diffusion process (1), the constraint (3) requires that $\sigma(q)$ satisfy the additional condition

$$\frac{1}{2} tr[\sigma(q)^T Diag(q)\bar{k}(q)Diag(q)\sigma(q)] \leq \chi \tag{5}$$

for all $q \in int(\mathscr{P}(X))$.

[12]Strong convexity, in this context, implies that $D(q'||q) \geq K|q'-q|^2$ for some constant $K > 0$. The assumption of homogeneity of degree one allows us to define differentiability in the usual way on $\mathbb{R}^{|X|}$.

[13]Technical footnote: we require only that (3) hold for all $t \in \mathbb{R}_+$, P-almost-everywhere. That is, the process q_t is indistinguishable from a process for which the constraint holds everywhere.

[14]For a proof, see Lemma 7 in the appendix.

In contrast, the global properties of the divergence D govern the constraint in the context of a pure jump process.

Example. K Markovian jump processes: in the context of the jump process (2), the constraint (3) requires that, for all $q \in \mathscr{P}(X)$,

$$\sum_{k=1}^{K} \psi_k(q) D(q + z_k(q) || q) \leq \chi. \tag{6}$$

We have specified the set of possible belief processes in this way to emphasize the connection between our approach in continuous time and the standard, discrete-time approach to rational inattention.[15] The constraint (3) implies a tradeoff between more frequent but less informative movements in beliefs and rarer but larger movements in beliefs. Suppose that the DM would like her beliefs to follow a jump process of the kind specified in (2). The DM can choose rare but informative signals (small $\psi_k(q)$, large $D(q + z_k(q)||q)$) or more frequent but less informative signals (larger $\psi_k(q)$, smaller $D(q + z_k(q)||q)$). In fact, there exists a limit in which jumps become very likely and very small ($|z_k| \to 0, \psi_k \to \infty$) and the stochastic process of beliefs and the information constraint for the jump process (2) converge to the stochastic process and constraint for a diffusion process. That is, the constraint (3) ensures continuity between the cost of a continuous belief process and the cost of a belief process with very small jumps.

Let $\mathscr{A}$ denote the set of feasible policies (i.e. filtered probability spaces defined on Ω, stochastic processes for beliefs consistent with (3), and stopping times), and let $\rho \geq 0$ denote the DM's rate of time preference. We will assume that at least one of ρ or κ is strictly positive, so that the DM faces some cost of delay.

Definition 1. The DM's problem given initial belief $\bar{q}_0 \in \mathscr{P}(X)$ is

$$V(\bar{q}_0) = \sup_{((\Omega, \mathscr{F}, \{\mathscr{F}_t\}, P), q, \tau) \in \mathscr{A}} E_0[e^{-\rho\tau} \hat{u}(q_\tau) - \kappa \int_0^\tau e^{-\rho s} ds].$$

We next discuss in more detail several features of our modeling approach.

[15] The working paper version of this paper (Hébert and Woodford [2019]) derives a version of our continuous-time problem by considering the limit of a sequence of discrete-time problems.

2.1 Remarks on the Model

Generality of the Beliefs Process. Our model allows the DM to choose from large space of possible beliefs processes, which we view as consistent with the spirit of the rational inattention paradigm. However, as we will show in our preliminary analysis below, the DM's problem can be restricted to a smaller and more tractable set of beliefs processes without reducing the utility achieved in the DM's problem.

Strictly Positive Utility. We assume in our model (following Zhong [2019]) that the utility function is strictly positive. In the $\rho = 0$, $\kappa > 0$ case, this assumption is unnecessary, and considering negative utilities would not change any results. In the $\rho > 0$, $\kappa = 0$ case, the value of never making a decision is zero. The economic implication of the assumption of strictly positive utility is that any action taken in finite time dominates never making a decision. This condition, which is stronger than necessary, ensures that optimal stopping times are well-behaved.

Discounting and the No-discounting Limit. Much of our analysis will focus on the case without discounting ($\rho = 0$), or on the limiting case in which $\rho \to 0^+$. We focus on these cases because they are more tractable, allowing us to obtain sharper results, and because we are motivated by settings in which the opportunity cost of time is large relative to the cost of the effects of discounting.

Consider for example a problem in which the maximum possible reward no more than \$2000. Assume a rate of time preference of 20% annualized (reflecting a very high degree of impatience) and a \$5/hour (less than the current US minimum wage) opportunity cost of time κ. The hourly cost of receiving the reward one hour later due to discounting is at most $\frac{20\%}{\text{year}} \times \frac{\$2000}{8760 \text{ hours per year}} \approx 0.05$ dollars per hour, roughly one hundred times smaller than the opportunity cost of time under these assumptions. Many decision problems, including in particular the kind of laboratory experiments we discuss in Section 5, involve even smaller stakes, and as a result in these problems we expect $\rho \hat{u}(q)$ to be small relative to κ. We will show a kind of continuity in the limit as ρ approaches zero, which we discuss in more detail below, justifying the approach of viewing the $\rho = 0$ case as an approximation of the case in which $\rho \hat{u}(q)$ is small relative to κ.

Information Constraints vs. Information Costs. We have described our model in terms of a constraint on rate at which information can be acquired. However, we would have

reached identical results had we instead treated information as having a utility cost. Both approaches are common in the rational inattention literature, and equivalent for our purposes, although they make different predictions in certain settings (e.g. with respect to the effect of "scaling up" the utility function u on behavior). In the working paper version of this paper (Hébert and Woodford [2019]), we discussed both primal (constraints) and dual (utility costs) problems, and provided some equivalence results.

In the case of no discounting ($\rho = 0$), whether the information is subject to a utility cost or constraint is irrelevant: the optimal policies are identical across the two cases. This property comes from the fact that the cost of delay is constant. In the case with discounting ($\rho > 0$), the cost of delay depends in part on the current level of the value function, which generates variation in the amount of information acquired with costly information, but not when information acquisition is constrained. Our results, however, are not sensitive to the differences between the optimal policies in these two cases.

Conditional vs. Unconditional Dynamics. The continuous time problem just described uses the "unconditional" dynamics for the beliefs q_t, meaning that beliefs are martingales. That is, by the usual Bayesian logic, the DM can never expect to revise her beliefs in any particular direction. The model can also be described in terms of the conditional dynamics for beliefs, which is to say how the beliefs evolve conditional on the true state $x \in X$. To illustrate this point, consider the Markovian diffusion example. In this example, conditional on the true state being $x \in X$, the DM's beliefs q_t follow a diffusion of the form[16]

$$dq_t \;=\; Diag(q_{t^-})\sigma(q_{t^-})\sigma(q_{t^-})^T e_x dt \;+\; Diag(q_{t^-})\sigma(q_{t^-})dB_{t|x}, \tag{7}$$

where e_x is a vector equal to one in the element corresponding to x and zero otherwise. Note that this implies that, if we write $\mu_{t^-|x}$ for the drift rate of $q_{t,x}$ in (7),

$$\mu_{t^-|x} \;=\; e_x^T Diag(q_{t^-})\sigma(q_{t^-})\sigma(q_{t^-})^T e_x \;\geq\; 0.$$

Thus, the DM will tend to assign more probability to the true state as evidence accumulates.

Likewise, consider the K Markovian jump processes example. In this example, condi-

[16]This expression follows from Bayes' rule and the Girsanov's theorem.

tional on the true state being $x \in X$,

$$dq_t = -\sum_{k=1}^{K} \psi_k(q_{t^-}) z_k(q_{t^-}) dt + \sum_{k=1}^{K} z_k(q_{t^-}) dJ^k_{t|x},$$

where $J^k_{t|x}$ is an independent Poisson process with intensity $\psi_k(q_{t^-})(1 + \frac{z_{k,x}(q_{t^-})}{q_{t^-,x}})$.[17] Jumps that increase the likelihood of the true state ($z_{k,x}(q_{t^-}) > 0$) are relatively more likely conditional on $x \in X$, consistent with the idea that the DM will tend to place more weight on the true state as time passes.

Studying the unconditional dynamics of beliefs is more convenient in most of our analysis; for this reason, we do not generalize our expressions of the conditional dynamics beyond the two examples just described. Moreover, these two examples are sufficient for the purposes of the example considered in Section 5 and to discuss the relationship between our model and DDM models.

Relation to DDM Models. In DDM models (see, e.g., Fudenberg et al. [2018]), a "decision variable" z_t is assumed to follow a process

$$dz_t = \delta_{|x} dt + \alpha dB_{t|x}, \tag{8}$$

where $\delta_{|x}$ is a drift that depends on $x \in X$, and $B_{t|x}$ is a Brownian motion conditional on $x \in X$. In the classic DDM, the decision variable z_t is assumed to be one-dimensional, and the DM is assumed to stop and choose from a set of two possible actions when z_t reaches one of the two ends of a line segment (each corresponding to one of the available actions). The classic DDM framework thus generates predictions about the joint distribution of actions, states, and decision times.

To understand the relationship between our optimizing model and DDM models, suppose that the DM in our model chooses a Markov diffusion process for beliefs, as in (1), and that this diffusion process travels along a line segment within the simplex. The conditional dynamics for beliefs will follow (7), and the belief process q_t in our model will have properties similar to those posited for the "decision variable" z_t in the DDM model. In particular, it will be diffusion process with a drift that depends on the true state x and an instantaneous variance that is independent of the state. In this case, discussed in Section 5, our model will generate predictions about the joint distribution of actions, states, and

[17] Again, this result follows from Bayes' rule and Girsanov's theorem.

decision times that closely resemble those of the DDM model.

Below, we establish conditions under which it will be optimal for the belief process to be a diffusion of this kind. Moreover, we establish conditions under which, in the case of a choice between only two possible actions, it is optimal for the DM in our model to choose a belief process that diffuses on a line until it reaches one of two stopping boundaries, as posited by the DDM.[18] That is, under certain conditions, the predictions of our model and those of the DDM model will coincide.

2.2 Preliminary Analysis

We first show that an optimal policy exists. This ensures that the questions we hope to address, such as when optimal policies involve only jumps or diffusions, have answers.

Lemma 1. *An optimal policy exists in the DM's problem.*

Proof. See the technical appendix, section C.3. □

We next show that the value function for our problem must satisfy a Hamilton-Jacobi-Bellman (HJB) equation. This is not trivial, because in our context, the value function need not be twice continuously-differentiable, and consequently the HJB equation cannot be derived in the usual fashion. We take an alternative approach using viscosity techniques to show that the value function is once continuously-differentiable, and that it is a solution to an HJB equation of a simpler problem.

To simplify our notation, we extend the definition of V to the set of positive measures ($\mathbb{R}_+^{|X|}$) by assuming homogeneity of degree one, and define the gradient of V, ∇V, in the usual way. Also, for any belief $q \in \mathscr{P}(X)$, let $Q(q)$ be the subset of $\mathscr{P}(X)$ consisting of all beliefs q' such that $q' \neq q, q' \ll q$ (the set for which $D(q'||q)$ is defined and non-zero).

Proposition 1. *Let $V(q)$ be the value function that solves the DM's problem (Definition 1). This value function is continuously differentiable on the interior of $\mathscr{P}(X)$ and the interior*

[18] It is well known that optimal Bayesian decision making would imply a process of this kind in the special case that (i) there are only two possible states x, so that the posterior necessarily moves on a line, and (ii) the only possible kind of information sampling is observation of a particular Brownian motion with state-contingent drift, so that the DM's only decision is when to stop observing and choose an action, as in Fudenberg et al. [2018]. The novelty of our result is that we allow a flexible choice of the kind of information that is sampled, subject to (3), and that our result applies regardless of the number of states in X.

of each face of $\mathscr{P}(X)$, and satisfies, for all $q \in \mathscr{P}(X)$,

$$\max\{\sup_{q' \in Q(q)} \frac{V(q') - V(q) - (q'-q)^T \cdot \nabla V(q)}{D(q'||q)} - \rho V(q) - \kappa, \quad \hat{u}(q) - V(q)\} = 0.$$

Proof. See the appendix, section B.2 □

This is the HJB equation of a restricted version of our problem in which the DM is constrained not to diffuse and to jump to only one destination (a process of the form (2) with $K = 1$). That is, imposing such a restriction on the belief dynamics does not reduce the DM's value function. Note that optimal policies may not exist in this restricted problem, if it is in fact strictly optimal to diffuse in the original problem; in such a case, a sequence of "pure jump" policies involving ever-smaller and more frequent jumps achieves the supremum. The useful general characterization of the value function in Proposition 1 allows us to establish further properties of optimal belief dynamics in a variety of special cases.

3 Preferences for Gradual and Discrete Learning

We next study the relationship between properties of the divergence D and properties of beliefs under optimal policies. We consider two cases: when there is a "preference for gradual learning" and when there is a "preference for discrete learning," terms we define below. With discounting, these two classes of divergences lead, respectively, to beliefs that evolve as a series of jumps and beliefs that jump only once. In the case of zero discounting, a preference for gradual learning leads to beliefs that diffuse, as in the DDM model.

3.1 Gradual Learning

We begin by defining what we call a "preference for gradual learning." This condition describes the relative costs of learning via jumps in beliefs vs. continuously diffusing beliefs, which are governed by the properties of the divergence D.

Definition 2. The divergence D exhibits a "*preference for gradual learning*" if, for all $q, q' \in \mathscr{P}(X)$ with $q' \ll q$,

$$D(q'||q) \geq (q'-q)^T \cdot \left(\int_0^1 (1-s)\bar{k}(sq' + (1-s)q)ds\right) \cdot (q'-q). \tag{9}$$

This preference is "strict" if the inequality is strict for all $q' \neq q$, and is "strong" if, for some $\delta > 0$ and some $m > 0$,

$$D(q'||q) \geq (1 + m|q' - q|^{\delta})(q' - q)^T \cdot (\int_0^1 (1-s)\bar{k}(sq' + (1-s)q)ds) \cdot (q' - q). \quad (10)$$

Note that, by the definition of $\bar{k}$, (9) holds with equality up to second order in $q' - q$. A preference for gradual learning requires that the difference of the higher-than-second-order terms be positive, a strict preference requires that they be strictly positive as q' approaches q, and a strong preference requires that they be of order $|q' - q|^{2+\delta}$.

One special case of particular interest involves Bregman divergences (such as the Kullback-Leibler divergence commonly used in the rational inattention literature). A Bregman divergence can be written, using some convex function $H : \mathscr{P}(X) \to \mathbb{R}$, as

$$D_H(q'||q) = H(q') - H(q) - (q' - q)^T \cdot \nabla H(q), \quad (11)$$

where $\nabla H(q)$ denotes the gradient. For a Bregman divergence, $\bar{k}(q)$ is the Hessian of $H(q)$, and (9) is an equality for all $q, q' \in \mathscr{P}(X)$.

Divergences exhibiting a (strict or strong) preference for gradual learning can be easily constructed from Bregman divergences. Suppose that

$$D(q'||q) = f(D_H(q'||q)),$$

where $f : \mathbb{R}_+ \to \mathbb{R}_+$ is a twice continuously-differentiable, strictly increasing, convex function with $f(0) = 0$, $f'(0) = 1$, and D_H is a Bregman divergence. The Hessian of D evaluated at $q' = q$ is the same as that of D_H, and by convexity

$$D(q'||q) \geq D_H(q'||q),$$

implying that D also exhibits a preference for gradual learning. This preference is strict if $f(\cdot)$ and $H(\cdot)$ are strictly convex, and strong if $H(\cdot)$ and $f(\cdot)$ are strongly convex.

Define $u_{max} = \max_{q \in \mathscr{P}(X)} \hat{u}(q)$ and $u_{min} = \min_{q \in \mathscr{P}(X)} \hat{u}(q)$. Our first main result is that when D exhibits a strong preference for gradual learning, under any optimal policy the probability of a jump of size greater than $(\frac{\rho(u_{max} - u_{min})}{m(\kappa + \rho u_{min})})^{\delta^{-1}}$ is zero. The result follows from (10), which in effect says that continuously moving beliefs from q to q' is less difficult (in the sense of the constraint (3)) than jumping. When $\rho = 0$, this is sufficient to show that

beliefs evolve continuously. When $\rho > 0$, the cost of delay depends in part on the level of the current value function, which encourages the DM to choose a belief process that jumps upwards as opposed to drifting upwards. The upper bound we derive in this case reflects a balancing of this consideration against the ease of having beliefs evolve gradually as opposed to discretely.

Proposition 2. *Define* $\Delta q_t = q_t - \lim_{s \uparrow t} q_s$. *If D exhibits a strong preference for gradual learning, then under any optimal policy,*

$$Pr\{\sup_{t \in \mathbb{R}_+} |\Delta q_t|^\delta > \rho \frac{(u_{max} - u_{min})}{m(\kappa + \rho u_{min})}\} = 0.$$

Proof. See the appendix, section B.5. □

The tightness of this bound depends in part on the relationship between ρu_{max} and κ. The former is an upper-bound on the cost of delay due to discounting, while the latter represents a direct opportunity cost of time. When $\rho = 0$ and $\kappa > 0$, this bound implies that the optimal belief process is a continuous martingale. More generally, when ρu_{max} is small relative to κ (with "small" being defined relative to the strength of the preference for gradual learning as parameterized by m and δ), the belief process will involve some combination of a continuous martingale and small jumps. In fact, a remarkable result from Zhong [2019] shows that with $\rho > 0$, the optimal policy involves only jumps outside of a nowhere-dense set. The combination of the Zhong [2019] result and Proposition 2 demonstrates that the optimal belief processes when ρu_{max} is strictly positive but small relative to κ involve small jumps. There is a kind of continuity between the $\rho > 0$ but small and $\rho = 0$ cases (assuming $\kappa > 0$). In the limit as $\rho \to 0^+$ holding fixed all of the other parameters, these jumps will become smaller and smaller, and the optimal belief process will eventually converge to a continuous martingale.[19]

In the particular case of no discounting ($\rho = 0, \kappa > 0$), we can reach stronger conclusions. An optimal beliefs process is in this case a continuous martingale, irrespective of the strength of the preference for gradual learning (m and δ). The following proposition provides a kind of upper hemi-continuity result in this case, showing that a continuous martingale belief process (and in fact a diffusion) continues to be optimal in the case of a (possibly non-strong or non-strict) preference for gradual learning. Recall that the matrix-valued function $\bar{k}(q)$ is defined in (4).

[19] We prove this continuity result in the technical appendix, section C.1.

Proposition 3. *If $\rho = 0$ and D exhibits a preference for gradual learning, then V is a viscosity solution (see e.g. Crandall et al. [1992]) to the HJB equation*

$$\max\{\sup_{\sigma \in \mathbb{R}^{|X|} \times \mathbb{R}^{|X|-1} : q^T \sigma = \vec{0}} tr[\sigma^T Diag(q)(\nabla^2 V(q) - \frac{\kappa}{\chi}\bar{k}(q))Diag(q)\sigma], \hat{u}(q) - V(q)\} = 0, \tag{12}$$

where $\nabla^2 V$ denotes the Hessian of V, and there exists an optimal policy such that q_t is a diffusion without jumps.

Proof. See the appendix, section B.6. □

Under an additional assumption on the matrix-valued function $\bar{k}$, we show that a preference for gradual learning is not only sufficient but necessary for beliefs to follow a diffusion process when $\rho = 0$. Specifically, we assume that $\bar{k}$ is "integrable," in the sense described by the following assumption.[20]

Assumption 1. *There exists a twice continuously-differentiable function $H : \mathbb{R}_+^{|X|} \to \mathbb{R}$ such that, for all q in the interior of the simplex,*

$$\bar{k}(q) = \nabla^2 H(q), \tag{13}$$

where $\nabla^2 H(q)$ denotes the Hessian of H evaluated at q and $\bar{k}(q)$ is define as in (4).

Any Bregman divergence has this property; as a result, the class of divergences satisfying this property includes the standard KL divergence and the "neighborhood-based" function that we introduce in Hébert and Woodford [2021]. Our earlier examples of divergences with a strong preference for gradual learning, which are not Bregman divergences themselves but were constructed by applying a convex function to a Bregman divergence, also satisfy this property. In these cases, the H function is the function used to define the Bregman divergence. This assumption is also automatically satisfied in the two state case, $|X| = 2$. However, this assumption imposes some restrictions if $|X| > 2$. It rules out, for example, the prior-invariant LLR cost functions of Pomatto et al. [2018] (a hypothetical H would have asymmetric third-derivative cross-partials). Note that $H(q)$ is convex, by the positive semi-definiteness of $\bar{k}(q)$, and homogenous of degree one.

[20] Mathematically, this assumption ensures that the integral $\int_0^1 (q' - \gamma(s))^T \cdot \bar{k}(\gamma(s)) \cdot \frac{d\gamma(s)}{ds} ds$ is the same for all differentiable paths of integration $\gamma : [0,1] \to \mathscr{P}(X)$ with $\gamma(0) = q$ and $\gamma(1) = q'$. That is, the straight-line path of integration used to define a preference for gradual learning (Definition 2) is without loss of generality.

Under this assumption, we demonstrate that a preference for gradual learning is necessary for beliefs to always result in a diffusion process. The key idea is that, if the DM always prefers diffusing to jumping, $\rho = 0$, and beliefs travel along a line segment in the simplex, then along that line segment (9) must hold. Assumption 1, combined with results we present in Section 4 below, allows us to construct utility functions such that it is optimal for the DM's beliefs to travel along any given line segment.[21]

Proposition 4. *Assume $\rho = 0$. If, given a divergence D, Assumption 1 is satisfied and, for all strictly positive utility functions $u_{a,x}$, there exists an optimal policy such that beliefs follow a diffusion process, then D exhibits a preference for gradual learning.*

Proof. See the appendix, section B.7. This is proven using Proposition 7 below. □

Summarizing, a preference for gradual learning is sufficient and, under an additional assumption, necessary to guarantee that an optimal belief process is a diffusion in the $\rho = 0$ case. A strong preference for gradual learning is sufficient to guarantee that jumps are bounded, and there is a kind of continuity between the $\rho > 0$ and $\rho = 0$ cases.

However, there are also circumstances in which large jumps are optimal. Zhong [2019] shows, in the particular case of $\rho > 0$ and Bregman divergence costs (equality in (9)), that the beliefs jump all the way to stopping points. This is striking in light of Proposition 3, which shows that with these same costs and $\rho = 0$, beliefs can follow a diffusion process. These results can be reconciled using results we will present in the next section: with Bregman divergence costs and $\rho = 0$, there are optimal policies that generate both pure diffusion and pure-jump belief processes.

3.2 Discrete Learning

We next provide conditions under which the DM jumps immediately to stopping beliefs, as a contrast to our previous gradual learning results. We define what we call a "preference for discrete learning" if the divergence D satisfies a kind of "chain rule" inequality.[22]

[21] The difficulty of extending this result (without our additional assumption, or with $\rho > 0$) is as follows. We know in these cases that if beliefs always diffuse or jump in small increments, then such behavior must be preferable to larger jumps within the continuation region of a given problem. But because we cannot construct explicit solutions in these cases, we cannot prove that this preference holds on the entire simplex.

[22] When this inequality holds with equality, the divergence is said to satisfy the chain rule property (Cover and Thomas [2012]).

Definition 3. The divergence D exhibits a "*preference for discrete learning*" if it satisfies, for all finite sets S, $\pi_s \in \mathscr{P}(S)$ and $q, q', \{q_s\}_{s\in S} \in \mathscr{P}(X)$ such that $\sum_{s\in S} \pi_s q_s = q'$ and $q' \ll q$,

$$D(q'||q) + \sum_{s\in S} \pi_s D(q_s||q') \geq \sum_{s\in S} \pi_s D(q_s||q). \tag{14}$$

Here, S is an arbitrary finite set; it is useful to think of each $s \in S$ as a signal realization, and to interpret $\{q_s\}$ as a set of posteriors consistent with a prior q'. If (14) holds, it is preferable to jump from q directly to the posteriors $\{q_s\}$ instead to the prior q'.

Bregman divergences satisfy (14) with equality (a result that follows from the definition (11)). One might expect that other classes of cost functions also exhibit a preference for discrete learning. However, as the following lemma demonstrates, under our regularity assumptions,[23] only the Bregman divergences exhibit a preference for discrete learning.[24]

Lemma 2. *The divergence D exhibits a preference for discrete learning if and only if D is a Bregman divergence.*

Proof. See the appendix, section B.8. The proof builds on Banerjee et al. [2005]. □

Consequently, if D exhibits a preference for discrete learning, it also exhibits a (non-strict) preference for gradual learning. In contrast, many cost functions exhibit a strict or strong preference for gradual learning and therefore do not exhibit a preference for discrete learning, and many others fall into neither category.[25]

If the cost function satisfies a preference for discrete learning, it is cheaper for the DM to jump to beliefs $\{q_s\}$ rather than visit the beliefs q'. Unsurprisingly, if this holds everywhere, it leads to optimal policies that stop immediately after jumping. We first show in the case of $\rho = 0$ that an optimal policy always involves jumping into the stopping region.

Proposition 5. *Define $\Delta q_t = q_t - \lim_{s\uparrow t} q_s$, and assume $\rho = 0$. If D exhibits a preference for discrete learning, then there exists an optimal Markovian pure jump process such that if $|\Delta q_t| > 0$, then $t = \tau$ (the DM stops immediately after any jump).*

[23] Our regularity assumptions are important here; it is possible that non-differentiable, non-Bregman divergences exhibiting a preference for discrete learning exist.

[24] It is known in the information geometry literature (see e.g. Amari and Nagaoka [2007]) that if (14) holds with equality, the divergence must be a Bregman divergence. A related result appears in Frankel and Kamenica [2019], in the context of their "order invariance" property. Our lemma extends these results by showing that the inequality in (14) is sufficient.

[25] For example, any strictly concave transformation of a Bregman divergence (as opposed to the convex transformations described previously) is not itself a Bregman divergence and does not exhibit a preference for gradual learning.

Proof. See the appendix, section B.9. This is proven using Proposition 7 below. □

The statement of Proposition 5 shows that if D is a Bregman divergence, is without loss of generality to assume that the DM stops immediately after a jump in beliefs. But in this case, there is also an optimal policy that diffuses (Proposition 3). This observation implies that the solutions to the HJB equations in Propositions 1 and 3 must be identical, despite one being written as controlling a diffusion process and the other a pure jump process. We revisit this observation in the next section.

Zhong [2019] (see appendix A.3 of that paper) presents a result that covers the $\rho > 0$ case.[26] With a preference for discrete learning, jumps will increase the value function, in the absence of a jump the value function will drift downwards, and stopping occurs immediately after the first jump. The intuition for these results comes from the observation that with discounting, delay is particularly costly when the value function is high. Zhong [2019] also shows that optimal policies do not involve diffusion (subject to some technical caveats).

Moving beyond the results of Zhong [2019], we provide an "only-if" result: if a divergence always results in immediate stopping after the first jump, then it must satisfy a preference for discrete learning. The intuition is that if it is always optimal to jump outside the continuation region, it cannot be less costly under the divergence D to jump to an intermediate point. Otherwise, there would be some utility function for which such behavior is optimal. To formalize this result, we say that the beliefs process q_t "does not diffuse" if the continuous part of the martingale q_t has zero quadratic variation.[27]

Proposition 6. *Define $\Delta q_t = q_t - \lim_{s \uparrow t} q_s$. Suppose the divergence D is such that, for all action spaces A, strictly positive utility functions $u_{a,x}$, and priors $\bar{q}_0 \in \mathscr{P}(X)$, there exists an optimal policy that does not diffuse on the interior of the continuation region outside of a nowhere-dense set and such that $|\Delta q_t| > 0$ implies $t = \tau$ (the DM stops after jumping). Then D exhibits a preference for discrete learning (i.e. is a Bregman divergence).*

Proof. See the appendix, section B.10. □

This result demonstrates that the jump-and-immediately-stop results of Proposition 5 and Zhong [2019] hold for all utility functions if and only if D is a Bregman divergence.

[26] The result from Zhong [2019] applies when $\kappa = 0$; but with $\rho > 0$, the $\kappa > 0$ problem is equivalent to a problem in which the utility function is shifted upwards by $\kappa\rho^{-1}$ and κ is set to zero (by Proposition 1).

[27] See e.g. theorem 4.18 of chapter I of Jacod and Shiryaev [2013] on the decomposition of martingales into a continuous martingale and discontinuous martingale.

Such cases are knife-edge, in that if one uses instead any strongly convex transformation of the Bregman divergence, then the optimal policy will involve bounded jumps (by Proposition 2) that converge to continuous processes as ρ becomes close to zero.

3.3 Gradual vs. Discrete Learning

We summarize the differences between gradual and discrete learning before proceeding. Note again that the technical appendix, section C.1, contains a formal result on the upper hemi-continuity of policies in the $\rho \to 0^+$ limit. With $\rho > 0$ and a sufficiently strong preference for gradual learning, the DM will optimally choose to have beliefs that jump in small increments. In the limit as $\rho \to 0^+$, these jumps will become infinitesimal, and in the $\rho = 0$ case the DM will optimally choose to have continuous beliefs. In contrast, with $\rho > 0$ and a preference for discrete learning, the DM will optimally choose to have beliefs that jump immediately into the stopping region. In the limit as $\rho \to 0^+$, this will continue to be case; however, when $\rho = 0$ and the DM has a preference for discrete learning, an optimal policy involving only diffusions also exists.

We interpret these results as follow. In the $\rho \to 0^+$ limit, which should be understood as assuming that the cost of delay due to discounting is dominated by the opportunity cost of time and which we view as empirically relevant, beliefs will either jump or diffuse, depending on whether the divergence D exhibits a strong preference for gradual learning or a preference for discrete learning. However, the value functions in these two cases will be identical. These results naturally lead to the question of whether these differences in belief dynamics lead to different predictions about the DM's behavior. We explore this question in the next two sections.

4 Equivalence to a Static RI Model

In this section, we show that a preference for gradual learning has the same implications as a preference for discrete learning, as far as the models' predictions regarding state-contingent choice behavior are concerned, in the limit in which $\rho = 0$. This is because in either case the state-contingent choice behavior predicted by our dynamic model of evidence accumulation is identical to that predicted by a static rational inattention model with a particular type of information-cost function. Two different dynamic models (defined by different divergences D) can be equivalent (in the sense of implying the same value function $V(q)$, and hence

leading to the same choice behavior) to the same static model; in particular, this will be true when both divergences are derived from the same entropy function $H(q)$, as shown below. In this case, the two dynamic models imply identical choice behavior. It further follows from our continuity results that even in the discounted case, predicted state-contingent choice behavior becomes identical in the two cases in the limit as $\rho \to 0^+$.

We first consider the case of a preference for gradual learning. The following result follows from an analysis of the HJB equation of Proposition 3 (the problem with $\rho = 0$ and a diffusion process for beliefs).

Proposition 7. *If $\rho = 0$, D exhibits a preference for gradual learning, and Assumption 1 holds, the value function is given by*

$$V(q_0) = \max_{\pi \in \mathscr{P}(A), \{q_a \in \mathscr{P}(X)\}_{a \in A}} \sum_{a \in A} \sum_{x \in X} \pi_a q_{a,x} u_{a,x} - \frac{\kappa}{\chi} \sum_{a \in A} \pi(a) D_H(q_a || q_0), \tag{15}$$

where the maximization is subject to the constraint that $\sum_{a \in A} \pi(a) q_a = q_0$, and D_H is the Bregman divergence associated with the entropy function H that is assumed to exist in Assumption 1.

This value function can be achieved by a homogenous diffusion process. Moreover, there exist maximizers π^* and q_a^* such that π^* indicates the unconditional probability of choosing the different actions in the continuous time problem, and such that for any a for which $\pi^*(a) > 0$, q_a^* is the belief the DM will hold when stopping and choosing that action.

Proof. See the appendix, section B.11. □

The problem stated in Proposition 7 is simply a static rational-inattention problem, in which the cost of the static information structure defined by the joint distribution of states and action choices is given by a uniformly posterior-separable cost function of the kind proposed by Caplin et al. [Forthcoming]. The mutual information cost function proposed by Sims is one such cost function. In this case, the entropy function H is the negative of Shannon's entropy, the corresponding Bregman divergence is the Kullback-Leibler divergence, and the information cost is mutual information. Thus Proposition 7 provides a foundation for this familiar kind of RI model, and hence for the predictions regarding stochastic choice obtained by Matêjka et al. [2015]. On the other hand, Proposition 7 also implies that other cost functions can be justified in a similar way. Indeed, any (twice-differentiable) uniformly posterior-separable cost function can be given such a justification, by choosing the $\bar{k}$ function defined by equation (13).

However, the same set of static models can also be justified by a dynamic model with discrete learning.

Corollary 1. *Assume $\rho = 0$ and that D exhibits a preference for discrete learning (i.e., is a Bregman divergence). Then the value function that solves the continuous time problem is the value function that solves the static rational inattention problem described in Proposition 7, with D in the place of D_H.*

Proof. This follows immediately from Lemma 2, Proposition 3, and Proposition 7. □

Given any uniformly posterior-separable cost function in a static rational inattention model, by setting D equal to the Bregman divergence associated with that cost function, we can justify that static model as the result of a dynamic model with a preference for discrete learning. We therefore conclude that models with a preference for gradual learning satisfying our integrability condition and models with a preference for discrete learning are indistinguishable from the perspective of their predictions about the joint distribution of states and actions.[28] In the next section, however, we show how information about stopping times can nonetheless be used to distinguish the models.

Finally, we should emphasize that in the case of gradual learning, our results depend on an additional assumption (Assumption 1); this integrability assumption will not hold in all cases. Consequently, equivalence with static models holds for all cost functions with a preference for discrete learning, but for only some cost functions in the case of a preference for gradual learning. And while all cost functions with a preference for discrete learning generate the same joint distribution of actions and states as some cost function with a preference for gradual learning, the reverse need not be true.

5 Implications for Response Times

Because our model is dynamic, it makes predictions not only about the joint distribution of actions and states, but also about the length of time that should be taken to reach a decision, and how this may vary depending on the action and the state. In the experimental literature on the accuracy of perceptual judgments, it is common to record the time taken for a subject to respond along with the response, as this is considered to give important

[28] In situations in which the static rational inattention problem does not itself have a unique solution, we have not ruled out the possibility that the models with discrete and gradual learning will make different predictions. However, we have no reason to believe this to be the case.

information about the nature of the decision process (e.g., Ratcliff and Rouder [1998]). In economic contexts as well, response times provide important information that can be used to discriminate between models, even when response times themselves are not what the economic analyst cares about. For example, Clithero [2018] and Alós-Ferrer et al. [2021] argue that preferences can more accurately be recovered from stochastic choice data when data on response times are used alongside observed choice frequencies.

Here we propose that data on response times can in principle be used to discriminate between alternative information-cost specifications. We focus on the zero-discounting limit, and show that cost functions that are equivalent in the sense of implying the same state-contingent choice probabilities — and hence the same value function — nevertheless make different predictions about the stopping time conditional on taking a particular action. Consequently, data on response times can inform us about whether there is a preference for gradual learning or for learning through discrete jumps.

5.1 A Two-State, Two-Action Example

We consider a simple example, motivated by an experiment reported by Kelly et al. [2021]. In the experiment, a stimulus is presented that is of one of two types, and the subject must report which of the two types was presented; the goal is to maximize the number of correct responses. We therefore assume there are two states ($X = \{\ell, r\}$), and two possible actions ($A = \{L, R\}$). Response L is the correct response when $x = \ell$, and R is correct when $x = r$. Since the DM's reward depends only on whether the response is correct or not, $u_{L,\ell} = u_{R,r} = u_{cor}$ and $u_{L,r} = u_{R,\ell} = u_{inc}$, with $u_{cor} > u_{inc}$.[29]

In this setting, we describe the predictions of our model when the divergence D exhibits a preference for discrete learning (i.e. is a Bregman divergence), and compare these predictions with those derived under the alternative assumption of a strong preference for gradual learning. We will compare divergences that are equivalent in the sense of the previous section: they will lead to the same predictions about the joint distribution of actions and states, but make different predictions about stopping times.

In either case, we assume that information costs are based on an entropy function $H(q)$, a homogeneous degree one function of (q_ℓ, q_r), twice continuously differentiable, and strongly convex when restricted to the unit simplex. We further assume that the function is symmetric, in the sense that $H(q_\ell, q_r) = H(q_r, q_\ell)$. One example of such a function

[29] Our characterizations of optimal behavior in this section can be generalized to apply to a more flexible class of multi-state, two-action examples, as shown in the appendix, section A.1.

is the negative of the Shannon entropy function,

$$H(q) \;=\; q_\ell \ln(\frac{q_\ell}{q_l+q_r}) + q_r \ln(\frac{q_r}{q_l+q_r}); \tag{16}$$

another is the "total information" (TI) cost of Bloedel and Zhong [2020],[30]

$$H(q) \;=\; (q_\ell - q_r)(\ln(q_\ell) - \ln(q_r)). \tag{17}$$

Any entropy function of this kind defines a Bregman divergence D_H, as in (11); we will call this "the PDL case."[31] On the other hand, if we let $D(q'||q) = f(D_H(q'||q))$, where f is a twice continuously differentiable, increasing, strongly convex function such that $f(0) = 0$ and $f'(0) = 1$, the information costs exhibit a strong preference for gradual learning; we will call this "the PGL case." We wish to compare the predictions regarding the distribution of response times in these two cases. We restrict our discussion to the $\rho \to 0^+$ limit, in which case the choice of the function f does not matter.

In this limit, the prediction regarding state-contingent response frequencies are identical in the two cases, and described by the static rational inattention problem defined in Proposition 7. The optimal stopping beliefs are easily characterized in the two-state case. For each action $a \in A$, let $q_a^* \in \mathscr{P}(X)$ be the posterior that maximizes

$$\sum_{x \in X} q_{a,x} u_{a,x} - \frac{\kappa}{\chi} H(q_a).$$

The solutions to these equations are uniquely defined, and satisfy $0 \leq q^*_{R,\ell} < 1/2 < q^*_{L,\ell} \leq 1$, and $q^*_{R,\ell} = 1 - q^*_{L,\ell}$. Then for any prior such that $q^*_{R,\ell} < q_{0,\ell} < q^*_{L,\ell}$, the unique solution to the static problem problem in Proposition 7 is given by[32]

$$\bar{q}_L \;=\; q^*_L, \qquad \bar{q}_R \;=\; q^*_R, \qquad \bar{\pi}_L \;=\; \frac{q_{0,\ell} - q^*_{R,\ell}}{q^*_{L,\ell} - q^*_{R,\ell}}.$$

[30] Desirable properties of this alternative to the Shannon measure of information costs are also discussed in Hébert and Woodford [2021].

[31] Recall that a cost function that satisfies our general regularity assumptions, and that exhibits a preference for discrete learning, is necessarily a Bregman divergence (Lemma 2). Thus our analysis of the PDL case is not very restrictive. Here we define the PDL case in terms of a particular entropy function so that we can define a corresponding PGL case for any PDL case.

[32] The fact that the stopping posteriors $\bar{q}_a$ are independent of the prior q_0 (for priors in the stated range) reflects the property of "locally invariant posteriors" discussed more generally by Caplin et al. [Forthcoming].

The accuracy rate (the likelihood of a correct response) in this problem is $\alpha \equiv q^*_{L,\ell} = q^*_{R,r}$.

While the stopping posteriors are the same in the two cases, the predicted dynamics of beliefs are different. In the PDL case, the posterior q_t will drift deterministically until (at some random time) it jumps to either q^*_L or q^*_R (at which time the decision is made). In the PGL case, instead, the posterior q_t diffuses randomly along the line segment that connects q^*_L and q^*_R, with the decision being made on the first occasion when one of the stopping posteriors is reached. These two models of belief dynamics imply different state-contingent distributions of response times, as we now illustrate with explicit calculations.

5.2 State-Contingent Distributions of Response Times

It is convenient to express the dynamics of beliefs in terms of a scalar state variable $\pi_{L,t} \in [0,1]$,[33] which corresponds to the posterior

$$q_t \;=\; q(\pi_{L,t}) \;\equiv\; q^*_R + \pi_{L,t}(q^*_L - q^*_R). \tag{18}$$

Here $\pi_{L,t}$ indicates the probability that the DM's eventual response will be L, conditional on the beliefs reached at time t; a decision is made when $\pi_{L,t}$ reaches 0 or 1. We define the value function $V(\pi_L)$ by substituting (18) into the $V(q)$ defined in Proposition 7, and let $\underline{\pi}_L$ be the value of the argument at which $V(\pi_L)$ reaches its minimum. In the symmetric case assumed here, $V(\pi) = V(1-\pi)$, so that $\underline{\pi}_L = 1/2$, as shown in the left panel of Figure 1.

5.2.1 The PDL case

We first consider belief dynamics in the PDL case. In this case, as emphasized by Zhong [2019], jumps must always increase the value function, and a deterministic drift of beliefs must steadily reduce the value function.[34] It follows that $\pi_{L,t}$ drifts toward $1/2$ deterministically until a jump to one of the stopping posteriors occurs.

If we start from a prior under which the state ℓ is more likely, then $\pi_{L,0} = \bar{\pi}_L > 1/2$.

[33] This representation of the belief dynamics continues to be possible in a more general case with many states (though only two possible actions), considered in the appendix, section A.1. This is because the DM's posterior q_t must always be a convex combination of the stopping posteriors $\bar{q}_R$ and $\bar{q}_L$, even when those posteriors lie in a high-dimensional space.

[34] Zhong [2019] shows in the $\rho > 0$ case that beliefs drift away from the currently most likely state; we consider the optimal policy in the $\rho = 0$ limit that is the limit of a sequence of such policies. The fact that the drift of beliefs must reduce the value function follows by applying the envelope theorem to the HJB equation of Proposition 1; see Zhong [2019].

(See the left panel of Figure 1 for an illustration.) In this case, the optimal policy is to jump towards $\pi_L = 1$ with the maximum possible intensity (with no jumps to the more distant stopping posterior at $\pi_L = 0$), and drift downwards if no jump occurs, as long as beliefs continue to satisfy $\pi_{L,t} > 1/2$. Constraint (6) implies that the Poisson arrival rate of jumps to $\pi_L = 1$ will be given by $\psi_L(\pi_{L,t}) = \chi / D_H(q_L^* || q(\pi_{L,t}))$. In order for q_t to be a martingale, the rate at which $\pi_{L,t}$ drifts downward deterministically in the absence of a jump must be

$$\mu(\pi_{L,t}) \;=\; -\frac{\chi(1-\pi_{L,t})}{D_H(q_L^* || q(\pi_{L,t}))}. \tag{19}$$

Integrating the differential equation for $\pi_{L,t}$ implied by this drift rate (and using the initial condition $\pi_{L,0}$ implied by the prior), we can compute the finite (and deterministic) time $\tilde{\tau} > 0$ at which $\pi_{L,\tilde{\tau}} = \underline{\pi}_L = 1/2$ in the absence of a jump.

Once beliefs reach $\pi_{L,\tilde{\tau}} = \underline{\pi}_L$, the DM randomizes between jumping to $\pi_L = 0$ and to $\pi_L = 1$ at equal Poisson rates,[35] where the common jump rate is the maximum one consistent with (6).

We summarize these results in Lemma 3 and Figure 1 below. We present the results in terms of response time *quantiles* $\hat{\tau}$ (as opposed to the response times τ). For each $\tau > 0$, we define the quantile $\hat{\tau} \equiv F(\tau)$, where $F(\cdot)$ is the cumulative distribution function of response times. Since there are no atoms in the response-time distributions in either the PDL or PGL cases, the distribution of quantiles $\hat{\tau}$ is a uniform distribution on $[0,1]$. In Figure 1, we report $g_L^x(\hat{\tau})$, which is the conditional likelihood of action L given that the true state is $x \in X$ and that the response time quantile is $\hat{\tau}$.

There are several advantages to expressing our models' testable predictions in terms of response-time quantiles rather than response times. First, we obtain quantitative predictions that are independent of time units, and thus independent of the numerical value of χ. Second, the response time observed in a laboratory experiment should not be identified with the decision time τ in our theoretical model. Instead, empirical estimation of stochastic models like the DDM always interprets the measured response time as an observation of $t_0 + \tau$, where t_0 (the "non-decision time," NDT) is a positive constant to be estimated.[36] The

[35] In the symmetric case considered here, $\underline{\pi}_L = 1/2$, and the two kinds of jumps occur at equal rates for times $t > \tilde{\tau}$. In the appendix, section A.2.2, we show how the characterization of belief dynamics here can be generalized to asymmetric cases as well, in which $\underline{\pi}_L$ need not equal 1/2. The rates at which the two kinds of jumps occur are then no longer equal, but instead have the relative values required in order for $\pi_{L,t}$ to be a martingale.

[36] For example, in Ratcliff and Rouder [1998] and Wagenmakers et al. [2007] this parameter is denoted T_{er}, while in Clithero [2018] it is written as *ndt*.

NDT may represent an unavoidable time lag between the experimenter's presentation of a stimulus to the subject and the beginning of the evidence-accumulation process, or a lag between the time τ at which the latent decision variable first reaches a stopping region and the subject's overt response. Predictions for the distributions of response-time quantiles $\hat{\tau}$ are instead independent of the value of t_0. Third, we can derive quantitative predictions for the response-time quantile distribution in the PDL case that do not depend on the functional form of $H(q)$, which is not true of the response-time distribution.

Lemma 3. *Consider a symmetric two-state/two-action problem of the kind described in the text, with a preference for discrete learning, and suppose that the prior q_0 is such that $1/2 < q_{0,\ell} < \bar{q}_{L,\ell}$ and that the accuracy rate is α. Then conditional on the true state x, the fraction of trials on which $\pi_L(t)$ reaches the value 1/2 prior to a decision is equal to $\hat{\tau}^x$, where*

$$\hat{\tau}^{\ell} = \left(\frac{\alpha}{\alpha - 1/2}\right)\left(\frac{q_{0,\ell} - 1/2}{q_{0,\ell}}\right), \qquad \hat{\tau}^{r} = \left(\frac{1-\alpha}{\alpha - 1/2}\right)\left(\frac{q_{0,\ell} - 1/2}{1 - q_{0,\ell}}\right). \tag{20}$$

The response-time quantile density function $g_L^x(\hat{\tau})$ is piecewise constant, equal to 1 for all $0 < \hat{\tau} < \hat{\tau}^x$, and equal to $q^*_{L,x}$ for all $\hat{\tau}^x < \hat{\tau} < 1$.

Proof. See the appendix, section A.2.1. □

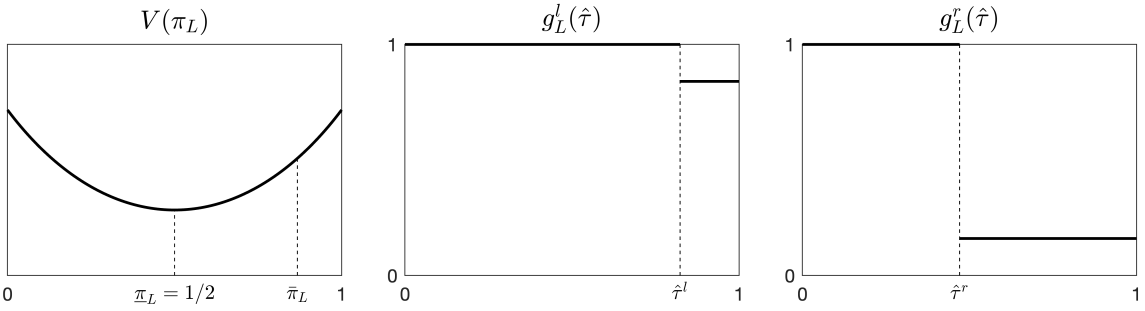

Figure 1: Predicted response-time quantile distributions with a preference for discrete learning. The first panel shows the value function $V(\pi_L)$, and the other two show the density functions $g_L^x(\hat{\tau})$ for the two possible states. Information costs are parameterized so that the predicted accuracy rate is $\alpha = 0.84$, and the prior is assumed to be one under which $q_{0,\ell} = 0.75$. The value function in the first panel is derived for a static RI problem in which $H(q)$ is the negative of the Shannon entropy function.

A complete quantitative description of the g_L^{ℓ} and g_L^r functions requires the numerical specification of only two quantities: the prior probability $q_{0,\ell}$ that the state is ℓ and the

accuracy rate α (which completely determines the stopping posteriors, given the symmetry of the problem). The latter quantity represents the only way in which the specification of $H(q)$ affects the predictions.

Figure 1 shows this solution in the case of a prior with $q_{0,\ell} = 0.75$ and an accuracy rate $\alpha = 0.84$, the values that accord with the data from the experiment in Kelly et al. [2021]. The first panel of the figure shows the value function, in the case that the cost function is the one implied by Shannon entropy (16), and χ/κ is specified so as to imply an accuracy rate of 0.84.[37] The other two panels show the functions $g_L^x(\hat{\tau})$ specified by Lemma 3.

If instead we assume a prior such that r is the more likely state ($q^*_{R,\ell} < q_{0,\ell} < 1/2$), we can obtain symmetric results.[38] The density functions shown in Figure 1 are also (with suitable relabeling) the predicted density functions g_R^r and g_R^ℓ in the case of a prior for which $q_{0,\ell} = 0.25$ (the same degree of prior uncertainty, but state r is more likely).

5.2.2 The PGL case and the DDM

We now consider the dynamics of $\pi_{L,t}$ under a strict preference for gradual learning. In this case, $\pi_{L,t}$ evolves as a diffusion on the line (18), starting from the initial condition $\pi_{L,0} = \bar{\pi}_L$. Hence the unconditional belief dynamics are of the form

$$dq_t \;=\; (q^*_L - q^*_R)\bar{\sigma}(\pi_{L,t})dB_t, \tag{21}$$

where $\bar{\sigma}(\pi_L)$ is scalar-valued (and equal to the largest quantity consistent with (5)) and dB_t is a one-dimensional Brownian motion. The belief dynamics conditional on the state x are also given by a diffusion, as specified in the following lemma.

Lemma 4. *Consider a two-action problem,*[39] *with a strong preference for gradual learning. Conditional on the state x, the univariate belief state $\pi_{L,t}$ evolves as*

$$d\pi_{L,t|x} \;=\; \frac{q^*_{L,x} - q^*_{R,x}}{q_x(\pi_{L,t})}\bar{\sigma}(\pi_{L,t})^2 dt \;+\; \bar{\sigma}(\pi_{L,t})dB_{t|x}, \tag{22}$$

[37] This is purely for purposes of illustration; while the cost function assumed in the first panel is consistent with the numerical assumptions in the other two panels, the predictions in the second and third panels, that we wish to test, do not depend on assuming Shannon entropy.

[38] See the appendix, section A.2.1, for further discussion. We can also characterize the densities $g_a^x(\hat{\tau})$ in the case of a symmetric prior, but in this case the solution is trivial (the two responses are equally likely at all quantiles), and there is also no difference between the predictions of the PDL and PGL cases. We accordingly focus on the asymmetric case.

[39] This result does not require that $|X| = 2$; see the appendix, section A.3.1.

where $q(\pi_L)$ is the function defined in (18), and $\bar{\sigma}(\pi_L)$ is the same function as in (21).

Proof. See the appendix, section A.3.1. The result follows directly from (7). □

This process resembles the internal "decision variable" specified by DDM models in mathematical psychology in several respects. The state variable $\pi_{L,t|x}$ diffuses until reaching one of two fixed boundaries (zero or one), which correspond to the two possible decisions. States for which L is relatively more likely ($\frac{q^*_{L,x}}{q^*_{R,x}}$ positive) feature upward drift, and the strength of this drift is stronger in states for which the relatively probability of choosing L is higher. The only difference between these dynamics and those posited by the DDM is that in general, neither the drift term nor the variance term in (22) is constant.

However, when $X = \{l, r\}$ and $H(q)$ is the TI entropy function (17), the belief dynamics implied by our model with a strict preference for gradual learning are exactly like those assumed in the standard DDM. Parameterize the belief state q_t by the implied log odds $z_t = \ln(q_{t,\ell}/q_{t,r})$, a smooth nonlinear transformation of $\pi_{L,t}$. Then, as shown in the appendix, section A.3.2, the evolution of the decision variable z_t is given by

$$dz_{t|G} = \chi dt + \sqrt{2\chi} dB_{t|G}, \qquad dz_{t|B} = -\chi dt + \sqrt{2\chi} dB_{t|B}. \tag{23}$$

These are exactly the dynamics postulated in the DDM, with the difference between the drifts associated with the two states determined by the information bound χ. A decision will be made when the variable z_t first reaches one or the other of two stopping values z^*_a, which are just the log-odds transformations of the stopping posteriors q^*_a determined by the solution to the static RI problem associated with the TI cost function.

We turn now to the predicted conditional distributions of response time quantiles in the PGL case. This amounts to studying the distribution of hitting times for a diffusion of the form (22), with initial condition $\pi_{L,0} = \bar{\pi}_L$ and absorbing boundaries at $\pi_L = 0$ and $\pi_L = 1$. As shown in the appendix, section A.3.3, we can compute these distributions by solving a partial differential equation subject to suitable boundary conditions. Figure 2 illustrates the solutions obtained for the response-time quantile density functions (using the same format as in Figure 1), in the case of the two alternative information cost functions (16) and (17).

In the PGL case (unlike the PDL case), the exact form of the cost function matters; because the belief state must diffuse from $\pi_{L,0}$ to one or the other of the boundaries, the local properties of the cost function over the entire line segment (18) are relevant, and not simply the aspects of the cost function that determine the two stopping posteriors. However, as the figure shows, in the case of our two example $H(q)$ functions, the predicted

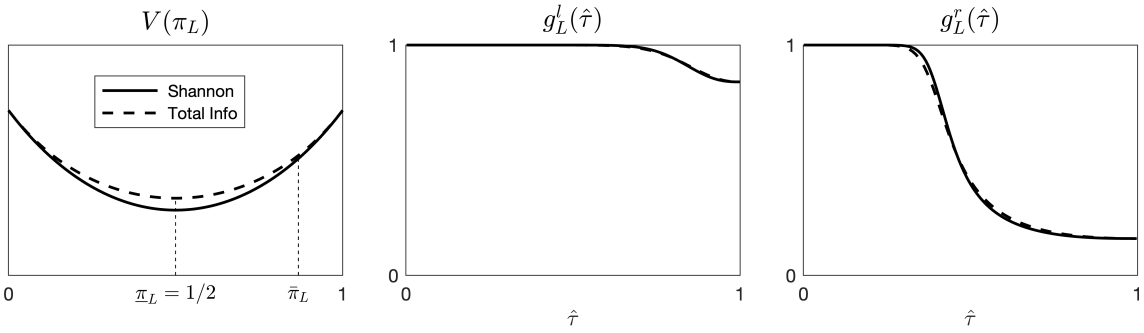

Figure 2: Predicted response-time distributions with a preference for gradual learning. The format is the same as in Figure 1. Each of the functions is shown for two alternative choices of the entropy function $H(q)$: negative Shannon entropy (solid lines, as in the first panel of Figure 1) and the entropy that results in a TI cost function (dashed lines). The accuracy rate α and the prior q_0 are parameterized as in Figure 1.

response-time quantile distributions are not very different, provided that the two models are parameterized so as to imply the same degree of accuracy. The differences between either of these PGL cases and the predictions for the PDL case in Figure 1 are more striking. In particular, when beliefs diffuse (rather than jumping), the relative frequency of L decisions as opposed to R decisions, conditional on a given state x, varies continuously with the response-time quantile $\hat{\tau}$, rather than being piece-wise constant and discontinuous.

As in the PDL case, the symmetry of the problem implies that if we reverse the prior probabilities of the two states, the predicted density functions $g_a^x(\hat{\tau})$ remain the same under a suitable relabeling. Thus Figures 1 and 2 provide the complete set of quantitative predictions that we wish to compare with the experimental data of [Kelly et al., 2021].

5.3 Empirical Evidence on Discrete vs. Gradual Learning

In the experiment of Kelly et al. [2021], subjects view a visual image of moving dots, and must decide whether the dominant direction of motion is leftward or rightward. Thus, as in the situation analyzed above, there are two possible responses L or R (indicating that the motion is leftward or rightward). Subjects' rewards in the experiment (and most likely any "psychic rewards" that they receive as well) depend only on whether a response is correct or not, and not on what the true direction was; hence the utilities should satisfy the symmetry property assumed above. In the trials of interest to us here, the subject also observes a color cue on each trial, before presentation of the visual image, which indicates that one direction

of motion is more likely than the other. Depending which cue is received on a given trial, the subject's prior should be either $q_{0,\ell} = 0.75$ or $q_{0,r} = 0.75$. (The cue is said to have a "75 percent validity" in either case.) Each cue is presented equally often.

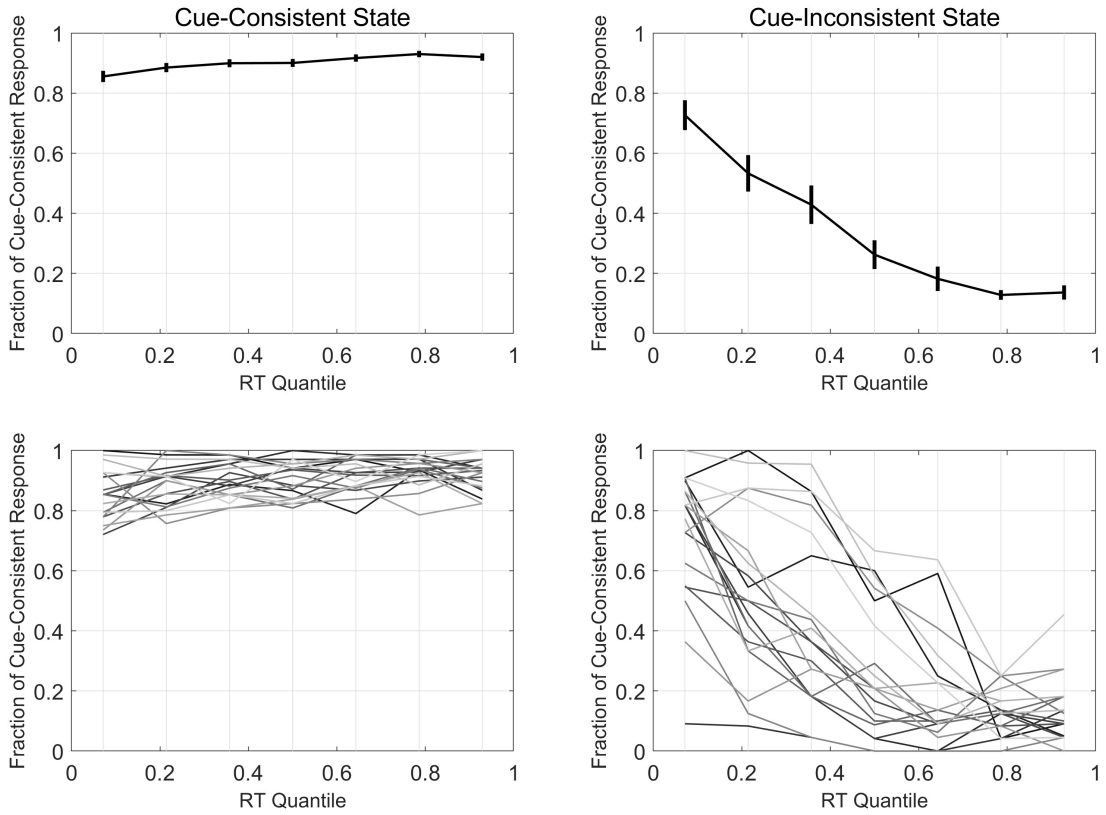

Figure 3: The relative frequency of cue-consistent and cue-inconsistent responses, as a function of the speed of the response, in the experiment of Kelly et al. [2021]. The curves shown in the left panels represent empirical versions of the function $g_L^{\ell}(\hat{\tau})$ shown in the middle column in Figures 1 and 2, while those in the right panel represent empirical versions of the function $g_L^{r}(\hat{\tau})$ shown in the right columns of the earlier figures. The top row presents estimates of the two curves obtained by pooling the data of all 20 subjects, while the bottom row shows the estimated curves for each of the individual subjects.

The data shown in Figure 3 indicate subjects' responses under what the authors call the "deadline" condition, which is the one under which subjects are under the greatest time pressure; this is the condition of most interest for our purposes, because the limited evidentiary basis for subjects' decisions is clearest in this case.[40] (Under the "deadline"

[40]Kelly *et al.* impose time pressure on their subjects by imposing a deadline, with a reward for correct responses only if they are made before the deadline. This fixed deadline acts as an additional cost of delay

condition, subjects' responses are correct only 84 percent of the time; average accuracy is much higher in the other experimental conditions.) The numerical predictions in Figures 1 and 2 have accordingly assumed a cue validity of 0.75 and an accuracy rate of 0.84.

The symmetry of the decision problem implies that in both the PDL and PGL cases, the predicted response-time quantile density functions $g_a^x(\hat{\tau})$ depend only on whether x is the state with higher prior probability (in which case the cue is "valid," in the terminology of Kelly *et al.*) and on whether a is the response indicated by the cue as more likely to be correct (i.e., the response is cue-consistent), and not on which cue was presented. Hence (as is also done in Kelly *et al.*) we pool the data from trials with cues of either type, and classify the trials according to whether the state is cue-inconsistent or cue-inconsistent, rather than according to whether the true state is ℓ or r. (This allows us a larger sample from the response-time distribution for each case.) Thus the left and right columns in Figure 3 correspond to the distributions for which theoretical predictions are presented in the middle column and right column, respectively, of Figures 1 and 2. (The vertical axis of each panel now plots the relative frequency of cue-consistent responses, rather than the relative frequency of L responses.)

The top row of Figure 3 (reproduced from Figure 3a of Kelly et al. [2021]) shows the empirical correlates of the two functions $g_L^\ell(\hat{\tau})$ and $g_L^r(\hat{\tau})$ graphed in Figures 1 and 2, when we pool the data of all 20 experimental subjects. Even when we pool the data of all subjects and sort the trials only on the basis of cue-consistency, we still have only a finite number, each with a specific response time; to estimate the conditional probability of a cue-consistent response, we must average over sufficiently wide ranges of quantiles. Thus in Figure 3 we group the responses into seven bins of approximately equal size: the 1/7 fastest responses, the 1/7 next-fastest, and so on. In the top row, the dot for each bin indicates the overall fraction of cue-consistent responses in that bin, as an estimate of the probability of a cue-consistent response; the vertical line indicates a range of estimates corresponding to this mean estimate plus or minus one standard error of measurement.[41]

above and beyond the subjects' opportunity cost of time; our model considers only the latter kind of cost. However, it is unlikely that subjects are able to estimate very accurately how much time they have left in the Kelly *et al.* experiment; this can be seen from the fact that they do not all respond at the same time, just before the deadline is reached. Thus we conjecture that the deadline has the effect of increasing subjects' perceived cost of continuing to deliberate by an amount that is relatively independent of the time already taken. We leave for future work an analysis of the consequences of a time cost that varies with the elapsed time in a setting in which subjects have limited awareness of the time elapsed.

[41] Here (following Kelly et al. [2021]) we treat the fraction of cue-consistent responses for each of the 20 subjects as an independent noisy observation of their common probability of giving a cue-consistent response, allowing a standard error to be computed for the estimate of that common probability.

The resulting estimates of the response-time quantile density functions do not look at all like the step functions shown in Figure 1. In particular, under the parameter values appropriate to the experiment, the discontinuity in the cue-inconsistent case (the right panel of Figure 3) should fall within the middle range of quantiles: one should observe a constant probability of cue-consistent responses (100 percent) in each of the first three bins, and another (much lower) constant probability in each of the last three bins, with an intermediate average probability in the central bin. Instead one sees what looks more like a steadily decreasing probability of cue-consistent responses the slower the response, as predicted by the PGL model and the DDM.

While it is common to fit pooled data of this kind to some version of the DDM, we cannot necessarily reject the PDL model simply on the basis of the curves shown in the top row of Figure 3. It is possible that the appearance of a gradually declining curve in the top right panel of the figure could reflect pooling of the data of individual subjects, each of whose response-time distribution was a step function of the kind predicted by the PDL model, but with very different values of the critical quantile $\hat{\tau}^r$, because of their differing information costs. In the second row of the figure, we consider this possibility by separately plotting the response frequencies by quantile for each of the 20 subjects.

We observe in the lower right panel that there are significant differences across subjects with regard to the fraction of responses on cue-inconsistent trials that are made too soon for the subject's response to be more likely to be correct than incorrect. Nonetheless, even when we disaggregate the data by subject, and allow for the possibility that the discontinuity in the response probability might occur earlier than the middle range of quantiles, it does not appear that a subject's probability of a cue-consistent response is constant once it drops below some very high value, as predicted by the PDL model. Instead, cue-inconsistent (i.e., correct) responses are more frequent in the case of the slowest responses, as predicted by the PGL model or the DDM.[42] Thus the results of Kelly et al. [2021] are more consistent with the predictions of the PGL model than those of the PDL model.

There are other differences between our model's predictions for the PDL and PGL cases

[42]For each subject, we let n be the largest quantity (less than or equal to 4) such that cue-consistent responses are more frequent than cue-inconsistent choices in each of the bins prior to bin n; thus if the subject's relative-frequency curve is a step function, the discontinuity may be inferred to occur in bin n. We find that for 15 of the subjects, the fraction of cue-consistent choices is lower on average in the last two bins (the slowest 2/7 of the subject's choices) than in bin $n+1$ (the first one in which all choices should be at percentiles greater than the one at which the discontinuity occurs). There are instead only two subjects for whom the inequality is reversed, making it unlikely that the difference between earlier and later decisions (among all those later than bin n) is due merely to random sampling from the same probability distribution in each bin.

that should in principle be testable. For example, as shown in the appendix, section A.2.2, it is possible to modify the payoffs in the decision problem in a way that makes no difference for the solution $(\bar{\pi}_L, \bar{q}_L, \bar{q}_R)$ of the static RI problem, but that adds a linear term to the value function $V(\pi_L)$. Because the second derivative of the value function is unchanged, such a change in the payoffs implies no change in the way that beliefs diffuse in the PGL case; but because the location of $\underline{\pi}_L$ changes, the predicted response-time distributions in the PDL case are different. Thus again it should be possible to discriminate between the PDL and PGL cases by looking at conditional distributions of response times (though not by looking at state-contingent response frequencies alone); but because we know of no such experiments, we leave this for further investigation elsewhere.

6 Discussion and Conclusion

We have proposed a continuous-time model of optimal evidence accumulation, and established conditions under which the state-contingent stochastic choices predicted by such a model coincide with those of a static rational inattention model. Our result provides both a potential interpretation for the use of certain types of information-cost functions in static rational inattention models, and a useful approach to solving for the predictions (including predictions about response times) of the dynamic model.

Our general framework is flexible enough to allow beliefs to evolve either as a continuous diffusion or in discrete jumps. We establish conditions under which beliefs necessarily evolve in only one of these ways. In particular, we establish conditions under which both the evolution of beliefs prior to a decision, and the stopping rule that determines the time taken for a decision and its accuracy, are similar to the assumptions of the drift-diffusion model in mathematical psychology. In this case, the DM's belief state can be represented as a diffusion on a line, the drift of which depends on the external state, and a decision is made at whatever time the belief state first reaches one of two time-invariant boundaries. Whether the conditions under which beliefs should evolve in this way are in fact characteristic of actual decision situations deserves further study; we show that at least in principle, it is possible to determine this on the basis of a study of the state-contingent joint distributions of responses and response times.

References

Carlos Alós-Ferrer, Ernst Fehr, and Nick Netzer. Time will tell: Recovering preferences when choices are noisy. *Journal of Political Economy*, 129(6):1828–1877, 2021.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.

Arindam Banerjee, Xin Guo, and Hui Wang. On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669, 2005.

Lawrence M Benveniste and Jose A Scheinkman. On the differentiability of the value function in dynamic models of economics. *Econometrica: Journal of the Econometric Society*, pages 727–732, 1979.

Alexander W Bloedel and Weijie Zhong. The cost of optimally-acquired information. *Unpublished Manuscript, November*, 2020.

Jerome R Busemeyer and James T Townsend. Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, 100: 432–459, 1993.

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

Andrew Caplin, Mark Dean, and John Leahy. Rationally inattentive behavior: Characterizing and generalizing Shannon entropy. *Journal of Political Economy*, Forthcoming.

Yeon-Koo Che and Konrad Mierendorff. Optimal dynamic allocation of attention. *American Economic Review*, 109(8):2993–3029, 2019.

Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

John A Clithero. Improving out-of-sample predictions using response times and a model of the decision process. *Journal of Economic Behavior & Organization*, 148:344–375, 2018.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Michael G Crandall, Hitoshi Ishii, and Pierre-Louis Lions. Users guide to viscosity solutions of second order partial differential equations. *Bulletin of the American Mathematical Society*, 27(1):1–67, 1992.

Mark Dean and Nathaniel Neligh. Experimental tests of rational inattention. *Unpublished manuscript*, June 2019.

Ernst Fehr and Antonio Rangel. Neuroeconomic foundations of economic choice — recent advances. *Journal of Economic Perspectives*, 25(4):3–30, 2011.

Eugene A Feinberg, Pavlo O Kasyanov, and Nina V Zadoianchuk. Fatou's lemma for weakly converging probabilities. *Theory of Probability & Its Applications*, 58(4):683–689, 2014.

Alexander Frankel and Emir Kamenica. Quantifying information and uncertainty. *American Economic Review*, 109(10):3650–80, 2019.

Drew Fudenberg, Philipp Strack, and Tomasz Strzalecki. Speed, accuracy, and the optimal timing of choices. *American Economic Review*, 108(12):3651–84, 2018.

Benjamin Hébert and Michael Woodford. Neighborhood-based information costs. *American Economic Review*, 111(10):3225–55, 2021.

Benjamin M Hébert and Michael Woodford. Rational inattention when decisions take time. Technical report, National Bureau of Economic Research w.p. 26415, 2019.

Jean Jacod and Albert Shiryaev. *Limit theorems for stochastic processes*, volume 288. Springer Science & Business Media, 2013.

Simon P Kelly, Elaine A Corbett, and Redmond G O'Connell. Neurocomputational mechanisms of prior-informed perceptual decision-making in humans. *Nature Human Behaviour*, 5(4):467–481, 2021.

Ian Krajbich, Bastiaan Oud, and Ernst Fehr. Benefits of neuroeconomics modeling: New policy interventions and predictors of preference. *American Economic Review*, 104(5): 501–506, 2014.

Filip Matêjka, Alisdair McKay, et al. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–98, 2015.

Stephen Morris and Philipp Strack. The Wald problem and the relation of sequential sampling and ex-ante information costs. *Unpublished manuscript*, February 2019.

Giuseppe Moscarini and Lones Smith. The optimal level of experimentation. *Econometrica*, 69(6):1629–1644, 2001.

Huyên Pham. Optimal stopping of controlled jump diffusion processes: a viscosity solution approach. In *Journal of Mathematical Systems, Estimation and Control*. Citeseer, 1998.

Huyên Pham. *Continuous-time stochastic control and optimization with financial applications*, volume 61. Springer Science & Business Media, 2009.

Luciano Pomatto, Philipp Strack, and Omer Tamuz. The cost of information. *arXiv preprint arXiv:1812.04211*, 2018.

Roger Ratcliff. Theoretical interpretations of speed and accuracy of positive and negative responses. *Psychological Review*, 92:212–225, 1985.

Roger Ratcliff and Jeffrey N Rouder. Modeling response times for two-choice decisions. *Psychological Science*, 9:347–356, 1998.

HL Royden and PM Fitzpatrick. Real analysis, 2010.

JF Schouten and JAM Bekker. Reaction time and accuracy. *Acta Psychologica*, 27:143–153, 1967.

Michael Shadlen and Daphna Shohamy. Decision making and sequential sampling from memory. *Neuron*, 90(5):927–939, 2016.

Christopher A Sims. Rational inattention and monetary economics. *Handbook of Monetary Economics*, 3:155–181, 2010.

Jakub Steiner, Colin Stewart, and Filip Matějka. Rational inattention dynamics: Inertia and delay in decision-making. *Econometrica*, 85(2):521–553, 2017.

Satohiro Tajima, Jan Drugowitsch, and Alexandre Pouget. Optimal policy for value-based decision-making. *Nature communications*, 7, 2016.

Satohiro Tajima, Jan Drugowitsch, Nishit Patel, and Alexandre Pouget. Optimal policy for multi-alternative decisions. *Nature Neuroscience*, 22:1503–1511, 2019.

Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

Eric-Jan Wagenmakers, Han L.J. van der Maas, and Raoul P.P.P. Grasman. An ez-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, 14(1):3–22, 2007.

Michael Woodford. Stochastic choice: An optimizing neuroeconomic model. *American Economic Review*, 104(5):495–500, 2014.

Michael Woodford. Modeling imprecision in perception, valuation, and choice. *Annual Review of Economics*, 12:579–601, 2020.

Weijie Zhong. Optimal dynamic information acquisition. *Unpublished manuscript*, January 2019.

A Response-Time Distributions in Binary-Choice Problems

Here we provide additional details of the calculations reported in section 5. We consider a problem in which there are two possible actions ($A = \{L, R\}$), and suppose that for each state, one or the other of the two actions is strictly preferable ($u_{L,x} \neq u_{R,x}$). We can then partition the state space into two disjoint subsets, X_L (the subset of states in which L is the "correct" response, meaning that it results in higher utility) and X_R (the subset for which R is the correct response). In the main text, we further specialize by assuming that utility depends only on whether the DM's response a is "correct" or not for the state x (as is true of the reward function in the experiment of Kelly et al. [2021]); but this hypothesis is not needed for the more general results here.

As in the main text, we consider information costs based on a strongly convex entropy function $H(q)$ (though here we do not impose a symmetry assumption). In the case of either PDL or PGL information costs, it follows from Proposition 7 that the predicted state-contingent response probabilities are the ones corresponding to the solution to the static RI problem stated in (15). For any prior $\bar{q}$, there will be a pair of optimal posteriors $\bar{q}_L$ and $\bar{q}_R$ given by the solution to this problem. (In general, these will depend on the prior. In the two-state case discussed in the main text, instead, they are independent of $\bar{q}$, for all priors that are not too extreme; but that is a special feature of the two-state case.) We shall assume that some information is acquired, which is to say that under the DM's actual prior q_0, $\bar{q}_L \neq \bar{q}_R \neq q_0$, and that all of these posteriors are on the interior of the simplex (to avoid technicalities). Under the optimal policy, beliefs will move (either diffusing or jumping/drifting) on the line segment connecting $\bar{q}_L$ and $\bar{q}_R$ in the simplex, which necessarily runs through q_0. Given this, it is possible, as in the main text, to parameterize beliefs at each point in time by a scalar variable π_{Lt}; the beliefs corresponding to a given value of π_{Lt} are again given by a function $q(\pi_{Lt})$ defined in (18).[43] In terms of this notation, the prior corresponds to $\pi_{L,0} = \bar{\pi}_L$, the probability of an L response in the solution to the static problem (15). At any subsequent time, π_{Lt} indicates the conditional probability of an eventual L response, given the state reached at time t; accordingly, the variable π_{Lt} must evolve as a martingale.

It further follows from Proposition 7 that the value function (in terms of the state vari-

[43] In the more general case discussed here, the stopping posteriors q_L^*, q_R^* referred to in (18) are understood to refer to the stopping posteriors $\bar{q}_L, \bar{q}_R$ that are optimal for a particular prior q_0.

able $\pi_{L,t}$) can be written as

$$V(\pi_{L,t}) = \pi_{L,t}V_L + (1-\pi_{L,t})V_R + \frac{\kappa}{\chi}H(q(\pi_{L,t})),$$

with $V_L = \sum_{x\in X}\bar{q}_{L,x}u_{L,x} - \frac{\kappa}{\chi}H(\bar{q}_L)$ and $V_R = \sum_{x\in X}\bar{q}_{R,x}u_{R,x} - \frac{\kappa}{\chi}H(\bar{q}_R)$. It follows by the strict convexity of $H(q)$ and the linearity of $q(\pi_L)$ that $V(\pi_{L,t})$ is strictly convex on $\pi_{L,t}\in[0,1]$. As a consequence of this convexity, there are three possible shapes of the value function $V(\pi_{L,t})$: it could be increasing on $[0,1]$, decreasing on $[0,1]$, or decreasing on $[0,\underline{\pi}_L)$ and increasing on $(\underline{\pi}_L,1]$ for some $\underline{\pi}_L\in(0,1)$. In the first two of these cases, we will say $\underline{\pi}_L = 0$ and $\underline{\pi}_L = 1$, respectively; thus in each case, $\underline{\pi}_L$ is the value of π_L at which $V(\pi_L)$ reaches its minimum. Alternative possible shapes for this function are illustrated in the first column of Figure 4. We show below that the shape of the value function is closely related to the properties of the stopping time distribution in the case of a preference for discrete learning.

A.1 General Results for the PDL Case

We first consider the PDL case, and as in the discussion in the main text, we first state our results for the case in which the minimum of the value function on the line segment, $V(\pi_L)$, occurs at some $\underline{\pi}_L < \bar{\pi}_L$, as illustrated in Figure 4. As noted in the main text, the results of Zhong [2019] imply that jumps must always increase the value function, while the value function must steadily decrease while beliefs drift (i.e., in the absence of a jump). Thus for any $\pi_{L,t} > \underline{\pi}_L$, the optimal policy is to jump towards $\pi_{L,t} = 1$ with the maximum possible intensity and drift downwards. Eventually, $\pi_{L,t}$ will drift downwards and equal $\underline{\pi}_L$, at which point the DM will randomize between jumping to $\pi_{L,t} = 1$ and $\pi_{L,t} = 0$ with unconditional probabilities $\underline{\pi}_L$ and $(1-\underline{\pi}_L)$. In the particular case of an upward-sloping value function ($\underline{\pi}_L = 0$), shown in the bottom row of Figure 4, the DM will choose R with certainty, immediately after reaching $\underline{\pi}_L$.

We first solve explicitly for the dynamics of beliefs during the first phase, while π_{Lt} remains in the interval $\underline{\pi}_L < \pi_{Lt} < \bar{\pi}_L$. The fact that π_{Lt} must be a martingale implies that the drift rate μ_t in the absence of a jump is determined by the instantaneous rate ψ_{Lt} at which jumps to $\bar{q}_L$ (i.e., to $\pi_{Lt} = 1$) occur: $\mu_t = -\psi_{Lt}(1-\pi_{Lt})$. The convexity of the value function implies that it is optimal to make the variability of the evolution as great as possible, consistent with the constraint on the rate of evidence accumulation; hence constraint (6) must hold with equality at each point in time. Since only one kind of jump occurs during this phase of the decision process, the constraint implies that the instantaneous rate at

which jumps of this kind occur is given by $\psi_L(\pi_{Lt}) = \chi / D_H(\bar{q}_L || q(\pi_{Lt}))$, as stated in the main text. This in turn implies the solution for the drift rate $\mu_t = \mu(\pi_{Lt})$ given by (19).

We can integrate the differential equation

$$\dot{\pi}_{Lt} \;=\; \mu(\pi_{Lt}), \tag{24}$$

starting from the initial condition $\pi_{L,0} = \bar{\pi}_L$, to determine the time $t = \tilde{\tau}$ at which the dynamics predict that $\pi_{L,t} = \underline{\pi}_L$, conditional on no jump having occurred before that time. If we let P^{early} be the unconditional probability of a jump occurring before time $\tilde{\tau}$, then the fact that q_t must be a martingale, together with the fact that the posterior reached at the end of the first phase must be $\bar{q}_L$ if a jump has occurred and $q(\underline{\pi}_L)$ otherwise, requires that

$$P^{early} \cdot \bar{q}_L + (1 - P^{early}) \cdot q(\underline{\pi}_L) \;=\; q_0.$$

This is a vector equation that must hold element-wise. In particular, we must have

$$P^{early} \cdot \bar{q}_{L,y} + (1 - P^{early}) \cdot q_y(\underline{\pi}_L) \;=\; q_{0,y}, \tag{25}$$

for each of the values $y \in \{\ell, r\}$, if we define

$$\bar{q}_{L,\ell} \;=\; \sum_{x \in X_L} \bar{q}_{L,x}, \qquad q_\ell(\pi_L) \;=\; \sum_{x \in X_L} q_x(\pi_L), \qquad q_{0,\ell} \;=\; \sum_{x \in X_L} q_{0,x},$$

and correspondingly use the subscript r to denote sums over the states in X_R. This can be solved for the value of P^{early} that is consistent with given values for $q_0, q(\underline{\pi}_L)$, and $\bar{q}_L$.

Let P_ℓ^{early} correspondingly denote the probability of a jump before time $\tilde{\tau}$, conditional on the state belonging to set X_L, and P_r^{early} the probability conditional on the state belonging to X_R. Bayes' rule requires that in order for $\bar{q}_L$ to be the posterior in the event of a jump before time $\tilde{\tau}$, it must be the case that

$$\bar{q}_{L,y} \;=\; \frac{q_{0,y} P_y^{early}}{P^{early}}$$

for either of the $y \in \{\ell, r\}$. Substituting the solution to (25) for P^{early} in this equation, and solving for the implied value of P_y^{early}, we obtain

$$P_y^{early} \;=\; \left(\frac{\bar{q}_{L,y}}{q_{0,y}} \right) \left(\frac{q_{0,y} - q_y(\underline{\pi}_L)}{\bar{q}_{L,y} - q_y(\underline{\pi}_L)} \right). \tag{26}$$

If time $\tilde{\tau}$ is reached with no jump, then from then on π_{Lt} remains unchanged at the value $\underline{\pi}_L$, until a jump occurs. In this second phase of the dynamics, jumps can occur to either $\bar{q}_L$ or $\bar{q}_R$, and the constant rates ψ_L, ψ_R at which the two types of jumps occur must be such that π_{Lt} is a martingale despite the absence of any drift. This requires that

$$\psi_L(\bar{q}_L - q(\underline{\pi}_L)) + \psi_R(\bar{q}_R - q(\underline{\pi}_L)) \;=\; 0,$$

which holds if and only if

$$\frac{\psi_L}{\psi_R} \;=\; \frac{\underline{\pi}_L}{1-\underline{\pi}_L}.$$

(Note that $\underline{\pi}_L < 1$ in the cases considered here.) This determines the relative rates at which the two kinds of jumps must occur, but not the absolute rates. However, the convexity of the value function again implies that these rates must be as large as possible, consistent with the constraint (6). Hence (6) must hold with equality, which requires that the total rate $\psi \equiv \psi_L + \psi_R$ at which jumps occur must be

$$\psi \;=\; \frac{\chi}{\underline{\pi}_L D_H(\bar{q}_L||q(\underline{\pi}_L)) + (1-\underline{\pi}_L) D_H(\bar{q}_R||q(\underline{\pi}_L))}.$$

Given this constant rate at which jumps occur at all times greater than $\tilde{\tau}$, the unconditional cumulative distribution function $F(\tau)$ for response times will be of the form

$$F(\tau) \;=\; 1 - (1 - P^{early}) e^{-\psi(\tau - \tilde{\tau})}$$

for all $\tau \geq \tilde{\tau}$.

The jump rates just calculated are unconditional ones; we can similarly compute a total jump rate ψ^y conditional on the class y to which the state belongs (i.e., on whether the correct response is L or R), and decompose the total jump rate ψ^y into a rate ψ^y_L of jumps to the response L and ψ^y_R of jumps to the response R. (All of these rates are constant jump rates for times after $\tilde{\tau}$.) Bayes' Rule requires that in order for the posterior to be $\bar{q}^*_a$ following a jump to response a, for either class of states y we must have

$$\bar{q}_{a,y} \;=\; \frac{q_y(\underline{\pi}_L)\psi^y_a}{\psi_a}.$$

This implies that

$$\psi^y_a \;=\; \psi_a \frac{\bar{q}_{a,y}}{q_y(\underline{\pi}_L)} \;=\; \frac{\underline{\pi}_a \bar{q}_{a,y}}{q_y(\underline{\pi}_L)} \cdot \psi \tag{27}$$

for $y \in \{\ell, r\}$ and $a \in \{L, R\}$, where $\underline{\pi}_R$ means $1 - \underline{\pi}_L$ (the unconditional probability of response R if there is no decision before time $\tilde{\tau}$). This allows us to compute the relative frequency with which different combinations of actual response a and correct response y should be observed, if the decision occurs after time $\tilde{\tau}$.

Now suppose that we separately compute distributions of response times for classes of trials that are classified (i) according to whether the state belongs to X_L or X_R (that is, according to what the correct response is on that trial), and (ii) according to the DM's response, as with the data presented in Figure 3. For each of the two possible sets of states (labeled by $y \in \{\ell, r\}$), we can compute a theoretical distribution of response times, which will have a cumulative distribution function $F^y(\tau)$. For either class of states y, this function can be decomposed into two parts, $F^y_L(\tau)$ and $F^y_R(\tau)$, where $F^y_a(\tau)$ indicates the probability of a jump to the response a before time τ. These are continuous, non-decreasing functions of τ, satisfying $F^y_L(\tau), F^y_R(\tau) \geq 0$ and $F^y_L(\tau) + F^y_R(\tau) = F^y(\tau)$ for all τ. Computing predictions for these functions requires an exact specification of the function H (and hence the divergence D_H). However, the predictions depend only on a finite number of parameters, rather than the complete details of the entropy function assumed, if we write them in terms of response-time *quantile* distributions instead of response times.

For either choice of the class of states y, we can define a response-time quantile $\hat{\tau}$ associated with any response time τ, using the mapping $\hat{\tau} = F^y(\tau)$. A generalized inverse (quantile) function can be defined by $\hat{F}^y(\hat{\tau}) = \inf\{\tau \geq 0 : F^y(\tau) \geq \hat{\tau}\}$. Then for any pair (y, a), we can define $G^y_a(\hat{\tau}) \equiv F^y_a(\hat{F}^y(\hat{\tau}))$. This function is defined for all quantiles $0 < \hat{\tau} < 1$, and identifies the cumulative number of a responses among the first $\hat{\tau}$ responses, conditional on the state belonging to the class y. (By definition, for any $\hat{\tau}$ we must have $G^y_L(\hat{\tau}), G^y_R(\hat{\tau}) \geq 0$ and $G^y_L(\hat{\tau}) + G^y_R(\hat{\tau}) \geq \hat{\tau}$, with equality wherever $\hat{F}^y(\hat{\tau})$ is strictly increasing.)

For either choice of the class of states y, let $\hat{\tau}^y \equiv F^y(\tilde{\tau}) = P^{early}_y$ be the response-time quantile at which the first phase ends, conditional on the state belonging to class y. Our results above then imply that for all quantiles $0 \leq \hat{\tau} \leq \hat{\tau}^y$, $G^y_R(\hat{\tau}) = 0$ and $G^y_L(\hat{\tau}) = \hat{\tau}$. For all quantiles $\hat{\tau}^y \leq \hat{\tau} \leq 1$, instead, we obtain

$$G^y_R \;=\; p^y_R \cdot (\hat{\tau} - \hat{\tau}^y), \qquad G^y_L \;=\; \hat{\tau}^y + p^y_L \cdot (\hat{\tau} - \hat{\tau}^y), \tag{28}$$

using the notation $p^y_a \equiv \psi^y_a / \psi$ for the probability of response a if the state belongs to the set y and no choice is made before time $\tilde{\tau}$. These functions are completely described by the quantities P^{early}_y given by (26) and ψ^y_a / ψ given by (27), regardless of any other properties

of the H function. Since we must have $(\psi_L^y/\psi) + (\psi_R^y/\psi) = 1$, there are only four degrees of freedom in the possible appearance of the functions.

Similar results can be obtained if we instead assume that $\underline{\pi}_L > \bar{\pi}_L$. In this case, the same kind of reasoning as above implies that for any $\pi_{L,t} < \underline{\pi}_L$, the optimal policy is to jump towards $\pi_{L,t} = 0$ (i.e., toward response R) with the maximum possible intensity, and otherwise to drift upwards. At a finite time $\tilde{\tau}$, $\pi_{L,t}$ will have risen enough to equal $\underline{\pi}_L$, after which point the DM will randomize between jumping to $\pi_{L,t} = 1$ and $\pi_{L,t} = 0$ with unconditional probabilities $\underline{\pi}_L$ and $(1 - \underline{\pi}_L)$. Analogs of all of the above formulas can be derived in the same way, simply reversing the roles of responses L and R and the classes of states ℓ and r. For example, instead of (26) we obtain

$$P_y^{early} \;=\; \left(\frac{\bar{q}_{R,y}}{q_{0,y}}\right)\left(\frac{q_{0,y} - q_y(\underline{\pi}_L)}{\bar{q}_{R,y} - q_y(\underline{\pi}_L)}\right), \tag{29}$$

and instead of (28), we obtain

$$G_L^y \;=\; p_L^y \cdot (\hat{\tau} - \hat{\tau}^y), \qquad G_R^y \;=\; \hat{\tau}^y + p_R^y \cdot (\hat{\tau} - \hat{\tau}^y)$$

in this alternative case.

Finally, in the special case in which $\underline{\pi}_L = \bar{\pi}_L$ exactly, there will be no "first phase" of the belief dynamics. (Equation (26) reduces to $P_y^{early} = 0$.) The optimal policy will immediately involve randomization between jumping to $\pi_{L,t} = 1$ and $\pi_{L,t} = 0$ with unconditional probabilities $\underline{\pi}_L$ and $(1 - \underline{\pi}_L)$. The conclusions obtained above concerning dynamics in "phase two" continue to hold in this limiting case, setting $\tilde{\tau}$ equal to zero. Thus for example (28 reduces to $G_a^y = p_a^y \cdot \hat{\tau}$ for $a = L, R$, and the equation holds for all $0 \leq \hat{\tau} \leq 1$.

A.2 Specialization to the Two-State Case

In the special case considered in the main text, there are only two states, one (state ℓ) in which L is the correct response, and another (state r) in which R is the correct response. In addition, motivated by the experimental design in Kelly et al. [2021], we assume certain symmetries that allow us to simplify the results obtained in the previous subsection. As explained in the text, these assumptions allow us to define the stopping posteriors $\bar{q}_a$ independently of the choice of prior, for any prior that is not too extreme.[44] In this special

[44]The prior must be some convex combination of the two stopping posteriors q_L^* and q_R^* that solve the maximization problem stated in the main text. In the case that there are only two states, the probability

case, the optimal stopping posteriors are denoted q_a^*. In addition, the symmetry assumptions imply that $q_{L,\ell}^* = q_{R,r}^* = \alpha$, for some $1/2 < \alpha < 1$. It follows that for any non-extreme prior (i.e., such that $q_{R,\ell}^* \leq q_{0,\ell} \leq q_{L,\ell}^*$), the overall *accuracy rate* (the fraction of correct responses) will be equal to α. Finally, the symmetry assumptions imply that $V(\pi_L)$ will have the symmetry $V(\pi_L) = V(1-\pi_L)$, with the consequence that this convex function must reach its unique minimum at $\underline{\pi}_L = 1/2$, as shown in the left panel of Figure 1. This means that the posterior $q(\underline{\pi}_L)$ will be one that assigns equal probability to the two states.

A.2.1 Proof of Lemma 3

Under these additional assumptions, the results derived in the previous subsection can be stated more simply. Substituting the values $\bar{q}_{L,\ell} = \alpha$, and $q_\ell(\underline{\pi}_L) = 1/2$, equation (26) reduces to equation (20) in the main text for the values of $\hat{\tau}^\ell, \hat{\tau}^r$. Similarly, substituting the values $\underline{\pi}_L = 1/2$ and $q_\ell(\underline{\pi}_L) = 1/2$ into (27), we obtain $p_a^y = \bar{q}_{a,y}$. Hence $p_a^y = \alpha$ for $(y,a) = (\ell,L)$ or (r,R), while it is equal to $(1-\alpha)$ for (r,L) or (ℓ,R).

The response-time quantile distributions are described in Lemma 3 (and similarly Figures 1 and 2, as well as Figure 4 below) in terms of density functions $g_a^y(\hat{\tau})$ rather than the cumulative distribution functions $G_a^y(\hat{\tau})$ defined in the previous subsection. The density function is given by

$$g_a^y(\hat{\tau}) \equiv \frac{\partial G_a^y(\hat{\tau})}{\partial \hat{\tau}}$$

at those values of $\hat{\tau}$ where the distribution function is differentiable.[45] Thus in the symmetric case, we have $g_L^\ell(\hat{\tau}) = 1$ for all $0 < \hat{\tau} < \hat{\tau}^\ell$, and $g_L^\ell(\hat{\tau}) = \alpha$ for all $\hat{\tau}^\ell < \hat{\tau} < 1$, where $\hat{\tau}^\ell$ is given in (20). Similarly, we have $g_L^r(\hat{\tau}) = 1$ for all $0 < \hat{\tau} < \hat{\tau}^r$, and $g_L^r(\hat{\tau}) = 1-\alpha$ for all $\hat{\tau}^r < \hat{\tau} < 1$, where $\hat{\tau}^r$ is also given in (20). This establishes the results stated in Lemma 3.

Similar methods can be applied if instead we assume a prior such that r is the more

simplex is simply a line segment, so that any prior must lie on the line that passes through the points q_L^* and q_R^*; the only special requirement is then that q_0 not be too extreme, in the sense that it lies between the two stopping posteriors. In the case that there are more than two states, instead, the set of priors that are convex combinations of any two stopping posteriors will be non-generic.

[45] These density functions can be interpreted as the conditional likelihood of action a given $y \in \{l,r\}$ and that the response time quantile is $\hat{\tau}$. These likelihoods are well-defined in the context of Lemma 3, because there are no atoms in the response times. More generally, if F^y is strictly increasing, the G_a^y functions are Lipschitz continuous, and hence the density functions are defined almost everywhere on the interval $[0,1]$, which is sufficient for our purposes. The relevant predictions of the model relate to integrals of these densities over intervals; we cannot expect to experimentally test a prediction about the value of the density at a single point. For example, in Figure 3, we compare the theoretical predictions (shown in Figure 1 in terms of density functions) with data on the number of observations in each of several response-time bins; each of these bins corresponds to 1/7 of the interval $[0,1]$.

likely state ($q^*_{R,\ell} < q_{0,\ell} < 1/2$). In this case, equation (29) implies that

$$\hat{\tau}^r \;=\; \left(\frac{\alpha}{\alpha - 1/2}\right)\left(\frac{1/2 - q_{0,\ell}}{1 - q_{0,\ell}}\right), \qquad \hat{\tau}^\ell \;=\; \left(\frac{1-\alpha}{\alpha - 1/2}\right)\left(\frac{1/2 - q_{0,\ell}}{q_{0,\ell}}\right),$$

instead of the formulas given in (20). The response-time quantile density functions are given in this case by $g^r_R(\hat{\tau}) = 1$ for all $0 < \hat{\tau} < \hat{\tau}^r$, and $g^r_R(\hat{\tau}) = \alpha$ for all $\hat{\tau}^r < \hat{\tau} < 1$, and similarly $g^\ell_R(\hat{\tau}) = 1$ for all $0 < \hat{\tau} < \hat{\tau}^\ell$, and $g^\ell_R(\hat{\tau}) = 1 - \alpha$ for all $\hat{\tau}^\ell < \hat{\tau} < 1$.

We observe that the predicted density function $g^r_R(\hat{\tau})$ when $q_{0,\ell} = 1 - v$ is mathematically identical to the predicted density function $g^\ell_L(\hat{\tau})$ when $q_{0,\ell} = v$, for any measure $1/2 < v < 1$ of the "validity" of the cues (i.e., of the informativeness of the priors). Similarly, the predicted density function $g^\ell_R(\hat{\tau})$ when $q_{0,\ell} = 1 - v$ is mathematically identical to the predicted density function $g^r_L(\hat{\tau})$ when $q_{0,\ell} = v$. Thus when testing these predictions in a symmetric case, we can pool the distribution of L responses conditional on a state of type ℓ when the prior is $q_{0,\ell} = v$ and the distribution of R responses conditional on a state of type r when the prior is $q_{0,\ell} = 1 - v$ (so that $q_{0,r} = v$), calling all of these the distribution of "cue-consistent responses" in the case of a "cue-consistent state." (The theoretical prediction for this distribution is shown by the graph of $g^\ell_L(\hat{\tau})$ in Figure 1.) Similarly, we can pool the distribution of L responses conditional on a state of type r when the prior is $q_{0,\ell} = v$ and the distribution of R responses conditional on a state of type ℓ when the prior is $q_{0,\ell} = 1 - v$, calling all of these the distribution of "cue-consistent responses" in the case of a "cue-inconsistent state." (The theoretical prediction for this distribution is shown by the graph of $g^\ell_L(\hat{\tau})$ in Figure 1.) This is the method used to compare the experimental data with our theoretical predictions in Figure 3.

A.2.2 Asymmetric Rewards in the PDL Case: Examples

In asymmetric cases, we continue to have qualitatively similar predictions. For example, when the prior implies that state ℓ is more likely, then $g^\ell_L(\hat{\tau}) = 1$ for all $0 < \hat{\tau} < \hat{\tau}^\ell$, and $g^\ell_L(\hat{\tau}) = p^\ell_L < 1$ for all $\hat{\tau}^\ell < \hat{\tau} < 1$; it is only the formulas determining the values of $\hat{\tau}^\ell$ and p^ℓ_L that differ.

Figure 4 illustrates how asymmetric rewards modify the predictions shown in Figure 1 of the main text. We continue to assume that there are only two states (ℓ, r), such that $u_{L,\ell} > u_{R,\ell}$ and $u_{R,r} > u_{L,r}$. However, we now consider the possibility that $u_{L,\ell} \neq u_{R,r}$ and $u_{L,r} \neq u_{R,\ell}$. As in Figure 1, we assume that $H(q)$ is given by Shannon entropy. In all rows

of the new figure, we also continue to assume that the utility differential between correct and incorrect responses is the same in both states:

$$u_{L,\ell} - u_{R,\ell} \;=\; u_{R,r} - u_{L,r} \;=\; \Delta \;>\; 0,$$

where the value of Δ is the one assumed in Figure 1. Since the maximization problems defining q_a^* for each of the possible actions depend only on these return differentials, the optimal stopping posteriors (q_L^*, q_R^*) are the same in all three rows of the figure, and remain the same as in Figure 1. Finally, in all three rows of the new figure, we continue to assume the same prior (with $1/2 < q_{0,\ell} < q_{L,\ell}^*$) as in Figure 1. This means that the value of $\bar{\pi}_L$ in the solution to the static RI problem defined in (15) is the same in all three rows, and the same as in Figure 1.

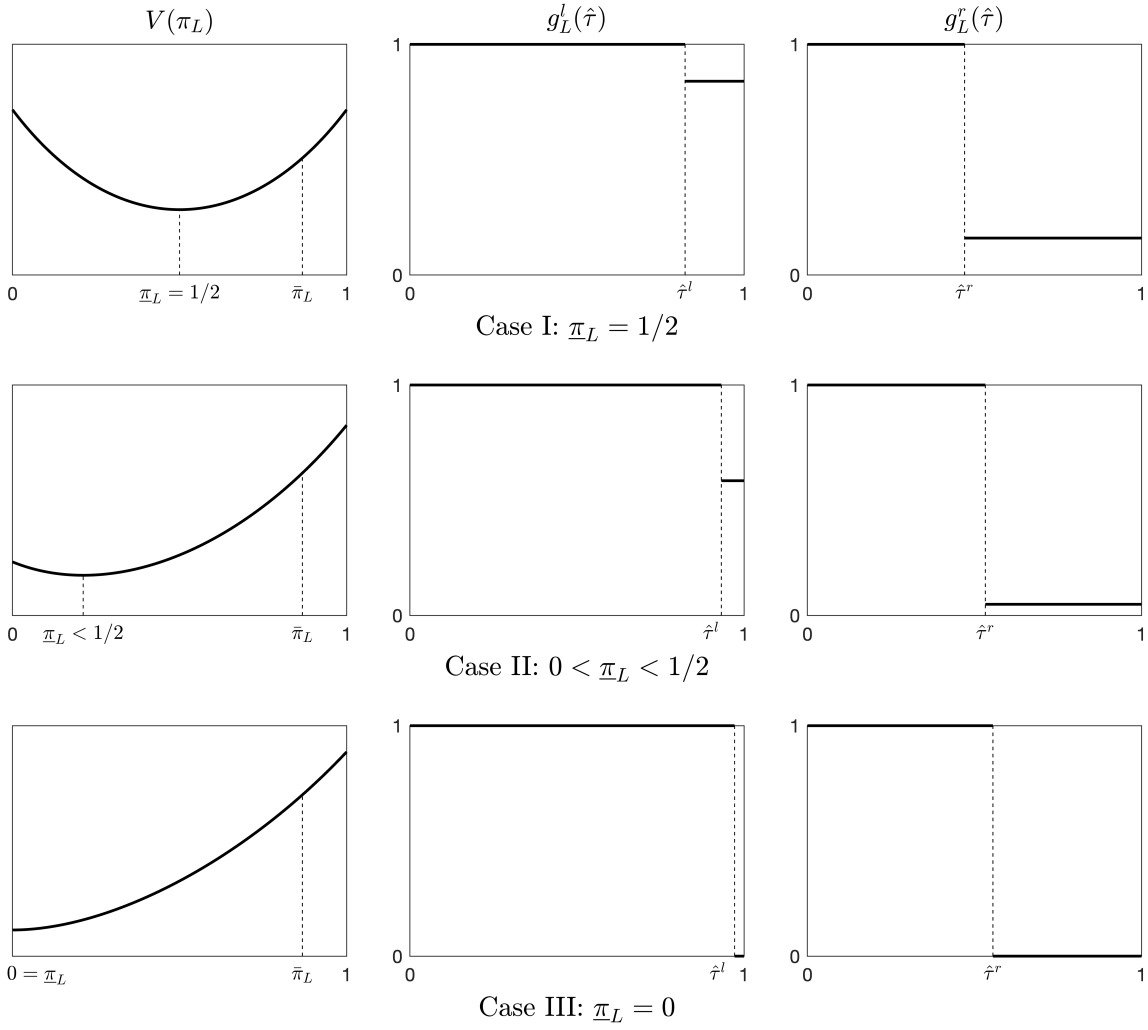

Figure 4: Predicted response-time distributions with a preference for discrete learning, under three different possible assumptions about the degree of asymmetry of payoffs. Each row shows the value function $V(\pi_L)$, the function $g_L^x(\hat{\tau})$ for states x in which L is the correct response ("ℓ states"), and the function $g_L^x(\hat{\tau})$ for states x in which R is the correct response ("r states"), under a particular assumption about the relative payoffs in ℓ and r states. In the first row (Case I), the rewards for correct or incorrect responses are the same in ℓ and r states; in the lower rows, the utilities associated with r states are made progressively lower relative those associated with ℓ states. In all numerical calculations shown, $H(q)$ is assumed to be the negative of the Shannon entropy function, and parameters are chosen as in Figures 1 and 2 of the main text. (Figure 1 in the main text corresponds to the first row of this figure.)

The top row of the figure (which corresponds to Figure 1 in the main text) makes the further assumption that the DM's reward depends only on whether the response is correct

for the state, so that $u_{L,\ell} = u_{R,r} > u_{R,\ell} = u_{L,r}$. In this case, as discussed above, the value function is symmetric around the value $\underline{\pi}_L = \frac{1}{2}$. In the other two rows, we continue to assume the same utility differential between the correct and incorrect responses in each state, but we no longer assume that the reward for a correct response is the same in states ℓ and r. If we let $u(L,\ell) - u(R,r) = u(R,\ell) - u(L,r) = \delta$, then $\delta = 0$ corresponds to Case I, shown in the top row. If instead $0 < \delta < \Delta$, we have an asymmetric value function and $0 < \underline{\pi}_L < 1/2$ (Case II), as shown in the second row of the figure. (The numerical solution shown in the second row is for the case $\delta = \Delta/2$.) Finally, if $\delta \geq \Delta$, the value function is monotonically increasing and $\underline{\pi}_L = 0$ (Case III), as shown in the bottom row. (The numerical solution shown is for the case in which $\delta = \Delta$ exactly.) We obtain similarly asymmetric solutions if $\delta < 0$, but with the roles of states ℓ and r reversed.

Each row of the figure shows the value function $V(\pi_L)$ and the response-time quantile density functions $g_L^y(\hat{\tau})$ for a particular value of δ. Increasing the utility of state ℓ relative to the utility of being in state r, while preserving unchanged the utility differential between correct and incorrect responses in both states, is of no consequence for the solution $(q_L^*, q_R^*, \bar{\pi}_L)$ to the static RI problem associated with a given prior; but it does change the expected utility $V(q_0)$ implied by that prior. Hence the value function $V(q)$, and correspondingly the transformed value function $V(\pi_L)$, depend on the value of the parameter δ. Specifically, we have

$$V(\pi_L) \;=\; V_0(\pi_L) + \delta\pi_L,$$

where $V_0(\pi_L)$ is the symmetric value function shown in the first row of the figure. For arbitrary δ, this continues to be a strictly convex function, with its unique local minimum at the point $\underline{\pi}_L$ implicitly defined by the first-order condition

$$V_0'(\underline{\pi}_L) \;=\; -\delta.$$

Because of the convexity of $V_0(\pi_L)$, the solution for $\underline{\pi}_L$ is a monotonically decreasing function of δ (for all $\delta < -V_0'(0)$), with $\underline{\pi}_L < 1/2$ for all $\delta > 0$, as shown for Case II in the figure. When $\delta \geq -V_0'(0)$, the value function is monotonically increasing over the entire interval $[0,1]$, as shown for Case III in the figure; in this case, $\underline{\pi}_L = 0$. Under the assumption of Shannon entropy, as assumed in the figure, $V_0'(0) = -\Delta$, so that Case III is reached when $\delta \geq \Delta$.

Because the optimal stopping posteriors are independent of the value of δ, we continue to have $q_{L,\ell}^* = \alpha, q_{R,\ell}^* = 1 - \alpha$, regardless of the value of δ (and α continues to measure the

predicted overall accuracy rate). Equation (26) therefore implies that

$$\hat{\tau}^{\ell} = \left(\frac{\alpha}{\alpha - \underline{q}_{\ell}}\right)\left(\frac{q_{0,\ell} - \underline{q}_{\ell}}{q_{0,\ell}}\right), \qquad \hat{\tau}^{r} = \left(\frac{1-\alpha}{\alpha - \underline{q}_{\ell}}\right)\left(\frac{q_{0,\ell} - \underline{q}_{\ell}}{1 - q_{0,\ell}}\right),$$

using the shorthand $\underline{q}_{\ell} \equiv q_{\ell}(\underline{\pi}_L) = \underline{\pi}_L\alpha + (1-\underline{\pi}_L)(1-\alpha)$. This generalizes (20), which corresponds to the case $\underline{q}_{\ell} = 1/2$ (i.e., $\underline{\pi}_L = 0$). Similarly, equation (27) implies that

$$p_L^{\ell} = \frac{\underline{\pi}_L\alpha}{\underline{\pi}_L\alpha + (1-\underline{\pi}_L)(1-\alpha)}, \qquad p_L^{r} = \frac{\underline{\pi}_L(1-\alpha)}{\underline{\pi}_L(1-\alpha) + (1-\underline{\pi}_L)\alpha},$$

generalizing the expressions given in Lemma 3 for the case $\underline{q}_{\ell} = 1/2$.

Figure 4 illustrates how the predicted response-time quantile density functions change as we vary the size of δ (i.e., the degree of asymmetry of the rewards available in the two different states). As δ increases, $\underline{\pi}_L$ and hence $\underline{q}_{\ell}$ monotonically decrease; in the limit as δ approaches Δ (Case III is reached), $\underline{\pi}_L$ approaches zero and hence $\underline{q}_{\ell}$ approaches a lower bound of $1-\alpha$. As a result, both $\hat{\tau}^{\ell}$ and $\hat{\tau}^{r}$ increase monotonically, though they remain less than 1 even when $\underline{\pi}_L = 0$ (Case III). And both p_L^{ℓ} and p_L^{r} decrease monotonically, each approaching zero as $\underline{\pi}_L \to 0$. Hence as Case III is approached, the response-time quantile density functions approach limiting distributions (shown in the bottom row of the figure) in which all responses before quantile $\hat{\tau}^{y}$ are L responses and all responses thereafter are R responses.[46]

Thus varying δ changes the predicted distribution of response times (and response-time quantiles) in the case of a preference for discrete learning, even though it has no effect on the solution to the static RI problem, and thus no effect on predicted state-contingent response probabilities. The conclusion is different in the case of a strict preference for gradual learning, as we now discuss.

[46] In the case that $\underline{\pi}_L = 0$ (Case III), we can't actually describe the response-time quantile distribution using a density function, since the distribution of response times has an atom at $\tilde{\tau}$. However, quantiles remain well-defined, and the quantile distribution functions $G_a^x(\hat{\tau})$ remain continuous (and differentiable everywhere except at $\hat{\tau} = \hat{\tau}^{y}$) for all values $\underline{\pi}_L > 0$. Moreover, there are well-defined limiting density functions as $\underline{\pi}_L \to 0$. These limiting density functions are the ones shown for "Case III" in the bottom row of the figure.

A.3 Response Times in the PGL Case

A.3.1 Proof of Lemma 4

We now consider the dynamics of the belief state $\pi_{L,t}$, conditional on the true state being $x \in X$, under a strict preference for gradual learning. In this subsection, as in subsection A.1, we do not restrict ourselves to the case $|X| = 2$, assumed in the main text; nor do we impose the symmetry assumptions made there.

We note first that the information constraint for a diffusion process, (5), will bind in any solution to the HJB equation (12). Because diffusion takes place on a line (18), the unconditional belief dynamics must be of the form[47]

$$dq_t \;=\; (\bar{q}_L - \bar{q}_R)\bar{\sigma}(\pi_{L,t})dB_t,$$

where $\bar{\sigma}(\pi_L)$ is scalar-valued and dB_t is a one-dimensional Brownian motion. This is a diffusion of the kind assumed in (1), where the matrix $\sigma(q_t)$ in that expression is now an $|X|$-vector, with an element corresponding to each state $x \in X$ given by

$$\sigma_x(q(\pi_{L,t})) \;=\; \frac{(\bar{q}_{L,x} - \bar{q}_{R,x})}{q_x(\pi_{L,t})} \tag{30}$$

at any point on the line segment between $\bar{q}_R$ and $\bar{q}_L$. Substituting this expression for $\sigma(q(\pi_{L,t}))$, we see that the constraint (5) holds with equality if and only if

$$\bar{\sigma}(\pi_{L,t})^2 \;=\; \frac{2\chi}{(\bar{q}_L - \bar{q}_R)^T \nabla^2 H(q(\pi_{L,t}))(\bar{q}_L - \bar{q}_R)}.$$

We can similarly use (7) to obtain a diffusion that describes the conditional dynamics of beliefs. If we use $dq_{t|x}$ to denote the evolution of the posterior q_t conditional on the true state being x, and $dq_{t,x'|x}$ for the evolution of $q_{t,x'}$ (the posterior probability of some state x') conditional on the true state being x, then substitution of (30) into (7) yields

$$\begin{aligned} dq_{t,x'|x} &= q_{t-,x'}\sigma_{x'}(q_{t-})\sigma_x(q_{t-})dt + q_{t-,x'}\sigma_{x'}(q_{t-})dB_{t|x} \\ &= (\bar{q}_{L,x'} - \bar{q}_{R,x'})\frac{(\bar{q}_{L,x} - \bar{q}_{R,x})}{q_{t,x}}\bar{\sigma}(\pi_{L,t})^2 dt + (\bar{q}_{L,x'} - \bar{q}_{R,x'})\bar{\sigma}(\pi_{L,t})dB_{t|x}. \end{aligned}$$

[47] Here, as in subsection A.1, we use the more general notation $\bar{q}_L, \bar{q}_R$ to refer to the stopping posteriors that are optimal for a given prior q_0, rather than assuming that these must coincide with the posteriors q_L^*, q_R^* defined for the two-state case in the main text.

These dynamics for $q_{t|x}$ imply that the posterior remains always on the line segment connecting $\bar{q}_R$ to $\bar{q}_L$. In fact, one observes that they can be written in the form

$$dq_{t|x} \;=\; (\bar{q}_L - \bar{q}_R) d\pi_{L,t|x},$$

where the state-contingent dynamics of the coordinate $\pi_{L,t}$ conditional on true state x follow the diffusion specified in (22). This establishes Lemma 4.

A.3.2 DDM Dynamics as a Special Case

We have noted in the main text that under a particular assumption about information costs, the belief dynamics implied by our model with a strict preference for gradual learning are exactly like those assumed in the standard DDM. Let us suppose again that there are only two states, $X = \{\ell, r\}$, as in the main text, and that utility depends only on whether the response is correct for the state or not, also as in the main text. We further assume in this subsection that $H(q)$ is the "total information" (TI) cost function specified in (17). Note that in this special case we have

$$Diag(q)\nabla^2 H(q) Diag(q) \;=\; \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

at any posterior q in the interior of the probability simplex.

In the two-state case, the posterior can be represented by a single number, $q_{t,\ell}$, and the unconditional belief dynamics (until a decision is made) can be written in the form

$$dq_{t,\ell} \;=\; \tilde{\sigma}(q_{t,\ell}) dB_t,$$

where $\tilde{\sigma}(q_\ell)$ is a scalar-valued function. This is a diffusion of the general form (1), where $\sigma(q)$ is a vector with elements $\sigma_\ell(q) = \tilde{\sigma}(q_\ell)/q_\ell$, $\sigma_r(q) = -\tilde{\sigma}(q_\ell)/q_r$. It follows that the maximum rate of information accumulation consistent with (5) is given by

$$\tilde{\sigma}(q_\ell) \;=\; \sqrt{2\chi} q_\ell (1 - q_\ell).$$

From this it follows, using (7), that the state-contingent belief dynamics are of the form

$$dq_{t,\ell|x} \;=\; \mu_x(q_{t-,\ell}) dt + \tilde{\sigma}(q_{t-,\ell}) dB_{t|x} \tag{31}$$

for $x \in \{\ell, r\}$, where

$$\mu_\ell(q_\ell) \;=\; 2\chi q_\ell (1-q_\ell)^2, \qquad \mu_r(q_\ell) \;=\; -2\chi q_\ell^2 (1-q_\ell).$$

Since $\pi_{L,t}$ is an affine transformation of $q_{t,\ell}$, the dynamics of the belief state $\pi_{L,t}$ still have a non-constant drift and instantaneous variance in this case. But suppose that we instead parameterize the belief state by the posterior log odds, $z_t \equiv \ln(q_{t,\ell}/q_{t,r})$. The is just a smooth nonlinear transformation $z_t = Z(q_{t,\ell})$ of the posterior probability of state ℓ. We can then use Ito's lemma together with (31) to show that the state-contingent dynamics of this variable are given by equation (23) in the main text.

A.3.3 Conditional Response-Time Distributions: Numerical Approach

We next consider the implications of a strict preference for gradual learning for the predicted distribution of stopping times. These can be derived via standard dynamic programming arguments. In the numerical results reported in the text, we assume that there are only two states (ℓ, r), and that utility depends only on whether the response is correct for the state. We present results for two possible information cost functions (associated with different specifications of the entropy function $H(q)$), the Shannon entropy (16) and the TI entropy (17), both of which satisfy the symmetry property assumed in the main text. Finally, in Figure 2 as in Figure 1, we present the response-time quantile distributions for a case in which $1/2 < q_{0,\ell} < q^*_{L,\ell}$. We therefore maintain this assumption in the discussion below.

As in subsection (A.3.1), we describe the dynamics of beliefs conditional on true state x in terms of the evolution of the coordinate $\pi_{L,t|x}$, specified in (22). For either state x, let $\phi^x_L(\pi_L, s)$ be the probability of hitting the boundary $\pi_{L,t} = 1$ or at or before time s (and before reaching the other decision boundary), if at time t no decision has been made and $\pi_{L,0} = \pi_L$. This function has the same form regardless of the time t, owing to the Markovian property of the optimal belief dynamics, and must satisfy the partial differential equation

$$\frac{1}{\bar{\sigma}(\pi_L)^2}\phi^x_{L,s}(\pi_L, s) \;=\; \frac{q^*_{L,x} - q^*_{R,x}}{q_x(\pi_L)}\phi^x_{L,\pi}(\pi_L, s) + \frac{1}{2}\phi^x_{L,\pi\pi}(\pi_L, s), \tag{32}$$

where $\phi^x_{L,s}$, $\phi^x_{L,\pi}$, and $\phi^x_{L,\pi\pi}$ are the first and second-order partial derivatives with respect to s, π, and π-twice. The associated boundary conditions are $\phi^x_L(1,s) = 1$, $\phi^x_L(0,s) = 0$, and $\phi^x_L(\pi_L, 0) = 0$ for all $\pi \in (0,1)$.

By definition, $F_L^x(\tau) = \phi_L^x(\bar{\pi}_L, \tau)$. Consequently, solving the PDE (32) allows us to compute F_L^x. The same PDE, with different boundary conditions, can also be used to compute F_R^x. From these two functions, we can compute the corresponding cumulative distribution functions $G_a^x(\hat{\tau})$ as a function of the quantile $\hat{\tau} = F^x(\tau)$. Finally, (numerical) differentiation of the cumulative distribution functions allows us to compute the corresponding density functions $g_a^x(\hat{\tau})$. The numerical solutions for the two functions $g_L^x(\hat{\tau})$ are shown in Figure 2 in the main text, for each of the two entropy functions that we consider.

Note that our results for the PGL case differ from those in the PDL case in that the predictions for the PGL case depend on the specific entropy function that is assumed (as illustrated by the two different cases considered in the figure), even given our calibrated values for $q_{0,\ell}$ and α. Because of this dependence of the precise predictions for the PGL case on the entropy function, we do not consider in the text how well the observed response-time distributions match specific numerical predictions for that case. We content ourselves with the observation that the model prediction for the PGL case is a smoothly decreasing density function rather than a step function, and (at least qualitatively) the experimental data seem more consistent with this prediction.

B Proofs of Main Results

B.1 Useful Lemmas

B.1.1 Dynamic Programming Principle

We begin by pointing out that a standard dynamic programming principle holds in our environment.

Lemma 5. *(Dynamic Programming Principle) Under any feasible policy* $((\Omega, \tilde{\mathscr{F}}, \{\tilde{\mathscr{F}}_t\}, \tilde{P}), \tilde{q}, \tilde{\tau}) \in \mathscr{A}$*, for any* $t \in \mathbb{R}_+$ *and* $\omega \in \Omega$ *with* $t < \tilde{\tau}(\omega)$*, and any stopping time* $\tau_1 \in [t, \tilde{\tau}]$*,*

$$V(\tilde{q}_t(\omega)) \geq E^{\tilde{P}}[e^{-\rho(\tau_1 - t)} V(\tilde{q}_{\tau_1}) - \kappa \int_t^{\tau_1} e^{-\rho(s-t)} ds | \tilde{\mathscr{F}}_t](\omega).$$

If this policy is an optimal policy, equality must hold.

Proof. By contradiction: suppose that for some ω, t, policy, and stopping time τ_1, and $\varepsilon > 0$,

$$V(\tilde{q}_t) + \varepsilon = E^{\tilde{P}}[e^{-\rho(\tau_1 - t)} V(\tilde{q}_{\tau_1}) - \kappa \int_t^{\tau_1} e^{-\rho(s-t)} ds | \tilde{\mathscr{F}}_t].$$

By Definition 1 (redefining the time variable), this policy must satisfy

$$E^{\tilde{P}}[e^{-\rho(\tilde{\tau}-t)}\hat{u}(\tilde{q}_{\tilde{\tau}}) - \kappa\int_t^{\tilde{\tau}} e^{-\rho(s-t)}ds|\tilde{\mathscr{F}}_t] \leq V(\tilde{q}_t),$$

and therefore

$$\varepsilon \leq E^{\tilde{P}}[e^{-\rho(\tau_1-t)}\{e^{-\rho(\tilde{\tau}-\tau_1)}\hat{u}(\tilde{q}_{\tilde{\tau}}) - V(\tilde{q}_{\tau_1})) - \kappa\int_{\tau_1}^{\tilde{\tau}} e^{-\rho(s-\tau_1)}ds|\tilde{\mathscr{F}}_t].$$

By iterated expectations, there must exist some $\omega \in \Omega$ such that

$$V(\tilde{q}_{\tau_1}) + \varepsilon e^{\rho(\tau_1-t)} \leq E^{\tilde{P}}[e^{-\rho(\tilde{\tau}-\tau_1)}\hat{u}(\tilde{q}_{\tilde{\tau}}) - \kappa\int_{\tau_1}^{\tilde{\tau}} e^{-\rho(s-\tau_1)}ds|\tilde{\mathscr{F}}_{\tau_1}].$$

Redefining the time variables generates a contradiction of Definition 1.

Now suppose the policy is optimal but equality does not hold. There must be some $\omega \in \Omega$ such that

$$E^{\tilde{P}}[e^{-\rho(\tilde{\tau}-\tau_1)}\hat{u}(\tilde{q}_{\tilde{\tau}}) - V(\tilde{q}_{\tau_1})) - \kappa\int_{\tau_1}^{\tilde{\tau}} e^{-\rho(s-\tau_1)}ds|\tilde{\mathscr{F}}_{\tau_1(\omega)}](\omega) > 0,$$

contradicting Definition 1. □

B.1.2 A Convexity Lemma

As preparation for the proof of Proposition 1, we first derive a lemma that is useful in simplifying the optimization problem stated in Definition 1. Starting from any belief $q \in \mathscr{P}(X)$, consider a deviation from the optimal policy that involves either jumping in one direction or in exactly the opposite direction, with the intensities of the two possible jumps balanced so as to imply that beliefs will be a martingale even if they do not change in the absence of a jump; the policy is maintained until a jump occurs, or some fixed amount of time passes. (Suppose that the jumps in each direction are small enough to be feasible, and that the intensities with which they occur are chosen so that (6) binds. Then this represents a feasible policy.) If no jump has occurred by the fixed time, one then follows the optimal policy starting from beliefs q from then onward. Such a deviation from the optimal policy cannot possibly increase the value function relative to the one achieved by the optimal policy. This allows us to establish the following result.

Lemma 6. *For any $q \in \mathscr{P}(X)$, $\alpha \in (0,1)$, and $z \in \mathbb{R}^{|X|}$ such that $q \pm z \in \mathscr{P}(X)$ and $q \pm z \ll q$,*

$$\chi^{-1}(\rho V(q) + \kappa)(\alpha D(q + (1-\alpha)z||q) + (1-\alpha)D(q - \alpha z||q)) \geq \\ \alpha V(q + (1-\alpha)z) + (1-\alpha)V(q - \alpha z) - V(q).$$

Proof. The result holds trivially for $z = \vec{0}$. Suppose $z \neq \vec{0}$.

Consider a $K = 2$ Poisson process, with jump directions $z_1 = (1-\alpha)z$ and $z_2 = -\alpha z$ and intensities $\psi_1 = \alpha\bar{\psi}$ and $\psi_2 = (1-\alpha)\bar{\psi}$, where

$$\bar{\psi} = \frac{\chi}{\alpha D(q + (1-\alpha)z||q) + (1-\alpha)D(q - \alpha z||q)}.$$

By assumption, $\bar{\psi}$ is strictly positive and finite. Observe by construction under this policy that q_t does not drift and this this policy is feasible.

Suppose the DM chooses this policy starting from beliefs q until h units of time have passed or a jump occurs. If a jump occurs before h time has passed, suppose the DM gathers no information until h time has passed, and that after time h the DM resumes her optimal policies.

By Lemma 5, the discounted expected utility of such a strategy must be less than the utility achieve by an optimal strategy, which yields

$$V(q) \geq e^{-\rho h}\{\alpha V(q + (1-\alpha)z)(1 - e^{-\bar{\psi}h}) + (1-\alpha)V(q - \alpha z)(1 - e^{-\bar{\psi}h}) + e^{-\bar{\psi}h}V(q)\} \\ - \kappa \int_0^h e^{-\rho s} ds.$$

We can rewrite this as

$$(\frac{\kappa}{\rho} + V(q))(e^{\rho h} - 1)e^{\bar{\psi}h} \geq (\exp(\bar{\psi}h) - 1)(\alpha V(q + (1-\alpha)z) + (1-\alpha)V(q - \alpha z) - V(q)).$$

Taking the limit as $h \to 0^+$,

$$(\kappa + \rho V(q))\frac{1}{\bar{\psi}} \geq \alpha V(q + (1-\alpha)z) + (1-\alpha)V(q - \alpha z) - V(q).$$

We can write the expression as

$$\chi^{-1}(\kappa+\rho V(q))(\alpha D(q+(1-\alpha)z||q)+(1-\alpha)D(q-\alpha z||q)) \geq \\ \alpha V(q+(1-\alpha)z)+(1-\alpha)V(q-\alpha z)-V(q),$$

which is the result. □

B.1.3 A Characterization of the Constraint

In the proof of Lemma 10 below, we rely on the following lemma, which translates the constraint (3) into a constraint on the characteristics of the martingale.

Lemma 7. *Suppose a beliefs process is a quasi-left-continuous martingale. Then the process is a semi-martingale with characteristics* (B,C,ν)*, where* $B_t = 0$*,*

$$C_t = \int_0^t \sigma_s \sigma_s^T ds$$

and

$$\nu(\omega;dt,dz) = \psi_t(dz;\omega)dt,$$

where $\sigma_s\sigma_s^T$ *is a predictable, symmetric positive-definite matrix-valued process and* $\psi_t(dz;\omega)$ *is a predictable positive measure on* $\mathbb{R}^{|X|}$ *for each* $(\omega,t) \in \Omega \times \mathbb{R}_+$*. If the beliefs process satisfies (3), then the process is indistinguishable from one for which, for all* $\omega \in \Omega$ *and* $s \in \mathbb{R}_+$*,*

$$\frac{1}{2}tr[\sigma_s\sigma_s^T \bar{k}(q_{s^-})] + \int_{\mathbb{R}^{|X|}\setminus\{0\}} D(q_{s^-}+z||q_{s^-})\psi_s(dz) \leq \chi. \tag{33}$$

Proof. See the appendix, section C.2. □

B.2 Proof of Proposition 1

We begin by proving, using Lemma 6, that the value function is locally Lipschitz-continuous.

Lemma 8. *The value function* $V(q)$ *is locally Lipschitz-continuous on the interior of the simplex and the interior of each face of the simplex.*

Proof. See the technical appendix, section C.7. □

We next prove that $V(q)$ is continuously differentiable on the interior of the simplex. The argument adapts lemma 1 of Benveniste and Scheinkman [1979] to the Lipschitz-continuous setting using the generalized derivatives approach of Clarke [1990].

Lemma 9. *The value function $V(q)$ is continuously differentiable on the interior of the simplex and the interior of each face of the simplex.*

Proof. See the technical appendix, section C.8. □

Armed with this differentiability result, let us revisit Lemma 6. Defining $z = \frac{1}{1-\alpha}\bar{z}$ and $\varepsilon = \frac{\alpha}{1-\alpha}$,

$$\chi^{-1}(\rho V(q)+\kappa)(D(q+\bar{z}||q)+\varepsilon^{-1}D(q-\varepsilon\bar{z}||q)) \geq$$
$$V(q+\bar{z})-V(q)+\varepsilon^{-1}(V(q-\varepsilon\bar{z})-V(q)).$$

Note that this holds for all q in the interior of the simplex, $\bar{z} \in \mathbb{R}^{|X|}$, and $\varepsilon > 0$ such that $q+\bar{z} \ll q$ and $q-\varepsilon\bar{z} \ll q$. Considering the limit as $\varepsilon \to 0^+$, and assuming $\bar{z} \neq \vec{0}$ and hence that $D(q+\bar{z}||q) > 0$,

$$\chi^{-1}(\rho V(q)+\kappa) \geq \frac{V(q+\bar{z})-V(q)-\bar{z}^T \cdot \nabla V(q)}{D(q+\bar{z}||q)}.$$

This result can be rephrased as: for all q in the interior of the simplex,

$$\sup_{q' \in \mathscr{P}(X)\setminus\{q\}: q' \ll q} \frac{V(q')-V(q)-(q'-q)^T \cdot \nabla V(q)}{D(q'||q)} \leq \chi^{-1}(\rho V(q)+\kappa). \tag{34}$$

We next argue, via a viscosity solution approach, that

$$\sup_{q' \in \mathscr{P}(X)\setminus\{q\}: q' \ll q} \frac{V(q')-V(q)-(q'-q)^T \cdot \nabla V(q)}{D(q'||q)} = \chi^{-1}(\rho V(q)+\kappa) \tag{35}$$

on the intersection of the interior of the simplex and the continuation region. We begin by proving that V is a viscosity sub-solution of the HJB associated with the original problem. The proof adapts the approach of Pham [2009] to our setting; that textbook is also a useful reference on viscosity solutions in an HJB context. Let $\mathbb{S}_{|X|,(|X|-1)}$ be the set of $|X| \times (|X|-1)$ matrices and $\mathscr{M}_+(\mathbb{R}^{|X|})$ be the space of positive measures on $\mathbb{R}^{|X|}$.

Lemma 10. *Let $\phi : \mathbb{R}_+^{|X|} \to \mathbb{R}$ be a function that is homogenous of degree one, twice continuously-differentiable on the interior of the simplex, and satisfies $\phi(q) \geq V(q)$ for all $q \in \mathscr{P}(X)$ and $\phi(q_0) = V(q_0)$ for some q_0 on the interior of the simplex. Then*

$$\max\{\sup_{\sigma_0,\psi_0 \in \mathscr{A}(q_0)} \frac{1}{2} tr[\sigma_0 \sigma_0^T \nabla^2 \phi(q_0)] + \int_{\mathbb{R}^{|X|}\setminus\{0\}} (\phi(q_0+z) - \phi(q_0) - z^T \cdot \nabla\phi(q_0))\psi_0(dz) \\ -\rho V(q_0) - \kappa, \hat{u}(q_0) - V(q_0)\} \geq 0, \tag{36}$$

where $A(q_0)$ is the set of $(\sigma,\psi) \in \mathbb{S}_{|X|,(|X|-1)} \times \mathscr{M}_+(\mathbb{R}^{|X|})$ satisfying

$$\frac{1}{2} tr[\sigma_0 \sigma_0^T \bar{k}(q_0)] + \int_{\mathbb{R}^{|X|}\setminus\{0\}} D(q_0+z||q_0)\psi_0(dz) \leq \chi$$

and such that $q_0 + z \in \mathscr{P}(X)$ for all $z \in supp(\psi_0)$.

Proof. See the technical appendix, section C.9. Analogous results can be derived for each face of the simplex. □

Now define the test function

$$\phi(q;q_0,\alpha) = \alpha D(q||q_0) + V(q_0) + (q-q_0)^T \cdot \nabla V(q_0)$$

for some $\alpha \in (0, \chi^{-1}(\rho V(q_0) + \kappa)$, given any q_0 on the relative interior of the simplex such that $V(q_0) > \hat{u}(q_0)$. By the twice continuously-differentiability of D, this test function is twice continuously-differentiable in q, and by construction, it satisfies $\phi(q_0;q_0,\alpha) = V(q_0)$. Noting, by the homogeneity of degree one of V and of D in its first argument, that $V(q_0) = q_0^T \cdot \nabla V(q_0)$, this function is homogenous of degree one.

It also satisfies

$$\frac{1}{2} tr[\sigma_0 \sigma_0^T \nabla^2 \phi(q_0)] + \int_{\mathbb{R}^{|X|}\setminus\{0\}} (\phi(q+z) - \phi(q) - \nabla\phi(q)) d\psi_0(z) = \\ \frac{\alpha}{2} tr[\sigma_0 \sigma_0^T \bar{k}(q)] + \alpha \int_{\mathbb{R}^{|X|}\setminus\{0\}} D(q+z||q) d\psi_0(z),$$

and therefore

$$\sup_{\sigma_0,\psi_0 \in \mathscr{A}(q_0)} \frac{1}{2} tr[\sigma_0 \sigma_0^T \nabla^2 \phi(q_0)] + \int_{\mathbb{R}^{|X|}\setminus\{0\}} (\phi(q+z) - \phi(q) - \nabla\phi(q)) d\psi_0(z) = \alpha\chi$$

and thus (36) cannot hold as

$$\alpha\chi < \rho V(q_0) + \kappa.$$

We therefore conclude that there exists some $q_\alpha \in \mathscr{P}(X) \setminus \{q_0\}$ with $q_\alpha \ll q$ (because q, being in the interior, has full support) such that

$$\alpha D(q_\alpha || q_0) + V(q_0) + (q - q_0)^T \cdot \nabla V(q_0) < V(q_\alpha).$$

Considering a sequence of α converging to $\chi^{-1}\rho V(q_0) + \kappa$ from below yields

$$\sup_{q' \in \mathscr{P}(X)\setminus\{q\}: q' \ll q} \frac{V(q') - V(q) - (q' - q)^T \cdot \nabla V(q)}{D(q'||q)} \geq \chi^{-1}(\rho V(q) + \kappa).$$

Combining this with (34) proves that (35) holds for all q_0 in the interior of the simplex such that $V(q_0) > \hat{u}(q_0)$.

Repeating the argument for each face extends the result to the interior of each face of the simplex. At the extreme points of the simplex, $V(q) = \hat{u}(q)$ (as it is impossible for beliefs to move away from the extreme points, and hence stopping is optimal), and the result extends vacuously. It follows that for all $q \in \mathscr{P}(X)$, either $V(q_0) = \hat{u}(q_0)$ or (35) holds, proving the result.

B.3 Additional Lemma

The following lemma shows that the value function's curvature is limited by the possibility of diffusing along a line.

Lemma 11. *For all $q, q' \in \mathscr{P}(X)$ such that $q' \ll q$ and $q' \neq q$,*

$$V(q') - V(q) - (q' - q) \cdot \nabla V(q) \leq$$
$$(q' - q)^T \cdot \left(\int_0^1 (1-s)\chi^{-1}(\rho V(sq' + (1-s)q) + \kappa)(\bar{k}(sq' + (1-s)q)ds) \cdot (q' - q).\right.$$

Proof. Assume q and q' are in the interior of the simplex.

By Proposition 1,

$$V(q_2) - V(q_1) - (q_2 - q_1)^T \cdot \nabla V(q_1) \leq \chi^{-1}(\rho V(q_1) + \kappa) D(q_2 || q_1)$$

for any q_1, q_2 on the line segment connecting q and q'. Applying this in reverse,

$$(q_2-q_1)^T\cdot(\nabla V(q_2)-\nabla V(q_1))\le\chi^{-1}(\rho V(q_1)+\kappa)D(q_2||q_1)+\chi^{-1}(\rho V(q_2)+\kappa)D(q_1||q_2).$$

Let $q_1=q+\frac{m}{n}s(q'-q)$ and $q_2=q+\frac{m+1}{n}s(q'-q)$ for some integers m,n such that $0\le m<n$ and $s\in[0,1]$. It follows that

$$\begin{aligned}
&s(q'-q)^T\cdot(\nabla V(q+s(q'-q))-\nabla V(q))=\\
&s(q'-q)^T\cdot\sum_{m=0}^{n-1}(\nabla V(q+\frac{m+1}{n}s(q'-q))-\nabla V(q+\frac{m}{n}s(q'-q)))\le\\
&n\chi^{-1}\sum_{m=0}^{n-1}\{(\rho V(q+\frac{m}{n}s(q'-q))+\kappa)D(q+\frac{m+1}{n}s(q'-q)||q+\frac{m}{n}s(q'-q))\}+\\
&n\chi^{-1}\sum_{m=0}^{n-1}\{(\rho V(q+\frac{m+1}{n}s(q'-q))+\kappa)D(q+\frac{m}{n}s(q'-q)||q+\frac{m+1}{n}s(q'-q))\}.
\end{aligned}$$

Apply Taylor's theorem (a first-order Taylor expansion, using the Lagrange form of the remainder):

$$\begin{aligned}
&(n)^2D(q+\frac{m+1}{n}s(q'-q)||q+\frac{m}{n}s(q'-q))=\\
&\frac{1}{2}s^2(q'-q)^T\cdot\nabla_1^2D(q+\frac{m+c_{m,n,s}}{n}s(q'-q)||q+\frac{m}{n}s(q'-q))\cdot(q'-q)
\end{aligned}$$

for some $c_{m,n,s}\in[0,1]$, where ∇_1^2 denotes the Hessian with respect to the first argument.

Define, for $r\in[0,1)$,

$$\begin{aligned}
f_n(r,s)=&\frac{\chi^{-1}}{2}(\rho V(q+\frac{\lfloor nr\rfloor}{n}s(q'-q))+\kappa)s^2\\
&\times(q'-q)^T\cdot(\nabla_1)^2D(q+\frac{\lfloor nr\rfloor+c_{\lfloor nr\rfloor,n,s}}{n}s(q'-q)||q+\frac{\lfloor nr\rfloor}{n}s(q'-q))\cdot(q'-q).
\end{aligned}$$

Note that $f_n(r,s)$ is constant on any interval $[\frac{m}{n},\frac{m+1}{n})$ with integer m,n such that $0\le m<n$.

Consequently,

$$n\sum_{m=0}^{n-1}\{(\rho V(q+\frac{m}{n}s(q'-q))+\kappa)D(q+\frac{m+1}{n}s(q'-q)||q+\frac{m}{n}s(q'-q))\}=$$
$$n^{-1}\sum_{m=0}^{n-1}f_n(\frac{m}{n},s)=\int_0^1 f_n(r,s)dr.$$

By the continuity of the second derivative of D, and the boundedness of the value function, $f_n(r,s)$ is bounded uniformly on (n,r).

By the dominated convergence theorem,

$$\liminf_{n\to\infty} n\sum_{m=0}^{n-1}\{(\rho V(q+\frac{m}{n}s(q'-q))+\kappa)D(q+\frac{m+1}{n}s(q'-q)||q+\frac{m}{n}s(q'-q))\}=$$
$$\liminf_{n\to\infty}\int_0^1 f_n(r,s)dr=$$
$$\frac{\chi^{-1}}{2}(q'-q)^T\cdot\{\int_0^1 s^2(\rho V(q+rs(q'-q))+\kappa)\bar{k}(q+rs(q'-q))dr\}\cdot(q'-q).$$

Similarly, define, for $r\in[0,1)$,

$$g_n(r,s)=\frac{\chi^{-1}}{2}(\rho V(q+\frac{\lfloor nr\rfloor+1}{n}s(q'-q))+\kappa)s^2$$
$$\times(q'-q)^T\cdot(\nabla_1)^2D(q+\frac{\lfloor nr\rfloor+\hat{c}_{\lfloor nr\rfloor,n,s}}{n}s(q'-q)||q+\frac{\lfloor nr\rfloor+1}{n}s(q'-q))\cdot(q'-q)$$

for some $\hat{c}_{m,n,s}\in[0,1]$.

By an identical argument,

$$\liminf_{n\to\infty}\int_0^1 g_n(r,s)dr=\frac{\chi^{-1}}{2}(q'-q)^T\cdot\{\int_0^1 s^2(\rho V(q+rs(q'-q))+\kappa)\bar{k}(q+rs(q'-q))dr\}\cdot(q'-q).$$

It follows that

$$(q'-q)^T\cdot(\nabla V(q+s(q'-q))-\nabla V(q))\leq$$
$$\chi^{-1}(q'-q)^T\cdot\{\int_0^1 s(\rho V(q+rs(q'-q))+\kappa)\bar{k}(q+rs(q'-q))dr\}\cdot(q'-q).$$

Integrating,

$$V(q')-V(q)-(q'-q)^T\cdot\nabla V(q)=(q'-q)^T\cdot\int_0^1(\nabla V(q+s(q'-q))-\nabla V(q))ds.$$
$$\leq(q'-q)^T\cdot\{\int_0^1\int_0^1 s\chi^{-1}(\rho V(q+rs(q'-q))+\kappa)\bar{k}(q+rs(q'-q))drds\}\cdot(q'-q)$$

and

$$\int_0^1\int_0^1 s(\rho V(q+rs(q'-q))+\kappa)\bar{k}(q+rs(q'-q))drds=$$
$$\int_0^1\int_0^s(\rho V(q+l(q'-q))+\kappa)\bar{k}(q+l(q'-q))dlds=$$
$$\int_0^1(1-l)(\rho V(q+l(q'-q))+\kappa)\bar{k}(q+l(q'-q))dl,$$

which is the result.

This result extends immediately to q' on the boundary of the simplex by continuity, and to each face of the simplex by repeating the argument on each face. □

B.4 Additional Lemma

Lemma 12. *Let $u_{max}=\max_{q\in\mathscr{P}(X)}\hat{u}(q)$ and $u_{min}=\min_{q\in\mathscr{P}(X)}\hat{u}(q)$. If D exhibits a strong preference for gradual learning, then*

$$\frac{V(q')-V(q)-(q'-q)^T\cdot\nabla V(q)}{D(q'||q)}<\chi^{-1}(\rho V(q)+\kappa) \tag{37}$$

for all $q,q'\in\mathscr{P}(X)$ such that $q'\ll q$, $q'\neq q$, and

$$|q'-q|^\delta>\frac{\rho(u_{max}-u_{min})}{m(\kappa+\rho u_{min})}.$$

Proof. By contradiction: suppose the reverse inequality holds for some q' satisfying this condition. Then by Lemma 11 and the definition of a strong preference for gradual learning,

$$\frac{D(q'||q)}{1+m|q'-q|^\delta}(\rho u_{max}+\kappa)\chi^{-1}\geq V(q')-V(q)-(q'-q)\cdot\nabla V(q)\geq\chi^{-1}(\rho u_{min}+\kappa)D(q'||q),$$

which yields $\frac{\rho(u_{max}-u_{min})}{\kappa+\rho u_{min}} \geq m|q'-q|^{\delta}$, a contradiction. □

B.5 Proof of Proposition 2

Define the function

$$\phi(q;q_0) = \chi^{-1}(\rho V(q_0)+\kappa)D(q||q_0)+V(q_0)+(q-q_0)^T \cdot \nabla V(q_0).$$

By the HJB equation (Proposition 1),

$$\phi(q;q_0) \geq V(q),$$

with equality if $q = q_0$.

By the Lemma 12, for any $\varepsilon > 0$, there exists a $\delta_\varepsilon > 0$ such that

$$\phi(q;q_0) \geq V(q)+\delta_\varepsilon$$

for all q such that

$$|q-q_0| \geq \varepsilon + (\frac{\rho(u_{max}-u_{min})}{m(\kappa+\rho u_{min})})^{\delta^{-1}}.$$

Take as given the times $t \geq 0$ and $t+h$ for some $h \in h > 0$. Define $\tau_h = \min\{\tau, t+h\}$ as the minimum of the optimal stopping time and $t+h$.

In what follows, let $E^c_{t^-}[X]$ denote the $\mathscr{F}_{t^-}$ conditional expectation of $\mathbf{1}\{\tau \geq t\}X$ under P. By the bounds above,

$$\begin{aligned}
&\frac{1}{h}E^c_{t^-}[e^{-\rho(\tau_h-t)}\phi(q_{\tau_h};q_{t^-}) - \kappa\int_t^{\tau_h} e^{-\rho(s-t)}ds - V(q_{t^-})] \geq \\
&\quad \frac{1}{h}E^c_{t^-}[e^{-\rho(\tau_h-t)}V(q_{\tau_h}) - \kappa\int_t^{\tau_h} e^{-\rho(s-t)}ds - V(q_{t^-})] + \\
&\quad \frac{\delta_\varepsilon}{h}E^c_{t^-}[\mathbf{1}\{|q_{\tau_h}-q_{t^-}| \geq \varepsilon + (\frac{\rho(u_{max}-u_{min})}{m(\kappa+\rho u_{min})})^{\delta^{-1}}\}].
\end{aligned}$$

By the dynamic programming principle (Lemma 5), under an optimal policy (which exists by Lemma 1),

$$E^c_{t^-}[e^{-\rho(\tau_h-t)}V(q_{\tau_h}) - \kappa\int_t^{\tau_h} e^{-\rho(s-t)}ds - V(q_{t^-})] = 0.$$

Likewise, by the martingale property of q_t,

$$E^c_{t^-}[(q_{\tau_h} - q_{t^-})^T \cdot \nabla V(q_{t^-})] = 0.$$

Both of these should be understood as holding P-a.e. on the strict continuation region, and everywhere outside this region, qualifications that also apply to the equations that follow.

Consequently, we must have

$$\frac{1}{h}E^c_{t^-}[e^{-\rho(\tau_h - t)}\phi(q_{\tau_h}; q_{t^-}) - \kappa\int_t^{\tau_h} e^{-\rho(s-t)}ds - V(q_{t^-})] \geq$$
$$\frac{\delta_\varepsilon}{h}E^c_{t^-}[\mathbf{1}\{|q_{\tau_h} - q_{t^-}| \geq \varepsilon + (\frac{\rho(u_{max} - u_{min})}{m(\kappa + \rho u_{min})})^{\delta^{-1}}\}].$$

Note that,

$$\frac{\kappa}{h}E^c_{t^-}[\int_t^{\tau_h} e^{-\rho(s-t)}ds] \geq \frac{\kappa}{h}(\int_t^{t+h} e^{-\rho(s-t)}ds)E^c_{t^-}[\mathbf{1}\{\tau_h \geq t+h\}],$$

which yields, by $h^{-1}\int_t^{t+h} e^{-\rho(s-t)}ds \geq e^{-\rho h}$,

$$\frac{\kappa}{h}E^c_{t^-}[\int_t^{\tau_h} e^{-\rho(s-t)}ds] \geq \kappa e^{-\rho h}E^c_{t^-}[\mathbf{1}\{\tau \geq t+h\}].$$

We adopt the convention that $q_s = q_\tau$ for all $s \geq \tau$. It follows that

$$\frac{1}{h}E^c_{t^-}[e^{-\rho(\tau_h - t)}(\phi(q_{\tau+h}; q_{t^-}) - V(q_{t^-})) - \kappa e^{-\rho h}\mathbf{1}\{\tau \geq t+h\} - (1 - e^{-\rho(\tau_h - t)})V(q_{t^-})] \geq$$
$$\frac{\delta_\varepsilon}{h}E^c_{t^-}[\mathbf{1}\{|q_{t+h} - q_{t^-}| \geq \varepsilon + (\frac{\rho(u_{max} - u_{min})}{m(\kappa + \rho u_{min})})^{\delta^{-1}}\}].$$

By $\phi(q; q_0) \geq V(q) > 0$ and $1 \geq e^{-\rho(\tau_h - t)}$,

$$\frac{1}{h}E^c_{t^-}[\phi(q_{\tau+h}; q_{t^-}) - V(q_{t^-}) - \kappa e^{-\rho h}\mathbf{1}\{\tau \geq t+h\} - (1 - e^{-\rho(\tau_h - t)})V(q_{t^-})] \geq$$
$$\frac{\delta_\varepsilon}{h}E^c_{t^-}[\mathbf{1}\{|q_{t+h} - q_{t^-}| \geq \varepsilon + (\frac{\rho(u_{max} - u_{min})}{m(\kappa + \rho u_{min})})^{\delta^{-1}}\}].$$

By the definition of ϕ and the martingale property of q_t,

$$\frac{1}{h}E^c_{t^-}[\phi(q_{\tau+h};q_{t^-}) - V(q_{t^-}) - \kappa e^{-\rho h}\mathbf{1}\{\tau \geq t+h\} - (1-e^{-\rho(\tau_h - t)})V(q_{t^-})] = \\ \frac{1}{\chi h}(\rho V(q_{t^-}) + \kappa)E^c_{t^-}[e^{-\rho(\tau_h - t)}D(q_{t+h}||q_{t^-})] - \\ \frac{1}{h}(1 - E^c_{t^-}[e^{-\rho(\tau_h - t)}])V(q_{t^-}) - \kappa e^{-\rho h}E^c_{t^-}[\mathbf{1}\{\tau \geq t+h\}].$$

By the non-negativity of D,

$$\frac{1}{\chi h}(\rho V(q_{t^-}) + \kappa)E^c_{t^-}[e^{-\rho(\tau_h - t)}D(q_{t+h}||q_{t^-})] \leq \frac{1}{\chi h}(\rho V(q_{t^-}) + \kappa)E^c_{t^-}[D(q_{t+h}||q_{t^-})]$$

and by the non-negativity of V and $1 \geq e^{-\rho h} + \rho h e^{-\rho h}$,

$$\frac{1}{h}(1 - E^c_{t^-}[e^{-\rho(\tau_h - t)}])V(q_{t^-}) \geq E^c_{t^-}[\mathbf{1}\{\tau \geq t+h\}]e^{-\rho h}\rho V(q_{t^-}).$$

Therefore

$$\frac{1}{\chi}(\rho V(q_{t^-}) + \kappa)E^c_{t^-}[\frac{1}{h}D(q_{t+h}||q_{t^-}) + \chi e^{-\rho h}\mathbf{1}\{\tau \geq t+h\}] \geq \\ \frac{\delta_\varepsilon}{h}E^c_{t^-}[\mathbf{1}\{|q_{t+h} - q_{t^-}| \geq \varepsilon + (\frac{\rho(u_{max} - u_{min})}{m(\kappa + \rho u_{min})})^{\delta^{-1}}\}].$$

By the definition of the constraint (3), for any ω such that $\tau(\omega) \geq t$,

$$\limsup_{h \downarrow 0} \frac{1}{h}E_{t^-}[D(q_{t+h}||q_{t^-})](\omega) \leq \chi,$$

from which it follows (by $\delta_\varepsilon > 0$) that

$$\limsup_{h \downarrow 0} f_h(\omega, t) = 0,$$

where

$$f_h(\omega,t) = \frac{1}{h}E_{t^-}[\mathbf{1}\{|q_{t+h} - q_{t^-}| \geq \varepsilon + (\frac{\rho(u_{max} - u_{min})}{m(\kappa + \rho u_{min})})^{\delta^{-1}}\}](\omega).$$

Note by convention that $q_{t+h} = q_{t^-}$ if $\tau < t$, and consequently this limit result holds irrespective of whether $\tau(\omega) \geq t$.

Observe that, by the Markov and Burkholder-Davis-Gundy inequalities,

$$f_h(\omega,t) \leq \frac{1}{h}\frac{E_{t^-}[|q_{t+h}-q_{t^-}|^2](\omega)}{(\varepsilon+(\frac{\rho(u_{max}-u_{min})}{m(\kappa+\rho u_{min})})^{\delta^{-1}})^2} \leq 4\frac{1}{h}\frac{E_{t^-}[<q,q>_{t+h}-<q,q>_{t^-}](\omega)}{(\varepsilon+(\frac{\rho(u_{max}-u_{min})}{m(\kappa+\rho u_{min})})^{\delta^{-1}})^2},$$

where $<q,q>$ denotes the quadratic variation. As a consequence of the constraint and the strong convexity of D (with associated constant $K>0$), $<q,q>_{t+h}-<q,q>_{t^-} \leq \frac{\chi h}{K}$, and consequently

$$f_h(\omega,t) \leq \frac{4\chi}{K(\varepsilon+(\frac{\rho(u_{max}-u_{min})}{m(\kappa+\rho u_{min})})^{\delta^{-1}})^2}.$$

Observe that

$$\begin{aligned} E_0[\sum_{k=0}^{n-1} h_n f_{h_n}(\omega,kh_n)] &= E_0[\sum_{k=0}^{n-1} \mathbf{1}\{|q_{h_n(k+1)}-q_{(h_nk)^-}| \geq \varepsilon+(\frac{\rho(u_{max}-u_{min})}{m(\kappa+\rho u_{min})})^{\delta^{-1}}\}] \\ &\geq E_0[\max_{k\in\{0,\ldots,n-1\}} h_n f_{h_n}(\omega,kh_n)]. \end{aligned}$$

Fix some $T>0$, and define the sequence $h_n = n^{-1}T$. Define $\mu_n(r)$ as a collection of point masses with mass h_n on $r=\frac{k}{n}T$ for each $k\in\{0,\ldots,n-1\}$. By definition,

$$\int_0^T f_{h_n}(\omega,r)d\mu_n(r) = \sum_{k=0}^{n-1} h_n f_{h_n}(\omega,kh_n).$$

By the reverse Fatou's lemma with varying measures,

$$\limsup_{n\to\infty} E_0[\int_0^T f_{h_n}(\omega,r)d\mu_n(r)] \leq E_0[\int_0^T (\limsup_{n\to\infty} f_{h_n}(\omega,r))dr] = 0.$$

Likewise, we can write

$$\max_{k\in\{0,\ldots,n-1\}} h_n f_{h_n}(\omega,kh_n) = \sup_{r\in[0,T)} h_n f_{h_n}(\omega,\lfloor\frac{r}{T}n\rfloor h_n)$$

and observe by the dominated convergence theorem that

$$0 = \lim_{n\to\infty} E_0[\sup_{r\in[0,T)} h_n f_{h_n}(\omega,\lfloor\frac{r}{T}n\rfloor h_n)] = E_0[\lim_{n\to\infty}\sup_{r\in[0,T)} h_n f_{h_n}(\omega,\lfloor\frac{r}{T}n\rfloor h_n)]$$

and

$$E_0[\lim_{n\to\infty}\sup_{r\in[0,T)} h_n f_{h_n}(\omega, \lfloor\frac{r}{T}n\rfloor h_n)] \geq E_0[\sup_{r\in[0,T)}\lim_{n\to\infty} h_n f_{h_n}(\omega, \lfloor\frac{r}{T}n\rfloor h_n)] \geq 0,$$

which yields

$$E_0[\sup_{r\in[0,T)} \mathbf{1}\{|q_t(\omega) - q_{t^-}(\omega)| \geq \varepsilon + (\frac{\rho(u_{max}-u_{min})}{m(\kappa+\rho u_{min})})^{\delta^{-1}}\}] = 0,$$

as q_{t^-} is left-continuous and q_t is right-continuous. Since this must hold for all $T > 0$, P-a.e.,

$$\sup_{t\in\mathbb{R}_+} \mathbf{1}\{|q_t(\omega) - q_{t^-}(\omega)| \geq \varepsilon + (\frac{\rho(u_{max}-u_{min})}{m(\kappa+\rho u_{min})})^{\delta^{-1}}\} = 0,$$

and therefore

$$Pr\{\sup_{t\in\mathbb{R}_+} |q_t(\omega) - q_{t^-}(\omega)| \geq \varepsilon + (\frac{\rho(u_{max}-u_{min})}{m(\kappa+\rho u_{min})})^{\delta^{-1}}\} = 0.$$

Because this holds for all $\varepsilon > 0$,

$$Pr\{\sup_{t\in\mathbb{R}_+} |q_t(\omega) - q_{t^-}(\omega)| > (\frac{\rho(u_{max}-u_{min})}{m(\kappa+\rho u_{min})})^{\delta^{-1}}\} = 0,$$

which is the result.

B.6 Proof of Proposition 3

We first prove the claim concerning the HJB equation (which involves proving the viscosity sub- and super- solution properties) and then argue for the existence of an optimal diffusion process.

Viscosity Sub-Solution By Proposition 1, anywhere $V(q_0) > \hat{u}(q_0)$ there exists a vector $\{v \in \mathbb{R}^{|X|} : |v| = 1 \,\&\, v^T q_0 = 0\}$ such that either, for some $\varepsilon > 0$ with $q_0 + \varepsilon Diag(q_0)v \in \mathscr{P}(X)$,

$$\frac{V(q_0 + \varepsilon Diag(q_0)v) - V(q_0) - \varepsilon v^T \cdot Diag(q_0) \cdot \nabla V(q_0)}{D(q_0 + \varepsilon Diag(q_0)v||q_0)} = \chi^{-1}\kappa,$$

or

$$\lim_{\varepsilon\to 0^+}\sup\frac{V(q_0+\varepsilon Diag(q_0)v)-V(q_0)-\varepsilon v^T\cdot Diag(q_0)\cdot\nabla V(q_0)}{D(q_0+\varepsilon Diag(q_0)v||q_0)}=\chi^{-1}\kappa.$$

We begin by proving that the latter must in fact hold under a preference for gradual learning. Suppose not; then for some $\delta>0$, $\bar{\varepsilon}>0$ and all $\alpha\in(0,\bar{\varepsilon})$,

$$\alpha(V(q_0+\varepsilon Diag(q_0)v)-V(q_0))-\alpha\chi^{-1}\kappa D(q_0+\varepsilon Diag(q_0)v||q_0)\geq$$
$$(V(q_0+\alpha\varepsilon Diag(q_0)v)-V(q_0))-\chi^{-1}\kappa D(q_0+\alpha\varepsilon Diag(q_0)v||q_0)+\delta$$

This can be written as

$$V(q_0+\varepsilon Diag(q_0)v)-V(q_0)-\frac{1}{\alpha}(V(q_0+\alpha\varepsilon Diag(q_0)v)-V(q_0))\geq$$
$$\chi^{-1}\kappa D(q_0+\varepsilon Diag(q_0)v||q_0)-\frac{1}{\alpha}\chi^{-1}\kappa D(q_0+\alpha\varepsilon Diag(q_0)v||q_0)+\delta.$$

Considering the limit as $\alpha\to 0^+$, and applying Lemma 11 and a preference for gradual learning,

$$\chi^{-1}\kappa D(q_0+\varepsilon Diag(q_0)v||q_0)\geq V(q_0+\varepsilon Diag(q_0)v)-V(q_0)-\varepsilon v^T Diag(q_0)\nabla V(q_0)$$
$$\geq\chi^{-1}\kappa D(q_0+\varepsilon Diag(q_0)v||q_0)+\delta$$

a contradiction. We must therefore have

$$\lim_{\varepsilon\to 0^+}\sup\varepsilon^{-2}(V(q_0+\varepsilon Diag(q_0)v)-V(q_0)-\varepsilon v^T\cdot Diag(q_0)\cdot\nabla V(q_0))=$$
$$\frac{1}{2}\frac{\kappa}{\chi}v^T Diag(q_0)\bar{k}(q_0)Diag(q_0)v.$$

for some $v\in\mathbb{R}^{|X|}:|v|=1\,\&\,v^Tq=0$. Consequently, any twice continuously-differentiable test function satisfying

$$\phi(q_0)=V(q_0)$$

and $\phi(q)\geq V(q)$ must satisfy $\nabla\phi(q_0)=V(q_0)$ and

$$v^T Diag(q_0)(\nabla^2\phi(q_0)-\frac{\kappa}{\chi}\bar{k}(q_0))Diag(q_0)v\geq 0$$

which is the viscosity sub-solution property, as we must have

$$\max\{\max_{\{v\in\mathbb{R}^{|X|}:|v|=1\,\&\,v^Tq=0\}} v^T Diag(q_0)(\nabla^2\phi(q_0)-\frac{\kappa}{\chi}\bar{k}(q_0))Diag(q_0)v, \hat{u}(q_0)-\phi(q_0)\}\geq 0.$$

Viscosity Super-Solution By Proposition 1, for any vector $\{v\in\mathbb{R}^{|X|}:|v|=1\,\&\,v^Tq_0=0\}$,

$$\lim_{\varepsilon\to 0^+}\frac{V(q_0+\varepsilon Diag(q_0)v)-V(q_0)-\varepsilon v^T\cdot Diag(q_0)\cdot\nabla V(q_0)}{D(q_0+\varepsilon Diag(q_0)v||q_0)}\leq\chi^{-1}\kappa,$$

and therefore for all such v,

$$\lim_{\varepsilon\to 0^+}\varepsilon^{-2}(V(q_0+\varepsilon Diag(q_0)v)-V(q_0)-\varepsilon v^T\cdot Diag(q_0)\cdot\nabla V(q_0))\leq \frac{1}{2}\frac{\kappa}{\chi}v^T Diag(q_0)\bar{k}(q_0)Diag(q_0)v.$$

Consequently, any twice continuously-differentiable test function satisfying

$$\phi(q_0)=V(q_0)$$

and $\phi(q)\leq V(q)$ must satisfy

$$v^T Diag(q_0)(\nabla^2\phi(q_0)-\frac{\kappa}{\chi}\bar{k}(q_0))Diag(q_0)v\leq 0,$$

and by $V(q_0)\geq\hat{u}(q_0)$ we must have

$$\max\{\max_{\{v\in\mathbb{R}^{|X|}:|v|=1\,\&\,v^Tq=0\}} v^T Diag(q_0)(\nabla^2\phi(q_0)-\frac{\kappa}{\chi}\bar{k}(q_0))Diag(q_0)v, \hat{u}(q_0)-\phi(q_0)\}\leq 0.$$

Diffusion Process Consider a version of the DM's problem in which the DM is restricted to choose processes of the form

$$dq_t=Diag(q_t)\sigma_t dB_t,$$

subject to the constraint

$$\frac{1}{2}tr[\sigma_t^T Diag(q_t)\bar{k}(q_t)Diag(q_t)\sigma_t]\leq\chi,$$

and the requirement that a solution to the resulting SDE exist, as in the example given in the text. Call the associated value function V^R. By standard arguments (see, e.g., Pham [2009]), V^R is the unique viscosity solution to the HJB equation described in this proposition, and hence $V^R = V$ and the optimal policies implementing V^R also implement V.

B.7 Proof of Corollary 4

We begin by observing that Proposition 7 characterizes the solution to the HJB equation of Proposition 3 (irrespective of whether D exhibits a preference for gradual learning or not). The only place gradual learning is used in the proof of Proposition 7 is to show that

$$\lim_{h\to 0^+} h^{-1} E_{t^-}[H(q_{t+h}) - H(q_{t^-})] \leq \lim_{h\to 0^+} h^{-1} E_{t^-}[D(q_{t+h}||q_{t^-})]$$

for any feasible policy; but if policies are restricted to continuous martingales, this equation holds (with equality) by Ito's lemma and Assumption 1.

Now consider in particular utility functions with only two actions, L and R (all other action in A are dominated by those two and hence will never occur with positive probability). Using the first-order conditions for the static problem, we have, assuming interior solutions,

$$u_L - \frac{\kappa}{\chi}\nabla H(q_L^*(q_0)) = u_R - \frac{\kappa}{\chi}\nabla H(q_R^*(q_0))$$

and

$$\pi_L^*(q_0)q_L^*(q_0) + (1-\pi_L^*(q_0))q_R^*(q_0) = q_0.$$

Now pick any q_0, q_L, q_R such that $q_0 = \pi q_L + (1-\pi)q_R$ for some $\pi \in (0,1)$. Set

$$u_L = \frac{\kappa}{\chi}\nabla H(q_L)) - \frac{\kappa}{\chi}\nabla H(q_0) + K\iota$$

and

$$u_R = \frac{\kappa}{\chi}H_q(q_R)) - \frac{\kappa}{\chi}H_q(q_0) + K\iota$$

for some K such that both u_L and u_R are strictly positive, where ι is a vector of ones. Observe that if the solution is interior, q_L, q_R, and π are optimal policies.

If the solution is not interior, stopping must be optimal. By the convexity of H,

$$q_L^T \cdot u_L - \frac{\kappa}{\chi}H(q_L) + \frac{\kappa}{\chi}H(q_0) + \frac{\kappa}{\chi}(q_L - q_0)^T H_q(q_0)) - q_0^T \cdot u_L = \\ \frac{\kappa}{\chi}(q_L - q_0)^T H_q(q_L) - \frac{\kappa}{\chi}H(q_L) + \frac{\kappa}{\chi}H(q_0) \geq 0,$$

and likewise for q_R. It follows that the q_0 is in the continuation region, and therefore that (q_L, q_R, π) are indeed optimal policies in the static problem.

By the "locally invariant posteriors" property described by Caplin et al. [Forthcoming], it follows that for any $q = \alpha q_L + (1-\alpha)q_R$ with $\alpha \in [0,1]$, (q_L, q_R, α) are optimal policies given initial prior q_0.

As in the proof of Theorem 7, this implies that the value function is twice-differentiable on the line segment between q_L and q_R, with

$$(q_L - q_0)^T \cdot \nabla^2 V(q) \cdot (q_L - q_0) = \frac{\kappa}{\chi}(q_L - q_0)^T \bar{k}(q)(q_L - q_0)$$

for all q on that line segment (this is a slight abuse of notation, as $V(q)$ may not be twice-differentiable in all directions, but is guaranteed to be twice-differentiable in the relevant direction). Integrating,

$$V(q_L) - V(q_0) - (q_L - q_0)^T \cdot \nabla V(q_0) = \\ \frac{\kappa}{\chi}(q_L - q_0)^T \cdot \left(\int_0^1 (1-s)\bar{k}(sq_L + (1-s)q_0)ds\right) \cdot (q_L - q_0) = \\ \frac{\kappa}{\chi}H(q_L) - \frac{\kappa}{\chi}H(q_0) - \frac{\kappa}{\chi}(q_L - q_0)^T \cdot \nabla H(q_0)).$$

By the sub-optimality of jumping directly from q_0 to q_L, it must be the case that

$$V(q_L) - V(q_0) - (q_L - q_0)^T \cdot \nabla V(q_0) \leq \frac{\kappa}{\chi}D(q_L||q_0)$$

and therefore a preference for gradual learning holds between the points q_0 and q_L.

This argument can be repeated for all (q_0, q_L) in the relative interior of the simplex. By the convexity of D and H, we can extend the result to the entirety of the simplex by continuity, proving that a preference for gradual learning must hold.

B.8 Proof of Lemma 2

Recall the definition of a preference for discrete learning: for all $q, q', \{q_s\}_{s\in S}$ with $q' \ll q$ and $\sum_{s\in S} \pi_s q_s = q'$,

$$D(q'||q) + \sum_{s\in S} \pi_s D(q_s||q') \geq \sum_{s\in S} \pi_s D(q_s||q)$$

Therefore, for all $z \in \mathbb{R}^{|X|}$ with support on the support of q' and ε sufficiently small,

$$D(q'||q' + \varepsilon z) + \sum_{s\in S} \pi_s D(q_s||q') \geq \sum_{s\in S} \pi_s D(q_s||q' + \varepsilon z).$$

At $\varepsilon = 0$, this inequality is satisfied by construction. Differentiating the left-hand side (using the assumption that D is differentiable),

$$\frac{\partial}{\partial \varepsilon}[D(q'||q' + \varepsilon z) + \sum_{s\in S} \pi_s D(q_s||q')]|_{\varepsilon=0} = 0,$$

because $D(q'||q' + \varepsilon z)$ is minimized at $\varepsilon = 0$. It follows that the inequality requires that

$$\sum_{s\in S} \pi_s \frac{\partial}{\partial \varepsilon} D(q_s||q' + \varepsilon z)|_{\varepsilon=0} = 0,$$

as otherwise the inequality would be violated for some sufficiently small ε.

By step 1 in the proof of theorem 4 of Banerjee et al. [2005], it follows immediately that

$$D(q'||q) = H(q') - H(q) - (q' - q)^T \cdot \nabla H(q)$$

for some convex function H, where ∇H denotes the gradient. Note that theorem 4 of Banerjee et al. [2005] is stated as requiring that

$$\sum_{s\in S} \pi_s D(q_s||q' + \varepsilon z)$$

be minimized at $\varepsilon = 0$ for all z, but step 1 of the proof in fact only requires that $\varepsilon = 0$ correspond to a critical value for all z. Step 2 of that proof relaxes slightly the regularity conditions, but we have simply assumed these. Minimization is only required to establish the last step of the proof, step 3, which proves strict convexity of H. Strict convexity of $H(q)$ on the support of q follows in our setting immediately from our assumptions on D.

B.9 Proof of Proposition 5

Because D is a Bregman divergence, it satisfies a preference for gradual learning, and the value function described in Proposition 7 is the value function for the DM's problem.

That value function can be implemented in the following way. Let $\pi^* \in \mathscr{P}(A)$ and $\{q_i^* \in \mathscr{P}(X)\}_{i=1}^{|A|}$ be optimal policies in the static problem described in Proposition 7, given some arbitrary assignment of the actions to the numbers $\{1,2,\ldots,|A|\}$. Consider the dynamic K jumps example policy, with $K = |A|$, $z_k = q_k^* - q_0$, and

$$\psi_k = \pi_k^* \frac{\chi}{-H(q_0) + \sum_{i=1}^{|A|} \pi_i^* H(q_i^*)}.$$

Observing that $\sum_{k=1}^{K} z_k \psi_k = 0$, under such a policy beliefs do not drift, and that the policy is feasible, as

$$\sum_{k=1}^{k} \psi_k D(q_i^* || q_0) = \chi$$

and the resulting stochastic process statisfies standard existence conditions (as its coefficients are constant). Assume the DM immediately stops after the first jump. The utility achieved is

$$\begin{aligned} E_0[\hat{u}(q_\tau) - \kappa\tau] &= \sum_{k=1}^{K} \pi_k^* \hat{u}(q_k^*) - \kappa \int_0^\infty e^{-s\sum_{k=1}^{K} \psi_k} ds \\ &= \sum_{k=1}^{K} \pi_k^* \hat{u}(q_k^*) - \frac{\kappa}{\chi}(-H(q_0) + \sum_{i=1}^{|A|} \pi_i^* H(q_i^*)), \end{aligned}$$

which is the value function of Proposition 7. It follows that this policy is an optimal policy.

B.10 Proof of Proposition 6

We divide this proof into three steps. First, we establish necessary optimality conditions. Second, we construct a utility function for which a particular set of policies is optimal. Third, we show that the optimality of this set of policies implies a preference for discrete learning.

Step 1: Necessary Optimality Conditions Under the assumption that there is no continuous martingale component of q_t (note that q_t is equivalent to a purely discontinuous

martingale by the assumption that it does not diffuse outside of a nowhere-dense set), by Lemma 7, we can characterize the martingale q_t entirely by the predictable compensator

$$\nu(\omega; dt, dz) = \psi_t(dz; \omega)dt$$

such that

$$\int_{\mathbb{R}^{|X|}\setminus\{\vec{0}\}} D(q_{t^-} + z||q_{t^-}))\psi_t(dz) \leq \chi.$$

Because the martingale q_t is of finite variation, we have, for any stopping time τ,

$$\begin{aligned} E_t[e^{-\rho\tau}V(q_\tau)] - e^{-\rho t}V(q_t) &= E_t[\int_t^\tau \int_{\mathbb{R}^{|X|}\setminus\{\vec{0}\}} e^{-\rho l}(V(q_{l^-}+z) - V(q_{l^-}) - z^T \cdot \nabla V(q_{l^-}))\psi_l(dz)dl] \\ &\quad - E_t[\int_t^\tau e^{-\rho l}\rho V(q_{l^-})dl] \\ &= E_t[\kappa \int_t^\tau e^{-\rho l}dl] \end{aligned}$$

and consequently by Proposition 1,

$$\int_{\mathbb{R}^{|X|}\setminus\{\vec{0}\}} (V(q_{l^-}+z) - V(q_{l^-}) - z^T \cdot \nabla V(q_{l^-}) - \chi^{-1}(\rho V(q_{l^-}) + \kappa)D(q_{l^-}+z||q_{l^-}))\psi_l(dz) = 0.$$

By assumption, this must hold from any initial q_t in the continuation region.

It follows that there must exist some $z^*(q_{l^-}) \in \mathbb{R}^{|X|} \setminus \{\vec{0}\}$ such that

$$V(q_{l^-}+z^*(q_{l^-})) - V(q_{l^-}) - z^*(q_{l^-})^T \cdot \nabla V(q_{l^-}) = \chi^{-1}(\rho V(q_{l^-}) + \kappa)D(q_{l^-}+z^*(q_{l^-})||q_{l^-}), \tag{38}$$

and moreover that by the immediate stopping result that

$$V(q_{l^-}+z^*(q_{l^-})) = \hat{u}(q_{l^-}+z^*(q_{l^-})),$$

and for all feasible z,

$$V(q_{l^-}+z) - V(q_{l^-}) - z^T \cdot \nabla V(q_{l^-}) \leq \chi^{-1}(\rho V(q_{l^-}) + \kappa)D(q_{l^-}+z||q_{l^-}). \tag{39}$$

To facilitate what follows, we write these conditions in the following manner, akin to a

static rational inattention problem:

$$0 = \sup_{\mu \in int(\mathscr{P}(\{1,2,3\})), \{q_i \in \mathscr{P}(X)\}_{i\in\{1,2,3\}} : \sum_{i=1}^3 \mu_i q_i = q_{l^-}} \frac{\mu_1 \hat{u}(q_1) + \mu_2 V(q_2) + \mu_3 V(q_3) - V(q_{l^-}) - \chi^{-1}(\rho V(q_{l^-}) + \kappa) \sum_{i=1}^3 \mu_i D(q_i || q_{l^-})}{\mu_1}. \tag{40}$$

Choosing $\mu_3 = \mu_2 = \frac{1}{2}(1-\mu_1)$ and $q_3 = q_2 = q_{l^-} - \frac{\mu_1}{1-\mu_1} z^*(q_{l^-})$ is feasible for π_1 sufficiently small and achieves (38) in the limit as $\mu_1 \to 0^+$. The numerator is always weakly negative by (39), and hence (40) must hold.

Step 2: Construct a utility function with certain optimal policies Let us take as given any interior $q, q', q_1, q_2 \in \mathscr{P}(X)$ and $\pi \in (0,1)$ such that

$$\pi q_1 + (1-\pi) q_2 = q',$$

and construct a utility function such that $z = q_1 - q$ and $z = q_2 - q$ are both optimal policies from q, meaning that

$$\begin{aligned} \hat{u}(q_1) - V(q) - (q_1 - q)^T \cdot \nabla V(q) &= \chi^{-1}(\rho V(q) + \kappa) D(q_1 || q), \\ \hat{u}(q_2) - V(q) - (q_2 - q)^T \cdot \nabla V(q) &= \chi^{-1}(\rho V(q) + \kappa) D(q_2 || q), \end{aligned}$$

and for which $V(q) > \hat{u}(q)$ and

$$V(q') \leq V(q).$$

The basic idea behind this proof is to construct the utility function in such a way as to ensure that the value function is the solution to a static rational inattention problem, in that the optimal policy is to jump to one of three beliefs with intensities such that beliefs do not drift.

Define, for some $\xi = (0,1)$, an interior $q_3 \in \mathscr{P}(X)$ such that

$$\xi q_3 + (1-\xi) q' = q.$$

Note that such a q_3 exists by the assumption that q is in the interior of the simplex.

Let $v \in \mathbb{R}^{|X|}$ be a vector and let k_1, k_2, k_3, K be constants. Define

$$\theta = \chi^{-1}(\rho K + \kappa).$$

Suppose there are three actions, and let their utilities satisfy, for $a \in A = \{1,2,3\}$,

$$u_a = \theta \nabla_1 D(q_a||q) + v + |X|^{-1} \iota k_a,$$

where $u_a \in \mathbb{R}^{|X|}$ are the payoffs associated with action a, $\nabla_1 D(q_a||q)$ is the gradient with respect to the first argument and $\iota \in \mathbb{R}^{|X|}$ is a vector of ones. This gradient exists by the differentiability of D in its first argument and the assumption that q_a is interior. Define

$$k_a = \theta D(q_a||q) - \theta q_a^T \cdot \nabla_1 D(q_a||q) + K - q^T v$$

so that

$$\theta D(q_a||q) = q_a^T \cdot u_a - K - (q_a - q)^T v.$$

Note that, to satisfy the requirement that $u_{a,x}$ be positive, we will require that K be sufficiently large given v (we provide an explicit expression below).

Observe that, for any a, a', that

$$\begin{aligned} q_a^T (u_a - u_{a'}) = \theta D(q_a||q) + K - (q - q_a)^T v \\ - \theta D(q_{a'}||q) - K + (q - q_{a'})^T v \\ - (q_a - q_{a'})^T \cdot u_{a'}, \end{aligned}$$

and by the convexity of D that

$$q_a^T (u_a - u_{a'}) \geq 0,$$

and therefore $\hat{u}(q_a) = q_a^T u_a$.

By the convexity of D, for any $q'' \ll q$ and any $a \in \{1,2,3\}$,

$$\theta D(q''||q) \geq \theta D(q_a||q) + (q'' - q_a)^T \cdot \theta \nabla_1 D(q_a||q),$$

which is

$$\theta D(q''||q) \geq \max_{a \in \{1,2,3\}} (q'')^T u_a - (q'' - q)^T \cdot v - K. \tag{41}$$

By the strict convexity of D, this inequality must be strict for all $q'' \in \{q_1, q_2, q_3\}$, and must

be an equality for $q'' \in \{q_1, q_2, q_3\}$. Note that this implies $K > \hat{u}(q)$.

Let us now consider the "static rational inattention problem"

$$\max_{\mu \in \mathscr{P}(A), \{\hat{q}_i \in \mathscr{P}(X)\}_{i \in A}} \sum_{i \in A} \mu_i \{\hat{u}(\hat{q}_i) - \theta D(\hat{q}_i || q)\}$$

subject to $\sum_{i \in A} \mu_i \hat{q}_i = q$. By the "Lagrangian lemma" of Caplin et al. [Forthcoming] applied to the vector v, the above conditions show that $\mu^* = (\pi(1-\xi), (1-\pi)(1-\xi), \xi)$ and $\hat{q}_i^* = q_a$ are optimal, noting by construction that

$$\pi(1-\xi)q_1 + (1-\pi)(1-\xi)q_2 + \xi q_3 = q.$$

Note by construction that the maximized value is $K = \sum_{a \in A} \mu_a^* \{\hat{u}(q_a) - \theta D(q_a || q)\}$. Note also that the optimal policy is unique (up to a permutation of the assignment of i to A) by the strictness of (41) for $q'' \notin \{q_1, q_2, q_3\}$ and the uniqueness of the weights μ^* satisfying $\sum_{i \in A} \mu_i^* q_a = q$.

Consider the value function associated with this utility function, $V(q''; v, K)$. We must have, by sub-optimality, for any $q'' \ll q$,

$$V(q'') - V(q; v, K) - (q'' - q)^T \cdot \nabla V(q; v, K) \leq \chi^{-1}(\rho V(q; v, K) + \kappa) D(q'' || q). \tag{42}$$

Applying this to $q'' \in \{q_1, q_2, q_3\}$ and using $V(q'') \geq \hat{u}(q'')$,

$$K - V(q; v, K) - (q_a - q)^T \cdot (\nabla V(q; v, K) - v) \leq \chi^{-1} \rho (V(q; v, K) - K) D(q_a || q). \tag{43}$$

Summing by μ^*, we find that $V(q; v, K) \geq K$, and consequently $V(q; v, K) > \hat{u}(q)$.

Now consider any policy in the static problem, $(\mu \in int(\mathscr{P}(A)), \{\hat{q}_i \in \mathscr{P}(X)\}_{i \in A})$. Observe that, by (41) and (42),

$$\sum_{i \in A} \mu_i (V(\hat{q}_i) - \hat{u}(\hat{q}_i)) + K - V(q; v, K) \leq \sum_{i \in A} \mu_i \chi^{-1} \rho (V(q; v, K) - K) D(\hat{q}_i || q).$$

strictly if $\hat{q}_i \notin \{q_1, q_2, q_3\}$ for any $i \in A$.

Using this equation and $\hat{u}(\hat{q}_1) \leq V(\hat{q}_1)$, we have

$$\frac{\mu_1\hat{u}(\hat{q}_1)+\mu_2 V(\hat{q}_2;v,K)+\mu_3 V(\hat{q}_3;v,K)-V(q;v,K)-\chi^{-1}(\rho V(q)+\kappa)\sum_{i=1}^3\mu_i D(\hat{q}_i||q)}{\mu_1} \leq$$
$$\frac{\sum_{i=1}^3\mu_i\{V(\hat{q}_i;v,K)-\hat{u}(\hat{q}_i)+K-V(q;v,K)-\chi^{-1}\rho(V(q;v,K)-K)D(\hat{q}_i||q)\}}{\mu_1}+$$
$$\frac{-K+\sum_{i=1}^3\mu_i\{\hat{u}(\hat{q}_i)-\theta D(\hat{q}_i||q)\}}{\mu_1},$$

and therefore by $\mu_1 \in (0,1]$ and $-K+\sum_{i=1}^3\mu_i\{\hat{u}(\hat{q}_i)-\theta D(\hat{q}_i||q)\} \leq 0$,

$$\frac{\mu_1\hat{u}(\hat{q}_1)+\mu_2 V(\hat{q}_2;v,K)+\mu_3 V(\hat{q}_3;v,K)-V(q;v,K)-\chi^{-1}(\rho V(q)+\kappa)\sum_{i=1}^3\mu_i D(\hat{q}_i||q)}{\mu_1} \leq$$
$$-K+\sum_{i=1}^3\mu_i\{\hat{u}(\hat{q}_i)-\theta D(\hat{q}_i||q)\} \leq 0.$$

Consequently, the sequence of policies $(\mu_n \in int(\mathscr{P}(A)), \{\hat{q}_{i,n} \in \mathscr{P}(X)\}_{i\in A})$ achieving

$$\lim_{n\to\infty}\frac{\mu_{1,n}\hat{u}(\hat{q}_{1,n})+\mu_{2,n}V(\hat{q}_{2,n};v,K)+\mu_{3,n}V(\hat{q}_{3,n};v,K)-V(q;v,K)}{\mu_1}-$$
$$\frac{\chi^{-1}(\rho V(q;v,K)+\kappa)\sum_{i=1}^3\mu_{i,n}D(\hat{q}_{i,n}||q)}{\mu_1}=0$$

(which exists by (40)) must achieve

$$\lim_{n\to\infty}-K+\sum_{i=1}^3\mu_{i,n}\{\hat{u}(\hat{q}_{i,n})-\theta D(\hat{q}_{i,n}||q)\}=0.$$

By the boundedness of the simplex, this sequence has a convergent subsequence, and by the uniqueness (up to a permutation) of the optimal policy in the "static problem," this convergent subsequence must converge to some permutation of $\mu^*, \{q_1,q_2,q_3\}$. Supposing without loss of generality that $\lim_{n\to\infty}\hat{q}_{1,n}=q_1$,

$$\mu_1\hat{u}(q_1)+\mu_2 V(q_2;v,K)+\mu_3 V(q_3;v,K)-V(q;v,K)-\chi^{-1}(\rho V(q;v,K)+\kappa)\sum_{i=1}^3\mu_i^* D(q_i||q)=0$$

and that $\hat{u}(q_1)=V(q_1;v,K)$. It follows immediately that jumping to $z_a=q_a-q$ with probability μ_a^* is an optimal policy of the dynamic problem, and by the uniqueness of

the optimal policy in the "static problem," this must be the only optimal policy. By the assumption of immediate stopping, $\hat{u}(q_2) = V(q_2; v, K)$ and $\hat{u}(q_3) = V(q_3; v, K)$.

Therefore,

$$K - V(q; v, K) - \chi^{-1}\rho(V(q; v, K) - K)\sum_{i=1}^{3} \mu_i^* D(q_i||q) = 0,$$

which yields $V(q; v, K) = K$. Plugging this into (43),

$$(q_a - q)^T \cdot (\nabla V(q; v, K) - v) \geq 0,$$

implying that q is a local minima of $V(q; v, K) - v^T \cdot q$ over the set

$$\{\tilde{q} \in \mathscr{P}(X) : \exists \hat{\pi} \in \mathscr{P}(A)\, s.t. \sum_{a \in A} \hat{\pi}_a q_a = \tilde{q}\},$$

and thus that $(q_a - q)^T \cdot (\nabla V(q; v, K) - v) = 0$.

This result holds regardless of the values of v, K. Choose

$$v = -\theta \nabla_1 D(q'||q),$$

and by sub-optimality of jumping to q' from q we have

$$V(q'; v, K) \leq V(q; v, K) + (\pi_1 q_1 + (1 - \pi)q_2 - q)^T \cdot \nabla V(q; v, K) + \theta D(q'||q),$$

recalling that $\pi_1 q_1 + (1 - \pi)q_2$. Using $(q_a - q)^T \cdot (\nabla V(q; v, K) - v) = 0$, this is

$$V(q'; v, K) \leq V(q; v, K) - \theta(q' - q)\nabla_1 D(q'||q) + \theta D(q'||q).$$

By the convexity of D,

$$V(q'; v, K) \leq V(q; v, K),$$

as required.

To establish positive utilities, choose for some $\varepsilon > 0$

$$\begin{aligned} -K = \min_{x \in X, a \in \{1,2,3\}} & e_x^T \cdot (\theta \nabla_1 D(q_a||q) - \theta \nabla_1 D(q'||q)) + \theta D(q_a||q) \\ & - \theta q_a^T \cdot \nabla_1 D(q_a||q) - \theta q^T \cdot \nabla_1 D(q'||q) - \varepsilon, \end{aligned}$$

which ensures that

$$\min_{x\in X, a\in A} u_{a,x} = \varepsilon.$$

Step 3: Prove the inequality We begin by proving that a preference for discrete learning exists for two-signal alphabets, and assuming that all of the relevant elements of the simplex are interior. We then extend the result to prove the full preference for discrete learning.

Proof by contradiction: suppose there exists an interior $q, q', q_1, q_2 \in \mathscr{P}(X)$ and $\pi \in (0,1)$ such that

$$\pi q_1 + (1-\pi) q_2 = q'$$

and

$$D(q'|q) + \pi D(q_1|q') + (1-\pi) D(q_2||q') < \pi D(q_1|q) + (1-\pi) D(q_2||q).$$

By the results of the previous step, there exists an action space A and utility function u such that $z = q_1 - q$ and $z = q_2 - q$ are both optimal policies from q, and for which

$$V(q') \leq V(q),$$

where V denotes the value function given those utilities (i.e. the $V(q; v, K)$ in step 2 above, for the particular values of v, K chosen above).

Then we must have, for $a \in \{1,2\}$,

$$V(q_a) - V(q) - (q_a - q)^T \cdot \nabla V(q) = (\rho V(q) + \kappa) D(q_a||q),$$

$$V(q') - V(q) - (q' - q)^T \cdot \nabla V(q) \leq (\rho V(q) + \kappa) D(q'||q),$$

$$V(q_a) - V(q') - (q_a - q')^T \cdot \nabla V(q') \leq (\rho V(q') + \kappa) D(q_a||q') \leq \theta (\rho V(q) + \kappa) D(q_a||q'),$$

Putting these together,

$$(\rho V(q) + \kappa)(D(q'||q) + D(q_a||q') - D(q_a||q)) \geq -(q_a - q')^T \cdot [\nabla V(q') - \nabla V(q)].$$

Summing over $a \in \{1,2\}$ weighted by π and $(1-\pi)$, and using $(\rho V(q) + \kappa) > 0$,

$$D(q'|q) + \pi D(q_1|q') + (1-\pi) D(q_2||q') \geq \pi D(q_1|q) + (1-\pi) D(q_2||q),$$

a contradiction.

We conclude that for all interior $q, q', q_1, q_2 \in \mathscr{P}(X)$ and $\pi \in (0,1)$,

$$D(q'|q) + \pi D(q_1|q') + (1-\pi)D(q_2||q') \geq \pi D(q_1|q) + (1-\pi)D(q_2||q).$$

The result extends immediately to more than two $\{q_s\}$ by adding this expression for different pairs. The result extends to the boundary of the simplex by continuity.

B.11 Proof of Proposition 7

Define $\phi(q_t)$ as the static value function in the statement of the theorem (we will prove that it is equal to $V(q_t)$, the value function of the dynamic problem). We first show that any strategy for the DM achieves weakly less utility than $\phi(q_0)$. We then show that $\phi(q_t)$ satisfies the HJB equation of Proposition 3 (at least in a viscosity sense), and construct a diffusion strategy with the properties described that achieves the value $\phi(q_0)$.

Step 1: Show that all other feasible policies achieve a lower utility First, we verify that alternative policies achieve less utility than $\phi(q_0)$. Observe that for any feasible process, by the definition of gradual learning and Assumption 1, we must have

$$\lim_{h \to 0^+} h^{-1} E_{t^-}[H(q_{t+h}) - H(q_{t^-})] \leq \lim_{h \to 0^+} h^{-1} E_{t^-}[D(q_{t+h}||q_{t^-})] \leq \chi,$$

and consequently

$$E_0[\hat{u}(q_\tau) - \kappa\tau] \leq E_0[\hat{u}(q_\tau) - \frac{\kappa}{\chi} H(q_\tau) + \frac{\kappa}{\chi} H(q_0)].$$

Let $a^*(q)$ be a selection from $\arg\max_{a \in A} \sum_{x \in X} u_{a,x} q_x$. We can write this as

$$E_0[\hat{u}(q_\tau) - \kappa\tau] \leq \sum_{a \in A} \pi_a E_0[q_\tau^T \cdot u_a - \frac{\kappa}{\chi} H(q_\tau) + \frac{\kappa}{\chi} H(q_0) | a^*(q_\tau) = a],$$

where $\pi_a = E_0[\mathbf{1}\{a^*(q_\tau) = a\}]$. By the convexity of H,

$$E_0[q_\tau^T \cdot u_a - \frac{\kappa}{\chi} H(q_\tau) + \frac{\kappa}{\chi} H(q_0) | a^*(q_\tau) = a] \leq q_a^T \cdot u_a - \frac{\kappa}{\chi} H(q_a) + \frac{\kappa}{\chi} H(q_0),$$

where

$$q_a = E_0[q_\tau | a^*(q_\tau) = a].$$

By the martingale property of beliefs, we must have $\sum_{a\in A}\pi_a q_a = q_0$. We conclude that

$$E_0[\hat{u}(q_\tau)-\kappa\tau] \leq \max_{\pi\in\mathscr{P}(A),\{q_a\in\mathscr{P}(X)\}_{a\in A}} \sum_{a\in A}\pi_a\{q_a^T\cdot u_a - \frac{\kappa}{\chi}H(q_a)+\frac{\kappa}{\chi}H(q_0)\},$$

which is the result.

Step 2: $\phi(q_t)$ satisfies the HJB equation in a viscosity sense We begin by observing, by the homogeneity of degree one of D in its first argument, that

$$(q')^T\cdot\nabla_1^2 D(q'||q)=\vec{0},$$

and consequently

$$q^T\cdot\nabla^2 H(q) = q^T\cdot\bar{k}(q)=\vec{0},$$

and therefore converse of Euler's homogenous function theorem applies. That is, $\nabla H(q_t)$ is homogenous of degree zero, and $H(q_t)$ is homogeneous of degree one.

We start by showing that the function $\phi(q_t)$ is twice-differentiable in certain directions. Substituting the definition of a Bregman divergence into the statement of theorem,

$$\phi(q_0) = \max_{\pi\in\mathscr{P}(A),\{q_a\in\mathscr{P}(X)\}_{a\in A}} \sum_{a\in A}\sum_{x\in X}\pi(a)u_{a,x}q_{a,x}+\frac{\kappa}{\chi}H(q_0)-\frac{\kappa}{\chi}\sum_{a\in A}\pi(a)H(q_a),$$

subject to the constraint ($\sum_{a\in A}\pi_a q_a = q_0$). Define a new choice variable, $\hat{q}_a=\pi(a)q_a$. By definition, $\hat{q}_a\in\mathbb{R}_+^{|X|}$, and the constraint is $\sum_{a\in A}\hat{q}_a = q_0$. By the homogeneity of H, the objective is

$$\sum_{a\in A}u_a^T\cdot\hat{q}_a+\frac{\kappa}{\chi}H(q_0)-\frac{\kappa}{\chi}\sum_{a\in A}H(\hat{q}_a),$$

where $u_a\in\mathbb{R}^{|X|}$ is the vector of $\{u_{a,x}\}_{x\in X}$. Any choice of $\hat{q}_a$ satisfying the constraint can be implemented by some choice of π and q_a in the following way: set $\pi(a)=\iota^T\hat{q}_a$, and (if $\pi(a)>0$) set

$$q_a=\frac{\hat{q}_a}{\pi(a)}.$$

If $\pi(a)=0$, set $q_a=q_0$. By construction, the constraint will require that $\pi(a)\leq 1$, $\sum_{a\in A}\pi(a)=1$, and the fact that the elements of q_a are weakly positive will ensure $\pi(a)\geq 0$. Similarly, $\iota^T q_a = 1$ for all $a\in A$, and the elements of q_a are weakly greater than zero. Therefore, we can implement any set of $\hat{q}_a$ satisfying the constraint $\sum_{a\in A}\hat{q}_a=q_0$.

Rewriting the problem in Lagrangian form,

$$\phi(q_0) = \max_{\{\hat{q}_a \in \mathbb{R}^{|X|}\}_{a\in A}} \min_{\xi \in \mathbb{R}^{|X|}, \{\nu_a \in \mathbb{R}_+^{|X|}\}_{a\in A}} \sum_{a\in A} u_a^T \cdot \hat{q}_a + \frac{\kappa}{\chi} H(q_0)$$
$$- \frac{\kappa}{\chi} \sum_{a\in A} H(\hat{q}_a) + \xi^T (q_0 - \sum_{a\in A} \hat{q}_a) + \sum_{a\in A} \nu_a^T \hat{q}_a.$$

Observe that $\phi(q_0)$ is convex in q_0. Suppose not: for some $q = \lambda q_0 + (1-\lambda)q_1$, with $\lambda \in (0,1)$, $\phi(q) < \lambda\phi(q_0) + (1-\lambda)\phi(q_1)$. Consider a relaxed version of the problem in which the DM is allowed to choose two different $\hat{q}_a$ for each a. Because of the convexity of H, even with this option, the DM will set both of the $\hat{q}_a$ to the same value, and therefore the relaxed problem reaches the same value as the original problem. However, in the relaxed problem, choosing the optimal policies for q_0 and q_1 in the original problem, scaled by λ and $(1-\lambda)$ respectively, is feasible. It follows that $\phi(q) \geq \lambda\phi(q_0) + (1-\lambda)\phi(q_1)$. Note also that $\phi(q_0)$ is bounded on the interior of the simplex. It follows by Alexandrov's theorem that is is twice-differentiable almost everywhere on the interior of the simplex.

By the convexity of H, the objective function is concave, and the constraints are affine and a feasible point exists. Therefore, the KKT conditions are necessary. The objective function is continuously differentiable in the choice variables and in q_0, and therefore the envelope theorem applies. We have, by the envelope theorem,

$$\nabla\phi(q_0) = \frac{\kappa}{\chi}\nabla H(q_0) + \xi,$$

and the first-order conditions (for all $a \in A$ with $\hat{q}_a \neq \vec{0}$),

$$u_a - \frac{\kappa}{\chi}\nabla H(\hat{q}_a) - \xi + \nu_a = 0. \tag{44}$$

If $\hat{q}_a = \vec{0}$, we must have $q^T(u_a - \xi) \leq \frac{\kappa}{\chi} H(q)$ for all q, meaning that $u_a - \kappa$ is a sub-gradient of $H(q)$ at $q = 0$. In this case, we can define $\nu_a = \vec{0}$ and observe that the first-order condition holds. Define $\hat{q}_a(q_0)$, $\xi(q_0)$, and $\nu_a(q_0)$ as functions that are solutions to the first-order conditions and constraints.

We next prove the "locally invariant posteriors" property described by Caplin et al. [Forthcoming]. Consider an alternative prior, $\tilde{q}_0 \in \mathscr{P}(X)$, such that

$$\tilde{q}_0 = \sum_{a\in A} \alpha(a)\hat{q}_a(q_0)$$

for some $\alpha(a) \geq 0$. Conjecture that $\hat{q}_a(\tilde{q}_0) = \alpha(a)\hat{q}_a(q_0)$, $\xi(\tilde{q}_0) = \xi(q_0)$, and $\nu_a(\tilde{q}_0) = \nu_a(q_0)$. By the homogeneity property,

$$\nabla H(\alpha(a)\hat{q}_a(q_0)) = \nabla H(\hat{q}_a(q_0)),$$

and therefore the first-order conditions are satisfied. By construction, the constraint is satisfied, the complementary slackness conditions are satisfied, and $\hat{q}_a$ and ν_a are weakly positive. Therefore, all necessary conditions are satisfied, and by the concavity of the problem, this is sufficient. It follows that the locally invariant posteriors property is verified.

Consider a perturbation

$$q_0(\varepsilon; z) = q_0 + \varepsilon z,$$

with $z \in \mathbb{R}^{|X|}$, such that $q_0(\varepsilon; z)$ remains in $\mathscr{P}(X)$ for some $\varepsilon > 0$. If z is in the span of $\hat{q}_a(q_0)$, then there exists a sufficiently small $\varepsilon > 0$ such that the above conjecture applies. In this case that ξ is constant, and therefore $\nabla\phi(q_0(\varepsilon; z))$ is directionally differentiable with respect to ε. If $q_0(-\varepsilon; z) \in \mathscr{P}(X)$ for some $\varepsilon > 0$, then $\nabla\phi$ is differentiable (let ∇_z denote the gradient with respect to z), with

$$\nabla_z \nabla\phi(q_0) = \frac{\kappa}{\chi} \nabla^2 H(q_0) \cdot z,$$

proving twice-differentiability in this direction. This perturbation exists anywhere the span of $\hat{q}_a(q_0)$ is strictly larger than the line segment connecting zero and q_0 (in other words, all $\hat{q}_a(q_0)$ are not proportional to q_0). Within this region, the strict convexity of $H(q_0)$ in all directions orthogonal to q_0 implies that, as required of the continuation region,

$$\phi(q_0) > \max_{a \in A} u_a^T \cdot q_0.$$

Outside of this region, all $\hat{q}_a(q_0)$ are proportional to q_0, implying that

$$\phi(q_0) = \max_{a \in A} u_a^T \cdot q_0,$$

as required for the stopping region.

Now consider an arbitrary perturbation z such that $q_0(\varepsilon; z) \in \mathbb{R}_+^{|X|}$ and $q_0(-\varepsilon; z) \in \mathbb{R}_+^{|X|}$

for some $\varepsilon > 0$. Observe that, by the constraint,

$$\varepsilon z = \sum_{a \in A} (\hat{q}_a(\varepsilon; z) - \hat{q}_a(q_0)).$$

It follows that

$$(\xi^T(q_0(\varepsilon; z)) - \xi^T(q_0))\varepsilon z = \sum_{a \in A} (\xi^T(q_0(\varepsilon; z)) - \xi^T(q_0))(\hat{q}_a(\varepsilon; z) - \hat{q}_a(q_0)).$$

By the first-order condition,

$$(\xi^T(q_0(\varepsilon; z)) - \xi^T(q_0))(\hat{q}_a(\varepsilon; z) - \hat{q}_a(q_0)) =$$
$$[\frac{\kappa}{\chi}\nabla H(\hat{q}_a(q_0)) - \frac{\kappa}{\chi}\nabla H(\hat{q}_a(\varepsilon; z)) + \nu_a^T(q_0(\varepsilon; z)) - \nu_a^T(q_0)](\hat{q}_a(\varepsilon; z) - \hat{q}_a(q_0)).$$

Consider the term

$$(\nu_a^T(q_0(\varepsilon; z)) - \nu_a^T(q_0))(\hat{q}_a(\varepsilon; z) - \hat{q}_a(q_0)) = \sum_{x \in X} (\nu_a^T(q_0(\varepsilon; z)) - \nu_a^T(q_0))e_x e_x^T(\hat{q}_a(\varepsilon; z) - \hat{q}_a(q_0)).$$

By the complementary slackness condition,

$$(\nu_a^T(q_0(\varepsilon; z)) - \nu_a^T(q_0))(\hat{q}_a(\varepsilon; z) - \hat{q}_a(q_0)) = -\nu_a^T(q_0(\varepsilon; z))\hat{q}_a(q_0) - \nu_a^T(q_0)\hat{q}_a(\varepsilon; z) \leq 0.$$

By the convexity of H,

$$\frac{\kappa}{\chi}(\nabla H(\hat{q}_a(q_0)) - \nabla H(\hat{q}_a(\varepsilon; z)))(\hat{q}_a(\varepsilon; z) - \hat{q}_a(q_0)) \leq 0.$$

Therefore,

$$(\xi^T(q_0(\varepsilon; z)) - \xi^T(q_0))\varepsilon z \leq 0.$$

Thus, anywhere ϕ is twice differentiable (almost everywhere on the interior of the simplex),

$$\nabla^2 \phi(q) \preceq \frac{\kappa}{\chi}\nabla^2 H(q) = \bar{k}(q),$$

with equality in certain directions. Therefore, it satisfies the HJB equation almost every-

where in the continuation region. Moreover, by the convexity of ϕ,

$$\frac{\kappa}{\chi}(\nabla H(q_0(\varepsilon;z)) - \nabla H(q_0))^T \varepsilon z \geq (\nabla \phi(q_0(\varepsilon;z)) - \nabla \phi(q_0))^T \varepsilon z \geq 0,$$

implying that the "Hessian measure" (see Villani [2003]) associated with $\nabla^2 \phi$ has no pure point component. This implies that ϕ is continuously differentiable.

Step 3: Show this value function can be achieved Next, we show that there is a strategy for the DM in the dynamic problem which can implement this value function. Suppose the DM starts with beliefs q_0, and generates some $\hat{q}_a(q_0)$ as described above. As shown previously, this can be mapped into a policy $\pi(a, q_0)$ and $q_a(q_0)$, with the property that

$$\sum_{a \in A} \pi(a, q_0) q_a(q_0) = q_0.$$

Claim: it is without loss of generality to assume that the set $A^* = \{a \in A : \pi(a, q_0) > 0\}$ satisfies $|A^*| \leq |X|$. To see this, note that if $|A^*| > |X|$, there must exist some $a_0 \in A^*$ such that, for some weights $w_a \in \mathbb{R}^{|A^*|-1}$,

$$(q_{a_0} - q_0) = \sum_{a \in A^* \setminus \{a_0\}} w_a (q_a - q_0),$$

as either $\{q_a - q_0\}_{a \in A^* \setminus \{a_0\}}$ forms a basis on the tangent space of the simplex or itself contains a redundant basis vector. By optimality, we must have

$$u_{a_0}^T q_{a_0} - \frac{\kappa}{\chi} H(q_{a_0}) = \sum_{a \in A^* \setminus \{a_0\}} w_a \{u_a^T q_a - \frac{\kappa}{\chi} H(q_a)\}.$$

If $q_{a_0} = q_0$, the policy

$$\tilde{\pi}(a, q_0\} = \begin{cases} 0 & a \notin A^* \setminus \{a_0\} \\ \frac{\pi(a,q_0)}{1-\pi(a_0,q_0)} & a \in A^* \setminus \{a_0\} \end{cases}$$

is also optimal (with the same choices of $\{q_a\}_{a \in A}\}$). If not, we must have $w \neq \vec{0}$.

We will construct a policy such that, for all times t,

$$q_t = \sum_{a \in A^*} \pi_t(a) q_a(q_0)$$

for some $\pi_t(a) \in \mathscr{P}(A^*)$. Let $\mathscr{C}$ (which will be the continuation region) be the set of q_t such that a $\pi_t \in \mathscr{P}(A^*)$ satisfying the above property exists and $\pi_t(a) < 1$ for all $a \in A^*$. The associated stopping rule will be the stop whenever $\pi_t(a) = 1$ for some $a \in A^*$.

For all $q_t \in \mathscr{C}$, there is a linear map from $\mathscr{P}(A^*)$ to $\mathscr{C}$, which we will denote $Q(q_0)$:

$$Q(q_0)\pi_t = q_t.$$

Let us suppose the DM chooses a process satisfying this equation and such that

$$d\pi_t = \sigma_\pi(\pi_t)\bar{\sigma}_\pi dB_t,$$

where $\sigma_{\pi,t}(\pi_t)$ is a bounded and continuous function (specified below) and $\bar{\sigma}_\pi$ is a full rank $|A^*| \times |X|$ matrix. Note that a weak solution to this SDE with initial condition $\pi_0 = \pi(a, q_0)$ exists,[48] and consequently this policy is feasible provided that the constraint (3) is satisfied.

We must have

$$Q(q_0)d\pi_t = Diag(q_t)\sigma_t dB_t,$$

which implies that

$$Q(q_0)\sigma_\pi(\pi_t)\bar{\sigma}_\pi = Diag(Q(q_0)\pi_t)\sigma_t$$

Define $\tilde{\phi}(\pi_t) = \phi(Q(q_0)\pi_t)$. As shown above,

$$Q^T(q_0)\nabla^2\phi(q_t)Q(q_0)$$

exists everywhere in Ω, and therefore

$$\tilde{\phi}(\pi_t) - \frac{\kappa}{\chi}H(Q(q_0)\pi_t)$$

is a martingale. We specify $\sigma_\pi(\pi_t)$ to respect the constraint,

$$\frac{1}{2}tr[\sigma_t\sigma_t^T Diag(q_t)\bar{k}(q_t)Diag(q_t)] = \chi > 0.$$

This can be rewritten as

$$\sigma_\pi(\pi_t) = \left(\frac{\chi}{\frac{1}{2}tr[\bar{\sigma}_\pi\bar{\sigma}_\pi Q^T(q_0)\bar{k}(Q(q_0)\pi_t))Q(q_0)]}\right)^{\frac{1}{2}}.$$

[48] See e.g. theorem 2.34 of chapter III of Jacod and Shiryaev [2013].

This function is continuous by the twice continuous-differentiability of D, and bounded above by the strong convexity of D.

Under the stopping rule described previously, the boundary will be hit a.s. as the horizon goes to infinity. As a result, by the martingale property described above, initializing $\pi_0(a) = \pi(a, q_0)$,

$$\tilde{\phi}(\pi_0) = E_0[\tilde{\phi}(\pi_\tau) - \frac{\kappa}{\chi}H(Q(q_0)\pi_\tau) + \frac{\kappa}{\chi}H(Q(q_0)\pi_0)].$$

By Ito's lemma,

$$\frac{\kappa}{\chi}H(Q(q_0)\pi_\tau) - \frac{\kappa}{\chi}H(Q(q_0)\pi_0) = \int_0^\tau \kappa dt = \kappa\tau.$$

By the value-matching property of ϕ, $\tilde{\phi}(\pi_\tau) = \hat{u}(Q(q_0)\pi_\tau)$. It follows that, as required,

$$\phi(q_0) = \tilde{\phi}(\pi_0) = E_0[\hat{u}(q_\tau) - \kappa\tau].$$

C Technical Appendix

C.1 Upper Hemi-Continuity of Policies

In this subsection, we show a form of upper hemi-continuity for optimal policies with respect to the limit as $\rho \to 0^+$.

Lemma 13. *Fix a utility function u, divergence D, constant $\chi > 0$, and cost of delay $\kappa > 0$. Consider a sequence of rates of time preference converging to zero, $\rho_n \to 0$, and let $((\Omega, \mathscr{F}_n, \{\mathscr{F}_{n,t}\}, P_n), q_n, \tau_n)$ be an associated sequence of optimal policies for each ρ_n. There exists a policy $((\Omega, \mathscr{F}^*, \{\mathscr{F}_t^*\}, P^*), q^*, \tau^*)$ that is optimal when $\rho = 0$ such that a subsequence of (q_n, τ_n) converges in law to (q^*, τ^*).*

Proof. See the appendix, section C.6. □

C.2 Proof of Lemma 7

The beliefs process q_t is a semi-martingale; therefore, there exists characteristics (B, C, ν) such that

$$B_t = \int_0^t b_s dA_s,$$

$$C_t = \int_0^t \hat{\sigma}_s \hat{\sigma}_s^T dA_s$$

and

$$\nu(\omega; dt, dz) = K_t(dz; \omega) dA_t,$$

for predictable processes b_s, σ_s and a transition kernel K, and an increasing, predictable process A that is continuous with respect to the Lebesgue measure on $\mathbb{R}_+$.[49] Because A is continuous with respect to the Lebesgue measure, we can define

$$\sigma_s \sigma_s^T = \hat{\sigma}_s \hat{\sigma}_s^T \frac{dA_s}{ds}$$

and

$$\psi_t(dz; \omega) = K_t(dz; \omega) \frac{dA_t}{ds}.$$

Because $q_t - q_0$ is a martingale, $B = 0$.[50]

[49] See proposition 2.9 of chapter II of Jacod and Shiryaev [2013].

[50] Informally, the characteristic B can be thought of as the drift of the semi-martingale; see definition 2.6 of chapter II of Jacod and Shiryaev [2013].

Lastly, let us prove that the stated constraint is satisfied if and only if (3) is satisfied (both up to an evanesce). Define the non-negative (by the convexity of D) family of stochastic processes

$$f_{t,s}(\omega) = \int_{\mathbb{R}^{|X|}\setminus\{\vec{0}\}} (D(q_{t^-}(\omega)+z||q_{s^-}(\omega)) - D(q_{t^-}(\omega)||q_{s^-}(\omega)) - z^T\cdot\nabla_1 D(q_{t^-}(\omega)||q_{s^-}(\omega)))\psi_t(dz;\omega) + \frac{1}{2}tr[\sigma_t(\omega)\sigma_t^T(\omega)\nabla_1^2 D(q_{t^-}(\omega)||q_{s^-}(\omega))].$$

To simplify notation, we treat $\sigma_{t,x}\sigma_{t,x'}\nabla^2_{1,x,x'}D(q_{t^-}||q_{s^-})$ as zero for any x or x' with $q_{t^-,x}=0$ or $q_{t^-,x'}=0$ (as σ_s will never be such that beliefs move off the boundary of the simplex), and likewise define the integral over $\mathbb{R}^{|X|}\setminus\{\vec{0}\}$ as zero outside of the support of ψ_t (as beliefs will never jump off the boundary of the simplex). These conventions allow the formula above to be applied regardless of whether beliefs are on the interior or edge of the simplex.

Note, by the definition of $\bar{k}$ and the divergence,

$$D(q_{t^-}||q_{t^-}) = z^T\cdot\nabla_1 D(q_{t^-}||q_{t^-}) = 0$$

and

$$\nabla_1^2 D(q_{t^-}||q_{t^-}) = \bar{k}(q_{t^-}),$$

and thus

$$f_{t,t}(\omega) = \int_{\mathbb{R}^{|X|}\setminus\{\vec{0}\}} D(q_{t^-}(\omega)+z||q_{t^-}(\omega))\psi_t(dz;\omega) + \frac{1}{2}tr[\sigma_t(\omega)\sigma_t^T(\omega)\bar{k}(q_{t^-}(\omega)].$$

By the twice continuous differentiability of D, the function

$$D(q_{t^-}(\omega)+z||q_{s^-}(\omega)) - D(q_{t^-}(\omega)||q_{s^-}(\omega)) - z^T\cdot\nabla_1 D(q_{t^-}(\omega)||q_{s^-}(\omega))$$

is bounded uniformly on z such that $q_t^-(\omega)+z \ll q_t^-(\omega)$ (i.e. the support of ψ_t) and s such that $|q_{s^-}(\omega)-q_{t^-}(\omega)| \leq \frac{1}{2}\min_{x\in X} q_{t^-,x}(\omega)$ (i.e. such that q_s does not lie near the boundary of the simplex). By the left-continuity of q_{t^-}, such a condition must hold for all

s sufficiently close to t; consequently, by the dominated convergence theorem,

$$\lim_{s\uparrow t} f_{t,s}(\omega) = f_{t,t}(\omega).$$

Applying Ito's lemma for semi-martingales,[51] the process

$$M_{t,s}(\omega) = \int_s^t f_{r,s}(\omega)dr - D(q_t(\omega)||q_{s^-}(\omega))$$

is a local martingale for any $s \in \mathbb{R}_+$ and satisfies $M_{s,s} = 0$ P-a.s.

Suppose (3) holds. We first use the following lemma to show that $\frac{1}{h}E_{t_1}[D(q_{t+h}||q_{t^-})]$ is bounded uniformly in t.

Lemma 14. *Fix some $\bar{h} > 0$ and suppose (3) holds. For any $t_1, t_2 \in \mathbb{R}_+$ with $t_2 > t_1$, there exists a constant $B > 0$ such that, for all $t \in [t_1, t_2]$,*

$$\sup_{h\in(0,\bar{h}]} \frac{1}{h}E_{t_1^-}[D(q_{t+h}||q_{t^-})] \leq B, \, P-a.e.$$

Proof. By contradiction: if this lemma does not hold, there must exist some $t_1, t \in \mathbb{R}_+$ with $t \geq t_1$ and P-positive measure subset of Ω such that, for all ω in this subset,

$$\sup_{h\in(0,\bar{h}]} \frac{1}{h}E_{t_1^-}[D(q_{t+h}||q_{t^-})](\omega) = \infty.$$

By the boundedness of D ($D(q'||q) \leq \bar{D}$ for all $q, q' \in \mathscr{P}(X)$ with $q' \ll q$), for any $\varepsilon > 0$,

$$\sup_{h\in(\varepsilon,\bar{h}]} \frac{1}{h}E_{t_1^-}[D(q_{t+h}||q_{t^-})](\omega) \leq \frac{\bar{D}}{\varepsilon}.$$

Consequently, we must have, for some sequence $\varepsilon_n \to 0^+$,

$$\lim_{n\to\infty} \sup_{h\in(0,\varepsilon_n]} \frac{1}{h}E_{t_1^-}[D(q_{t+h}||q_{t^-})](\omega) = \infty,$$

contradicting (3). □

It follows by the reverse Fatou lemma that (P-a.s., a caveat that applies to everything

[51] See e.g. theorem 2.42 of chapter II of Jacod and Shiryaev [2013].

that follows)

$$\frac{1}{h}\int_{t_1}^{t_2} \limsup_{h\downarrow 0} E_{t_1^-}[D(q_{t+h}||q_{t^-})]dt \geq \limsup_{h\downarrow 0} \frac{1}{h}\int_{t_1}^{t_2} E_{t_1^-}[D(q_{t+h}||q_{t^-})]dt.$$

By the martingale property of $M_{t,s}$,

$$\int_{t_1}^{t_2} E_{t_1^-}[D(q_{t+h}||q_{t^-})]dt = \int_{t_1}^{t_2} E_{t_1^-}[\int_t^{t+h} f_{s,t}ds]dt.$$

By Tonelli's theorem and the non-negativity of f, interchanging the integrals,

$$\int_{t_1}^{t_2} E_{t_1^-}[\int_t^{t+h} f_{s,t}ds]dt = E_{t_1^-}[\frac{1}{h}\int_{t_1}^{t_2+h}\int_{\max\{s-h,t_1\}}^{s} f_{s,r}drds].$$

By the non-negativity of f,

$$E_{t_1^-}[\frac{1}{h}\int_{t_1}^{t_2+h}\int_{\max\{s-h,t_1\}}^{s} f_{s,r}drds] \geq E_{t_1^-}[\frac{1}{h}\int_{t_1}^{t_2}\int_{\max\{s-h,t_1\}}^{s} f_{s,r}drds].$$

Observe that, for all $s > t_1$, and $h \in (0, s-t_1)$,

$$\frac{1}{h}\int_{\max\{s-h,t_1\}}^{s} f_{s,r}dr \geq \inf_{r\in[s-h,s)} f_{s,r}.$$

Therefore,

$$\liminf_{h\downarrow 0} \frac{1}{h}\int_{\max\{s-h,t_1\}}^{s} f_{s,r}dr \geq f_{s,s}.$$

Consequently, by Fatou's lemma,

$$\liminf_{h\downarrow 0} E_{t_1^-}[\frac{1}{h}\int_{t_1}^{t_2+h}\int_{\max\{s-h,t_1\}}^{s} f_{s,r}drds] \geq E_{t_1^-}[\frac{1}{h}\int_{t_1}^{t_2} f_{s,s}ds].$$

Combining these results yields

$$\chi(t_2-t_1) \geq E_{t_1^-}[\int_{t_1}^{t_2} f_{s,s}ds].$$

By the Lebesgue differentiation theorem and iterated expectations, for almost all $t_1 \in (0,\infty)$

and all $t_0 < t_1$,

$$\chi \geq E_{t_0}[f_{t_1,t_1}].$$

Considering the limit as $t_0 \uparrow t_1$, $\chi \geq E_{t_1^-}[f_{t_1,t_1}]$. By the predictability of the characteristics, $f_{t,t}$ is $\mathscr{F}_{t^-}$-measurable, and therefore, for almost all $t \in (0,\infty)$, P-a.s.,

$$f_{t,t}(\omega) \leq \chi.$$

There is a version of (σ,ψ) for which this holds everywhere (such a (σ,ψ) generates indistinguishable characteristics).

C.3 Proof of Lemma 1

Let us suppose we are given a sequence achieving the supremum. For all $n \in \mathbb{N}$, let $q_{t,n}$ be a martingale and τ_n be a stopping time defined on $(\Omega,\mathscr{F}_n,\{\mathscr{F}_{t,n}\},P_n)$, such that $q_{0,n} = \bar{q}_0$ and the constraint (3) is satisfied, and suppose that

$$V(\bar{q}_0) = \lim_{n\to\infty} E^{P_n}[e^{-\rho\tau_n}\hat{u}(q_{n,\tau_n}) - \kappa\int_0^{\tau_n} e^{-\rho s}ds|\mathscr{F}_{0,n}],$$

where $V(\bar{q}_0)$ is the value function of the DM's problem.

Our proof will consider separately the $\rho > 0, \kappa \geq 0$ and $\rho = 0, \kappa > 0$ cases. In what follows, we will adopt the convention that beliefs remain constant after the DM chooses to stop. Both cases will use the following three lemmas.

Lemma 15. *Let $((\Omega,\mathscr{F}_n,\{\mathscr{F}_{t,n}\},P_n),q_n,\tau_n) \in \mathscr{A}$ be a sequence of feasible policies. Then the sequence $\{q_n\}$ is tight.*[52]

Proof. See the technical appendix, section C.4. □

Lemma 16. *Let $((\Omega,\mathscr{F}_n,\{\mathscr{F}_{t,n}\},P_n),q_n,\tau_n) \in \mathscr{A}$ be a sequence of feasible policies, and suppose that the sequence $\{(q_n,y_n)\}$ is tight, where $y_{n,t}(\omega) = \mathbf{1}\{t \leq \tau(\omega)\}$. Then there exists a feasible policy $((\Omega,\mathscr{F}^*,\{\mathscr{F}^*_t\},P^*),q^*,\tau^*) \in \mathscr{A}$ such that a subsequence of (q_n,y_n) converges in law to (q^*,y^*), where $y^*_t(\omega) = \mathbf{1}\{t \leq \tau^*(\omega)\}$.*

Proof. See the technical appendix, section C.5. □

[52] Tightness, in this context, means that the laws of q_n are relatively compact in the weak* topology associated with $\mathbb{D}(\mathscr{P}(X))$ (the space of all càdlàg functions $\mathbb{R}_+ \to \mathscr{P}(X)$, endowed with the Skorokhod topology).

By Lemma C.4, the sequence $\{q_n\}$ is tight. The sequence of step functions $y_{n,t}(\omega) = \mathbf{1}\{t \leq \tau_n(\omega)\}$ is càdlàg and trivially tight, and consequently the sequence $\{(q_n, y_n)\}$ is tight.[53] Consequently, by Lemma 16, a candidate optimal policy $((\Omega, \mathscr{F}^*, \{\mathscr{F}_t^*\}, P^*), q^*, \tau^*) \in \mathscr{A}$ exists; consider only the subsequence that converges to this optimal policy in the sense of Lemma 16.

Define the process $v_n = (q_n, 2|X|^{\frac{1}{2}} y_n) \in \mathbb{D}(\mathbb{R}^{|X|+1})$ and $v^* = (q^*, 2|X|^{\frac{1}{2}} y^*) \in \mathbb{D}(\mathbb{R}^{|X|+1})$. Note that v_n converges in law to v^*. By the boundedness of the simplex, $|\Delta q_{n,t}| < |X|^{\frac{1}{2}}$, and consequently $|\Delta v_{n,t}(\omega)| > |X|^{\frac{1}{2}}$ if and only if $\tau_n(\omega) = t$, and likewise for v^*. It follows that $\tau_n(\omega) = \inf\{t \in \mathbb{R}_+ : |\Delta v_{n,t}(\omega)| > |X|^{\frac{1}{2}}\}$ and likewise for τ^*, and that $|\Delta v_{n,t}(\omega)| \neq |X|^{\frac{1}{2}}$ for all (ω, t), and likewise for v^*.

The mapping from $\alpha \in \mathbb{D}(\mathbb{R}^{|X|+1})$ to $\tau_\alpha = \inf\{t \in \mathbb{R}_+ : |\Delta\alpha(\omega)| > |X|^{\frac{1}{2}}\}$ is continuous in the Skorohod topology, and the mapping from α to $\alpha_{\tau_\alpha} \in \mathbb{R}^{|X|+1}$ is continuous wherever $\tau_\alpha < \infty$.[54] Applying these continuity results for the processes v_n and v^*, we show that the value function is achieved by showing that (τ_n, q_{τ_n}) converges in law to $(\tau^*, q^*_{\tau^*})$, which sufficient to prove the result. We first consider the $\rho > 0$ case, then the $\rho = 0, \kappa > 0$ case.

C.3.1 The $\rho > 0$ case

In this case, the result follows immediately from the convergence in law of v_n to v^* and this continuity. Specifically, by proposition 3.15 of chapter VI of Jacod and Shiryaev [2013], the function

$$g(\tau_n, q_{\tau_n}) = e^{-\rho\tau_n}\hat{u}(q_{\tau_n}) + \kappa\frac{e^{-\rho\tau_n}}{\rho}$$

converges in law to the function $g(\tau^*, q_{\tau^*})$. This function is bounded and continuous, and consequently

$$V(q_0) = \lim_{n\to\infty} E^{P_n}[g(\tau_n, q_{\tau_n}) - \frac{\kappa}{\rho}|\mathscr{F}_{0,n}] = E^{P^*}[g(\tau^*, q^*_{\tau^*}) - \frac{\kappa}{\rho}|\mathscr{F}^*_0],$$

which is the result.

[53] See theorem 3.21 of chapter VI of Jacod and Shiryaev [2013] for the necessary and sufficient conditions for tightness in this context.

[54] See e.g. proposition 2.7 of chapter VI of Jacod and Shiryaev [2013].

C.3.2 The $\rho = 0, \kappa > 0$ case

In this case, the function

$$g(\tau_n, q_{\tau_n}) = \hat{u}(q_{\tau_n}) - \kappa\tau_n$$

is unbounded below, and we must therefore demonstrate that the stopping time is never infinite.

We observe that τ_n is bounded in expectation. Because the sequence of policies achieves the supremum, for any $\varepsilon > 0$, there exists an $n_\varepsilon \in \mathbb{N}$ such that, for all $n \geq n_\varepsilon$,

$$E^{P_n}[\hat{u}(q_{\tau_n,n}) - \kappa\tau_n | \mathscr{F}_{0,n}] > V(q_0) - \varepsilon.$$

Consequently, we must have, for all $n \geq n_\varepsilon$, by $V(q_0) \geq u_{min}$,

$$E^{P_n}[\tau_n | \mathscr{F}_{0,n}] < \frac{u_{max} - u_{min} + \varepsilon}{\kappa}.$$

From this bound, it follows that, for any $T > 0$,

$$\frac{u_{max} - u_{min} + \varepsilon}{\kappa T} > E^{P_n}[\mathbf{1}\{\tau_n > T\} | \mathscr{F}_{0,n}],$$

From this, it follows that the laws of τ_n are tight; by the convergence in law of τ_n to τ^*, we have

$$\frac{u_{max} - u_{min} + \varepsilon}{\kappa T} > E^{P^*}[\mathbf{1}\{\tau^* > T\} | \mathscr{F}_{0,n}].$$

Note as well that $Pr\{\tau_n = \infty\} = Pr\{\tau^* = \infty\} = 0$, and consequently v_{τ_n} converges in law to v^*_τ, and hence that (τ_n, q_{τ_n}) converges in law to (τ^*, q_{τ^*}).

The function g is continuous and bounded above; consequently, by this convergence in law,

$$V(q_0) = \limsup_{n\to\infty} E^{P_n}[g(\tau_n, q_{\tau_n}) | \mathscr{F}_{0,n}] \leq E^{P^*}[g(\tau^*, q_{\tau^*})],$$

which is the result.

C.4 Proof of Lemma 15

To demonstrate tightness, it is sufficient to show that the predictable quadratic variation of $q_{n,t,x}$ (for some $x \in \{1, \ldots, |X|\}$, $< q_{n,x}, q_{n,x} >_t$), is C-tight (tightness as defined for a

continuous process).[55] By the constraint (3) and the strong convexity of D, we must have, for some $K > 0$,

$$\lim_{h \to 0^+} Kh^{-1} E^{P_n}[|q_{n,t^-+h} - q_{n,t^-}|^2 | \mathscr{F}_{t^-,n}] \leq \chi,$$

which implies

$$K^{-1}\chi t - \sum_{x \in X} < q_{n,x}, q_{n,x} >_t$$

is an increasing process. Trivially, the sequence of processes $y_{n,t} = K^{-1}\chi t$ is C-tight, and therefore $\sum_{j=1}^{|X|+1} < q_{n,j}, q_{n,j} >_t$ is C-tight,[56] and consequently $q_{t,n}$ is tight.

C.5 Proof of Lemma 16

By the definition of tightness, the laws of (q_n, y_n) lie in a relatively compact subset of $\mathscr{P}(\mathbb{D}(\mathbb{R}^{|X|+1}))$, where $\mathbb{D}(\mathbb{R}^{|X|+1})$ is endowed with the Skorohod topology and $\mathscr{P}(\mathbb{D}(\mathbb{R}^{|X|+1}))$ with the weak* topology. Consequently, a convergent subsequence of the laws exist. Let $\mathscr{L}^n$ be the law of (q_n, y_n), and let P^* be the limit of a convergent subsequence. In what follows, consider only this convergent subsequence.

Our goal is to construct a feasible policy under which the law of (q^*, y^*) is P^*. This requires constructing the appropriate filtered probability space and then showing that the constraint (3) is satisfied.

We have defined $\Omega = \mathbb{D}(\mathbb{R}^{|X|+1})$; let (q, y) denote the canonical process on this space, let $\mathscr{F}_t^*$ be the natural filtration associated with the canonical process, and let $\mathscr{F}^* = \lim_{t \to \infty} \mathscr{F}_t^*$. In what follows, consider the probability spaces $(\Omega, \mathscr{F}^*, \{\mathscr{F}_t^*\}, \mathscr{L}^n)$ and $(\Omega, \mathscr{F}^*, \{\mathscr{F}_t^*\}, P^*)$. Note that, by construction, the laws of (q_n, y_n) under $(\Omega, \mathscr{F}^n, \{\mathscr{F}_t^n\}, P^n)$ are identical to the laws of the canonical process under $(\Omega, \mathscr{F}^*, \{\mathscr{F}_t^*\}, \mathscr{L}^n)$.

Construct τ^* from the canonical process (q, y) by $\tau^*(\omega) = \inf\{t \in \mathbb{R}_+ : y_t(\omega) \geq 1\}$. The canonical process is adapted to $\{\mathscr{F}_t^*\}$, and consequently τ^*, as constructed, is a stopping time. Identify the process (q^*, y^*) with the canonical process.

Because they satisfy the constraint (3), the processes q_n are quasi-left-continuous.[57] By the continuity of the functions $g_{a,t}(q_n) = \max\{|\Delta q_{n,t}| - a, 0\}$ for any $a > 0$ and $t \in \mathbb{R}_+$,[58]

[55] See theorem 4.13 of chapter VI of Jacod and Shiryaev [2013] for the sufficiency claim, and definition 3.25 of chapter VI of Jacod and Shiryaev [2013] for a definition of C-tightness.

[56] See Proposition 3.35 of chapter VI of Jacod and Shiryaev [2013].

[57] This follows essentially immediately from the fact that $D(q'||q)$ is strictly positive for any $q' \neq q$. For a definition of quasi-left-continuity, see e.g. definition 2.25 of chapter I of Jacod and Shiryaev [2013].

[58] That such functions are continuous in the Skorohod topology is shown in e.g. corollary 2.8 of chapter VI of Jacod and Shiryaev [2013].

and the convergence in law of the q_n to q^*,

$$E^{P^*}[\max\{|\Delta q_t| - a, 0\}] = 0$$

for all $t \in \mathbb{R}_+$ and $a > 0$, from which it follows that q is also quasi-left-continuous under P^*.

Because the processes q_n satisfies the constraint (3) under $(\Omega, \mathscr{F}^n, \{\mathscr{F}_t^n\}, P^n)$, and (q_t^n, y_t^n) is $\mathscr{F}_t^n$-measurable, the canonical process q satisfies (3) under $(\Omega, \mathscr{F}^*, \{\mathscr{F}_t^*\}, \mathscr{L}^n)$ (i.e. the constraint can be conditioned down from $\mathscr{F}_t^n$ to the natural filtration of (q_t^n, y_t^n)).

By Lemma 7, the process q under $\mathscr{L}^n$ is characterized by the functions $\sigma_{n,t}$ and $\psi_{n,t}$ described in that lemma. Define the function

$$f_{n,t,s}(\omega) = \frac{1}{2} tr[\sigma_{n,t}(\omega)\sigma_{n,t}^T(\omega)\nabla_1^2 D(q_{n,t^-}(\omega)||q_{n,s^-}(\omega))] +$$
$$\int_{\mathbb{R}^{|X|}\setminus\{\vec{0}\}} (D(q_{n,t^-}(\omega)+z||q_{n,s^-}(\omega)) - D(q_{n,t^-}(\omega)||q_{n,s^-}(\omega)) - z^T \cdot \nabla_1 D(q_{n,t^-}(\omega)||q_{n,s^-}(\omega)))\psi_{n,t}(dz;\omega)$$

and the special case

$$f_{n,t,t}(\omega) = \frac{1}{2} tr[\sigma_{n,t}(\omega)\sigma_{n,t}^T(\omega)\bar{k}(q_{n,t^-}(\omega))]$$
$$+ \int_{\mathbb{R}^{|X|}\setminus\{\vec{0}\}} D(q_{n,t^-}(\omega)+z||q_{n,t^-}(\omega))\psi_{n,t}(dz;\omega).$$

As shown in Lemma 7 (by Ito's lemma), for any $s \in \mathbb{R}_+$, $n \in \mathbb{N}$, and $h > 0$,

$$E_{s^-}^{\mathscr{L}^n}[D(q_{s+h}||q_{s^-})] = E_{s^-}^{\mathscr{L}^n}[\int_s^{s+h} f_{n,t,s}dt],$$

and $f_{n,t,t} \leq \chi$.

Claim. $f_{n,t,s}$ is uniformly bounded $P^* - a.s.$ and satisfies, $P^* - a.s.$, for any $c \in [0,1]$

$$\lim \sup_{m\to\infty, \omega'\to\omega} f_{\hat{n}_m, s+ch_m, s}(\omega') \leq \chi.$$

Proof. See below. □

Using this claim, we prove by contradiction that (3) must hold. Suppose not: for some $s \in \mathbb{R}_+$, $\lim\sup_{h\downarrow 0} \frac{1}{h} E_{s^-}^{P^*}[D(q_{s+h}||q_{s^-})] > \chi$. In this case, by definition there must exist some

$\varepsilon > 0$ and sequence of strictly positive $h_m \to 0$ such that

$$\lim_{m\to\infty} \frac{1}{h_m} E_{s^-}^{P^*}[D(q_{s+h_m}||q_{s^-})] \geq \chi + 2\varepsilon.$$

By the convergence of $\mathscr{L}^n$ to P^* in the weak* topology, and the continuity and boundedness of D,[59] for each h_m, there exists an n_m such that, for all $n \geq n_m$,

$$|E_{s^-}^{P^*}[D(q_{s+h_m}||q_{s^-})] - E_{s^-}^{\mathscr{L}^n}[D(q_{s+h_m}||q_{s^-})]| < \varepsilon h_m.$$

Define $\hat{n}_m = \max\{n_m, m\}$. We conclude that if $\limsup_{h\downarrow 0} \frac{1}{h} E_{s^-}^{P^*}[D(q_{s+h}||q_{s^-})] > \chi$, we must have

$$\lim_{m\to\infty} \frac{1}{h_m} E_{s^-}^{\mathscr{L}^{\hat{n}_m}}[D(q_{s+h_m}||q_{s^-})] \geq \chi + \varepsilon.$$

Using the claim above, it follows by the reverse Fatou's lemma with weakly converging measures (Feinberg et al. [2014]) that

$$\lim_{m\to\infty} E_{s^-}^{\mathscr{L}^{\hat{n}_m}}[\int_0^1 f_{\hat{n}_m, s+ch_m, s} dc] \leq \chi.$$

Observing by a change of variable that

$$\int_0^1 f_{\hat{n}_m, s+ch_m, s}(\omega) dc = \frac{1}{h_m} \int_s^{s+h_m} f_{\hat{n}_m, t, s}(\omega) dt,$$

it follows that

$$\lim_{m\to\infty} \frac{1}{h_m} E_{s^-}^{\mathscr{L}^{\hat{n}_m}}[D(q_{s+h_m}||q_{s^-})] \leq \chi,$$

and consequently that (3) must hold.

To conclude the proof, we prove the claim by showing that $f_{n,t,s}$ is uniformly bounded ($P^* - a.s.$) and that the limit condition is satisfied.

For any $q \in \mathscr{P}(X)$, let $A(q)$ be the set of $v \in \mathbb{R}^{|X|}$ and $a \in (0, |X|^{\frac{1}{2}}]$ such that $q + av \in \mathscr{P}(X)$ and $q + av \ll q$.

Define the function

$$F(a, v; q_0, q_1) = \frac{D(q_1 + av||q_0) - D(q_1||q_0) - av^T \cdot \nabla_1 D(q_1||q_0)}{D(q_1 + av||q_1)}$$

[59] Note that $q_{s+h_m} \ll q_{s^-}$ for all $s \in \mathbb{R}_+$ and $h_m \geq 0$.

for $(a,v) \in A(q_1)$ and $q_1 \ll q_0$. Extend the definition of this function to $a = 0$ by continuity:

$$F(0,v;q_0,q_1) = \lim_{a \downarrow 0} F(a,v;q_0,q_1) = \frac{v^T \cdot \nabla_1^2 D(q_1||q_0) \cdot v}{v^T \cdot \bar{k}(q_1) \cdot v}.$$

Note that the closure of $A(q)$, $\bar{A}(q)$, is a compact set, by the compactness of the simplex.

Moreover, by the twice-continuous differentiability of D on the simplex and its faces, $F(\cdot)$ is continuous on the set $(q_0,q_1,a,v) \in \mathscr{P}(X) \times \mathscr{P}(X) \times [0,|X|^{\frac{1}{2}}] \times \mathbb{R}^{|X|}$ such that $q_1 \ll q_0$ and $(a,v) \in \bar{A}(q_1)$. This set is compact, and consequently $F(\cdot)$ is uniformly bounded.

Fix any $s > 0$. By the quasi-left-continuity of q^*, $q_s(\omega) = q_{s^-}(\omega)$ holds P^*-a.s. In what follows, consider any $\omega \in \Omega$ such that this holds. By the right-continuity with left-limits property of $q(\omega)$ and the assumption that $q_{s^-}(\omega) = q_s(\omega)$, there exists an $h_\omega > 0$ such that $q(\omega)$ is continuous on $t \in [s - h_\omega, s + h_\omega]$. Because $q(\omega)$ is continuous on this interval, if $q(\omega') \to q(\omega)$ in the Skorohod topology, $q(\omega')$ converges to $q(\omega)$ uniformly on this interval.[60] As a consequence of this convergence and the continuity of F, for any $(a,v) \in A(q_{n,s^-}(\omega))$,

$$\lim_{m \to \infty, \omega' \to \omega} F(a,v;q_{s^-}(\omega'), q_{(s+h_m)^-}(\omega')) = 1. \tag{45}$$

Using this result, we will show that the claim holds. To do this, we consider the characteristics (σ_n, ψ_n) defined from q_n via Lemma 7.

The characteristics $\sigma_{\hat{n}_m}$ satisfy

$$\frac{1}{2} tr[\sigma_{\hat{n}_m,s+h_m}(\omega') \sigma^T_{\hat{n}_m,s+h_m}(\omega') \bar{k}(q_{(s+h_m)^-}(\omega'))] \leq \chi.$$

By the strong convexity of D (implying $\bar{k} \succeq KI$, where I is the identity matrix), the sequence $\sigma_{\hat{n}_m,s+h_m}(\omega') \sigma^T_{\hat{n}_m,s+h_m}(\omega')$ is uniformly bounded in the matrix norm. It therefore has a convergent subsequence; let l index this subsequence, and define $\Sigma^*_s(\omega)$ as its limit.

Observe, by the twice continuous differentiability of D, that

$$\lim_{l \to \infty, \omega' \to \omega} \frac{1}{2} tr[\sigma_{\hat{n}_l,s+h_l}(\omega') \sigma^T_{\hat{n}_l,s+h_l}(\omega') \nabla_1^2 D(q_{(s+h_l)^-}(\omega')||q_{s^-}(\omega'))] = \frac{1}{2} tr[\Sigma^*_s(\omega) \bar{k}(q_{s^-}(\omega))].$$

[60] See e.g. proposition 1.17 of chapter VI of Jacod and Shiryaev [2013].

Now consider the measures on $\mathbb{R}^{|X|}$ defined by $v_{n,t}(dz;\omega) = D(q_{n,t^-}(\omega)+z||q_{n,t^-}(\omega))\psi_{n,t}(dz;\omega)$, observing that, by assumption,

$$\int_{\mathbb{R}^{|X|}} v_{n,t}(dz;\omega) \leq \chi.$$

Let $\mu_{n,t}(da,dv;\omega)$ be the measure on $\bar{A}(q_{t^-}(\omega)) \subset \mathbb{R}^{|X|+1}$ induced from $v_{n,t}$ by $z \to (|z|, \frac{z}{|z|})$. The measures μ_t are tight, and consequently there is a convergent subsequence of h_l such that $\mu_{n,s+h_l}(\cdot;\omega)$ converges in the weak* topology to some $\mu^*_s(\cdot;\omega)$. Pass to this subsequence, which we continue to index by l.

By the definition of $F(\cdot)$,

$$\int_{\mathbb{R}^{|X|}\setminus\{\vec{0}\}} \{D(q_{(s+h_l)^-}(\omega)+z||q_{s^-}(\omega)) - D(q_{(s+h_l)^-}(\omega)||q_{s^-}(\omega)) - z^T\cdot\nabla_1 D(q_{(s+h_l)^-}(\omega)||q_{s^-}(\omega)))\}\times$$
$$\psi_{\hat{n}_l,s+h_l}(dz;\omega) =$$
$$\int_{\mathbb{R}^{|X|+1}} F(a,v;q_{s^-}(\omega),q_{(s+h_l)^-}(\omega))\mu_{\hat{n}_l,s+h_l}(da,dv;\omega).$$

By the uniform boundedness of F and the reverse Fatou's lemma with weakly converging measures (Feinberg et al. [2014]), using (45),

$$\lim\sup_{l\to\infty,\omega'\to\omega} \int_{\mathbb{R}^{|X|+1}} F(a,v;q_{s^-}(\omega'),q_{(s+h_l)^-}(\omega'))\mu_{\hat{n}_l,s+h_l}(da,dv;\omega') \leq$$
$$\int_{\mathbb{R}^{|X|+1}} \mu^*_s(da,dv;\omega).$$

By Lemma 7,

$$\frac{1}{2}tr[\sigma_{\hat{n}_l,s+h_l}(\omega')\sigma^T_{\hat{n}_l,s+h_l}(\omega')\bar{k}(q_{(s+h_l)^-}(\omega'))] +$$
$$\int_{\mathbb{R}^{|X|+1}} \mu_{\hat{n}_l,s+h_l}(da,dv;\omega') \leq \chi$$

for all l, and consequently taking limits yields

$$\frac{1}{2}tr[\Sigma^*_s(\omega)\bar{k}(q_{s^-}(\omega))] + \int_{\mathbb{R}^{|X|+1}} \mu^*_s(da,dv;\omega) \leq \chi.$$

It follows immediately that

$$\lim \sup_{l\to\infty,\omega'\to\omega} f_{\hat{n}_l,s+h_l,s}(\omega') \leq \chi.$$

Since this must hold for any convergent subsequence, it follows that

$$\lim \sup_{m\to\infty,\omega'\to\omega} f_{\hat{n}_m,s+h_m,s}(\omega') \leq \chi.$$

By the uniform boundedness of F (call the uniform upper bound $\bar{F}$)

$$\begin{aligned} f_{n,t,s}(\omega) \leq \bar{F} \int_{\mathbb{R}^{|X|+1}} \mu_{n,t}(da,dv;\omega) \\ + \bar{F}\frac{1}{2} tr[\sigma_{n,t}(\omega)\sigma_{n,t}^T(\omega)\bar{k}(q_{t^-}(\omega))], \end{aligned}$$

which yields $f_{n,t,s}(\omega) \leq \bar{F}\chi$, proving the result.

C.6 Proof of Lemma 13

This structure of this proof is similar to the proof of the existence of optimal policies (Lemma 1), and will refer to lemmas used in that proof (see the technical appendix, section C.3).

We will construct a limit policy, $((\Omega,\mathscr{F}^*,\{\mathscr{F}_t^*\},P^*),q^*,\tau^*) \in \mathscr{A}$, from a subsequence of the optimal policies associated with each value of $\rho_n > 0$, and then show that this limit policy achieves the value function when $\rho = 0$. We begin by showing that the value function $V^*(q_0)$ associated with this limit policy when $\rho = 0$ satisfies, for this given subsequence,

$$\lim \sup_{n\to\infty} V(q_0;\rho_n) \leq V^*(q_0),$$

where $V(q_0;\rho_n)$ denotes the value function under the optimal policy associated with ρ_n.

By Lemma C.4, the sequence $\{q_n\}$ is tight. The sequence of step functions $y_{n,t}(\omega) = \mathbf{1}\{t \leq \tau_n(\omega)\}$ is càdlàg and trivially tight, and consequently the sequence $\{(q_n,y_n)\}$ is tight.[61] Consequently, by Lemma 16, a candidate limit policy $(\Omega,\mathscr{F}^*,\{\mathscr{F}_t^*\},P^*),q^*,\tau^*) \in \mathscr{A}$ exists; consider only the subsequence that converges to this limit policy in the sense of

[61] See theorem 3.21 of chapter VI of Jacod and Shiryaev [2013] for the necessary and sufficient conditions for tightness in this context.

Lemma 16.

Define the process $v_n = (q_n, 2|X|^{\frac{1}{2}} y_n) \in \mathbb{D}(\mathbb{R}^{|X|+1})$ and $v^* = (q^*, 2|X|^{\frac{1}{2}} y^*) \in \mathbb{D}(\mathbb{R}^{|X|+1})$. Note that v_n converges in law to v^*. By the boundedness of the simplex, $|\Delta q_{n,t}| < |X|^{\frac{1}{2}}$, and consequently $|\Delta v_{n,t}(\omega)| > |X|^{\frac{1}{2}}$ if and only if $\tau_n(\omega) = t$, and likewise for v^*. It follows that $\tau_n(\omega) = \inf\{t \in \mathbb{R}_+ : |\Delta v_{n,t}(\omega)| > |X|^{\frac{1}{2}}\}$ and likewise for τ^*, and that $|\Delta v_{n,t}(\omega)| \neq |X|^{\frac{1}{2}}$ for all (ω, t), and likewise for v^*.

The mapping from $\alpha \in \mathbb{D}(\mathbb{R}^{|X|+1})$ to $\tau_\alpha = \inf\{t \in \mathbb{R}_+ : |\Delta\alpha(\omega)| > |X|^{\frac{1}{2}}\}$ is continuous in the Skorohod topology, and the mapping from α to $\alpha_{\tau_\alpha} \in \mathbb{R}^{|X|+1}$ is continuous wherever $\tau_\alpha < \infty$.[62] Applying these continuity results for the processes v_n and v^*, we show that the value function $V^*(q_0)$ is achieved by showing that (τ_n, q_{τ_n}) converges in law to $(\tau^*, q^*_{\tau^*})$.

Let us first observe that, under an optimal policy, $Pr\{\tau_n = \infty\} = 0$. Suppose not, and there exists some $\varepsilon > 0$ such that $Pr\{\tau_n = \infty\} = \varepsilon$. Then there exists some time T_n such that $Pr\{\tau_n \geq T_n\} \leq \varepsilon(1 + \frac{u_{min}}{u_{max}})$, and consequently

$$E^{P_n}[e^{-\rho(\tau_n - T_n)}\hat{u}(q_{\tau_n}) - \kappa \int_{T_n}^{\tau_n} e^{-\rho_n s} ds | \mathscr{F}_{T_n,n}, \tau_n > T_n] \leq \varepsilon \frac{u_{min}}{u_{max}} u_{max} - \varepsilon \frac{\kappa}{\rho_n} < u_{min},$$

contradicting $V(q;\rho_n) \geq u_{min}$. It follows that $Pr\{\tau_n = \infty\} = 0$.

Now observe that we must have

$$E^{P_n}[e^{-\rho\tau_n}\hat{u}(q_{\tau_n}) - \kappa \int_0^{\tau_n} e^{-\rho_n s} ds | \mathscr{F}_{0,n}] \geq u_{min},$$

and consequently

$$\frac{u_{max} - u_{min}}{\kappa} \geq \frac{1}{\rho_n} E^{P_n}[1 - e^{-\rho_n \tau_n} | \mathscr{F}_{0,n}].$$

By Markov's inequality,

$$Pr\{\tau_n \geq T\} \leq \frac{\frac{1}{\rho_n} E^{P_n}[1 - e^{-\rho_n \tau_n} | \mathscr{F}_{0,n}]}{\frac{1}{\rho_n}(1 - e^{-\rho_n T})} \leq \frac{\frac{u_{max} - u_{min}}{\kappa}}{\frac{1}{\rho_n}(1 - e^{-\rho_n T})}.$$

By the weak convergence of τ_n to τ^*,

$$\liminf_{n \to \infty} Pr\{\tau_n > T\} \geq Pr\{\tau^* > T\},$$

[62] See e.g. proposition 2.7 of chapter VI of Jacod and Shiryaev [2013].

from which it follows that

$$Pr\{\tau^* > T\} \leq \frac{u_{max} - u_{min}}{\kappa T}.$$

Therefore, $Pr\{\tau^* = \infty\} = 0$. It follows by the aforementioned results that (τ_n, q_{τ_n}) converges in law to $(\tau^*, q^*_{\tau^*})$.

The function

$$g(\tau_n, q_{\tau_n}) = \hat{u}(q_n) - \kappa\tau_n$$

is continuous and bounded above. Consequently, by this convergence in law,

$$\limsup_{n\to\infty} V(q_0;\rho_n) = \limsup_{n\to\infty} E^{P_n}[g(\tau_n, q_{\tau_n})|\mathscr{F}_{0,n}] \leq E^{P^*}[g(\tau^*, q_{\tau^*})] = V^*(q_0).$$

By Lemma (1), an optimal policy exists for $\rho = 0$. Let $(\Omega, \mathscr{F}^+, \{\mathscr{F}^+_t\}, P^+), q^+, \tau^+) \in \mathscr{A}$ be such an optimal policy. This policy feasible when $\rho > 0$; consequently, we must have

$$\begin{aligned} V(q_0;\rho_n) &\geq E^{P^+}[e^{-\rho_n\tau^+}\hat{u}(q_{\tau^+}) - \kappa\int_0^{\tau^+} e^{-\rho_n s}ds|\mathscr{F}^+_0] \\ &\geq E^{P^+}[(1-\rho_n\tau^+)\hat{u}(q_{\tau^+}) - \kappa\tau^+|\mathscr{F}^+_0]. \end{aligned}$$

Observing that

$$\begin{aligned} E^{P^+}[\tau^+\hat{u}(q_{\tau^*})|\mathscr{F}^+_0] &\leq \frac{u_{max}}{\kappa}(E^{P^+}[\hat{u}(q_{\tau^+})|\mathscr{F}^+_0] - V(q_0;0)) \\ &\leq \frac{u_{max}}{\kappa}(u_{max} - u_{min}), \end{aligned}$$

we have

$$V(q_0;\rho_n) \geq V(q_0;0) - \rho_n\frac{u_{max}}{\kappa}(u_{max} - u_{min}).$$

It follows that

$$V^*(q_0) = \limsup_{n\to\infty} V(q_0;\rho_n) \geq V(q_0;0),$$

and consequently the candidate limit policy is an optimal policy.

C.7 Proof of Lemma 8

Let q_t be any point on the interior of $\mathscr{P}(X)$, and let $B_\delta = \{q' \in \mathscr{P}(X) : |q' - q_t| \leq \delta\}$ a $\delta > 0$ ball around q_t. Choose some $\bar{\delta} > 0$ such that $B_{3\bar{\delta}}$ is contained in interior of the

simplex. We will prove that V is Lipschitz-continuous on $B_{\bar{\delta}}$.

Choose $\bar{z} \in \mathbb{R}^{|X|} \setminus \{\vec{0}\}$ such that $|\bar{z}| \leq \bar{\delta}$, and apply Lemma 6, defining $z = \frac{1}{\alpha}\bar{z}$ and $\varepsilon = \frac{1-\alpha}{\alpha}$,

$$\chi^{-1}(\rho V(q_t) + \kappa)(\varepsilon^{-1} D(q_t + \varepsilon\bar{z}||q_t) + D(q_t - \bar{z}||q_t)) \geq \\ \varepsilon^{-1}(V(q_t + \varepsilon\bar{z}) - V(q_t)) + V(q_t - \bar{z}) - V(q_t). \tag{46}$$

for all $\varepsilon \in (0,1)$.

Define $\bar{u} = \max_{a \in A, x \in X} u_{a,x}$ and note that $0 < V(q) \leq \bar{u}$ for all q. Note also that D is (twice) continuously-differentiable in its first argument and $D(q||q) = 0$. Taking limits,

$$\limsup_{\varepsilon \to 0^+} \varepsilon^{-1}(V(q_t + \varepsilon\bar{z}) - V(q_t)) \leq \bar{u} + \chi^{-1}(\rho\bar{u} + \kappa) D(q_t - \bar{z}||q_t).$$

Now apply Lemma 6 at $q = q_t + \varepsilon\bar{z}$, defining $z = -\frac{1}{\alpha}\bar{z}$ and $\varepsilon = \frac{1-\alpha}{\alpha}$,

$$\chi^{-1}(\kappa + \rho V(q_t + \varepsilon\bar{z}))(\varepsilon^{-1} D(q_t||q_t + \varepsilon\bar{z}) + D(q_t + (1+\varepsilon)\bar{z}||q_t + \varepsilon\bar{z})) \geq \\ \varepsilon^{-1}(V(q_t) - V(q_t + \varepsilon\bar{z})) + V(q_t + (1+\varepsilon)\bar{z}) - V(q_t + \varepsilon\bar{z}),$$

for all $\varepsilon \in (0,1)$. By the convexity of D,

$$\varepsilon^{-1} D(q_t||q_t + \varepsilon\bar{z}) + \bar{z} \cdot \nabla_1 D(q_t||q_t + \varepsilon\bar{z}) \leq \varepsilon^{-1} D(q_t + \varepsilon\bar{z}||q_t + \varepsilon\bar{z}),$$

where ∇_1 denotes the gradient with respect to the first argument, and the inequality can be written as

$$\chi^{-1}(\kappa + \rho V(q_t - \varepsilon\bar{z}))(D(q_t + (1+\varepsilon)\bar{z}||q_t + \varepsilon\bar{z}) - \bar{z} \cdot \nabla_1 D(q_t||q_t + \varepsilon\bar{z})) \geq \\ \varepsilon^{-1}(V(q_t) - V(q_t + \varepsilon\bar{z})) + V(q_t + (1+\varepsilon)\bar{z}) - V(q_t + \varepsilon\bar{z}),$$

By the continuity of the gradient and the arguments above,

$$\liminf_{\varepsilon \to 0^+} \varepsilon^{-1}(V(q_t + \varepsilon\bar{z}) - V(q_t)) \geq -\bar{u} - \chi^{-1}(\rho\bar{u} + \kappa) D(q_t + \bar{z}||q_t).$$

Define

$$K = \max_{q' \in B_{\bar{\delta}}} \bar{u} + \chi^{-1}(\rho\bar{u} + \kappa) D(q'||q_t),$$

noting that D is finite on the interior of the simplex and hence by the compactness of $B_{\bar{\delta}}$, a finite maximum exists. We conclude that the Dini derivatives in the direction $\bar{z}$ are bounded by K. It follows (see, e.g., Royden and Fitzpatrick [2010] section 6.2) that V is locally Lipschitz continuous on $B_{\bar{\delta}}$.

Repeating the argument for each face of the simplex, using balls defined only the support of q_t, extends the result to all non-extreme points of the simplex.

C.8 Proof of Lemma 9

Note: this proof refers heavily to results from Clarke [1990].

By Lemma 8, V is locally Lipschitz on the interior of the simplex and on the interior of each face.

Let q_t be any point on the interior of $\mathscr{P}(X)$, and let $B_\delta = \{q' \in \mathscr{P}(X) : |q' - q_t| \leq \delta\}$ a $\delta > 0$ ball around q_t. Choose some $\bar{\delta} > 0$ such that $B_{4\bar{\delta}}$ is contained in interior of the simplex. We will prove that V is continuously differentiable on $B_{\bar{\delta}}$.

Choose $\bar{z} \in \mathbb{R}^{|X|} \setminus \{\vec{0}\}$ such that $|\bar{z}| \leq \bar{\delta}$, and apply Lemma 6, defining $z = \frac{\nu}{\alpha}\bar{z}$ and $\varepsilon = \frac{1-\alpha}{\alpha}$, to $q = q_t + \hat{z}$ for some $\hat{z} \in \mathbb{R}^{|X|}$ such that $|\hat{z}| < \bar{\delta}$,

$$\chi^{-1}(\rho V(q_t) + \kappa)(\varepsilon^{-1} D(q_t + \hat{z} + \varepsilon\bar{z} || q_t + \hat{z}) + D(q_t + \hat{z} - \bar{z} || q_t + \hat{z})) \geq$$
$$\varepsilon^{-1}(V(q_t + \hat{z} + \varepsilon\bar{z}) - V(q_t + \hat{z})) + V(q_t + \hat{z} - \bar{z}) - V(q_t + \hat{z}).$$

for all $\varepsilon \in (0,1)$ (which ensures that $q_t + \hat{z} + \varepsilon\bar{z} \in B_{3\bar{\delta}}$).

By the convexity of D,

$$\varepsilon^{-1} D(q_t + \hat{z} + \varepsilon\bar{z} || q_t + \hat{z}) - \bar{z} \cdot \nabla_1 D(q_t + \hat{z} + \varepsilon\bar{z} || q_t + \hat{z}) \leq \varepsilon^{-1} D(q_t + \hat{z} || q_t + \hat{z}),$$

where ∇_1 denotes the gradient with respect to the first argument, and the inequality can be written as

$$\chi^{-1}(\rho V(q_t) + \kappa)(\bar{z} \cdot \nabla_1 D(q_t + \hat{z} + \varepsilon\bar{z} || q_t + \hat{z}) + D(q_t + \hat{z} - \bar{z} || q_t + \hat{z})) \geq$$
$$\varepsilon^{-1}(V(q_t + \hat{z} + \varepsilon\bar{z}) - V(q_t + \hat{z})) + V(q_t + \hat{z} - \bar{z}) - V(q_t + \hat{z}).$$

Considering the limits

$$\lim_{\nu\to 0^+} \sup_{\hat{z}\in\mathbb{R}^{|X|}:|\hat{z}|<\nu,\varepsilon\in(0,\nu)} \chi^{-1}(\rho V(q_t)+\kappa)(\bar{z}\cdot\nabla_1 D(q_t+\hat{z}+\varepsilon\bar{z}||q_t+\hat{z})+$$

$$D(q_t+\hat{z}-\bar{z}||q_t+\hat{z}))-\varepsilon^{-1}(V(q_t+\hat{z}+\varepsilon\bar{z})-V(q_t+\hat{z}))-V(q_t+\hat{z}-\bar{z})+V(q_t+\hat{z})\geq 0,$$

we have

$$\chi^{-1}(\rho V(q_t)+\kappa)D(q_t-\bar{z}||q_t)\geq V(q_t-\bar{z})-V(q_t)+V^\circ(q_t;\bar{z}),$$

where

$$V^\circ(q_t;\bar{z})=\lim_{\nu\to 0^+}\sup_{\hat{z}\in\mathbb{R}^{|X|}:|\hat{z}|<\nu,\varepsilon\in(0,\nu)}\varepsilon^{-1}(V(q_t+\hat{z}+\varepsilon\bar{z})-V(q_t+\hat{z}))$$

is the Clarke generalized derivative in the direction $\bar{z}$, which exists by proposition 2.1.1 of Clarke [1990] and the local Lipschitz property.

By proposition 2.1.2 of Clarke [1990], a generalized gradient exists; let $x(q)\in\partial V(q)\subseteq\mathbb{R}^{|X|}$ denote a selection of such gradients with the property that

$$|x(q)-x(q_t)|\leq K|q-q_t|$$

for some $K>0$ and all $q\in B_{\bar{\delta}}$, which is possible by proposition 2.1.5 of Clarke [1990]. By proposition 2.1.2 of Clarke [1990],

$$V^\circ(q_t;\bar{z})\geq\bar{z}^T\cdot x(q_t),$$

and therefore

$$\chi^{-1}(\rho V(q_t)+\kappa)D(q_t-\bar{z}||q_t)\geq V(q_t-\bar{z})-V(q_t)+\bar{z}^T\cdot x(q_t).$$

Apply this equation in the opposite direction of $\bar{z}$, scaled by some $\varepsilon\in(0,1)$, for some point $q_t+\hat{z}$, again for some $\hat{z}\in\mathbb{R}^{|X|}$ such that $|\hat{z}|<\bar{\delta}$. We have

$$\chi^{-1}(\rho V(q_t)+\kappa)\varepsilon^{-1}D(q_t+\hat{z}+\varepsilon\bar{z}||q_t+\hat{z})+\bar{z}^T x(q_t+\hat{z})\geq\varepsilon^{-1}(V(q_t+\hat{z}+\varepsilon\bar{z})-V(q_t+\hat{z})),$$

and by the convexity of D as above,

$$\chi^{-1}(\rho V(q_t)+\kappa)\bar{z}^T\cdot\nabla_1 D(q_t+\hat{z}+\varepsilon\bar{z}||q_t+\hat{z})+\bar{z}^T x(q_t+\hat{z})\geq\varepsilon^{-1}(V(q_t+\hat{z}+\varepsilon\bar{z})-V(q_t+\hat{z}))$$

It follows, taking the limit superior as above, that

$$\bar{z}^T x(q_t) \geq V^\circ(q_t;\bar{z}).$$

This can only hold if $V^\circ(q_t;\bar{z}) = \bar{z}^T x(q_t)$, and as this must hold for all $\bar{z}$, $\partial V(q_t)$ is a singleton. Applying this argument to all $q \in B_{\bar{\delta}}$, it follows by proposition 2.2.4 of Clarke [1990] and the unnumbered corollary following that proposition that V is continuously differentiable on $B_{\bar{\delta}}$. Repeating this argument for all q_t on the interior of the simplex, it follows that V is continuously differentiable on the interior of the simplex. By identical arguments, V is continuously differentiable on each face of the simplex.

C.9 Proof of Lemma 10

Proof by contradiction: suppose

$$\sup_{\sigma_0,\psi_0 \in \mathscr{A}(q_0)} \frac{1}{2} tr[\sigma_0 \sigma_0^T \nabla^2 \phi(q_0)] + \int_{\mathbb{R}^{|X|}\setminus\{0\}} (\phi(q_0+z) - \phi(q_0) - z^T \cdot \nabla\phi(q_0)) \psi_0(dz) < \rho\phi(q_0) + \kappa$$

and $V(q_0) > \hat{u}(q_0)$.

Step 1: Prove this inequality must hold in some neighborhood around q_0. We must have, for some $\varepsilon > 0$,

$$\sup_{\sigma_0,\psi_0 \in \mathscr{A}(q_0)} \frac{1}{2} tr[\sigma_0 \sigma_0^T \nabla^2 \phi(q_0)] + \int_{\mathbb{R}^{|X|}\setminus\{0\}} (\phi(q_0+z) - \phi(q_0) - z^T \cdot \nabla\phi(q_0)) \psi_0(dz) \leq \rho\phi(q_0) + \kappa - \varepsilon$$

and

$$V(q_0) \geq \hat{u}(q_0) + \varepsilon.$$

Consider diffusion-only policies of the form

$$\sigma_0 \sigma_0^T = \frac{vv^T}{\chi^{-1} \frac{1}{2} v^T \bar{k}(q_0) v}$$

for some vector $v \in \mathbb{R}^{|X|}$ with $|v| = 1$. We must have

$$\max_{v \in \mathbb{R}^{|X|}: |v|=1} \frac{v^T \nabla^2 \phi(q_0) v}{v^T \bar{k}(q) v} \leq \chi^{-1}(\rho \phi(q_0) + \kappa - \varepsilon).$$

Now consider policies without diffusion and for which ψ_0 is a point mass on av, where $v \in \mathbb{R}^{|X|}$ with $|v| = 1$ and $a \in (0, |X|^{\frac{1}{2}}]$. Note that $|q' - q_0| \leq |X|^{\frac{1}{2}}$ for any $q' \in \mathscr{P}(X)$. For such policies,

$$\sup_{a,v \in (0,|X|^{\frac{1}{2}}] \times \mathbb{R}^{|X|}: |v|=1 \,\&\, q_0 + av \in \mathscr{P}(X)} F(q_0, a, v) \leq \chi^{-1}(\rho \phi(q_0) + \kappa - \varepsilon)$$

where

$$F(q_0, a, v) = \frac{\phi(q_0 + av) - \phi(q_0) - av \cdot \nabla \phi(q_0)}{D(q_0 + av || q_0)}.$$

Define

$$F(q_0, 0, v) = \lim_{a \to 0^+} F(q_0, a, v) = \frac{v^T \nabla^2 \phi(q_0) v}{v^T \bar{k}(q) v}$$

to combine these two conditions, which yields

$$\max_{a,v \in [0,|X|^{\frac{1}{2}}] \times \mathbb{R}^{|X|}: |v|=1 \,\&\, q_0 + av \in \mathscr{P}(X)} F(q_0, a, v) \leq \chi^{-1}(\rho \phi(q_0) + \kappa - \varepsilon).$$

Now observe that $F(q_0, a, v)$ is continuous in its arguments, and that the correspondence

$$\Gamma(q_0) = \{a, v \in [0, |X|^{\frac{1}{2}}] \times \mathbb{R}^{|X|} : |v| = 1 \,\&\, q_0 + av \in \mathscr{P}(X)\}$$

is a closed and bounded subset of $\mathbb{R}^{|X|+1}$ (and hence compact-valued), and is upper hemi-continuous.

It follows by the theorem of the maximum that

$$F^*(q_0) = \max_{a,v \in [0,|X|^{\frac{1}{2}}] \times \mathbb{R}^{|X|}: |v|=1 \,\&\, q_0 + av \in \mathscr{P}(X)} F(q_0, a, v)$$

is continuous in q_0.

Hence, there exists some $\delta > 0$ such that for all $q \in \mathscr{P}(X)$ with $|q - q_0| < \delta$,

$$F^*(q) \leq \chi^{-1}(\rho \phi(q_0) + \kappa - \frac{\varepsilon}{2}).$$

It follows that for all such q and all $(\sigma_0, \psi_0) \in \mathscr{A}(q)$,

$$\frac{1}{2}tr[\sigma_0\sigma_0^T\nabla^2\phi(q)]+$$
$$\int_{\mathbb{R}^{|X|}\setminus\{0\}}(\phi(q+z)-\phi(q)-z^T\cdot\nabla\phi(q))\psi_0(dz)\leq$$
$$-\chi^{-1}(\rho\phi(q)+\kappa-\frac{\varepsilon}{2})(\frac{1}{2}tr[\sigma_0\sigma_0^T\bar{k}(q)]+\int_{\mathbb{R}^{|X|}\setminus\{0\}}D(q+z||q)\psi_0(dz))\leq\rho\phi(q)+\kappa-\frac{\varepsilon}{2}.$$

By the continuity of V and $\hat{u}$, there exists a $\delta_2>0$ such that for all $|q-q_0|<\delta_2$,

$$V(q)-\hat{u}(q)\geq\frac{\varepsilon}{2}.$$

Consequently, for $|q-q_0|<\min\{\delta,\delta_2\}$, both inequalities hold.

Step 2: Apply Ito's Lemma Suppose the DM initially holds beliefs $q_t=q_0$. Let $\tau_h=\min\{\{\inf_{s\in[t,t+h]}:|q_s-q_0|\geq\min\{\delta,\delta_2\}\},h\}$, which is to say the stopping time associated with $h>0$ units of time passing or exiting the region just described, whichever comes first. Note that this region lies within the continuation region under the optimal policy, by construction.

Under the optimal policy (by Lemma 5),

$$V(q_t)=E_t[e^{-\rho(\tau_h-t)}V(q_{\tau_h})-\kappa\int_t^{\tau_h}e^{-\rho(s-t)}ds],$$

and therefore by $\phi(q)\geq V(q)$ and $\phi(q_0)=V(q_0)$,

$$\phi(q_0)\leq E_t[e^{-\rho(\tau_h-t)}\phi(q_{\tau_h})-\kappa\int_t^{\tau_h}e^{-\rho(s-t)}ds].$$

Recall by Lemma 7 that for any feasible beliefs process (and hence for any optimal policy), the beliefs process is a (semi-)martingale described by the σ_s and ψ_t defined in that lemma.

By Ito's lemma for semi-martingales[63],

[63] See e.g. theorem 2.42 of chapter II of Jacod and Shiryaev [2013].

$$\hat{\phi}_s = e^{-\rho s}\phi(q_s) - e^{-\rho t}\phi(q_t) + \frac{1}{2}\int_t^s e^{-\rho l}\{\rho\phi(q_{l^-}) - \frac{1}{2}tr[\sigma_l\sigma_l^T\nabla^2\phi(q_{l^-})]\}dl$$
$$- \int_t^s e^{-\rho l}\int_{\mathbb{R}^{|X|}\setminus\{\vec{0}\}}(\phi(q_{l^-}+z) - \phi(q_{l^-}) - z^T\cdot\nabla\phi(q_{l^-}))\psi_l(dz)dl$$

is a local martingale.

Note by the quasi-left-continuity of q_t that beliefs cannot jump by $|z| > \delta$ with positive probability at any time t, and hence $Pr\{\tau_h > t\} > 0$. By the martingale property of $\hat{\phi}_s$,

$$\begin{aligned}E_t[e^{-\rho\tau_h}\phi(q_{\tau_h})] - e^{-\rho t}\phi(q_t) = {} & E_t[\frac{1}{2}\int_t^{\tau_h} e^{-\rho l}tr[\sigma_l\sigma_l^T\nabla^2\phi(q_{l^-})]dl] \\ & + E_t[\int_t^{\tau_h} e^{-\rho l}\int_{\mathbb{R}^{|X|}\setminus\{\vec{0}\}}(\phi(q_{l^-}+z) - \phi(q_{l^-}) - z^T\cdot\nabla\phi(q_{l^-}))\psi_l(dz)dl] \\ & - E_t[\int_t^{\tau_h} e^{-\rho l}\rho\phi(q_{l^-})dl],\end{aligned}$$

which yields, by the fact that $|q_s - q_0| < \delta$ for all $l \in [t, \tau_h)$,

$$\kappa E_t[\int_t^{\tau_h} e^{-\rho(s-t)}ds] \leq E_t[e^{-\rho\tau_h}\phi(q_{\tau_h})] - e^{-\rho t}\phi(q_t) \leq (\kappa - \frac{\varepsilon}{2})E_t[\int_t^{\tau_h} e^{-\rho(s-t)}ds],$$

a contradiction by the observation that $Pr\{\tau_h > t\} > 0$.

We conclude that

$$\sup_{\sigma_0,\psi_0\in\mathscr{A}(q_0)}\frac{1}{2}tr[\sigma_0\sigma_0^T\nabla^2\phi(q_0)] + \int_{\mathbb{R}^{|X|}\setminus\{0\}}(\phi(q_0+z) - \phi(q_0) - z^T\cdot\nabla\phi(q_0))\psi_0(dz) \geq \rho\phi(q_0) + \kappa.$$