NBER WORKING PAPER SERIES

RATIONAL INATTENTION WHEN DECISIONS TAKE TIME

Benjamin M. Hébert
Michael Woodford

Working Paper 26415
http://www.nber.org/papers/w26415

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2019, Revised August 2021

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, Pietro Ortoleva, 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 for excellent research assistance, and the NSF for research support. We would particularly like to thank Philipp Strack and Doron Ravidfor discussing an earlier version of the paper, and Simon Kelly for sharing data from Kelly et al. (2021). 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 portions appeared in Benjamin Hébert's Ph.D. dissertation at Harvard University. All remaining errors are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Rational Inattention when Decisions Take Time
Benjamin M. Hébert and Michael Woodford
NBER Working Paper No. 26415
October 2019, Revised August 2021
JEL No. D8,D83

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 demonstrate that 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.

Benjamin M. Hébert
Graduate School of Business
Stanford University
655 Knight Way
Stanford, CA 94305
and NBER
bhebert@stanford.edu

Michael Woodford
Department of Economics
Columbia University
420 W. 118th Street
New York, NY 10027
and NBER
mw2230@columbia.edu

A technical appendix is available at http://www.nber.org/data-appendix/w26415

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 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 then demonstrate that the resulting state-dependent stochastic choice probabilities of our continuous-time model are equivalent to those of a static RI model. Any cost function for a static RI model in the uniformly posterior-separable family (in the terminology of Caplin et al. [2019]) can be interpreted using our framework. This result offers both a justification for using such cost functions in static RI problems and a means of connecting those cost functions to dynamic processes for beliefs, and in particular to data on decision times.

We focus our analysis on a limit in which decision times are short relative to the rate of time preference. 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). We also 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]). 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, as functions of model parameters including the opportunity cost of time. The key to this characterization is a demonstration that in a broad class of cases, both the stopping regions and the ex ante choice probabilities for any given initial prior are the same as in a static RI problem with an appropriately chosen static information cost function. 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 of particular types. 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. [2019] 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 [forthcoming], 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 diffusion (in the former case) or a pure jump process (in the latter). In section 4 we demonstrate that the state-dependent choice probabili-

[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.

ties predicted by our continuous-time model (in both the diffusion case and the jump case) are equivalent to those predicted by a static rational inattention model with a uniformly posterior-separable cost function. In section 5 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 $\bar{q}_0 \in \mathscr{P}(X)$. We allow the DM to choose any filtered probability space $(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_{t \in \mathbb{R}_+}, P)$ and stochastic process $q : \Omega \times \mathbb{R}_+ \to \mathscr{P}(X)$, such that q_t is a càd1àg $\{\mathscr{F}_t\}$-martingale and $q_0 = \bar{q}_0$, subject the constraint specified below.

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 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.

These two examples could also be combined, to generate a jump-diffusion process. And the quantities σ_t and $z_{k,t}$ can 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.[8]

We assume that DM is subject to a constraint on how fast her beliefs can evolve, specified in terms of a "posterior-separable" cost function (as in the static rational inattention problems considered by Caplin et al. [2019]). Posterior-separable cost functions 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 $q' \ll q$.[9] 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 $R_+^{|X|} \times \mathscr{P}(X)$ by assuming the function to be homogenous of degree one. We also assume that it is strongly convex in its first argument and twice continuously-differentiable in both arguments.[10]

We require the DM's belief process to satisfy

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

[8] We show, however, that Markovian optimal policies exist.

[9] 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.

[10] Strong convexity, in this context, implies that $D(q'||q) \geq m|q' - q|^2$ for some constant $m > 0$.

where $\chi > 0$ is a finite constant. This constraint can be understood as the continuous-time analog of requiring that $E_{t-h}[D(q_t||q_{t-h})] \leq \chi h$ 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 $E^P[\cdot|\mathscr{F}_t]$. We illustrate the implications of this constraint in the context of our examples; these implications follow from Ito's lemma (for a proof, see Lemma 6 in the appendix).[11]

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

$$\frac{1}{2}tr[\sigma(q)^T Diag(q)\bar{k}(q)Diag(q)\sigma(q)] \leq \chi \tag{4}$$

for all $q \in \mathscr{P}(X)$, where $\bar{k}(q)$ is an $|X| \times |X|$ matrix defined on the interior of the simplex,

$$\bar{k}_{x,x'}(q_{t^-}) = \frac{\partial^2 D(q||q_{t^-})}{\partial q_x \partial q_{x'}}|_{q=q_{t^-}}, \tag{5}$$

and extended to the boundary by continuity.

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 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.[12] 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)

[11] Technical footnote: we require only that (3) hold for all $(\omega, t) \in \Omega \times \mathbb{R}^+$ outside of an evanescent set i.e. that the process q_t is indistinguishable from a process for which the constraint holds everywhere.

[12] 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.

converge to the stochastic process and constraint for a diffusion process (1). 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, 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.

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.

Discounting and Strictly Positive Utility. Much of our analysis will focus on the case without discounting ($\rho = 0$), or on the limiting case in which $\rho \to 0^+$. Many decisions are made over short periods of time (seconds or minutes). With conventional rates of time preference, $\rho\tau$ should be extremely close to zero. As we will demonstrate, in the $\rho = 0$, $\kappa > 0$ case, the model is tractable and we are able (under certain additional assumptions) to characterize the value function. Consequently, provided that behavior is continuous in the limit as ρ approaches zero, holding fixed $\kappa > 0$ (and we will show that it is), we believe that it is reasonable to focus on the predictions when $\rho = 0$.

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.

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 the cost of information as entering the utility function. 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 cost is treated as 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 when information costs are utility costs, but not when information costs are constraints. Our results, however, are not sensitive to the differences between the optimal policies in these two cases.

Conditional vs. Unconditional Dynamics and the DDM Model. 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. In contrast, DDM models (see, e.g., Fudenberg et al. [2018]) are usually expressed in terms of the conditional dynamics of beliefs. A "decision variable" z_t is assumed to follow a process

$$dz_t = \delta_{|x} dt + \alpha dB_{t|x}, \tag{7}$$

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).

To understand the relationship between our optimizing model and DDM-style models, suppose that the DM chooses a diffusion process for beliefs, as in (1). (We establish con-

ditions below under which this will be optimal.) Conditional on the true state being $x \in X$, the DM's beliefs q_t follow a diffusion of the form the process[13]

$$dq_t \;=\; Diag(q_{t^-})\sigma(q_{t^-})\sigma(q_{t^-})^T e_x\, dt \;+\; Diag(q_{t^-})\sigma(q_{t^-})dB_{t|x}, \tag{8}$$

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 (8),

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

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

Thus if the DM chooses the kind of gradual evidence accumulation described by (1), the belief process q_t in our model has properties similar to those posited for the "decision variable" z_t in the DDM model: it is a diffusion process with a drift that depends on the true state x, and an instantaneous variance that is independent of the state. 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.[14]

2.2 Preliminary Analysis

We begin by showing that optimal policies exist. A key concern is the possibility of sequences of policies that involve increasingly frequent but small jumps and converge in the limit to diffusions. In this case, the stochastic processes for beliefs will converge to a continuous martingale, even though no martingale in the sequence is continuous. Nevertheless, the constraint in (3) is continuous in this limit, and the limiting policy is feasible.

Lemma 1. *There exists a set of optimal policies in the DM's problem.*

Proof. See the appendix, section A.1. □

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

[14] 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.

This result ensures that the questions we hope to address, such as when optimal policies involve jumps or diffusions, in fact have answers.

Next we 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 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 A.3 □

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. These two classes of divergences lead, respectively, to beliefs that move in small increments and beliefs that move in large increments. 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 $m > 0$,

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

Note that, to second order, $D(q'||q) = (q'-q)^T \bar{k}(q)(q'-q) + o(|q'-q|^2)$. A preference for gradual learning requires that 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

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

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.

We begin our analysis with a lemma, showing that the value function's curvature is limited by the possibility of diffusing along a line. Note that this lemma holds regardless of whether D exhibits a preference for gradual learning.

Lemma 2. *For all $q, q' \in \mathscr{P}(X)$ such that $q' \ll q$ and $q' \neq q$,*

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

Proof. See the appendix, section A.4. □

Lemma 2 and the HJB equation in Proposition 1 together show that the curvature of the value function ($V(q') - V(q) - (q' - q) \cdot \nabla V(q)$) is limited by both the possibility of directly jumping from q to q' and the possibility of attempting to diffuse from q to q'. In the case of a strong preference for gradual learning, the bound arising from the possibility of diffusing is tighter for sufficiently large values of $|q' - q|$. The intuition behind this result comes from a "race" between the strong preference for gradual learning of the divergence, which makes jumping as opposed to diffusing increasingly costly as $|q' - q|$ becomes large, and the potentially increasing cost of delay under a diffusion policy ($\rho V(sq' + (1-s)q)$ vs. $\rho V(q)$). Because the value function is bounded, the cost of delay can increase only so much, and consequently for $|q' - q|$ sufficiently large, the diffusion bound must be tighter than the direct jump bound.

Lemma 3. *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*

$$\frac{V(q') - V(q) - (q'-q)^T \cdot \nabla V(q)}{D(q'||q)} < \chi^{-1}(\rho V(q) + \kappa) \tag{12}$$

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 2 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. □

A consequence of this result is that when D exhibits a strong preference for gradual learning, there exists an optimal policy such that the probability of a jump of size greater than $(\frac{\rho(u_{max}-u_{min})}{m(\kappa+\rho u_{min})})^{\delta^{-1}}$ is zero.[15]

Proposition 2. *Define $\Delta q_t = q_t - \lim_{s \uparrow t} q_s$ and $\Delta V_t = V(q_t) - \lim_{s \uparrow t} V(q_s)$. If D exhibits a strong preference for gradual learning, then there exists an optimal policy such that*

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

and such that all jumps increase the value function ($\Delta V_t \geq 0$, almost surely strictly wherever $|\Delta q_t| > 0$).

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

The optimal policy in this case features upward (in the sense of the value function) jumps and downward drift. The fact that jumps only increase and never decrease the value function is a consequence of the exponential discounting. Exponential discounting can be

[15] We conjecture that a stronger result here is possible (at the expense of additional technicalities)– that the probability of such large jumps is zero under any optimal policy.

thought of as a penalty for delay that is increasing in the current level of the value function. For this reason, drifting upward and jumping downward is sub-optimal, because the former causes information to be acquired at a time when the cost of delay is high, and the latter acquires information at a time when the cost of delay is high rather than waiting for the cost of delay to decrease.[16]

In the particular case of no discounting ($\rho = 0, \kappa > 0$), we can reach stronger conclusions. The sub-optimality of a jump, (12), must hold for all $q' \neq q$. Consequently, an optimal diffusion policy exists. The following proposition extends this result to the case of a (possibly non-strong or non-strict) preference for gradual learning.[17] Recall that the matrix-valued function $\bar{k}(q)$ is defined in (5).

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*

$$\max\{\sup_{\sigma \in \mathbb{R}^{|X|} \times \mathbb{R}^{|X|-1} : q^T \sigma = \vec{0}} tr[\sigma^T Diag(q)(\nabla^2 V(q) - \frac{\kappa}{\chi}\bar{k}(q))Diag(q)\sigma], \hat{u}(q) - V(q)\} = 0, \tag{13}$$

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 A.6. □

Under an additional assumption (described in the next section), a preference for gradual learning is not only sufficient but necessary for beliefs to follow a diffusion process in the $\rho = 0$ case. In particular, we will demonstrate that if, for all utility functions, an optimal belief process in the continuous time limit is a diffusion, then the divergence must exhibit a preference for gradual learning. However, to make this statement, we must be able to characterize the belief dynamics, which we are able to do given an additional assumption. We therefore postpone our proof of necessity to the next section.[18]

Lastly, let us note that there is a kind of continuity between the $\rho > 0$ but small and $\rho = 0$ cases (assuming $\kappa > 0$). As ρ converges towards zero, with a strong preference for

[16] This intuition is reminiscent of a related result in Zhong [2019], discussed below.

[17] As before, to avoid technicalities, we do not prove a stronger claim that we conjecture holds: that in the case of a strict preference for gradual learning and $\rho = 0$, all policies involve diffusions.

[18] 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 be certain that this preference holds on the entire simplex.

gradual learning, the magnitude of jumps becomes increasingly small, and in the limit no jumps occur. Let us note also a remarkable result from Zhong [2019], which shows that with $\rho > 0$, the optimal policy involves jumps outside of a nowhere-dense set. These two results are compatible: the jumps in this case are small but not infinitesimal.

Zhong [2019] also 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, a result we restate below. 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.[19]

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 a 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,[20] only the Bregman divergences exhibit a preference for discrete learning.

Lemma 4. *The divergence D exhibits a preference for discrete learning if and only if D is a Bregman divergence.*

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

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

Proof. See the appendix, section A.7. 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 (e.g. if they have a strict preference for gradual learning in some parts of the parameter space and discrete learning in others).

If the cost function satisfies a preference for discrete learning, it its 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 4. *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 policy that does not diffuse and such that if $|\Delta q_t| > 0$, then $t = \tau$ (the DM stops immediately after any jump).*

Proof. See the appendix, section A.8. This is proven using Proposition 7 below. □

The statement of Proposition 4 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.

We next restate a result of Zhong [2019] (see appendix A.3 of that paper) that covers the $\rho > 0$ case.[21] With a preference for discrete learning, as with a preference for gradual learning, jumps will increase the value function. The intuition is essentially the same as the gradual learning case, and comes from the observation that with discounting, delay is particularly costly when the value function is high. However, unlike the gradual learning case, in which jumps are of bounded size, with a preference for discrete learning jumps are always immediately following by stopping. Zhong [2019] also shows that optimal policies do not involve diffusion (subject to some technical caveats).

Proposition 5. *(Zhong [2019]) Define $\Delta q_t = q_t - \lim_{s\uparrow t} q_s$ and $\Delta V_t = V(q_t) - \lim_{s\uparrow t} V(q_s)$ and assume $\rho > 0$. If D exhibits a preference for discrete learning, then in any optimal*

[21] 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).

policy, if $|\Delta q_t| > 0$, then $t = \tau$ (the DM stops after jumping) and $\Delta V_t > 0$ (jumps increase the value function). In any optimal policy, diffusion occurs only on a nowhere-dense set.

Moving beyond the results of Zhong [2019], we provide the "only-if" result: if a divergence always results in large jumps and immediate stopping, 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.[22]

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 A.9. □

Combining this result with Theorem 6, we have demonstrated that the jump-and-immediately-stop result of Zhong [2019] holds for all utility functions if and only if D is a Bregman divergence. 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 diffusion processes as ρ becomes close to zero.

3.3 Gradual vs. Discrete Learning

We summarize the differences between gradual and discrete learning before proceeding. With $\rho > 0$ and a 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 the DM will optimally choose to have beliefs that diffuse. In contrast, with $\rho > 0$ and a preference for discrete learning, the DM will optimally choose 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.

[22] 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.

We interpret these results as follow. In the $\rho \to 0^+$ limit, which we view as empirically relevant (as most decision-making experiments involve small time periods), 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 might 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 The Equivalence of Static and Dynamic Models

In this section, we analyze the $\rho = 0$ continuous time model under a preference for gradual learning and a preference for discrete learning. The main result of this section is that, both with a preference for gradual learning and a preference for discrete learning (under an integrability assumption in the case of a preference for gradual learning), the value function with $\rho = 0$ is equivalent to a static rational inattention problem with a uniformly posterior-separable cost function (i.e. a cost function defined from a Bregman divergence, see Caplin et al. [2019] or (16) below). Moreover, any twice continuously-differentiable uniformly posterior-separable cost function can be justified through either of these routes. Our equivalence result extends to policies as well, in the sense that the joint distribution of actions and states induced by optimal policies in the continuous time model is also optimal in the static model, and vice-versa.

This result has several implications. First, it demonstrates that both jump and diffusion-based models are tractable and that the value functions can be characterized without directly solving the associated partial differential equation. Second, it provides a micro-foundation for the uniformly posterior-separable cost functions that have been emphasized in the literature. Third, it proves that the two approaches are equivalent in terms of the predicted joint distribution of states $x \in X$ and actions $a \in A$. That is, any joint distribution of (x,a) that could be observed under discrete learning could be observed under gradual learning.

On this last point, however, we do not mean to imply that the diffusion and jump processes are equivalent. Both of them endogenously will result in the same joint distribution of actions and states, but will have different predictions about the joint distribution of actions, states, and stopping times. As a consequence, considering stopping times can help differentiate the two models, and we consider this in the next section.

Our results in the case of gradual learning depend on an additional integrability assump-

tion that does not hold generically. Consequently, equivalence with static models holds for all cost functions with a preference for discrete learning but only some cost functions for a preference for gradual learning, and 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, but the reverse is not true.

4.1 Gradual Learning

To prove our equivalence result, we restrict our attention to information-cost matrix functions that are "integrable," in the sense described by the following assumption.[23]

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,*

$$\bar{k}(q) = \nabla^2 H(q), \tag{15}$$

where $\nabla^2 H(q)$ *denotes the Hessian of* H *evaluated at* q *and* $\bar{k}(q)$ *is define as in (5).*

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 [forthcoming]. 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). We refer to the function H as the "entropy function," for reasons that will become clear below. Note that $H(q)$ is convex, by the positive semi-definiteness of $\bar{k}(q)$, and homogenous of degree one.

The problem we are analyzing is the HJB equation of Proposition 3 (the problem with $\rho = 0$ and a diffusion process for beliefs). We describe our equivalence result below.

[23] 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.

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

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

subject to the constraint that $\sum_{a \in A} \pi(a) q_a = q_0$, where D_H is the Bregman divergence associated with the entropy function H that is defined by Assumption 1, and this value function can be achieved by a pure diffusion process.

There exist maximizers π^* and q_a^* such that π^* is the unconditional probability, in the continuous time problem, of choosing a particular action, and q_a^*, for all a such that $\pi^*(a) > 0$, is the unique belief the DM will hold when stopping and choosing that action.

Proof. See the appendix, Section section A.10. □

Let S be a set of possible signal realizations, and let $p : X \to \mathscr{P}(S)$ be a "signal structure" that defines the conditional distribution of signal realizations in each state $x \in X$. Taken as given a prior $q_0 \in \mathscr{P}(X)$, and let $\hat{\pi}_s(p, q_0)$ and $\hat{q}_s(p, q_0)$ denote the unconditional signal probabilities and posteriors, respectively, given the prior and signal structure. Our continuous time problem is equivalent to a static rational inattention problem in which the DM chooses S and p, given a prior $q_0 \in \mathscr{P}(X)$, with a particular uniformly posterior-separable (UPS) cost function,

$$C(p, q_0; S) = \frac{\kappa}{\chi} \sum_{s \in S} \hat{\pi}_s(p, q_0) D_H(\hat{q}_s(p, q_0) || q_0), \tag{16}$$

and with the signal space S identified with the set of possible actions A. The equivalence between our (seemingly complex) continuous time model and this static model renders the former tractable, both in the special cases in which analytic solutions to the static model are available and computationally (because the static model is straightforward to study numerically). The cost scalar $\frac{\kappa}{\chi}$ parametrizes the tradeoff between stopping and acquiring more information, which is governed by the rate at which information can be acquired (χ) and the cost of delay (κ).

The mutual information cost function proposed by Sims is one example of a UPS 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 defined by (16) is mutual information. Thus Proposition 7 provides a foundation

for the standard static rational inattention model, and hence for the same predictions regarding stochastic choice as are obtained by Matêjka et al. [2015]. On the other hand, Proposition 7 also implies that other cost functions can also be justified. Indeed, any (twice continuously-differentiable) uniformly posterior-separable cost function (16) can be given such a justification, by choosing the $\bar{k}$ function defined by equation (15).

We conclude that all continuous time models with gradual learning that also satisfy our integrability condition are equivalent to a static model with a uniformly posterior-separable cost function, and that any such static model can be justified from some model with gradual learning. We next show that the same set of static models can be justified from a model with discrete learning. Before proceeding, however, we observe that this result allows us to demonstrate that a preference for gradual learning is necessary for beliefs to always result in a diffusion process, provided that Assumption 1 holds.

Corollary 1. *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 A.11. □

4.2 Discrete Learning

The result with a preference for discrete learning is an immediate corollary of Lemma 4 and the preceding Proposition 7 (the result with gradual learning).

Corollary 2. *Assume $\rho = 0$ and that D exhibits and 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. Immediate from Lemma 4, 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.[24] In the next section, we begin to explore how information about stopping times can be used to distinguish the models.

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 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 information about the nature of the decision process (e.g., Ratcliff and Rouder [1998]).

Here we propose that data on response times can in principle be used to discriminate between alternative information-cost specifications. We will show that divergences that are equivalent in the sense of implying the same state-contingent choice probabilities — and hence the same value function in the case that discounting is negligible — 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. Interestingly, it is possible to distinguish between these two hypotheses even when (as in the problems considered here) actions are taken only infrequently.[25]

5.1 The Two-Action Case

To illustrate this possibility, we consider a simple example, in which there are two possible actions ($A = \{L, R\}$). We will consider the $\rho \to 0^+$ limit and impose Assumption 1. We compare behavior with a divergence D exhibiting a strict preference for gradual learning to the behavior generated by the Bregman divergence D_H (as defined in Proposition 7), which exhibits a preference for discrete learning. By Proposition 7, these two divergences will generate identical value functions $V(q)$; but with a strict preference for gradual learning, beliefs will diffuse, whereas with a preference for discrete learning beliefs will jump.

[24] 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 is the case.

[25] It would obviously be easier to tell whether beliefs evolve continuously or in discrete jumps in a case where the DM is required to continuously adjust some response variable that can provide an indicator of her current state of belief.

With both of these divergences, by Proposition 7, beliefs will move on a line; this is a consequence of the "locally invariant posteriors" property of Caplin et al. [2019] and the fact that there are only two actions. Let q_L^* and q_R^* be the optimal posteriors given the prior $\bar{q}_0$ in the static rational inattention problem described in Proposition 7, and let π_L^* be the optimal unconditional probability of action L. We will assume some information is acquired, which is to say $q_L^* \neq q_R^* \neq \bar{q}_0$, and that all of these beliefs 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 q_L^* and q_R^* in the simplex, which necessarily runs through $\bar{q}_0$.

For this reason, it is convenient in the two-action case to express the dynamics of beliefs in terms of the state variable $\pi_{L,t} \in [0,1]$, which corresponds to the beliefs

$$q_t \;=\; q(\pi_{L,t}) \;\equiv\; q_R^* + \pi_{L,t}(q_L^* - q_R^*), \tag{17}$$

with $\pi_{L,0} = \pi_L^*$. When $\pi_{L,t}$ reaches one, the DM chooses L; when $\pi_{L,t}$ reaches zero, the DM chooses R. Our interest is in characterizing the conditional (on the true state $x \in X$) likelihood of stopping and choosing L or R at each time t.

Before proceeding, let us observe from Proposition 7 that the value function (in terms of the state variable $\pi_{L,t}$) 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 $V_L = \sum_{x \in X} q_{L,x}^* u_{L,x} - \frac{\kappa}{\chi} H(q_L^*)$ and $V_R = \sum_{x \in X} q_{R,x}^* u_{R,x} - \frac{\kappa}{\chi} H(q_R^*)$. 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 $[0,\tilde{\pi}_L)$ and increasing on $(\tilde{\pi}_L, 1]$ for some $\tilde{\pi}_L \in (0,1)$. In the first two of these cases, we will say $\tilde{\pi}_L = 0$ and $\tilde{\pi}_L = 1$, respectively; thus in each case, $\tilde{\pi}_L$ is the value of π_L at which $V(\pi_L)$ reaches its minimum. We will 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.

5.2 Distributions of Stopping Times

We begin by introducing some notation with which to describe our models' predictions regarding the distribution of observed response times. For any time τ, let $F_a^x(\tau)$ be the

cumulative probability of a decision a by time τ, conditional on the state being x. For the state x, and either action a, this is a right-continuous non-decreasing function, with a maximum equal to the overall probability of choosing a in state x. The sum $F^x(\tau) = F^x_L(\tau) + F^x_R(\tau)$ is the cumulative distribution function of decision times when the state is x.

It will be useful to state our theoretical predictions, not in terms of these distributions for the decision time τ, but rather in terms of corresponding distributions for the response-time quantile $\hat{\tau}$. (The quantile $\hat{\tau}$ of the response time τ is the fraction of all responses in that state for which the response time is no greater than τ.) This has two key advantages. First, it allows us to state predictions that are independent of time units. The state- and response-contingent distributions for τ depend on the values of both κ and χ; instead, the predicted distributions for $\hat{\tau}$ depend only on their ratio.

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.[26] The 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, as long as we assume that the NDT is a constant (or more precisely, that its variance is vanishingly small, as discussed below).

For any quantile $0 \leq \hat{\tau} \leq 1$, let $G^x_a(\hat{\tau})$ be the fraction of all decisions in state x for which the decision is a and the response-time quantile is no greater than $\hat{\tau}$. In the case of any $\hat{\tau}$ such that $\hat{\tau} = F^x(\tau)$ for some τ, we define, for either action a,

$$G^x_a(\hat{\tau}) \;=\; F^x_a(\tau). \tag{18}$$

In some cases, however, the theoretical distribution of decision times τ has an atom at some particular decision time $\bar{\tau}$. In this case, there is a jump in the c.d.f. at this point, $F^x(\bar{\tau}-) \;=\; \hat{\tau}_1 \;<\; \hat{\tau}_2 \;=\; F^x(\bar{\tau})$, which raises a question as to how $G^x_a(\hat{\tau})$ should be defined for quantiles $\hat{\tau}_1 \leq \hat{\tau} < \hat{\tau}_2$.

Let us suppose that, rather than a constant, the NDT on each trial is an independent

[26] 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*.

draw from a distribution with mean t_0, and a continuous distribution (albeit one with a vanishingly small variance).[27] The c.d.f. of the distribution of observed response times (counting the NDT) will then be continuous but steep around the value $t_0 + \bar{\tau}$. We can then define $G_a^x(\hat{\tau})$ for all $\hat{\tau}$ using (18), and define, in the case of an atom at $\bar{\tau}$,

$$G_a^x(\hat{\tau}) \;=\; F_a^x(\bar{\tau}-) + \left(\frac{\hat{\tau}-\hat{\tau}_1}{\hat{\tau}_2-\hat{\tau}_1}\right)\left(F_a^x(\bar{\tau}) - F_a^x(\bar{\tau}-)\right) \tag{19}$$

for all quantiles $\hat{\tau}_a \leq \hat{\tau} \leq \hat{\tau}_2$. This definition preserves the property that each $G_a^x(\hat{\tau})$ is a non-decreasing function and that $\sum_{a\in A} G_a^x(\hat{\tau}) \;=\; \hat{\tau}$.

The empirical correlate of the functions $G_a^x(\hat{\tau})$ can be computed using an experimental dataset in which on each trial, the true state x, the response a, and the response time have been recorded. An especially interesting feature of these functions is what they imply about how the relative probability of an L response as opposed to an R response varies with the rapidity of the decision (early decisions versus late decisions, as measured by the response-time quantile $\hat{\tau}$). For each state-response pair (x,a), let us define $g_a^x(\hat{\tau})$ as the right derivative of the function $G_a^x(\hat{\tau})$; we must have

$$g_a^x(\hat{\tau}) \;\geq\; 0, \qquad g_L^x(\hat{\tau}) + g_R^x(\hat{\tau}) \;=\; 1$$

for each quantile $\hat{\tau}$. The relative probability of an L response, conditional on state x, is then given by $g_L^x(\hat{\tau})$. In the discussion below, we focus on the predicted shapes of $g_L^x(\hat{\tau})$.

5.3 Response Times with a Preference for Discrete Learning

In the case of a preference for discrete learning, the functions $g_L^x(\hat{\tau})$ are simple to describe. In the previous sections, we have presented two relevant theoretical results. First, as discussed above, beliefs will drift on a line segment in the simplex, until jumping to either q_L^* or q_R^*. Second, as emphasized by Zhong [2019], jumps must always increase the value function, and the drift of beliefs will reduce the value function.[28]

Suppose, to simplify the exposition, that the minimum of the value function on the line segment, $V(\pi_L)$, occurs at some $\tilde{\pi}_L < \pi_L^*$. For any $\pi_{L,t} > \tilde{\pi}_L$, the optimal policy is to jump

[27] We assume that the random NDT on any given experimental trial is independent of both the true state x and the sequence of evidence collected on that trial, and hence also independent of the decision that is made.

[28] The fact that the drift of beliefs will reduce the value function follows by applying the envelope theorem to the HJB equation of Proposition 1; see Zhong [2019].

towards $\pi_{L,t} = 1$ with the maximum possible intensity and drift downwards. Eventually, $\pi_{L,t}$ will drift downwards and equal $\tilde{\pi}_L$, at which point the DM will randomize between jumping to $\pi_{L,t} = 1$ and $\pi_{L,t} = 0$ with unconditional probabilities $\tilde{\pi}_L$ and $(1 - \tilde{\pi}_L)$. In the particular case of an upward-sloping value function ($\tilde{\pi}_L = 0$), the DM will choose R with certainty after reaching $\tilde{\pi}_L$.

Regardless of the state, in the absence of a jump, $\pi_{L,t}$ will reach $\tilde{\pi}_L$ at a predictable time, as beliefs drift downwards at a rate determined by the constraint (6),

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

Let $\tilde{\tau}$ be the time at which $\pi_{L,t} = \tilde{\pi}_L$ in the absence of a jump.

The unconditional likelihood of a jump prior to that time is determined by the constraint (6); the conditional likelihood is then pinned down by Bayes' rule. Suppose that the true state is x, and let $\tilde{q} = q(\tilde{\pi}_L)$ be the beliefs the DM will hold if there has been no jump before the time $\tilde{\tau}$. By Bayes' rule, if her posterior must be q_L conditional on jumping before time $\tilde{\tau}$, the probability of such a jump must satisfy

$$\frac{Pr\{\sup_{t\in[0,\tilde{\tau})} |\Delta\pi_{L,t}| > 0 \,|x\}}{Pr\{\sup_{t\in[0,\tilde{\tau})} |\Delta\pi_{L,t}| > 0\}} \;=\; \frac{q^*_{L,x}}{\bar{q}_{0,x}}.$$

Moreover, by the martingale property of the unconditional belief process,

$$Pr\{\sup_{t\in[0,\tilde{\tau})} |\Delta\pi_{L,t}| > 0\} q^*_{L,x} + (1 - Pr\{\sup_{t\in[0,\hat{\tau})} |\Delta\pi_{L,t}| > 0\})\tilde{q}_x \;=\; \bar{q}_{0,x},$$

which yields

$$\hat{\tau}^x \;=\; Pr\{\sup_{t\in[0,\tilde{\tau})} |\Delta\pi_{L,t}| > 0|x\} \;=\; \frac{q^*_{L,x}}{\bar{q}_{0,x}}\left(\frac{\bar{q}_{0,x} - \tilde{q}_x}{q^*_{L,x} - \tilde{q}_x}\right) \;=\; \frac{q^*_{L,x}}{q_x(\pi^*_L)}\frac{\pi^*_L - \tilde{\pi}_L}{1 - \tilde{\pi}_L}.$$

Intuitively, the likelihood of a jump conditional on x increases as the relative likelihood of L given x increases. We can now observe that $Pr\{\sup_{t\in[0,\tilde{\tau})} |\Delta\pi_{L,t}| > 0|x\} = \hat{\tau}^x$ is also the likelihood of a decision before time $\tilde{\tau}$. Consequently, the quantile at which $\pi_{L,t} = \tilde{\pi}_L$ is $\hat{\tau}_x$, and $g^x_L(\hat{\tau}) = 1$ for all $\hat{\tau} < \hat{\tau}^x$.

After time $\tilde{\tau}$, $\pi_{L,t} = \tilde{\pi}_L$ until a jump occurs, and the relative likelihoods of jumping to L and R are constant over time. Consequently, $g^x_L(\hat{\tau})$ is constant on $\hat{\tau} \geq \hat{\tau}^x$, and determined

again by Bayes' rule:

$$g_L^x(\hat{\tau}) \;=\; \frac{q^*_{L,x}}{\tilde{q}_x}\left(\frac{\tilde{q}_x - q^*_{R,x}}{q^*_{L,x} - q^*_{R,x}}\right) \;=\; \frac{\tilde{\pi}_L q^*_{L,x}}{q_x(\tilde{\pi}_L)}.$$

We thus obtain strong predictions about the functional form for $g_L^x(\hat{\tau})$: it is equal to one for all $\hat{\tau} < \hat{\tau}^x$ and constant for $\hat{\tau} \geq \hat{\tau}^x$. States in which L is more likely ($\frac{q^*_{L,x}}{q^*_{R,x}}$ larger) feature larger quantiles $\hat{\tau}^x$ and higher values of $g_L^x(\hat{\tau})$ for $\hat{\tau} \geq \tau^x$. Note that if we had instead assumed $\pi^*_L < \tilde{\pi}_L$, we would reach similar conclusions with the roles of R and L reversed.

Figure 1 below illustrates these results. The first row considers the case of a symmetric value function, $\tilde{\pi}_L = \frac{1}{2}$, the second an asymmetric case in which $\tilde{\pi}_L \in (0, \frac{1}{2})$, and the third the case of a monotonically increasing value function ($\tilde{\pi}_L = 0$). The first column plots the value function. The second column considers $g_L^G(\hat{\tau})$ for a state G with $q^*_{L,G} > q^*_{R,G}$, and the third column considers $g_L^B(\hat{\tau})$ for a state B with $q^*_{L,B} < q^*_{R,B}$.

We obtain even stronger predictions in the case that H is the Shannon entropy function, $H(q) = \sum_x q_x \ln q_x$, as assumed in many models of rational inattention following Sims [2010]. In this case, it is well-known that in the solution to the static rational inattention problem, the optimal information structure does not distinguish between states except to the extent that the difference between them is payoff-relevant.[29] For example, suppose that, as in many perceptual experiments, the reward $u_{a,x}$ for a particular response a in a state x (say, when a stimulus of type x is presented) depends only on whether a was the "correct" or "incorrect" response in that state, and let X_a be the subset of states for which a is the correct response. Then because the utility differential $u_{L,x} - u_{R,x}$ is the same for all states $x \in X_L$, it follows that $q^*_{L,x}/q^*_{R,x}$ will be the same for every state $x \in X_L$; and similarly for every state $x \in X_R$. It then follows from our formulas above that both $\hat{\tau}^x$ and g_L^x (the constant value for all $\hat{\tau} \geq \hat{\tau}^x$) will be the same quantities for all $x \in X_L$ ("G states"), and will similarly be the same quantities for all $x \in X_R$ ("B states"). Thus there will only be two functions $g_L^x(\hat{\tau})$, the two functions shown as $g_L^G(\hat{\tau})$ and $g_L^B(\hat{\tau})$ in Figure 1.

The numerical examples shown in Figure 1 are calculated for an example of this kind. $H(q)$ is Shannon entropy, and $u_{a,x}$ depends only on whether x belongs to X_L or X_R (so that we can write $u_{a,G}$ in the case of a "G state" and $u_{a,B}$ in the case of a "B state"). In the top row (Case I), we further assume that the reward depends only on whether a response is correct or not (so that $u_{L,G} = u_{R,B} > u_{R,G} = u_{L,B}$). In this case, the symmetry of the problem results in a symmetric value function, as shown, with $\tilde{\pi}_L = 1/2$. If we write $q^*_{a,G} = \sum_{x \in X_L} q^*_{a,x}$ for

[29] This follows from the property that Caplin et al. [2019] call "invariance under compression."

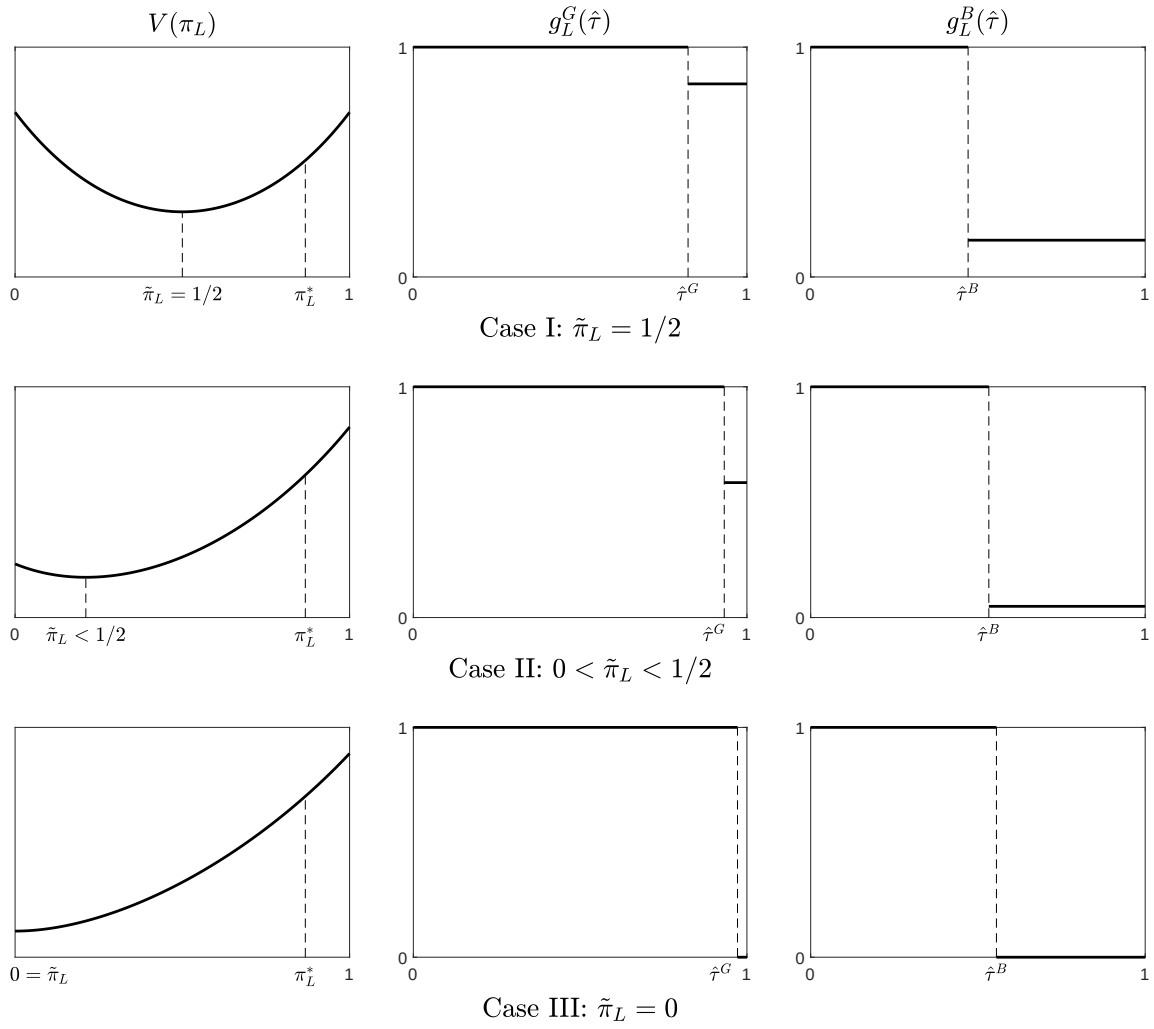

Figure 1: Predicted response-time distributions with a preference for discrete learning. Each row shows the value function $V(\pi_L)$, the function $g_L^x(\hat{\tau})$ for states x in which L is the correct response ("G states"), and the function $g_L^x(\hat{\tau})$ for states x in which R is the correct response ("B states"), under a particular assumption about the relative payoffs in G and B states. In the first row (Case I), the rewards for correct or incorrect responses are the same in G and B states; in the lower rows, the utilities associated with B states are made progressively lower relative those associated with G states. In all numerical calculations shown in this figure, $H(q)$ is assumed to be the Shannon entropy function. Information costs are parameterized so that the predicted accuracy rate ($Pr\{L|G\} = Pr\{R|B\}$) is $a = 0.84$, and the figures are drawn for the case of a prior $\bar{q}_{0,G} = 0.75$. These parameter choices are match the experimental data in Figure 3, but are also arbitrary, in the sense that the qualitative relationships illustrated in the graph will hold regardless of these choices, provided that the prior implies that $\tilde{\pi}_L < \pi_L^* < 1$.

the probability of a G state under the posterior q_a^*, and similarly $q_{a,B}^*$ for the probability of a B state, then the symmetry of the problem also implies that $q_{L,G}^* = q_{R,B}^*$, a quantity that we denote a (the predicted overall fraction of correct responses). Then since we must have

$$\frac{q_{L,x}^*}{q_{R,x}^*} = \frac{q_{L,G}^*}{q_{R,G}^*} = \frac{a}{1-a} \quad \forall x \in X_L, \qquad \frac{q_{L,x}^*}{q_{R,x}^*} = \frac{q_{L,B}^*}{q_{R,B}^*} = \frac{1-a}{a} \quad \forall x \in X_R,$$

the gradient of H (and hence the derivative of $V(\pi_L)$) at each of the stopping posteriors is completely determined by the value of a. Thus the value of a determines the value function $V(\pi_L)$, up to an additive constant. The value functions shown in Figure 1 assume that $a = 0.84$, the average accuracy rate in the experimental data shown in Figure 3.

In the second and third rows, we continue to assume the same utility differential between the correct and incorrect responses in any state as in the top row: $u_{L,G} - u_{R,G} = u_{R,B} - u_{L,B} = \Delta > 0$. However, we no longer assume that the reward for a correct response is the same in G and B states. If we let $u_{L,G} - u_{R,B} = u_{R,G} - u_{L,B} = \delta$, then $\delta = 0$ corresponds to Case I, shown in the top row. If instead $0 < \delta < \Delta$, we have an asymmetric value function and $0 < \tilde{\pi}_L < 1/2$ (Case II), as shown in the second row of the figure. (The numerical solution shown in the second row is for $\delta = \Delta/2$.) Finally, if $\delta \geq \Delta$, the value function is monotonically increasing and $\tilde{\pi}_L = 0$ (Case III), as shown in the bottom row. (The numerical solution shown is for $\delta = \Delta$.) We obtain similarly asymmetric solutions if $\delta < 0$, but with the roles of states G and B reversed.

The functions $g_L^x(\hat{\tau})$ shown in the other two panels of each row depend only on the accuracy rate a and the values of $\tilde{\pi}_L$, and π_L^* (which depends on the prior). In the cases shown in Figure 1, we assume a prior under which G states are more likely than B states, and let the prior probability of some G state be $\bar{q}_{0,G} = 0.75$ (matching the experiment of Kelly et al. [2021], discussed below); this corresponds to $\pi_L^* \approx 0.87$.[30]

The result that the function $g_L^x(\hat{\tau})$ for any state x is necessarily one of the two functions shown in Figure 1 depends on the special assumption that $H(q)$ corresponds to Shannon entropy. However, if we assume that there are only two states (that is, that it is possible to collect information directly about whether L or R is the correct response), then the numerical solution for the functions $g_L^x(\hat{\tau})$ for the two states $x = G, B$ does not depend on any details of the function $H(q)$, apart from a symmetry assumption: that $H(q_G, q_B) = H(q_B, q_G)$.

[30] If instead we were to assume that B states were more likely ex ante and that $\bar{q}_{0,B} = 0.75$, the function $g_L^B(\hat{\tau})$ would look like the function $g_L^G(\hat{\tau})$ shown in Figure 1, and the function $g_L^G(\hat{\tau})$ would look like the function $g_L^B(\hat{\tau})$ shown in Figure 1; hence we do not display additional figures for this case.

Under this assumption, the numerical values assumed for a, $\tilde{\pi}_L$, and $\bar{q}_{0,G}$ again completely determine the numerical specification of the functions $g_L^G(\hat{\tau})$ and $g_L^B(\hat{\tau})$, and will exactly match the ones shown in Figure 1.

5.4 Response Times with Gradual Learning and the DDM Model

Let us now consider instead 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. Note first that the constraint for a diffusion process, (4), will bind in any solution to the HJB equation (13). Because diffusion takes place on a line (17), the unconditional belief dynamics are of the form

$$dq_t \;=\; (q_L^* - q_R^*)\bar{\sigma}(\pi_{L,t})dB_t,$$

where $\bar{\sigma}(\pi_L)$ is scalar-valued and dB_t is a one-dimensional Brownian motion. The constraint (4) implies that

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

It then follows from (8) that the conditional dynamics of beliefs will be given by

$$dq_t \;=\; (q_L^* - q_R^*)\frac{q_{L,x}^* - q_{R,x}^*}{q_{t,x}}\bar{\sigma}(\pi_{L,t})^2 dt + (q_L^* - q_R^*)\bar{\sigma}(\pi_{L,t})dB_{t|x},$$

conditional on any true state x. From this, we can derive the conditional dynamics for the univariate belief state $\pi_{L,t}$:

$$d\pi_{L,t} \;=\; \frac{q_{L,x}^* - q_{R,x}^*}{q_x(\pi_{L,t})}\bar{\sigma}(\pi_{L,t})^2 dt + \bar{\sigma}(\pi_{L,t})dB_{t|x}. \tag{20}$$

This process resembles the conditional beliefs process of a DDM in several respects. The state variable $\pi_{L,t}$ diffuses until reaching one of two fixed boundaries (zero or one), which correspond to the two action choices. 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 of the standard DDM is that in general, neither the drift term nor the variance term in (20) is constant.

However, 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 that there are only two states, G and B, so that the posterior can be represented by a single real number, $q_{t,G}$, and the unconditional belief dynamics (until a decision is made) are of the form $dq_{t,G} = \sigma(q_{t,G})dB_t$, where $\sigma(q_G)$ is a scalar. And suppose that instead of the Shannon entropy function, $H(q)$ is what Bloedel and Zhong [2020] call a "total information" (TI) cost function:[31]

$$H(q) = (q_G - q_B)(\ln(q_G) - \ln(q_B)).$$

The maximum rate of information accumulation consistent with (4) is then given by

$$\sigma(q_G) \;=\; \sqrt{2\chi}q_G(1-q_G),$$

from which it follows, using (8), that the state-contingent belief dynamics are of the form

$$dq_{t,G} \;=\; \mu_x(q_{t,G})dt + \sigma(q_{t,G})dB_{t|x} \tag{21}$$

for $x = G, B$, where

$$\mu_G(q_G) \;=\; 2\chi q_G(1-q_G)^2, \qquad \mu_B(q_G) \;=\; -2\chi q_G^2(1-q_G).$$

Since $\pi_{L,t}$ is a linear transformation of $q_{t,G}$, 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, $z_t = \ln(q_{t,G}/q_{t,B})$. The is just a smooth nonlinear transformation $z_t = Z(q_{t,G})$ of the posterior probability of state G; we can then use Ito's lemma together with (21) to show that the state-contingent dynamics of this variable are given by

$$dz_t \;=\; \chi dt + \sqrt{2\chi}dB_{t|G}, \qquad dz_t \;=\; -\chi dt + \sqrt{2\chi}dB_{t|B}.$$

These are just the kind of dynamics postulated in the standard 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

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

q_a^* determined by the solution to the static rational inattention problem associated with the TI cost function.

We turn now to 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. Let $\phi_L^x(\pi_L, s)$ be the probability of hitting $\pi_{L,t} = 1$ or at or before time s (and before reaching the other decision boundary), if at time t no decision has been made and $\pi_{L,0} = \pi_L$. Note that 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

$$\frac{1}{\bar{\sigma}(\pi_L)^2}\phi_{L,s}^x(\pi_L, s) \;=\; \frac{q_{L,x}^* - q_{R,x}^*}{q_x(\pi_L)}\phi_{L,\pi}^x(\pi_L, s) + \frac{1}{2}\phi_{L,\pi\pi}^x(\pi_L, s), \tag{22}$$

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

By definition, $F_L^x(\tau) = \phi_L^x(\pi_L^*, \tau)$, and consequently solving this PDE allows us to compute F_L^x. The same PDE, with different boundary conditions, can be used to compute F_R^x, and from these we can compute the slope as a function of the quantile, $g_L^x(\hat{\tau})$.

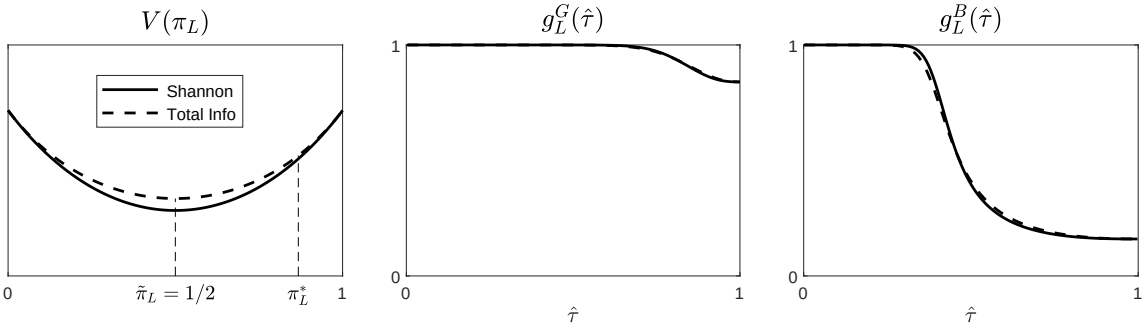

Figure 2: Predicted response-time distributions with a preference for gradual learning. The panels correspond to the same three columns as in Figure 1. Each of the functions is shown for two alternative choices for the entropy function $H(q)$: Shannon entropy (as in Figure 1, solid lines), and the entropy that results in a total information cost function (dashed lines). The accuracy rate a and prior probability of a G state are parameterized as in Figure 1; the value functions are drawn for the case of symmetric payoffs (Case I in Figure 1).

Figure 2 plots the recovered value of $g_L^x(\hat{\tau})$ for two cases with a strict preference for gradual learning. In one of these (shown by the solid lines), our numerical assumptions

are the same as in the first row of Figure 1, except that we now assume a strict preference for gradual learning. In particular, we again assume a function $H(q)$ given by Shannon entropy; we assume that rewards depend only on whether the DM's response is correct (so that the value function is symmetric); and we assume the same numerical values for a and $\bar{q}_{0,G}$ as in Figure 1. We show the figures only for the symmetric case (Case I) from Figure 1, because as just explained, the utilities enter into the calculation of the functions $g_L^x(\hat{\tau})$ in the case of a strict preference for gradual learning only through their effect on the values of the stopping posteriors q_a^*. Since the stopping posteriors are unaffected by a change in the value of δ (holding fixed the utility differential Δ), the predicted functions $g_L^x(\hat{\tau})$ under a strict preference for gradual learning will be the same in all three of the cases shown in Figure 1; hence we show only a single row in Figure 2.

As explained earlier, in the limit as $\rho \to 0^+$, the value function is the same in the case of a preference for gradual learning (PGL) as in the case of a preference for discrete learning (PDL); hence in the case of the cost function based on Shannon entropy (shown by the solid line), the value function shown in Figure 2 is identical to the one shown in the upper left panel of Figure 1. The dashed line shows the corresponding value function in the case of a two-state model with the "total information" cost function discussed above. In this alternative case, the model is again parameterized so as to imply the same values for a and $\bar{q}_{0,G}$ as in the Shannon case (and as in Figure 1), and the value function is again shown for the symmetric case (Case I in Figure 1). In the PDL model, the functions $g_L^x(\hat{\tau})$ would continue in this alternative case to be the ones shown in the top row of Figure 1, as discussed above. In the case of a strict preference for gradual learning, instead, the functions $g_L^x(\hat{\tau})$ are slightly different in the case of the total information cost function than in the case of the Shannon cost function. This is because in the PGL case the local curvature of the value function matters for the belief dynamics, and this is slightly different for the two cost functions, as shown in the left panel of Figure 2.

Nonetheless, there are not large quantitative differences between the distributions of response times implied by the two alternative cost functions, when these are parameterized to imply the same accuracy rate a. Since the belief dynamics and stopping times implied by the PGL model with the total information cost are identical to those of a particular parameterization of the DDM, we see that the distributions of response times implied by the PGL model in either case are similar to those implied by the DDM. The differences between the predictions of either version of the PGL model and the predictions for the PDL case are instead more notable.

5.5 Discriminating Between Discrete and Gradual Learning

A first important difference between the predictions of the discrete learning model and those of the gradual learning model concerns the effects of varying the parameter δ in Figure 1. In the gradual learning model, transformations of the utility function that shift utilities in a state contingent (but not action-contingent) fashion have no effect on behavior. To see this, suppose we shift the utility function $u_{a,x}$ to some $u'_{a,x} = u_{a,x} + v_x$. It is immediately, from Proposition 7, that the value function will be shifted by $\sum_{x \in X} q_x v_x$ for any $q \in \mathscr{P}(X)$. However, this will have no effect on behavior, as the level and slope of the value function did not enter our analysis above of the case of gradual learning; only the second derivatives of the value function at each point matter. In contrast, in the case of discrete learning, changes in the level and slope of the value function determine the location of the minimum, $\hat{\pi}_L$, and hence influence predicted behavior, as shown by a comparison of the different rows of Figure 1. This generates a testable difference between the two models: changes in payoffs conditional on states by not actions should affect behavior in the discrete learning case but not the gradual learning case. At present, however, we are not aware of any experiments that would test this proposition.

A second difference, of course, is that the response-time densities in the discrete learning case are step functions (as shown in Figure 1), whereas in the gradual learning case (and the DDM) they are sigmoid curves (as shown in Figure 2). This is also a difference in the models' predictions that should be testable. In fact, many experimental studies in the perceptual literature record distributions of response times, and these are often argued to be at least roughly consistent with the predictions of some version of the DDM (the parameters of which are estimated from such distributions). Figure 3 shows an example of the kind of data obtained in such studies, reported in Kelly et al. [2021]. This perceptual experiment is of particular interest for purposes of the present discussion, because the authors provide subjects with an informative cue which ought to give rise to an asymmetric prior of the kind assumed in Figures 1 and 2 (the case in which the PDL model, but not the PGL model, predicts that there should be a discontinuity in the function $g_L^x(\hat{\tau})$ at a critical quantile $\hat{\tau}^x$ for each state).

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 thus 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 whether the true direction was left or right; hence the utilities should satisfy the symmetry property assumed in "Case I" 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 therefore be either $\bar{q}_{0,G} = 0.75$ or $\bar{q}_{0,B} = 0.75$. (The cue is said to have a "75 percent validity" in either case.) Each type of cue is presented equally often.

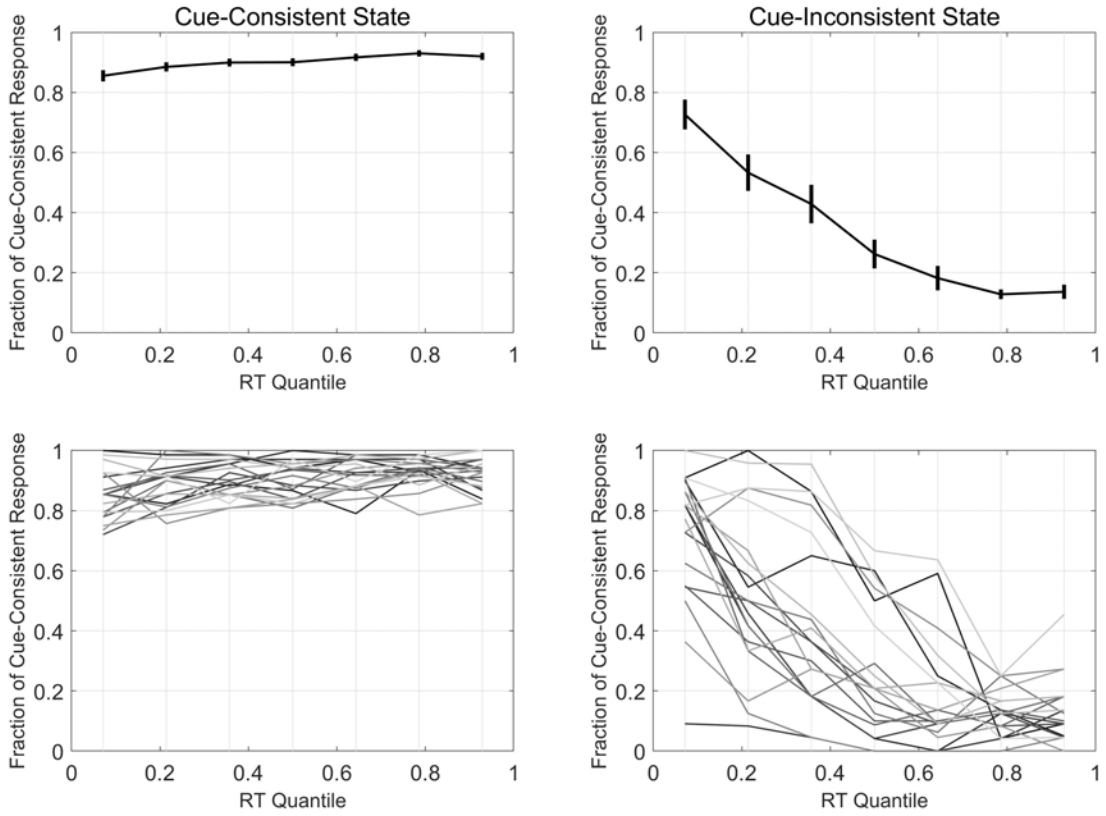

Figure 3: The relative frequency of cue-consistent and cue-inconsistent responses, as a function of the speed of the response, in the experiment of Kelly et al. [2021]. The curves shown in the left panels represent empirical versions of the function $g_L^G(\hat{\tau})$ shown in the middle column in Figures 1 and 2, while those in the right panel represent empirical versions of the function $g_L^B(\hat{\tau})$ 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. (Under the "deadline" condition, subjects' responses are correct only 84 percent of the time; average accuracy is much higher in the other experimental conditions.)

The PDL model makes sharp predictions about how the functions $g_L^x(\hat{\tau})$ should look in such an experiment, under the assumption of a cost function based on Shannon entropy, as proposed by Sims [2010] and Matêjka et al. [2015]. Because of the symmetry of the experimental setting as between the two possible directions of motion, the predictions should be those of Case I in Figure 1. As explained above, the predicted functions $g_L^x(\hat{\tau})$ then depend solely on the numerical values of two parameters, the accuracy rate a and the prior probability of a G state, $\bar{q}_{0,G}$. If we assume equal information costs for all of the experimental subjects (and hence a common value for a), then a should correspond to the overall accuracy rate observed on trials with informative cues (the value $a = 0.84$ assumed in Figure 1), while if we assume that subjects correctly understand the implications of the cues, on each trial $\bar{q}_{0,G}$ should be either 0.25 or 0.75 (the latter case being the one assumed in Figure 1).

Furthermore, for all trials on which the cue indicates that G states are more likely (as assumed in Figure 1), the function g_L^x should be the same for all G states (i.e., all trials on which the correct response would be L), and equal to the function shown in the middle panel of the top row of Figure 1. Moreover, this same function is what g_R^x should be like for all B states (all trials on which the correct response would be R), on trials on which the cue indicates that B states are more likely. In other words the function shown in the top middle panel is what g_a^x should be like whenever x is a *cue-consistent* state (the correct answer is in fact the one indicated by the cue as more likely to be correct) and a is the *cue-consistent* response. Hence we should be able to pool all of the trials on which the state is cue-consistent,[32] classifying them according to whether the response is cue-consistent on each trial, and estimate a single joint distribution of responses and response times, which should correspond to the function in the top middle panel of Figure 1. (Pooling the trials in this class, despite their differing priors and other differences in the exact stimuli presented, has the advantage of allowing a larger sample to be used to estimate the conditional probabilities of the two responses at different quantiles of the response-time distribution.) Similarly, we should be able to pool all of the trials on which the state is *cue-inconsistent*, classifying them according to whether the response is cue-consistent on each trial. The predicted conditional probability of a cue-consistent response on all such trials is

[32] Kelly et al. [2021] call these the "valid cue" trials, and pool their experimental data in exactly this way.

given by the function shown in the upper right panel of Figure 1.

The top row of Figure 3 (reproduced from Figure 3a of Kelly et al. [2021]) shows the empirical correlates of the two functions $g_L^G(\hat{\tau})$ and $g_L^B(\hat{\tau})$ in the top row of Figure 1, 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 of responses, each with a specific response time; in order to estimate the conditional probability of a cue-consistent response, it is necessary to 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.[33]

The resulting estimates of the functions $g_L^G(\hat{\tau})$ and $g_L^B(\hat{\tau})$ 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 (even under the Shannon entropy cost function) 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 $\hat{\tau}^B$, 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 do indeed observe in the lower right panel that there are significant differences across subjects with

[33]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.

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.[34]

Nonetheless, a subtle issue remains with regard to testing for the discontinuity predicted by the PDL model. In the case of the Shannon cost function, our model of optimal evidence accumulation implies that the fraction of correct responses at any time delay τ should be the same in any state x for which the cue-consistent response is incorrect; hence we should be able to estimate the function $g_a^x(\hat{\tau})$, and observe the predicted discontinuity, using data that pool all cue-inconsistent states x, as is done in Figure 3. This would also be true, even under a different cost function (such as the total information cost function), if we suppose that it is possible to obtain signals conditioned only upon payoff-relevant information — so that in an experiment where the subject's reward depends only on whether a response L or R is correct, there are only two states (G or B). But one might well assume a different kind of information costs, in which the cost of discriminating between two states x_1, x_2 depends on their perceptual similarity — which may correspond to their proximity in some "psychological space," as in the "neighborhood-based cost functions" discussed in Hébert and Woodford [forthcoming] — as a result of which states x with the same implications for payoffs may nonetheless result in different conditional distributions for responses and response times.[35] In the context of the Kelly et al. [2021] experiment, the states $x \in X$ can be distinguished not only based on the dominant direction of motion for the moving dots but also based on the nature of the randomly moving dots overlaid on top.

[34] 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.

[35] For evidence that this is the case, and hence that the Shannon cost function is empirically implausible, at least for perceptual tasks like those in the study of Kelly et al. [2021], see discussions in Dean and Neligh [2019] and Hébert and Woodford [forthcoming].

If we assume that the sets X_L and X_R each consist of many states, and that the entropy function $H(q)$ is not Shannon entropy, then the PDL model continues to predict that for each state x, the function $g_L^x(\hat{\tau})$ should be a step function of the form shown in Figure 1. However the critical stopping time at which the discontinuity occurs will in general differ for different states x that are perceptually distinct though payoff-equivalent. Hence if we pool the states for which a given response is correct, we could obtain an estimated relative-frequency curve that is progressively decreasing over some significant range of quantiles. In the absence of additional data — differentiated according to each of the perceptually distinguishable states x, and with a sufficient quantity of data for each x to allow a clear conclusion about how g_L^x varies with the quantile $\hat{\tau}$ — we cannot reach a conclusive judgment about whether response times in experiments like that of Kelly et al. [2021] are inconsistent with discrete learning. We can however with greater confidence rule out the joint hypothesis of a preference for discrete learning and Shannon information costs.

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.

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. *Unpublished manuscript*, February 2019.

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.

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*, forthcoming.

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. *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.

Dmitrii S Silvestrov. *Limit theorems for randomly stopped stochastic processes*. Springer Science & Business Media, 2012.

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 Proofs

A.1 Proof of Lemma 1

We will prove that from any sequence of policies achieving the value function in the limit, we can construct a feasible policy that achieves the value function. Consequently, because a sequence achieving the supremum exists (by definition), an optimal policy exists. We phrase the main result in terms of the lemma below to facilitate re-using the result in the proof of Proposition 2 below.

Lemma 5. *For all $n \in \mathbb{N}$, let $q_{t,n}$ be a martingale and τ_n be a stopping time defined on $(\Omega_n, \mathscr{F}_n, \{\mathscr{F}_{t,n}\}, P_n)$, 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. Then there exists a stochastic basis $(\Omega^*, \mathscr{F}^*, \{\mathscr{F}_t^*\}, P^*)$, martingale q_t^*, and stopping time τ^* such that $q_0^* = \bar{q}$, the constraint (3) is satisfied, and

$$V(\bar{q}_0) = E^{P^*}[e^{-\rho\tau^*}\hat{u}(q_{\tau^*}^*) - \kappa\int_0^{\tau^*} e^{-\rho s}ds|\mathscr{F}_0^*].$$

Moreover, there exists a sub-sequence such that the laws of $(q_{t,n}, \tau_n)$ converge in law to the law of (q_t^*, τ^*).

Proof. See the technical appendix, section B.1. □

We outline the steps of the proof of this lemma below.

1. Define a variable $x_n = f(\tau_n)$, and show it is integrable ($\rho = 0$) or bounded ($\rho > 0$)
2. Show the processes $q_{t,n}$ and x_n are tight, and converge in law to some (q_t^*, x^*).
3. Construct a stochastic basis such that $\tau^* = f^{-1}(x^*)$ is a stopping time and q_t^* is a martingale
4. Show that (q_t^*, τ^*) achieves the value function $V(\bar{q}_0)$.
5. Show that q_t^* is feasible.

The proof itself is quite technical and contains almost no economics. Interested readers are advised to have copies of Jacod and Shiryaev [2013] and Silvestrov [2012] at hand.

We should also point out that our results could be reformulated to avoid assuming the existence of optimal policies. Proposition 2 could be modified to claim that there exists a sequence of policies satisfying the stated conditions and achieving the value function in the limit, as opposed to stating that an optimal policy satisfies the stated conditions. Likewise, our proof that the value function must be a viscosity sub-solution to the associated HJB equation (Lemma 10) could be reformulated to avoid assuming the existence of maximizing policies (as in, for example, the proof in Pham [2009]). Of course, it is more straightforward, in light of the existence of optimal policies, to simplify the matters by referring to such policies.

In our proof and 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 6. *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*

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

Proof. By proposition 2.9 of chapter II of Jacod and Shiryaev [2013], 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 $\mathbb{R}_+$. 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

$$\psi_t(dz; \omega) = K_t(dz; \omega) \frac{dA_t}{ds}.$$

By theorem 2.21 of chapter II of Jacod and Shiryaev [2013], because $q_t - q_0$ is a martingale, $B = 0$.

Lastly, let us prove that the stated constraint is satisfied if (3) is satisfied. Applying Ito's lemma (see theorem 2.42 of chapter II of Jacod and Shiryaev [2013]), for any $\bar{h} \geq h$,

$$E_{(t-h)^-}[D(q_t||q_{(t-h)^-})] = \frac{1}{2} E_{(t-h)^-}[\int_{t-h}^{t} tr[\sigma_s \sigma_s^T \nabla_1^2 D(q_{s^-}||q_{(t-h)^-}) ds]$$
$$+ E_{(t-h)^-}[\int_{t-h}^{t} \int_{\mathbb{R}^{|X|} \setminus \{\vec{0}\}} (D(q_{s^-} + z||q_{(t-h)^-}) - D(q_{s^-}||q_{(t-h)^-}) - z^T \cdot \nabla_1 D(q_{s^-}||q_{(t-h)^-})) \psi_s(dz) ds].$$

By the predictability of the characteristics (which ensures left-continuity), the twice continuous-differentiability of D in its first argument, and the mean-value theorem,

$$\lim_{h \to 0^+} h^{-1} E_{(t-h)^-}[D(q_t||q_{(t-h)^-})] = \frac{1}{2} tr[\sigma_{t^-} \sigma_{t^-}^T \nabla_1^2 D(q_{t^-}||q_{t^-})]$$
$$+ \int_{\mathbb{R}^{|X|} \setminus \{\vec{0}\}} (D(q_{t^-} + z||q_{t^-}) - D(q_{t^-}||q_{t^-}) - z^T \cdot \nabla_1 D(q_{t^-}||q_{t^-})) \psi_t(dz),$$

where ∇_1 and ∇_1^2 denote the gradient and Hessian with respect to the first argument. 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

$$\nabla_1^2 D(q_{t^-}||q_{t^-}) = \bar{k}(q_{t^-})$$

which is the result. □

A.2 A useful 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 7. *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.

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. □

A.3 Proof of Proposition 1

We begin by proving, using Lemma 7, 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 B.2. □

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 B.3. □

Armed with this differentiability result, let us revisit Lemma 7. 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,

$$\sup_{q' \in \mathscr{P}(X)\setminus\{q\}: q' \ll q} \frac{V(q')-V(q)-(q'-q)^T \cdot \nabla V(q)}{D(q'||q)} \leq \chi^{-1}(\rho V(q)+\kappa). \tag{23}$$

We next argue, via a viscosity solution approach, that

$$\sup_{q' \in \mathscr{P}(X)\setminus\{q\}: q' \ll q} \frac{V(q')-V(q)-(q'-q)^T \cdot \nabla V(q)}{D(q'||q)} = \chi^{-1}(\rho V(q)+\kappa) \tag{24}$$

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

$$\max\{\sup_{\sigma_0,\psi_0 \in \mathscr{A}(q_0)} \frac{1}{2} tr[\sigma_0 \sigma_0^T \nabla^2 \phi(q_0)] + \int_{\mathbb{R}^{|X|}\setminus\{0\}} (\phi(q_0+z) - \phi(q_0) - z^T \cdot \nabla\phi(q_0))\psi_0(dz) \\ -\rho V(q_0) - \kappa, \hat{u}(q_0) - V(q_0)\} \geq 0, \tag{25}$$

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 B.4. 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 (25) 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 (23) proves that (24) 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 (24) holds, proving the result.

A.4 Proof of Lemma 2

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,

$$(q_2 - q_1)^T \cdot (\nabla V(q_2) - \nabla V(q_1)) \leq (\rho V(q_1) + \kappa) D(q_2||q_1) + (\rho V(q_2) + \kappa) D(q_1||q_2).$$

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 $0 \leq m < n$

and $s \in [0,1]$. It follows that

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

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

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

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

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

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

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). We have

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

and by the dominated convergence theorem,

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

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,

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

Therefore,

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

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.

A.5 Proof of Proposition 2

This proof is essentially a "verification" proof. We construct a sequence of sub-optimal policies that converge to the optimal policy, and then shows that such a sequence does not involve large jumps.

We being by constructing an ε-sub-optimal policy. Suppose the DM chooses a $K = 1$ jump process of the form

$$dq_t = -\frac{\chi}{D(q_{t^-} + z_\varepsilon^*(q_{t^-})||q_{t^-})} z_\varepsilon^*(q_{t^-})dt + z_\varepsilon^*(q_{t^-})dJ_t,$$

where J_t is a poisson process with intensity $\frac{\chi}{D(q_{t^-}+z_\varepsilon^*(q_{t^-})||q_{t^-})}$ and $z_\varepsilon^* : \mathscr{P}(X) \to \mathbb{R}^{|X|} \setminus \{\vec{0}\}$ is a feasible optimal policy constructed in the following manner.

1). Anywhere a maximizer of

$$\frac{V(q') - V(q) - (q'-q)^T \cdot \nabla V(q)}{D(q'||q)}$$

on $q' \in Q(q)$ exists, $z_\varepsilon^*(q)$ is a maximizer. By Proposition 1, in this case,

$$\frac{V(q + z_\varepsilon^*(q)) - V(q) - z_\varepsilon^*(q) \cdot \nabla V(q)}{D(q + z_\varepsilon^*(q)||q)} = \chi^{-1}(\rho V(q) + \kappa),$$

and by Lemma 3,

$$|z_\varepsilon^*(q)|^\delta \leq \frac{\rho(u_{max} - u_{min})}{m(\kappa + \rho u_{min})}.$$

Note that we must have, for any $\alpha \in (0,1]$,

$$V(q + \alpha z_\varepsilon^*(q)) - V(q) - \alpha z_\varepsilon^*(q) \cdot \nabla V(q) \leq D(q + \alpha z_\varepsilon^*(q)||q)\chi^{-1}(\rho V(q) + \kappa)$$

and consequently (by the strict convexity of D)

$$(V(q + \alpha z_\varepsilon^*(q)) - V(q)) < \alpha(V(q + z_\varepsilon^*(q)) - V(q)).$$

It follows that if $V(q+z_\varepsilon^*(q)) \leq V(q)$, then $V(q+\alpha z_\varepsilon^*(q)) \leq V(q)$ for all $\alpha \in [0,1]$. Applying Lemma 2 and a strong preference for gradual learning in this case,

$$V(q') - V(q) - (q'-q)\cdot \nabla V(q) < (\rho V(q) + \kappa) D(q + z_\varepsilon^*(q) || q),$$

a contradiction. We conclude that $V(q+z_\varepsilon^*(q)) > V(q)$.

2). Anywhere no such maximizer exists (diffusing is optimal), by Proposition 1 there must exist a sequence q_n' converging to q and such that

$$\lim_{n\to\infty} \frac{V(q_n') - V(q) - (q_n'-q)^T \cdot \nabla V(q)}{D(q_n'||q)} = \chi^{-1}(\rho V(q) + \kappa).$$

Choose any $z_\varepsilon^*(q)$ such that $|z_\varepsilon^*(q)| \leq \min\{\varepsilon, (\frac{\rho(u_{max}-u_{min})}{m(\kappa+\rho u_{min})})^{\delta^{-1}}\}$ and

$$\frac{V(q+z_\varepsilon^*(q)) - V(q) - z_\varepsilon^*(q)\cdot\nabla V(q)}{D(q+z_\varepsilon^*(q)||q)} \geq \chi^{-1}(\rho V(q) + \kappa - \varepsilon).$$

Assume the stopping policy is to stop immediately upon exiting the region $V(q) > \hat{u}(q)$. Let τ_ε be the associated stopping time under these policies, $\tau_\varepsilon = \inf\{t \geq 0 : V(q_t) \leq \hat{u}(q_t)\}$.

Under such policies,

$$\begin{aligned} d(e^{-\rho t}V(q_t)) = &-\rho e^{-\rho t}V(q_t)dt + \\ &- \frac{\chi e^{-\rho t}}{D(q_{t^-} + z^*(q_{t^-})||q_{t^-})} z^*(q_{t^-})^T \cdot \nabla V(q_{t^-})dt \\ &+ e^{-\rho t}(V(q_{t^-} + z^*(q_{t^-})) - V(q_{t^-}))dJ_t, \end{aligned}$$

and consequently

$$\begin{aligned} E_t[e^{-\rho(\tau-t)}V(q_\tau)] - V(q_t) = &E_t[\int_t^{\tau_\varepsilon} e^{-\rho(s-t)}\chi \frac{V(q_{s^-}+z^*(q_{s^-})) - V(q_{s^-}) - z^*(q_{s^-})^T\cdot\nabla V(q_{s^-})}{D(q_{s^-}+z^*(q_{s^-})||q_{s^-})} ds] \\ &- E_t[\int_t^{\tau_\varepsilon} e^{-\rho(s-t)}\rho V(q_{s^-})ds], \end{aligned}$$

which is

$$E_t[e^{-\rho(\tau_\varepsilon - t)}V(q_{\tau_\varepsilon})] - V(q_t) \geq (\frac{\kappa-\varepsilon}{\rho})E_t[1 - e^{-\rho(\tau_\varepsilon - t)}]. \tag{26}$$

Now observe that

$$dV(q_t) = -\frac{\chi}{D(q_{t^-} + z^*(q_{t^-})||q_{t^-})} z^*(q_{t^-})^T \cdot \nabla V(q_{t^-})dt + (V(q_{t^-} + z^*(q_{t^-})) - V(q_{t^-}))dJ_t,$$

and therefore by

$$\frac{V(q + z^*_\varepsilon(q)) - V(q) - z^*_\varepsilon(q) \cdot \nabla V(q)}{D(q + z^*_\varepsilon(q)||q)} \geq \chi^{-1}(\rho V(q) + \kappa - \varepsilon)$$

we have

$$E_t[V(q_{\tau_\varepsilon})] - V(q_{t_\varepsilon}) \geq (\rho u_{min} + \kappa - \varepsilon)E_t[\tau_\varepsilon - t].$$

Consequently, for $\varepsilon \in (0, \rho u_{min} + \kappa)$, we must have

$$E_t[\tau_\varepsilon - t] \leq \frac{u_{max} - u_{min}}{\rho u_{min} + \kappa - \varepsilon}. \tag{27}$$

The utility under this sub-optimal policy is

$$V^\varepsilon(q_t) = E_t[e^{-\rho(\tau_\varepsilon - t)} V(q_{\tau_\varepsilon})] - \frac{\kappa}{\rho} E_t[1 - e^{-\rho(\tau_\varepsilon - t)}].$$

By (26),

$$V^\varepsilon(q_t) \geq V(q_t) - \frac{\varepsilon}{\rho} E_t[1 - e^{-\rho(\tau_\varepsilon - t)}],$$

and by the inequality

$$1 - e^{-x} \leq x$$

for $x \geq 0$, it follows by (27) that

$$V^\varepsilon(q_t) \geq V(q_t) - \varepsilon \frac{u_{max} - u_{min}}{\rho u_{min} + \kappa - \varepsilon}.$$

Hence in the the limit as $\varepsilon \to 0^+$, the sequence of policies just constructed achieves the value function.

By Lemma 5, there exists a subsequence of these policies that converges in law to the law of an optimal policy. Let $q_{s,n}$ denote the stochastic process for beliefs along this subsequence, and let q^* denote the limit. Define $g(\Delta q_s) = \max\{|\Delta q_s| - (\frac{\rho(u_{max} - u_{min})}{m(\kappa + \rho u_{min})})^{\delta^{-1}}, 0\}$. By proposition 3.16 of section VI of Jacod and Shiryaev [2013], the law of $\hat{g}_{s,n} = g(\Delta q_{s,n})$

converges to the law of $\hat{g}_s = g(\Delta q_s^*)$, and by construction $\hat{g}_{s,n}$ is everywhere zero. It follows immediately that $\hat{g}_s$ is zero P-almost-everywhere.

We prove the second claim by a similar argument. By essentially the same argument, for any $\omega > 0$, $h_{t,n,\omega} = h_\omega(q_{t^-,n}, \Delta q_{t,n}) = (V(q_{t^-,n} + \Delta q_{t,n}) - V(q_{t^-,n}) \max\{|\Delta q_{t,n}| - \omega, 0\}$ converges in law to $h^*_{t,\omega} = h_\omega(q^*_{t^-}, \Delta q^*_t)$. Because the $h_{t,n,\omega}$ are everywhere positive for sufficiently large n (as the jumps approximating the diffusion become smaller than ω), strictly so wherever $|\Delta q_{t,n}| > \omega$, it follows that $h^*_{t,\omega}$ shares these properties P-almost-everywhere, for all $\omega > 0$.

A.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 2 and a preference for gradual learning,

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

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

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

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

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

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.

A.7 Proof of Lemma 4

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.

A.8 Proof of Theorem 4

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

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

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.

A.9 Proof of Theorem 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 6, 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

$$V(q_{l^-}+z^*(q_{l^-}))-V(q_{l^-})-z^*(q_{l^-})^T\cdot\nabla V(q_{l^-})=\chi^{-1}(\rho V(q_{l^-})+\kappa)D(q_{l^-}+z^*(q_{l^-})||q_{l^-}), \tag{28}$$

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,

$$V(q_{l^-}+z)-V(q_{l^-})-z^T\cdot\nabla V(q_{l^-})\leq\chi^{-1}(\rho V(q_{l^-})+\kappa)D(q_{l^-}+z||q_{l^-}). \tag{29}$$

To facilitate what follows, we write these conditions in the following manner, akin to a static rational inattention problem:

$$0=\sup_{\mu\in int(\mathscr{P}(\{1,2,3\})),\{q_i\in\mathscr{P}(X)\}_{i\in\{1,2,3\}}:\sum_{i=1}^3\mu_i q_i=q_{l^-}} \frac{\mu_1\hat{u}(q_1)+\mu_2V(q_2)+\mu_3V(q_3)-V(q_{l^-})-\chi^{-1}(\rho V(q_{l^-})+\kappa)\sum_{i=1}^3\mu_iD(q_i||q_{l^-})}{\mu_1}. \tag{30}$$

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 (28) in the limit as $\mu_1\to 0^+$. The numerator is always weakly negative by (29), and hence (30) 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

$$\theta D(q''||q) \geq \max_{a \in \{1,2,3\}} (q'')^T u_a - (q'' - q)^T \cdot v - K. \tag{31}$$

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. [2019] 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 (31) 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 $q'' \ll q$,

$$V(q'') - V(q; v, K) - (q'' - q)^T \cdot \nabla V(q; v, K) \leq \chi^{-1}(\rho V(q; v, K) + \kappa) D(q''||q). \tag{32}$$

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

$$K - V(q; v, K) - (q_a - q)^T \cdot (\nabla V(q; v, K) - v) \leq \chi^{-1} \rho (V(q; v, K) - K) D(q_a||q). \tag{33}$$

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 (31) and (32),

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

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 (30)) 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 $\lim_{n\to\infty}\hat{q}_{1,n} = q_1$,

$$\mu_1\hat{u}(q_1)+\mu_2V(q_2;v,K)+\mu_3V(q_3;v,K)-V(q;v,K)-\chi^{-1}(\rho V(q;v,K)+\kappa)\sum_{i=1}^{3}\mu_i^*D(q_i||q)=0$$

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 (33),

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

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.

A.10 Proof of Theorem 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

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

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,

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

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 $\hat{q}_a \neq \vec{0}$),

$$u_a - \frac{\kappa}{\chi}\nabla H(\hat{q}_a) - \xi + \nu_a = 0. \tag{34}$$

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. [2019]. 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 everywhere 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 2: 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 $w \neq \vec{0}$.

Now consider a policy that sets, for some $\varepsilon > 0$,

$$\tilde{\pi}(a, q_0\} = \begin{cases} 0 & a \notin A^* \\ \pi(a, q_0) - \varepsilon \sum_{a \in A^* \setminus \{a_0\}} w_a & a = a_0 \\ \pi(a, q_0) + \varepsilon w_a & a \in A^* \setminus \{a_0\} \end{cases}.$$

The maximum feasible ε, $\bar{\varepsilon}$, must set $\tilde{\pi}(a, q_0) = 0$ for some $a \in A^*$ and achieve the same utility (again with the same choices of $\{q_a\}_{a \in A}\}$). Repeating this argument, it is without loss of generality to suppose $|A^*| \leq |X|$.

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 such that

$$Q(q_0)d\pi_t = Diag(q_t)\sigma_t dB_t.$$

By the assumption that $|X| \geq |A^*|$, there exists a $|A^*| \times |X|$ matrix $\sigma_{\pi,t}$ such that

$$Q(q_0)\sigma_{\pi,t} = Diag(q_t)\sigma_t$$

and $d\pi_t = \sigma_{\pi,t}dB_t$. 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 scale $\sigma_{\pi,t}$ 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

$$\frac{1}{2}tr[\sigma_{\pi,t}\sigma_{\pi,t}^T Q^T(q_0)\bar{k}(Q(q_0)\pi_t))Q(q_0)] = \chi,$$

Note that we will always have $tr[\sigma_{\pi,t}\sigma_{\pi,t}^T] > 0$.

By the positive-definiteness of $\bar{k}$ in all directions except those constant in the support of $Q(q_0)\pi_t$, 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,

$$\phi(q_0) = \tilde{\phi}(\pi_0) = E_0[\hat{u}(q_\tau) - \kappa\tau].$$

A.11 Proof of Corollary 1

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

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

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. [2019], 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,

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

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 Technical Appendix

B.1 Proof of Lemma 5

Recall the outline of the proof steps:

1. Define a variable $x_n = f(\tau_n)$, and show it is integrable ($\rho = 0$) or bounded ($\rho > 0$)
2. Show the processes $q_{t,n}$ and x_n are tight, and converge in law to some (q_t^*, x^*).
3. Construct a stochastic basis such that $\tau^* = f^{-1}(x^*)$ is a stopping time and q_t^* is a martingale
4. Show that (q_t^*, τ^*) achieves the value function $V(\bar{q}_0)$.
5. Show that q_t^* is feasible.

Step 1: Define the variable $x_n = f(\tau_n)$, and show it is integrable/bounded. The purpose of this step is to deal with the following issue: when $\rho > 0$, it is not immediate that $E^{P_n}[\tau_n | \mathscr{F}_{0,n}] < \infty$.

Define

$$u_{max} = \max_{a \in A, x \in X} u_{a,x},$$

and define the random variable $x_n : \Omega_n \to \mathbb{R}_+$ by $x_n = f(\tau_n)$,

$$f(\tau) = \begin{cases} \tau & \rho = 0 \\ (1 - e^{-\rho\tau}) u_{max} & \rho > 0 \end{cases}$$

and observe that it is $\mathscr{F}_n$-measurable, and $x \in [0, u_{max}]$ if $\rho > 0$. We have

$$f^{-1}(x) = \begin{cases} x & \rho = 0 \\ -\rho^{-1} \ln(1 - \frac{x}{u_{max}}) & \rho > 0 \end{cases}$$

and by convention define $f^{-1}(u_{max}) = \infty$ if $\rho > 0$.

If $\rho = 0$, by the definition of limits, for any $\varepsilon > 0$, there exists an $n_\varepsilon \in \mathbb{N}$ such that, for all $n \geq n_\varepsilon$,

$$E^{P_n}[\hat{u}(q_{\tau_n,n}) - \kappa\tau_n | \mathscr{F}_{0,n}] \geq V(q_0) - \varepsilon.$$

Consequently, we must have (for all $n \geq n_\varepsilon$), by $V(q_0) \geq u_{min} = \min_{a \in A, x \in X} u_{a,x}$,

$$E^{P_n}[\tau_n | \mathscr{F}_{0,n}] \leq \frac{u_{max} - u_{min} + \varepsilon}{\kappa}, \tag{35}$$

In the $\rho > 0$ case, x_n is bounded.

Step 2: The processes $q_{t,n}$ and x_n are tight By theorem 4.13 of chapter VI of Jacod and Shiryaev [2013], it is sufficient to show that the predictable quadratic variation of $q_{t,n,j}$ (for some $j \in \{1, \ldots, |X|\}$, $< q_{t,n,j}, q_{t,n,j} >$, is C-tight (tightness as defined for a continuous process, see definition 3.25 of chapter VI of Jacod and Shiryaev [2013]). By the constraint (3) and the strong convexity of D, we must have, for some $m > 0$,

$$\lim_{h \to 0^+} m h^{-1} E^{P_n}[|q_t - q_{t-h}|^2 | \mathscr{F}_{t-h,n}] \leq \chi,$$

which implies

$$m^{-1}\chi t - \sum_{j=1}^{|X|+1} < q_{t,n,j}, q_{t,n,j} >$$

is an increasing process. Trivially, the sequence of processes $y_{n,t} = m^{-1}\chi t$ is C-tight (see Proposition 3.26 of chapter VI of Jacod and Shiryaev [2013]), and therefore by Proposition 3.35 $\sum_{j=1}^{|X|+1} < q_{t,n,j}, q_{t,n,j} >$ is C-tight, and consequently $q_{t,n}$ is tight. Note that this result also demonstrates that the processes $q_{t,n}$ are quasi-left-continuous.

Let us now show that the x_n variables are tight. If $\rho > 1$, this is immediate by $x_n \in [0, u_{max}]$. If $\rho = 0$ (and thus $\kappa > 0$), we show tightness via integrability. Recall the definition of tightness: the laws of τ_n are tight if, for any $\hat{\varepsilon} > 0$, there is a compact set $K_{\hat{\varepsilon}} \subset \bar{\mathbb{R}}_+$ such that, for all n, $E^{P_n}[\mathbf{1}\{x_n \in K_{\hat{\varepsilon}}\}] > 1 - \hat{\varepsilon}$. In the $\rho = 0$ case, by (35) above, for some $\varepsilon > 0$ and all $n \geq n_\varepsilon$,

$$\kappa^{-1}(u_{max} - u_{min} + 2\varepsilon) > E^{P_n}[\tau_n | \mathscr{F}_{0,n}] \geq E^{P_n}[\frac{u_{max} - u_{min} + 2\varepsilon}{\hat{\varepsilon}\kappa} \mathbf{1}\{x_n > \frac{u_{max} - u_{min} + 2\varepsilon}{\hat{\varepsilon}\kappa}\} | \mathscr{F}_{0,n}],$$

which proves tightness for $K_{\hat{\varepsilon}} = [0, \frac{u_{max} - u_{min} + 2\varepsilon}{\hat{\varepsilon}}]$ on the subsequence $n \geq n_\varepsilon$.

Let $\mathbb{D}(\mathscr{P}(X))$ be the space of all càdlàg functions $\mathbb{R}_+ \to \mathscr{P}(X)$, endowed with the Skorokhod topology; we have shown that a subsequence of $(q_{t,n}, x_n)$ is tight on $\mathbb{D}(\mathscr{P}(X)) \times \mathbb{R}_+$ endowed with the product topology. It follows by Prokhorov's theorem that there exists a subsequence of this subsequence such that the laws of $(q_{t,n}, x_n)$, $\mathscr{L}_n \in \mathscr{P}(\mathbb{D}(\mathscr{P}(X)) \times$

$\mathbb{R}_+$) converge weakly to some $\mathscr{L}^* \in \mathscr{P}(\mathbb{D}(\mathscr{P}(X)) \times \mathbb{R}_+)$. In what follows, consider only this subsequence of the subsequence.

Step 3: Construct the stochastic basis, martingale, and stopping time By the Skorokhod representation theorem, there exists an $(\Omega^*, \mathscr{F}^*, P^*)$ and random variables $(q^*_{t,n}, x^*_n)$ and (q^*_t, x^*) with laws $\mathscr{L}_n$ and $\mathscr{L}^*$ such that, P^*-almost-surely, $(q^*, x^*) = \lim_{n\to\infty}(q^*_n, x^*_n)$.

Define $\tau^* : \Omega^* \to \mathbb{R}_+ \cup \{\infty\}$ by $\tau^*(\omega) = f^{-1}(x^*(\omega))$, recalling that f^{-1} was defined above in step 1, and let the process y^*_t be defined by $y^*_t(\omega) = \mathbf{1}\{\tau^*(\omega) \leq t\}$, adopting the convention that $y^*_t(\omega) = 0$ if $\tau^*(\omega) = \infty$. Define $\{\mathscr{F}^*_t\}$ as the natural filtration of the process (q^*_t, y^*_t). Applying proposition 1.10 of section IX of Jacod and Shiryaev [2013], the process q^*_t is a martingale adapted to this filtration, and by construction τ^* is a stopping time (as y^*_t is $\mathscr{F}^*_t$-measurable).

Thus constructed, the collection $((\Omega^*, \mathscr{F}^*, \{\mathscr{F}^*_t\}, P^*), q^*_t, \tau^*)$ is our candidate optimal policy.

Step 4: The process (q^*_t, τ^*) achieves the value function The key step here is proving that the law of $q_{\tau_n,n}$ converges to the law of $q^*_{\tau^*}$. For this purpose, we rely on results from Silvestrov [2012]. We first consider the $\rho > 0$ case, which is more complex due to the possibility of not stopping, and then consider the $\rho = 0$ case.

Step 4a: the $\rho > 0$ case We start by proving that $q_{\min\{\tau_n,T\},n}$ converges in law to $q^*_{\min\{\tau^*,T\}}$ for any $T > 0$. It is immediate, from the boundedness and continuity of $\min\{f^{-1}(x), T\}$ and the continuous mapping theorem, that $(q_n, \min\{f^{-1}(x_n), T\})$ converges in law to $(q_n, \min\{f^{-1}(x^*), T\})$. In the the context of Silvestrov [2012], condition $\mathscr{A}_{20}$ holds.

By Definition 1.6.7 and Theorem 1.6.6 of Silvestrov [2012], the Skorokhod topology compactness condition $\mathscr{J}_4$ of Silvestrov [2012] must hold (this is essentially the tightness condition shown above). By the quasi-left-continuity of the $q_{t,n}$, for any $\delta > 0$ and $t \in [0,T]$, $E^{P_n}[\max\{|\Delta q_{t,n}| - \delta, 0\}|\mathscr{F}_{0,n}] = 0$, and by weak convergence $E^{P^*}[\max\{|\Delta q^*_t| - \delta, 0\}|\mathscr{F}_{0,n}] = 0$. It follows that q^*_t is quasi-left-continuous, and consequently that condition $\mathscr{C}_4$ of Silvestrov [2012] holds. By theorem 2.3.2 and lemma 2.3.1 of Silvestrov [2012], $(q_{\min\{\tau_n,T\},n}, \min\{\tau_n, T\})$ converges in law to $(q^*_{\min\{\tau^*,T\}}, \min\{\tau^*, T\})$.

Define

$$V_{n,T}(q_0) = E^{P_n}[e^{-\rho\min\{\tau_n,T\}}\hat{u}(q_{n,\min\{\tau_n,T\}}) - \frac{\kappa}{\rho}(1 - e^{-\rho\min\{\tau_n,T\}})|\mathscr{F}_{0,n}].$$

By the definition of convergence in law and the continuity of $\hat{u}$,

$$\lim_{n\to\infty} V_{n,T}(q_0) =$$
$$E^{P^*}[e^{-\rho \min\{\tau^*,T\}}\hat{u}(q_{\min\{\tau^*,T\}}) - \frac{\kappa}{\rho}(1-e^{-\rho\min\{\tau^*,T\}})|\mathscr{F}_0^*].$$

Note that

$$V_{n,T}(q_0) = E^{P_n}[e^{-\rho\tau_n}\hat{u}(q_{n,\tau_n}) - \int_0^{\tau_n} e^{-\rho s}\kappa ds|\mathscr{F}_{0,n}]$$
$$- e^{-\rho T}E^{P_n}[\mathbf{1}\{\tau_n > T\}(e^{-\rho(\tau_n - T)}\hat{u}(q_{n,\tau_n}) - \hat{u}(q_{n,T}) - \kappa\int_T^{\tau_n} e^{-\rho(s-T)}ds)|\mathscr{F}_{0,n}].$$

By $\hat{u}(q) \in [u_{min}, u_{max}]$ for u_{max} as defined above and $u_{min} = \min_{a\in A, x\in X} u_{a,x} > 0$, and $\kappa \geq 0$, we have

$$V_{n,T}(q_0) \geq E^{P_n}[e^{-\rho\tau_n}\hat{u}(q_{n,\tau_n}) - \int_0^{\tau_n} e^{-\rho s}\kappa ds|\mathscr{F}_{0,n}]$$
$$- (u_{max} - u_{min})e^{-\rho T}E^{P_n}[\mathbf{1}\{\tau_n > T\}|\mathscr{F}_{0,n}].$$

Recall (by assumption) that

$$\lim_{n\to\infty} E^{P_n}[e^{-\rho\tau_n}\hat{u}(q_{n,\tau_n}) - \int_0^{\tau_n} e^{-\rho s}\kappa ds|\mathscr{F}_{0,n}] = V(q_0),$$

and consequently

$$E^{P^*}[e^{-\rho\min\{\tau^*,T\}}\hat{u}(q_{\min\{\tau^*,T\}}) - \frac{\kappa}{\rho}(1-e^{-\rho\min\{\tau^*,T\}})|\mathscr{F}_0^*] \geq$$
$$V(q_0) - (u_{max} - u_{min})e^{-\rho T}.$$

Taking the limit as $T \to \infty$,

$$E^{P^*}[e^{-\rho\tau^*}\hat{u}(q_{\tau^*}) - \frac{\kappa}{\rho}(1-e^{-\rho\tau^*})|\mathscr{F}_0^*] \geq V(\bar{q}_0),$$

and hence this policy must be optimal if feasible.

Step 4b: the $\rho = 0$ case In this case we don't need to worry about never stopping, and hence don't need the previous limit approach. By step 2, (q_n, τ_n) converges in law to

(q^*, τ^*), which implies that Silvestrov [2012] condition $\mathscr{A}_{20}$ holds. By Definition 1.6.7 and Theorem 1.6.6 of Silvestrov [2012], the Skorokhod topology compactness condition $\mathscr{J}_4$ of Silvestrov [2012] must hold. By the quasi-left-continuity of the $q_{t,n}$, for any $\delta > 0$ and $t \in \mathbb{R}_+$, $E^{P_n}[\max\{|\Delta q_{t,n}| - \delta, 0\}|\mathscr{F}_{0,n}] = 0$, and by weak convergence $E^{P^*}[\max\{|\Delta q_t^*| - \delta, 0\}|\mathscr{F}_{0,n}] = 0$. It follows that q_t^* is quasi-left-continuous, and consequently that condition $\mathscr{C}_4$ of Silvestrov [2012] holds. By theorem 2.3.2 and lemma 2.3.1 of Silvestrov [2012], (q_{τ_n}, τ_n) converges in law to $(q^*_{\tau^*}, \tau^*)$. It is immediate that

$$\lim_{n\to\infty} E^{P_n}[\hat{u}(q_{\tau_n}) - \kappa\tau_n|\mathscr{F}_{0,n}] = V(\bar{q}_0) = E^{P^*}[\hat{u}(q^*_{\tau^*}) - \kappa\tau^*|\mathscr{F}^*_0].$$

Step 5: prove feasibility Here we rely on the following lemma:

Using this lemma, define for any $h > 0$

$$F_n(t,h) = E^{P_n}[D(q_{t,n}||q_{(t-h)^-,n})|\mathscr{F}_{(t-h)^-,n}],$$

and observe that $F_n(t,h) = 0$ be the quasi-left-continuity of $q_{t,n}$.

By Ito's lemma (see theorem 2.42 of chapter II of Jacod and Shiryaev [2013]), for any $\bar{h} \geq h$,

$$\begin{aligned} E^{P_n}[F_n(t,h)|\mathscr{F}_{t-\bar{h},n}] = \frac{1}{2}E^{P_n}[\int_{t-h}^{t} tr[\sigma_{s,n}\sigma_{s,n}^T\nabla_1^2 D(q_{s^-,n}||q_{(t-h)^-,n})ds|\mathscr{F}_{t-\bar{h},n}] \\ + E^{P_n}[\int_{t-h}^{t}\int_{\mathbb{R}^{|X|}\setminus\{\vec{0}\}} \{D(q_{s^-,n}+z||q_{(t-h)^-,n}) - D(q_{s^-,n}||q_{(t-h)^-,n}) - \\ z^T\cdot\nabla_1 D(q_{s^-,n}||q_{(t-h)^-,n})\}\psi_{s,n}(dz)ds|\mathscr{F}_{t-\bar{h},n}], \end{aligned}$$

where ∇_1 and ∇_1^2 denote the gradient and hessian with respect to the first argument. By the predictability of $\sigma_{s,n}$, $\psi_{s,n}$, and $q_{(t-h)^-,n}$,

$$\begin{aligned} E^{P_n}[\frac{\partial}{\partial h}F_n(t,h)|\mathscr{F}_{t-\bar{h},n}] = E^{P_n}[\frac{1}{2}tr[\sigma_{t-h,n}\sigma_{t-h,n}^T\bar{k}(q_{(t-h)^-,n})] \\ + \int_{\mathbb{R}\setminus\{0\}} D(q_{(t-h)^-,n}+z||q_{(t-h)^-,n})d\psi_{t-h,n}(z)|\mathscr{F}_{t-\bar{h},n}] \leq \chi, \end{aligned}$$

where we have used the definition

$$\nabla_1^2 D(q_{(t-h)^-,n}||q_{(t-h)^-,n}) = \bar{k}(q_{(t-h)^-,n})$$

and that

$$D(q_{(t-h)^-,n}||q_{(t-h)^-,n}) = 0, \nabla_1 D(q_{(t-h)^-,n}||q_{(t-h)^-,n}) = \vec{0}.$$

Considering the limit as $\bar{h} \to h$,

$$\frac{\partial}{\partial h} F_n(t,h) = \frac{1}{2} tr[\sigma_{t-h,n} \sigma_{t-h,n}^T \bar{k}(q_{(t-h)^-,n})] + \int_{\mathbb{R}\setminus\{0\}} D(q_{(t-h)^-,n} + z||q_{(t-h)^-,n}) d\psi_{t-h,n}(z) \leq \chi.$$

It follows that

$$F_n(t,h) \leq \chi h.$$

By the continuity of D and the convergence in law of $q_{t,n}$ to q_t^*, for any $t \in \mathbb{R}_+$ and $h > 0$,

$$\lim_{n\to\infty} E^{P_n}[F_n(t,h)|\mathscr{F}_{0,n}] = E^{P^*}[F(t,h)|\mathscr{F}_0^*]$$

where

$$F(t,h) = E^{P^*}[D(q_t^*||q_{(t-h)^-}^*)|\mathscr{F}_{(t-h)^-}^*],$$

and therefore, P^*-a.s. and for all $h > 0$,

$$h^{-1}(F(t,h) - F(t,0)) \leq \chi.$$

Because $F(t,h)$ is left-continuous by construction, this property must hold outside of an evanescent set (see the remark after 1.10 of chapter I of Jacod and Shiryaev [2013]). It follows immediately that

$$\lim\sup_{h\to 0^+} h^{-1}(F(t,h) - F(t,0)) \leq \chi,$$

which is the result.

Combining steps 4 and 5, the candidate optimal policy achieves a utility greater than or equal to $V(\bar{q}_0)$ and is feasible, and therefore optimal.

B.2 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 7, defining $z = \frac{1}{\alpha}\bar{z}$ and

$\varepsilon = \frac{1-\alpha}{\alpha}$,

$$\chi^{-1}(\rho V(q_t)+\kappa)(\varepsilon^{-1}D(q_t+\varepsilon\bar{z}||q_t)+D(q_t-\bar{z}||q_t)) \geq \\ \varepsilon^{-1}(V(q_t+\varepsilon\bar{z})-V(q_t))+V(q_t-\bar{z})-V(q_t). \tag{36}$$

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 7 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.

B.3 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 7, 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.

B.4 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,

$$V(q_t)=E_t[e^{-\rho(\tau_h-t)}V(q_{\tau_h})-\kappa\int_t^{\tau_h}e^{-\rho(s-t)}ds],$$

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 6 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. The following essentially restates Ito's lemma for (semi-)martingales.

Lemma 11. *For any twice continuously-differentiable function $\phi : \mathbb{R}^{|X|} \to \mathbb{R}$,*

$$\hat{\phi}_s = e^{-\rho s}\phi(q_s) - e^{-\rho t}\phi(q_t) + \frac{1}{2}\int_t^s e^{-\rho l}\{\rho\phi(q_{l^-}) - \frac{1}{2}tr[\sigma_l\sigma_l^T\nabla^2\phi(q_{l^-})]\}dl$$
$$- \int_t^s e^{-\rho l}\int_{\mathbb{R}^{|X|}\setminus\{\vec{0}\}}(\phi(q_{l^-}+z) - \phi(q_{l^-}) - z^T\cdot\nabla\phi(q_{l^-}))\psi_l(dz)dl$$

is a martingale.

Proof. See theorem 2.42 of chapter II of Jacod and Shiryaev [2013]. □

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 $\hat{\phi}_s$ defined in the above lemma,

$$E_t[e^{-\rho\tau_h}\phi(q_{\tau_h})] - e^{-\rho t}\phi(q_t) = E_t[\frac{1}{2}\int_t^{\tau_h} e^{-\rho l}tr[\sigma_l\sigma_l^T\nabla^2\phi(q_{l^-})]dl]$$
$$+ E_t[\int_t^{\tau_h} e^{-\rho l}\int_{\mathbb{R}^{|X|}\setminus\{\vec{0}\}}(\phi(q_{l^-}+z) - \phi(q_{l^-}) - z^T\cdot\nabla\phi(q_{l^-}))\psi_l(dz)dl]$$
$$- E_t[\int_t^{\tau_h} e^{-\rho l}\rho\phi(q_{l^-})dl],$$

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

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