Mulitdimensional and Selective Attention

Srijita Ghosh

December 15, 2017

Abstract

We consider a decision problem of a rational Bayesian DM who chooses multiple attributes to maximize payoff. The DM does not know the payoff function and faces a cognitive cost of attention. We do not assume the dimensions to be additively separable in the payoff function which introduces two novel features, one, the DM can learn about the correlation of the payoff and two, the DM can choose to pay attention to only a subset of attributes and infer about other inputs given his belief over possible payoff functions. We then characterize the optimal attention strategy when the cost of attention has two components, Shannon mutual entropy of the expected posterior and prior probability and cost of a choosing a partition of the state space, which reflects the complexty of the choice problem. We find the DM faces a complexity-precision trade-off which implies choosing to observe a strict subset of inputs, namely *selective attention* is payoff relevant. We also characterize the types of information, namely, conditional (e.g. agricultural extension services) and unconditional (e.g. social learning) on the DM's belief, that the DM would be willing to buy for different prior beliefs.

1 Introduction

We consider a rational producer who chooses multiple inputs to maximize his payoff from output where the payoff function is unknown to the DM. If learning is feasible, the optimal strategy for a rational producer would be to learn about the productivity of all inputs as long as all inputs are payoff relevant. But the following field experiment shows that the producer may not always do so.

Hanna, Mullainathan and Schwartzstein (2014) conducted a field experiment with seaweed farmers in Indonesia, where most of the farmers had been farming seaweeds for almost 15-20 years. In the baseline survey they found almost $97 - 98\%$ of farmers knew the optimal level of certain aspects of the production function, namely the distance between lines, distance between pods, the cycle length etc.

Whereas almost 87% of farmers did not know the optimal level of another important dimension, the pod size. In practice the farmers were using many different levels of pod size in their own lands and earning heterogenous payoff from different levels of pod size but they fail to notice the optimal pod size.

They further noted that on average for any farmer a switch from average pod size within his own farm to the optimal pod size would have yielded at least 7% increase in

annual income. This magnitude is significant compared to standard government subsidy programs for farmers in Indonesia. When asked, most the farmers said they didn't think that pod size is an important factor for productivity.

One of the questions we want to address in this paper is when would a rational decision maker (DM henceforth) in a multi-dimensional choice problem choose to pay attention to only a strict subset of dimensions. The choice of observing only a strict subset of dimensions has been referred as *selective attention* by Hanna et al (2014).

To answer when would *selective attention* be an optimal choice of a DM we formulate a multi-dimensional decision problem where the DM faces a constraint to update beliefs. We interpret this constraint as cognitive cost of paying attention.

Several papers have already analyzed this choice under attention constraint problem for one-dimensional decision problem. We extend that model to multiple dimension and analyze two key differences.

First, in a mutli attribute setting different attributes can be correlated to each other. If DM has a belief over the correlation then he can choose to learn only about one attribute and update his belief over joint payoff based on his belief about correlation. The additional choice of learning about correlation increases the set of attention strategies available to the DM.

Second, multiple dimensional action space generates a *natural* categorization along the dimensions. This implies the DM can now choose to pay attention after categorizing the data along the dimensions.

To capture these two features purpose we consider two types cost of attention as described in two different strands of literature, namely *Rational Inattention* and *Coarse Categorization*. In a RI model the cost of attention depends on the Blackwell informativeness of the attention strategy. The DM's optimal attention strategy depends on the payoff relevance of the strategy and his prior belief.

In the coarse categorization models the DM wants to predict the value of a variable of interest based on his observations. He can either face an exogenous cost which forces him to categorize the state space for updating his belief. Given this exogenous cost he then chooses to learn only about a coarser partition of the state space. This is known as coarse categorization.

The DM can also face an endogenous cost of choosing a finer partition. The endogenous cost comes from a bias-variance trade-off. If the categorization is fine then each category has very few data point which increases the bias, on the other hand if the category is too coarse the variance within the category increases which affect the precision of prediction. The cost in the coarse categorization is similar to a cost of paying attention one or more dimensions.

In the RI model the cost function does not depend on the number of dimensions wheras in the coarse categorization model conditional on a categorization the DM does not face any cost of updating beliefs. So we combine two types of cost in a model which we call *costlly partition* model to generate two features of the model.

For all these cost functions given a prior belief μ_0 we want to ask the following questions,

1. *What* would the DM choose to learn about, i.e, whether *selective attention* is optimal?

2. *How much* the DM would optimally choose to learn?

3. If the DM chooses to learn selectively, is it payoff relevant?

The third question, namely whether *selective attention* is payoff relevant or not is of interest to us. To consider the implication of the payoff relevance of the attention strategy we consider the following policy experiment.

Suppose we observe that the DM chooses only a strict subset of dimensions to pay attention to. If the an attention strategy is payoff relevant for him, then he may be willing to buy infromation about the unattended input for a positive price.

On the other hand if the attention strategy is not payoff relevant for him, i.e., the input is not relevant for maximize payoff given his prior belief thenhe would not be willing to pay for information about the unattended dimensions.

We find that the *selective attention* strategy is payoff relevant for only the *costly partition* model. In this context we consider two types of market for information, namely *conditional*, i.e, conditional on the period $t = 0$ belief of the DM and *unconditional*, i.e., information about the average payoff of an input snd characterize the DM's choice of information strategy based on his prior belief.

In the next section we discuss a problem discussed in context of agriculture sector in developing economies to show the relevance of market for infromation to the producers. The rest of the paper is organized as below: section 3 introduces the model, section 4 derives the attention policy for each type of cost function, section 5 reviews the literature and section 6 concludes.

2 An example from Agricultural Economics

This question is particularly relevant for agricultural sector in developing countries. There has been widespread agreement about development economists and policy maker around the world that farmers in developing economies do not produce at the optimal level which impacts the level of poverty in those economies. Many developing countries have chosen policies to increase the yield for the farmers by various methods, commonly referred together as *agricultural extension* policy.

As Haug(2007) summarizes the phases of agricultural extension programs, the initial emphasis of the programs was to inform the farmers about new technology and incentivize them to adopt the more productive technology. However, despite a high enough return from adopting a new technology the rate of actual adoption in different parts of world has been surprisingly low. Several field experiments have been conducted to analyze the reason and impact of low adoption. For a literature review see Janvry et al (2016).

However the World Development Report (2008) discusses that the gap in productivity arises from two types of gaps, one *technology gap*, i.e., the farmers do not know the existence of a new technology. The earlier *extension* programs were designed to bridge

this gap. However there is also a *management gap* which is the gap in the productivity between the best practice knowledge of the technology and the actual practice of the farmers. In recent times many policy makers have focused on *extension-plus* programs to bridge this gap.

Foster and Rosenzweig(2010) discusses the role of learning on technology adoption. They argue that there are two possible ways farmers update their belief, by experimenting on their own, i.e., private learning and learning from the experience from others, i.e, social learning. They document that the rate of learning is higher for farmers with higher level of education and argue that the difference is due to better ability of the educated farmers to learn. This implies to the possibility that the cost of learning not only depends on cost of experimentation but also cognitive abilities of the farmers.

For the social learning channel they show that the assumption that farmers understands the relationship between input use and profits, i.e, the structure of the production function is not realistic for complex technology. For example if the famer has to choose between varieties of HYV seeds that has different resistance to different types of pests, then it is unlikely that he would know the structure of the production function.

Then they show that the farmers learn less from social learning when they do not know the structure of the production function of their neighbours they are learning from. The impact of social learning decreases with the complexity of the technology. In general for any agricultural extension program this problem has been crucial for technology adoption. Even though there are externalities from learning and policies incentivize the farmers to learn they choose to not do so because of their lack of knowledge of the structure of the production function which makes it difficult for the famers to extract information from the experience of others.

Janvry et al(2016) analyze the constraints on optimal choice posed by demand side, mediating factor and supply side factors looking at the results of several field experiment. They conclude that the supply side constraints are important but less understood in the literature. From a case study in Eastern Indian state of Odisha they show that the learning takes place when the technology is sufficiently simple, the gain in payoff is sufficiently high and induces change is associated choice variables, i.e, fertlizer use and labor intensitivity of methods. They suggest that similar to developed countries agro dealer and commercial partners can serve as a sources of information by choosing appropriate interlinked contrcats.

Rivera nd Alex (2004) analyzes the impact of two extensio-plus programs, one public and one private in India. The public program KHDP (Kerala Horticulture Developmenet Programme) provided the farmers with access to credit, group marketing, processing technology for supply side support and participatory technology development. Besides relaxing the physical constraints the programme helped the farmers with pest and disease control and use of low cost inputs more efficiently. The impact of the program on profitability is significantly higher than the traditional extension program.

In the private program, a leadng tractor company in India set up over the counter service providing centers where the farmers can buy several services. One of the service is advisory and field supervision service where the field supervisors provide farmers guidance

on variety selection, land preparation, pest and disease management, and fertilizer use to help reduce cost of cultivation and to realize better yields. The success of the initial pilot program encouraged the company to roll out the program in other parts of India as well.

All these evidence can be summarized as follows:

1. Farmers do not always use the payoff maximizng practice for production
2. Learning is costly and farmers choose to learn when the structure of production technology is simple and change in payoff is significant
3. Farmers often know that they are not using the optimal production practice and willing to buy services that helps them increase productivity
4. Instead of learning about an average impact of a technology, farmers are more interested to learn the idiosyncratic impact of the technology on their own farm which is hard to obtain by social learning
5. Knowledge about the structure relating yields to attributes helps the farmer to learn
6. There is a positive correlation between eductaion and learning by the farmer which is mostly due to incresae in ability to learn with education

Even though there are field experiments to analyze the impact of different extension policies there are very few theoretical work analyzing the impact of selling information to a farmer. In this paper we suggest for different cost of attention what type of market for information would be efficient. Also we want to analyze the impact on the attention strategy of the farmer when he has the option to buy information.

3 Model

3.1 Environment

Let us consider a finite horizon discrete time environment where $t \in \{1, 2, \dots T\}$ and $T < \infty$. A rational DM faces a dynamic choice problem faced by . Each period the DM chooses two inputs to maximize production of an output. Let $\mathbf{A} = A_1 \times A_2$ denote the set of all possible input combination where each input A_i has only two levels a_{i1} and a_{i2}. A typical element in $\mathbf{A}$ is denoted by $\mathbf{a}_{ij} = (a_{1i}, a_{2j})$. Let $Y = \{0, 1\}$ denote the set of possible values of output.

Let $\pi : \mathbf{A} \to Y$ denote the payoff function that relates the choice of an input combination to output. We assume that at the beginning of period $t = 0$ the DM does not know the true payoff function. Let Ω denote the set of all possible payoff functions. We would consider Ω as the state space where a typical state ω denotes a particular payoff function. The main feature of the model is that the payoff depends on multiple dimensions which are possibly correlated and the exact nature of the correlation is unknown to the DM.

Since we do not retsrict the set of possible values of π in Ω, any correlation between the inputs can be expressed with an appropriate state ω. We further assume that the true state is chosen according to a data generating process(DGP) $\mu^* \in \Delta(\Omega)$ which is time invariant.

As there are only four possible input combinations $\mathbf{a}_{ij}$ and only two levels of ouput Y, the state space contains $2^4 = 16$ possible states. The list of all possible states are given in table 1. In general for any decision problem with m_i many levels of attribute A_i and k many levels of Y we have a states space with $k^{\Pi_i m_i}$ many states.

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}	ω_{11}	ω_{12}	ω_{13}	ω_{14}	ω_{15}	ω_{16}
$\pi(\mathbf{a}_{11})$	1	1	1	1	0	1	1	1	0	0	0	1	0	0	0	0
$\pi(\mathbf{a}_{12})$	1	1	1	0	1	1	0	0	1	1	0	0	1	0	0	0
$\pi(\mathbf{a}_{21})$	1	1	0	1	1	0	1	0	1	0	1	0	0	1	0	0
$\pi(\mathbf{a}_{22})$	1	0	1	1	1	0	0	1	0	1	1	0	0	0	1	0

Table 1: State space

Note that, ω_1 and ω_{16} are the two states where all input combination generate equal payoff, i.e., the payoff function is a constant. We would call these two states as *constant payoff* states. States $\{\omega_2, .., \omega_5\}$ are such that all but one input combinations give payoff of 1. Compare this set of states with $\{\omega_{12}, ..., \omega_{15}\}$ where only one input combination gives a payoff of 1. We would call the first set of states *good* states and the second state as *bad* states and for all these eight states both inputs are payoff relevant.

For the other six states either or both the inputs can be payoff relevant. We call them *middle* states as only two input combinations give payoff of 1 and the other two combinations give 0.

At the beginning of period $t = 0$ is DM enters with a prior belief over Ω denoted by $\mu_0 \in \Delta(\Omega)$. We assume that the true DGP $\mu^* \in supp(\mu_0)$, .i.e, all feasible states has a positive probability under this prior. For any other period let μ_t denote the belief over Ω at the beginning of period $t \in T$. We assume that in any period DM can choose to pay attention subject to a cost and learn about the true payoff function.

The timeline of the choice problem is as follows:

1. DM enters period t with belief μ_t
2. Given belief μ_t DM chooses an attention strategy and updates belief to μ_{t+1}
3. DM then chooses an input combination $\mathbf{a}_{ij}$ based on μ_{t+1} to maximize payoff
4. Payoff Y is realized and DM enters next period with belief μ_{t+1}.

Note that, we do not let the DM automatically update his belief based on his realized payoff. This is consistent with our assumption that attention is costly so any updating is a choice of the DM, i.e., learning is not automatic in this model even when the information is avaliable to the DM.

3.2 Partitions

The main objective of this model is to analyze the implication of the multiple dimensionality of the action space. But our definition of state space consists only of the payoff function and does not treat the different dimensions separately. Given the product space nature of the action space $\mathbf{A}$ we want to allow the DM to choose a projection of the $\mathbf{A}$ on one of the two dimension A_i.

To translate the attention choice in terms of action space to the state space we need to consider partitions of the state. When DM pays attention to a projection of the action space then he would not be able to differentiate some states. This would generate a partition of the state space. For example if we want to consider a scenario where the DM doesn't want to pay attention to input A_2, then he should not be able to distinguish those states in Ω that differs only in terms of payoff from A_2.

To formalize this notion, we define a set of auxiliary problems based on the partitions of the state space Ω and consider that the DM can choose from a menu of choice problems indexed by the partitions rather than the original decision problem based on Ω.

Let $\mathbb{P} = \{\mathcal{P}_i\}_{i\in\mathcal{I}}$ denote the set of all available partitions of Ω and $\mathcal{P}_i$ denote a typical element in $\mathbb{P}$ with $\mathcal{I}$ being the indexing set. Since Ω is finite $\mathcal{I}$ and $\mathbb{P}$ are also finite. We assume that in each period t given μ_t the DM can choose a different partition $\mathcal{P}_t$.

Let us define a *partition* i choice problem for period t using partition $\mathcal{P}_i$. The new state space is denoted by $\Omega_i \equiv \mathcal{P}_i$. Each block in the partition $\mathcal{P}_i$ is considered as a new state and is denoted by $\omega_j^i = \cup_k \omega_k$ such that $\omega_k \in \Omega$. The corresponding belief for each block is obtained by $\mu_t^i\left(\omega_j^i\right) = \sum_{\omega_k\in\omega_j^i} \mu_t\left(\omega_k\right)$.

The action space $\mathbf{A}$ and output Y remains same. Thus the expected payoff in state ω_j^i in partition $\mathcal{P}_i$ problem is given by,

$$E_{\mu_t}\left[\pi\left(\mathbf{a}_{ij}|\omega_j^i\right)\right] = \begin{cases} \frac{1}{\mu_t^i(\omega_j^i)}\sum_{\omega_k}\mu_t(\omega_k)\pi(\mathbf{a}_{ij}|\omega_k) & \text{if} \quad \mu_t^p(\omega_j^p) > 0 \\ 0 \quad \text{otherwise.} \end{cases} \tag{1}$$

Since μ_t^i is defined over ω_j^i, in the partition i problem the DM can only update μ_{t+1} over the partition, i.e, over ω_j^i and not for each $\omega_k \in \omega_j^i$ separately. This is the restriction on any attention strategy implied by the partition. The belief over original state space Ω can only be updated given this restrictions. If ω_p and ω_q are two states in ω_j^i with non-zero beliefs under μ_t, then after updating under the partition problem i we have,

$$\frac{\mu_t\left(\omega_p\right)}{\mu_t\left(\omega_q\right)} = \frac{\mu_{t+1}\left(\omega_p\right)}{\mu_{t+1}\left(\omega_q\right)}. \tag{PC}$$

We call equation PC the *partition consistency* condition.

We would denote a *partition i problem* as the auxiliary problem with state space Ω_i, the compressed beliefs over blocks of the partition μ_t^i with same action and payoff space that satisfies equation 1 and *partition consistency*. The partition consistency condition applies to the attention strategy but it implies that the DM can not have different choice probabilities conditional on states in the same block of the partition.

The set of all available partition $\mathbb{P}$ does not include all possible partitions of Ω. We would only consider the following partitions to be available to the DM. This assumption does not change any of the following analysis but we restrict our analysis to economically relevant partitions.

The avaliable partitions are the following:

1. DM only observes one level of one input
2. DM only observes one input combination
3. DM only observes one input
4. For a given level of one input the DM observes both levels of the other input
5. DM does not observe only one input

Let Ω denote the finest partition $\left\{\{\omega_i\}_{i=1}^{16}\right\}$, i.e., the original state space and Ω_0 denote the coarsest partition with no learning.

We further assume that when the DM observes only A_i and not observe A_j then to update his belief he consider a flat prior belief over the levels of input A_j, i.e., he assumes that both the two levels of input A_j are used with equal probabilities. In the next subsection we describe how each of this partition would look like under this assumption.

3.2.1 Examples of Partitions

Suppose DM only observes a_{11} and not A_2. Then his prior belief would give equal probabilities to $\mathbf{a}_{11}$ and $\mathbf{a}_{12}$. This implies he would not be able to distinguish states that are only different based on a_{12} or states where expected payoff from the two actions $\mathbf{a}_{11}$ and $\mathbf{a}_{12}$ are same. So the resulting partition is $\mathcal{P}_{a_{11}} = \big\{\{\omega_1, \omega_2, \omega_3, \omega_6\}, \{\omega_4, \omega_5, \omega_7, \omega_8, \omega_9, \omega_{10}, \omega_{12}, \omega_{13}\}, \{\omega_{11}, \omega_{14}, \omega_{15}, \omega_{16}\}\big\}$.

Suppose DM only observes $\mathbf{a}_{11}$ and no other input combinations. Then he can only distinguish states where $\mathbf{a}_{11}$ generates a different payoff. The resulting partition is $\mathcal{P}_{\mathbf{a}_{11}} = \big\{\{\omega_1, \omega_2, \omega_3, \omega_4, \omega_6, \omega_7, \omega_8, \omega_{12}\}, \{\omega_5, \omega_9, \omega_{10}, \omega_{11}, \omega_{13}, \omega_{14}, \omega_{15}, \omega_{16}\}\big\}$.

Suppose the DM only observes only A_1 and not input A_2 then the resulting partition is $\mathcal{P}_{A_1} = \big\{\{\omega_1\}, \{\omega_2, \omega_3\}, \{\omega_4, \omega_5\}, \{\omega_6\}, \{\omega_7, \omega_8, \omega_9, \omega_{10}\}, \{\omega_{11}\}, \{\omega_{12}, \omega_{13}\}, \{\omega_{14}, \omega_{15}\}, \{\omega_{16}\}\big\}$.

Suppose the DM observes A_2 given a_{11}, i.e., he observes $\mathbf{a}_{11}, \mathbf{a}_{12}$. Then he cannot distinguish states where the difference in payoff is only due to a_{12}, i.e., the resulting partition is $\mathcal{P}_{A_2|a_{11}} = \big\{\{\omega_1, \omega_2, \omega_3, \omega_6\}\{\omega_4, \omega_7, \omega_8, \omega_{12}\}, \{\omega_5, \omega_9, \omega_{10}, \omega_{13}\}, \{\omega_{11}, \omega_{14}, \omega_{15}, \omega_{16}\}\big\}$.

Finally suppose the DM does not observe only $\mathbf{a}_{21}$, then he cannot distinguish the pairs of states that are only different in terms of payoff from $\mathbf{a}_{21}$. The resulting partition would be $\mathcal{P}_{\neg\mathbf{a}_{21}} = \big\{\{\omega_1, \omega_3\}, \{\omega_2, \omega_6\}, \{\omega_4, \omega_8\}, \{\omega_5, \omega_{10}\}, \{\omega_7, \omega_{12}\}, \{\omega_9, \omega_{13}\}, \{\omega_{11}, \omega_{15}\}, \{\omega_{14}, \omega_{16}\}\big\}$.

Note that, $\mathcal{P}_{a_{11}}$ is coarser than $\mathcal{P}_{A_1}, \mathcal{P}_{A_2|a_{11}}$ and $\mathcal{P}_{\neg\mathbf{a}_{21}}$. However it is not coparable to $\mathcal{P}_{\mathbf{a}_{11}}$. Similarly $\mathcal{P}_{\mathbf{a}_{11}}$ is coarser than $\mathcal{P}_{A_2|a_{11}}$ and $\mathcal{P}_{\neg\mathbf{a}_{21}}$ but not comparable with $\mathcal{P}_{A_1}$. Finally $\mathcal{P}_{A_2|a_{11}}$ is coarser than $\mathcal{P}_{\neg\mathbf{a}_{21}}$.

In summary if one input combination is observed in both partitions then the partition with more blocks would be finer. But if only one input is observed, i.e., no specifc input combiantion is observed separately then the resulting partition can not ranked with the partitions where the DM observes a strict subset of input combinations. In the former the DM has more information about one input but does not know the payoff relevance of the other input and in the latter he does not altogether observe certain combinations. So for the observed combination he has more information.

3.3 Information structure:

Each period the DM has a choice to update his belief about Ω subject to some cost. Let S denote the set of signal alphabets such that $|S| \geq |\mathbf{A}|$. To update his belief the DM chooses a signal/information structure ν, given S, and updates his belief following Bayes rule, namely,

$$P_t\left(\omega_i|\nu\right) = \frac{P_t\left(\nu|\omega_i\right)\mu_t\left(\omega_i\right)}{\sum_j P_t\left(\nu|\omega_j\right)\mu_t\left(\omega_j\right)}, \tag{2}$$

where $P_t\left(\nu|\omega\right)$ denotes the probability of observing signal ν in state ω in period t. Following Matějka and McKay(2015) we claim that as long as choosing an information structure which is more Blackwell informative is more costly the DM would not choose two different signals to obtain the same posterior due to Blackwell sufficiency. Hece instead of explicitly using information/signal structures it is sufficient to look at the distribution of posteriors generated by signals.

If the DM chooses not to pay attention in period $t \in T$ then $\mu_t = \mu_{t+1}$, which implies learning is not automatic in this environment. Learning is costly in this environment and the utility cost can be interpreted of as a cognitive cost, either a cost of paying attention/mental accounting and/or memory.

The DM chooses an action to maximize his payoff net of the cost of attention. We assume that each belief state, i.e, a particular distribution of beliefs over Ω, has an unique action that maximizes the expected payoff given the belief. So we can directly consider the implied conditional choice probabilities of different actions for different states instead of beliefs over the states.

In other words, given a prior belief over states and information structure if the DM observes the signal alphabet s_1, then that generates a posterior belief over states and an unique action choice given this belief. Let $P_t\left(\mathbf{a}_{ij}\right)$ denote the unconditional probability of choosing action $\mathbf{a}_{ij}$ at the beginning of period t given the belief μ_t and let $P_t\left(\mathbf{a}_{ij}|\omega_k\right)$ denote the conditional probability of choosing action a_{ij} in state ω_k in period t.

Given his belief μ_t the DM chooses an information structure and updates his belief to $P\left(\mathbf{a}_{ij}|\omega_k\right)$ which generates the choice of an action $\mathbf{a}_{ij}$ in period t. The realized distribution $P_t(\mathbf{a}_{ij}|\omega_k)$ given the period t learning strategy becomes the unconditional choice probability $P_{t+1}(\mathbf{a}_{ij})$ in period $t+1$. Since the DM is Bayesian, if $P_t\left(\mathbf{a}_{ij}\right) = 0$ then $P_s(\mathbf{a}_{ij}) = 0$ for all $s > t$.

Let us define $\gamma_t = \Delta\left(\left\{P_t(\mathbf{a}_{ij}|\omega)\right\}_{\omega\in\Omega}\right)$ as the distribution over all the possible conditional probabilities given the unconditional choice probability and the information struc-

ture chosen in period t. The assumption of Bayes updating restricts γ to be such that the expectation over conditional choice probabilities generates the unconditional choice probability, that is for every $\mathbf{a}_{ij} \in \mathbf{A}$

$$P_t\left(\mathbf{a}_{ij}\right) = E_{\gamma_t}[P_t\left(\mathbf{a}_{ij}|\omega\right)].$$

The cost of information is defined over γ and denoted by $K : \Delta\left(\Delta\left(\Omega\right)\right) \to \mathbb{R}$. We assume the function $K\left(.\right)$ satisfies the following three properties, namely, weak monotonicity, weak convexity and costless inattention. Monotonicity implies if γ and γ' are two posterior distribution such that γ is more Blackwell informative than γ', then

$$K\left(\gamma\right) \geq K\left(\gamma'\right).$$

Weak convexity implies, if $\gamma' = \alpha \circ \gamma_1 + (1-\alpha) \circ \gamma_2$ then for all $\alpha \in (0,1)$

$$K\left(\gamma'\right) \leq \alpha K\left(\gamma_1\right) + \left(1-\alpha\right) K\left(\gamma_2\right).$$

And costless inattention is a particular normalization where,

$$K\left(P\left(a\right)\right) = 0$$

i.e., if the DM pays no attention and chooses according to his unconditional probabilities then he would incur no cost of learning.[1]

For any partition i problem the cost of learning is defined by the same function $K(\gamma)$, but for all $P(\mathbf{a}_{ij}|\omega)$ in the support of γ would satisfy the *partition consistency* condition for partition $\mathcal{P}_t$ given in equation PC.

3.4 Choice Problem

We assume the DM has 1 unit of indivisible land and uses some input combination $\mathbf{a}_{ij}$ in each period to maximize his expected payoff net of cost of learning. The choice problem faced by a DM consists of choosing a strategy specifying the choice of partition in each period and corresponding γ_t, given μ_t such that it maximizes his net payoff.

Definition 1. *A strategy of the DM consists of a sequence of partitions* $\{\mathcal{P}_t\}_t \in \mathbb{P}$ *and distributions of conditional choice probabilities* $\{\gamma_t\}_{t\in T} = \left\{\Delta\left(\left\{P_t\left(\mathbf{a}_{ij}|\omega_j^t\right)\right\}_{\omega^t\in\mathcal{P}_t}\right)\right\}_{t\in T}$ *given* μ_t *such that* $P_t\left(\mathbf{a}_{ij}|\omega_j^t\right) : \mathbb{P} \times \Delta\left(\Omega\right) \to \Delta(\mathbf{A})$ *follows Bayes rule and equation PC satisfies for all* $\omega_k^t \in \mathcal{P}_t$.

[1] Caplin and Dean (2015) argues that these three assumptions on the cost function can be guaranteed if the the behavior of the DM satisfies NIAS and NIAC conditions. The NIAC and NIAS conditions implies that the choice problem of a ration DM can be represented as a maximization of payoff minus a cost of learning.

The DM's problem is to choose a strategy $\left\{\mathcal{P}_t, P_t\left(a_{ij}|\omega^p\right)\right\}_{t\in T}$ to maximize his sum of net expected payoffs over all partition problems, i.e.,

$$V(\mu_0) = \max_{\mathcal{P}_t\in\mathbb{P}}\left\{\max_{\{\gamma_t\}_{t\in T}}\sum_{t\in T}\left[E\left(\pi\left(P_t\left(a_{ij}|\omega_j^t\right)\right)\right) - K(\gamma_t)\right]\right\} \tag{AP}$$

Proposition 1. *The solution to problem AP exists.*

The proof of proposition 1 is given in the appendix. Proposition 1 guarantees that if the behavior of the DM can be modeled as if he is optimizing over input combination and attention strategies subject to an attention cst given by $K(.)$ then the solution to his problem exists.

3.5 Selective Attention

As described in the introduction selective attention is an attention choice where the DM pays attention to only a strict subset of inputs, which means only one input in this $2\times 2\times 2$ model.

When the DM observes only one input say A_i then he cannot distinguish between states where the difference in payoff is only due to the other input A_j. This generates a particular partition of the state space Ω. If the DM does not observe the level of input A_j then the DM's has a flat prior over the levels of A_j, i.e, he beliefs $a_{j1} = p$ is chosen with probability p for $p\in[0,1]$. This implies the DM believes the probability of choosing a_{j1} is 0.5 which same as that of a_{j2}. The formal definition of the *selective attention partition* is given below,

Definition 2. *A selectively attentive partition $\mathcal{P}_i$ for input A_i is a partition that satisfies the following conditions:*

1. *If two states ω_k and ω_l belong to the same block ω^p then the payoff from choosing any levels of A_i conditional on not observing A_j, i.e, when the marginal distribution over A_j is uniform, is same. This implies for all possible choices of input A_i the DM can not distinguish the two states without observing the level of A_j.*

2. *If two states belong to different blocks then there exist a level of A_i say a_{ij} such that the payoff from a_{ij} is different for the two states conditional on not observing A_j, i.e, when the marginal distribution of A_j is uniform.*

The definition gives us the two following selectively attentive partitions respectively for two inputs. $\mathcal{P}_{A_1} = \left\{\{\omega_1\},\{\omega_2,\omega_3\},\{\omega_4,\omega_5\},\{\omega_6\},\{\omega_7,\omega_8,\omega_9,\omega_{10}\},\ \{\omega_{11}\},\{\omega_{12},\omega_{13}\},\right.$ $\left.\{\omega_{14},\omega_{15}\},\{\omega_{16}\}\right\}$ and $\mathcal{P}_{A_2} = \left\{\{\omega_1\},\{\omega_2,\omega_4\},\{\omega_3,\omega_5\},\{\omega_7\},\ \{\omega_6,\omega_8,\omega_9,\omega_{11}\},\ \{\omega_{10}\},\right.$ $\left.\{\omega_{12},\omega_{14}\},\{\omega_{13},\omega_{15}\},\{\omega_{16}\}\right\}$.

Let us define selective attention in terms of the selectively attentive strategies as follows:

Definition 3. *A DM is selectively attentive towards input A_i, if he updated his belief by only choosing partition $\mathcal{P}_{A_i}$ in any period t.*

In the following section we will consider other economically relevant sequences of partition such as sequential learning of the two inputs.

4 Costly Attention Problem

In the earlier section we have shown that the solution to the attention problem exists under general assumptions on the cost function $K(\gamma)$. But to solve for properties of the attention choice we will now introduce more specific cost functions that satisfies the conditions discussed in the last section.

The main objective of this model is to analyze the impact of multiple dimensions on attention allocation. Multiple dimensions introduce two features to an otherwise standard choice problem with costly attention. First, if the dimensions are correlated, i.e, learning about one gives information about other dimensions then the DM can find it optimal to choose to learn only about a subset of dimensions. Since more informative attention strategies are more costly, presence of correlation can increase the net payoff of the DM.

To analyze this feature of the model we borrow the cost function from the *Rational Inattention*(henceforth RI) literature. In the RI models the most common cost functions is the Shannon mutual entropy cost which we discuss in detail in the next section. In these models the cost function is the Shannon mutual entropy between the prior belief and the expected posterior belief conditional on the attention strategy.

Note that DM's belief over correlation between the dimensions can be represented by a belief μ over the state space Ω. In the Shannon model the attention strategy depends on the prior belief μ and the payoff Y, so expressing correlation as a prior belief is sufficient to solve the DM's problem who can update his belief over correlation of dimensions.

The other interesting feature of multiple dimension is that the action space is a product space. This helps the DM to categorize the state space by considering projection of the action space over specific dimensions which would simplify the DM's problem. This feature is not present in the RI literature because the cost function is not directly related to the action space. For the RI cost structure a multi-dimensional attention problem is identical to an unidimensional choice problem with state space Ω and four possible actions. In other words the "name" of the states do not affect the RI cost functions.

To capture the second feature of the model we consider the type of attention problems discussed in the *coarse categorization* literature. In a coarse categorization model, in equilibrium DM chooses the same action across all states in the same category. If DM is optimally choosing the categorization then he faces a trade-off between a finer partition which helps him to reduce the variance within a category and coarser partition which reduces the bias in predicting for a category.

To introduce this trade-off we assume that different partitions have a costs associated with them and a finer partition is more costly. In this reduced form version of the cost

of categorization the DM trades off the benefit of a finer partition of a more informative attention strategy (less variance within a single category/partition) and the cost of choosing a finer partition. In this framework we interpret the cost as a cognitive cost of memory related to the complexity of the problem.

For example in Schwarzstein(2014) model the DM can choose to pay attention to one of the dimension, say A_1, without paying any cost but for the other dimension he faces a cognitive cost. If the cost is sufficiently high compared then DM does not encode the other dimension A_2 and categorize all values to A_2 together.

But in these coarse categorization models the objective of the DM is to predict the value of the output and not to maximize payoff, as a result the DM would choose an attention strategy that does not depend on the expected payoff but the strategy is focused to make his predictions correct. This generates the opposite result of the RI models where the attention strategy depends on the expected payoff.

To illustrate the difference let us consider the following extreme example. Suppose the prior belief of the DM is given by $\mu_0(\{\omega_1, \omega_{16}\}) = 1$, i.e, only the *constant payoff* states are possible. A DM with RI cost function would choose to pay no attention because choosing any input combination generates same payoff and learning is costly. But a DM who wants to predict the output would still pay attention to know whether it is ω_1 or ω_{16}. In general in a coarse categorization model given a categorization the DM would learn perfectly about the states which is never the case with RI models.

Thus the two types of models shows the two possible exterme cost of attention function. In the next two sections we first discuss the behavioral implication for the two extreme assumptions on the cost function. Then we will introduce a new cost structure which we call *costly partition* model combining these two exterme cost functions.

For the *costly partition* model we ask the following questions, namely,

1. If and when the DM would choose a coarse categorization?

2. Given a categorization what is the optimal attention strategy for the DM?

3. How does the two types of costs interact in the attention choice, more specifically when the DM chooses a coarser categorization whether he increases his attention to compensate?

4. How does the cost of attention interacts with the dynamic nature of the problem?

The main feature of the mulitdimensional problem that we want to emphasize here is that the natural partitioning of the state space based on the dimensions relaxes the attention problem of the DM by giving him an extra choice. In addition to how much to learn about as in an usual RI model the DM can choose what to learn about or in which order to learn about as well. The *costly partition model* intends to discuss this feature of attention.

4.1 Shannon Cost

The *Shannon cost* is similar to the RI models except the DM chooses a partition first then decides how much to learn given his prior belief. The cost function is given by,

$$C(\mu_t, \lambda) = \lambda \left[E_{\mu_t} \left(H \left(P_t \left(\mathbf{a}_{ij} | \omega_k^p \right) \right) \right) - H \left(P_t \left(\mathbf{a}_{ij} \right) \right) \right] \qquad \text{(Shannon)}$$

where λ denotes the marginal cost of learning and H is the negative of the entropy function, namely,

$$H(\mathbf{p}) = \sum_{p \in \mathbf{p}} p \ln p.$$

Note that, first, the cost function is the same for all partitions and defined over the divergence of the prior and expected posterior beliefs over the state space. In a coarser partition, however, the DM's attention choice is constrained as it has to satisfy the *partition consistency condition*. This imply the DM would never choose a coarse categorization.

Second, the cost is linear in Shannon mutual entropy so breaking up the learning problem into many parts, given a partition does not help the DM. As a result the DM would choose to learn as quickly as possible. Combining the two results we get the following theorem:

Theorem 1. *1. The optimal learning strategy has the following features:*

i. The original state space Ω, i.e. the finest possible partition is weakly better that all other partitions in any period t.

ii. There is no learning in any period $t \geq 1$.

iii. The conditional choice probabilities follow weighted logistic rule, i.e,

$$P(\mathbf{a}_{ij}|\omega) = \frac{exp(\sum_{t \in T} \pi(\mathbf{a}_{ij}|\omega)/\lambda) P(\mathbf{a}_{ij})}{\sum_{\mathbf{a} \in \mathbf{A}} exp(\sum_{t \in T} \pi(\mathbf{a}|\omega)/\lambda) P(\mathbf{a})}$$

where $P(\mathbf{a})$ is the unconditional probability of choosing action $\mathbf{a}$ that follows Bayes rule,

$$P(\mathbf{a}) = \sum_{\omega} \mu_0(\omega) P(\mathbf{a}|\omega)$$

2. Selective learning of input A_i is an optimal strategy if and only if at least one of the two conditions hold true, namely,

i. there exists a level of input A_i say a_{ij} such that unconditional choice probability $P(a_{ij}, a_{-i1}) + P(a_{ij}, a_{-i2}) = 1$,

ii. the prior probability of constant payoff states ω_1 and ω_{16} sums up to one.

The conditions for selective attention for input A_i imply that the DM would choose to not pay any attention to A_{-i} only when the extpected payoff from doing so is zero. This happens because of the convexity of the Shannon cost function in the choice probabilities. As a result when DM is choosing not to pay attention to A_{-i}, he would not change his behavior even when the DM is offered information about A_{-i} for free.

4.2 Coarse Categorization

In the coarse categorization literature the usual assumption is that the DM can choose a coarser categorization for a lower cost or there is a fixed number of categories which is smaller than the full state space (see Fryer and Jackson(2008), Mullinathan, Schwarzstein and Schleifer(2008)).

We start with Schwarzstein(2014) model where the DM faces a coarse decision making problem. In this model DM starts with a hierarchical prior, i.e., a prior over models of production and conditional on each model a prior over the payoff functions.

The four different models of the production function are $M_{12}, M_{1\neg 2}, M_{\neg 12}, M_{\neg 1n\neg 2}$. M_{12} implies both inputs are payoff relevant, $M_{\neg 12}$ implies input A_1 is not payoff relevant and input A_2 is not, $M_{1\neg 2}$ implies input A_1 is payoff relevant but A_2 is not and finally $M_{\neg 1\neg 2}$ implies none of the two inputs are payoff relevant. In the framework of our environment these four models represent blocks of a partition over state space. If both inputs are payoff relevant then changing the level of input for one should affect the payoff for a given level of other input.

We can rewrite the belief over four models in terms of belief over partitions of Ω. The four models corresponds to the following partition, $M_{\neg 12} = \{\omega_7, \omega_{10}\}$, $M_{1\neg 2} = \{\omega_6, \omega_{11}\}$, $M_{\neg 1\neg 2} = \{\omega_1, \omega_{16}\}$, $M_{12} = \{\omega_2, \omega_3, \omega_4, \omega_5, \omega_8, \omega_9, \omega_{12}, \omega_{13}, \omega_{14}, \omega_{15}\}$.

Let μ_{ij} denote the prior belief over model M_{ij}. Schwarzstein(2014) assumes that the prior belief has independence structure, namely, we can find μ_1 and μ_2 such that $\mu_{12} = \mu_1\mu_2$, $\mu_{\neg 12} = (1-\mu_1)\,\mu_2$, $\mu_{1\neg 2} = \mu_1\,(1-\mu_2)$ and $\mu_{\neg 1\neg 2} = (1-\mu_1)\,(1-\mu_2)$. In terms of Ω in our model we can write $\mu_2 = 1 - \mu_0(\{\omega_1, \omega_6, \omega_{11}, \omega_{16}\})$. The attention constraint works as follows: the DM can always encode input A_1 but he can only encode A_2 if $\mu_2 > b$.

Using this encoding strategy or attention allocation function Schwarzstein(2014) finds as as $t \to \infty$ with a time invariant cognitive bound b, when $b \leq \mu_2$ the DM either learns perfectly about the model but when $b > \mu_2$ he pays selective attention to input A_1, i.e., chooses $\mathcal{P}_{A_1}$ and only learns about the conditional model of production.

Note that μ_2 is Schwarzstein(2014) model is same as $\mu_0(\{\omega_2, \omega_3, \omega_4, \omega_5, \omega_7, \omega_8, \omega_9\,, \omega_{10}, \omega_{12}, \omega_{13}, \omega_{14}, \omega_{15}\})$. So the conditional for selective attention under

$$\mu_0(\{\omega_2, \omega_3, \omega_4, \omega_5, \omega_7, \omega_8, \omega_9\omega_{10}, \omega_{12}, \omega_{13}, \omega_{14}, \omega_{15}\}) < b$$

The interesting feature of the model is that given the DM chooses to observe the a partition he learns perfectly about it so the only type of mistake he can make is by choosing a coarser partition. Also, since the cost function only depends on the partition

the margin of the choice is always determined by the value of μ_2 and not the composition of μ_2.

However, the attention strategy in this model is exogenously given and thus does not depend on payoff from different actions which is opposite of the Shannon model. The following remark shows the importance of payoff relevant attention strategy.

Remark 1. *Consider the following two scenarios, safe, where the probability of bad states are zero and risky, where the probability of the good states are zero. If in both cases the value of μ_2 remains same then the attention strategy would remain the same, i.e., in both cases he either chooses Ω or $\mathcal{P}$.*

Two major features of Schwarzstein(2014) model that are not present in our model are the following: first, the DM is not facing a prediction problem, rather a choice problem conditional on his information about payoff at the time of choice. Second, in our model we do not assume any exogenously given attention function rather we want to define a cost function that would generate an optimal attention allocation. This leads to our next example which introduces categorization in a relatively similar framework that is used in our model.

For the coarse categorization the cost function only applies to partitions and does not depend on information content of the attention strategy. In line with Schwarzstein(2014) we assume a cost function where the DM can observe one of input, say A_1 for free but there is a cost of paying attention to input A_2. So the cost function given by,

$$c(\mathcal{P}) = \begin{cases} 0 & \text{for} \quad \mathcal{P} = \mathcal{P}_{A_1} \\ \bar{c} & \text{for} \quad \mathcal{P} = \Omega \\ M & \text{ow} \end{cases} \tag{3}$$

where $M >> 1$. Given this cost function we know that the effective choice of the DM consists of only $\mathcal{P}_{A_1}$ and Ω. The DM would choose to pay attention to Ω if and only if the expected benefit from observing Ω is higher than the cost $\bar{c}$.

Here the only cost of attention is due to the choice of partition, i.e., given a partition there is no extra cost of learning. The information structure is as follows: given a partition $\mathcal{P}$ if the true state is ω which belongs to block b_i, then the DM observes a random signal that reveals one state at random in b_i. For example, under $\mathcal{P}_{A_1}$ if the true state is ω_2, the DM observes ω_2 and ω_3 with equal probability.

Given updated belief the DM chooses a distribution of actions to maximize his expected payoff. In the previous case the DM would thus choose only $\mathbf{a}_{11}$ and $\mathbf{a}_{12}$ with equal probability.

Result 1. *Given cost function in equation* 3

1. *Given a partition the DM would learn perfectly about all blocks/states in the partition.*

2. *The DM would choose $\mathcal{P}_{A_1}$ if and only if*

$$\mu_0(\{\omega_7, \omega_8, \omega_9, \omega_{10}, \omega_{12}, \omega_{13}, \omega_{14}, \omega_{15}\}) < 2\bar{c}$$

3. *There exists values of μ_2 such that DM would not be selectively attentive in risky scenario, but the would be selectively attentive in safe scenario.*

The difference between this result and Schwarzstein(2014) shows the impact of two major assumptions. First the DM is not facing a predictive task and second the DM optimally chooses the attention strategy. Since the DM does not need to predict the outcome for all possible actions in *good* states he does not need to learn about all input combinations and can still ensure a payoff of 1. This generates the difference in the condition of selective attention in the two cases.

The major difference between this model and Shannon model is that the DM learns perfectly about the chosen partition in this model which is never true in Shannon model. This further implies that the attention strategy would not change if the payoff in *good* states change only. The attention strategy is payoff dependent in a restricted sense, i.e., it depends only on the payoff in *medium* and *bad* states.

Finally in this paper we want to explore when a DM who faces a cost of attention would be willing to buy information that he has chosen to not attend to. Schewarzstein(2014) claims that the DM would never do so because even if he is offered informatio about A_2 to encode the information he would have to pay the same cost b.

However that would not be the case in our model The cost of partition is interpreted as not a cost of encoding but rather a cost of memory or storage of information. If the DM has too high a cost of storage or memory then he would be willing to buy information that he has unattended before. In the next section we explore this possibility and analyze the impact of a market for information.

4.3 Costly Partition

Finally we combine the two types of the cost from the last subsection and construct a new cost of attention function. The additive cost function has two components. First, Shannon mutual entropy cost given a partition $\mathcal{P}$,

$$C\left(P_t\left(a_{ij}\right), P_t\left(a_{ij}|\omega_k^p\right)\right) = \lambda\left[E_{\mu_t}\left(H\left(P_t\left(a_{ij}|\omega_k^p\right)\right)\right) - H\left(P_t\left(a_{ij}\right)\right)\right]$$

Second component is a cost of choosing a partition, $c(\mathcal{P})$. The cost function $c(.)$ satisfies the follwoing assumptions;

Assumption 1. *The cost function $c(\mathcal{P})$ is weakly monotone in the coarsening order on $\mathbb{P}$, i,e., if $\mathcal{P}$ is coarser than $\mathcal{Q}$ then,*

$$c(\mathcal{P}) \leq c(\mathcal{Q}).$$

Assumption 2. *Inattention is costless, i.e.,*

$$c(\Omega_0) = 0.$$

Combining two functions we get,

$$K(\gamma) = E_{\mu_t}(H(P_t(\mathbf{a}_{ij}|\omega_k^p))) - H(P_t(\mathbf{a}_{ij})) + c(supp(\mu_t)) \tag{CP}$$

Assumption 1 is similar to the coarse categorization models, where choosing a finer partition is costly. The implicit cost in those models are often due to increased bias in prediction whereas we interpret the cost as increased complexity. Assumption 2 is a normalization that ensures that if the DM chooses not to learn anything then he does not pay any cost.

The two assumptions imply that the DM may find it optimal to choose a coarser partition because of the lower cost of partition. But the cost of choosing a coarser partition is that of reduced informativeness. The result is summarized in the next proposition. to prove proposition 2 we first need to prove the following lemmas.

Lemma 1. *For any cost of partition function $c(\mathcal{P})$ the DM would never choose the same partition more than once under any optimal strategy.*

Lemma 2. *For any cost of partition function $c(\mathcal{P})$ if the DM chooses a partition $\mathcal{P}$ in period t then it is not optimal to choose a partition coarser or finer to $\mathcal{P}$ in any future period $s > t$. Also, If the DM chooses no learning in period t then he would not choose to learn in any future period $s > t$.*

Proposition 2. *For a given cost function $c(\mathcal{P})$ satisfies assumption 1 and 2 and a marginal cost of attention λ there exists a set of prior belief μ_0 such that choosing the finest partition Ω is not optimal. Moreover, choosing a partition coarser than Ω is payoff relevant for the DM, i.e., the posterior distribution generated by the coarser partition is less information than that of Ω.*

The major difference between this model and the Shannon model is that choosing a coarser partition than Ω is payoff relevant for the DM here wheras under Shannon model since the cost function does not change with the partitions, Ω is always weakly better strategy, i.e, choosing a coarser partition is optimal only when it is payoff equivalent to the original stae space Ω. This observation generates the following implication:

Implication 1. *Suppose the DM chooses a coarser partition and does not observe some input combination $\mathbf{a}_{ij}$, then there exists a price $p > 0$ for which he would be willing to buy information about $\mathbf{a}_{ij}$.*

This is different from both the RI models and Schwarzstein(2014) where the DM would never buy information about an input he has not previously attended.

In the rest of the section we will characterize the optimal startegy of the DM given the cost function $c(\mathcal{P})$ and marginal cost λ in terms of prior belief μ_0.

Lemma 3. *If more than one partition is chosen in an optimal strategy then the DM would choose the partition that generates more informative posterior earlier.*

Proof of this lemma is in the Appendix. Orthogonality of the chosen partitions along with linearity of Shannon cost function in posterior beliefs ensures the result.

Lemma 4. *Suppose the DM chooses not to observe a subset of input combinations in period $t = 0$. If in any period $t \geq 1$ he chooses to observed previously unattended input combination then he would do so with the sequence of partition for which total cost of aprtition $\sum_{t\geq 1} c(\mathcal{P}_1)$ is lowest.*

Proof of lemma 4 is given in the Appendix. One implication of this lemma is that the DM would never choose to observe same input combination twice, however if he observes some level of only one input he may be willing to observe an input combination that contains that level of that input.

The intuition behind the proof as follows: if DM does not observe a set of input/input combinations in an earlier period then the participation consistency condition would apply for only the unattended combination. If in a later period the DM decides to observe already observed input combination he would optimally choose not to learn more about them.

This result holds true by LIP of Shannon cost and the assumption that all possible signal structures are available to the DM. Thus for all other entropy functions the result would remain true. In general if the cost of updating function is UPS (uniform posterior separable) then the result goes through. These two lemmas together characterize the possible sequence of partitions for any optimal strategy.

First we consider the possible strategies when the optimal attention strategy involves only one period and does not involve observing Ω.There are two types of partition, first, observe only one input and assume that other input is drawn from an uniform distribution and second observe a strict subset of input combinations. If we divide the five types of partitions according to this criterion, we get the following classification:

1. Does not observe both inputs simultaneously
   i. $\mathcal{P}_{a_{ij}}$: only observe one level of one input A_i, say a_{ij}
   ii. *Selective Attention*: $\mathcal{P}_{A_i}$, i.e, observe only input A_i
2. Observe both inputs simultaneously
   i. $\mathcal{P}_{\mathbf{a}_{ij}}$: observe only one input combination $\mathbf{a}_{ij}$,
   ii. $\mathcal{P}_{A_j|a_{ik}}$: observe only two input combinations $(\mathbf{a}_{k1}, \mathbf{a}_{k2})$ or $(\mathbf{a}_{1k}, \mathbf{a}_{2k})$,
   iii. $\mathcal{P}_{\neg\mathbf{a}_{ij}}$: does not observe input combination $\mathbf{a}_{ij}$.

Note that the first group of partitions can not be ranked with the second group of partitions but both groups are arranged according to coarseness of partition. For simplicity we assume that all partitions of a given types has same cost of partition, e.g., $c(\mathcal{P}_{a_{i1}}) = c(\mathcal{P}_{a_{i2}})$ for $i = 1, 2$. This assumption along with monotonicity of cost functions implies the follwing inequalities,

$$c(\mathcal{P}_{a_{ij}}) \leq c(\mathcal{P}_{A_i})$$
$$c(\mathcal{P}_{\mathbf{a}_{ij}}) \leq c(\mathcal{P}_{A_j|a_{ik}}) \leq c(\mathcal{P}_{\neg\mathbf{a}_{ij}}).$$

Next, we consider the possible strategies when the optimal strategies involves more than one period. There are three possible strategies available to the DM, first, always observe only one input at a time, always observe two inputs, i.e, some input combinations at a time and third observe one input in one period and both inputs in another period.

Note that lemma 3 ensures that it is not optimal for the DM to observe both inputs before at $t = 0$ and then only one input in next period(s). This is true because observing only one input is preferred when the prior belief about states where only one input combination is payoff relevant is higher. In that case by lemma 3 he should use the partition with one input only in the earlier period.

This gives the possible more than one period strategies:

1. **Choose $\mathcal{P}_{\mathbf{a}_{ij}}$ in period $t = 0$**: by lemma 3 the only possible period $t \geq 1$ strategy is to choose $\mathcal{P}_{ij}$ in period $t = 1$ and no learning thereafter.

2. **Choose $\mathcal{P}_{A_j|a_{ik}}$ in period $t = 0$**: there are three types of partition he can choose in period $t \geq 1$; observe the two other input combinations by choosing $\mathcal{P}_{A_j|a_{il}}$ where $i \neq j, k \neq l$, observe one unattended input combination at a time for next two periods or just one input combination in $t = 1$ and finally, observe $\mathcal{P}_{ail}$ in period $t = 1$ where $k \neq l$, i.e, observe one level l on input A_i which is not observed in period $t = 0$.

3. **Choose $\mathcal{P}_{\mathbf{a}_{ij}}$ in period $t = 0$**: there are two types of possible strategies, either observe other input combinations, i.e., $\mathcal{P}_{\mathbf{a}_{kl}}$ where $i \neq k$ and $j \neq l$ in subsequent periods or observe only one input a_{ij} in subsequent periods. This strategy can involve at most four periods of learning.

4. **Choose $\mathcal{P}_{A_i}$ in period $t = 0$**: there are two types of strategies available in subsequent periods;

   i. *Sequantial Attention*: observe A_j, i.e., choose $\mathbf{P}_{A_j}$ in period $t = 1$

   ii. *Conditional Sequential Attention*: observe A_j given a_{ik} in period $t = 1$ or observe $\mathbf{a}_{k1}$(or $\mathbf{a}_{1k}$) and $\mathbf{a}_{k2}$(or $\mathbf{a}_{2k}$) in periods $t = 1$ and $t = 2$ subsequently.

5. **Choose $\mathcal{P}_{a_{ij}}$ in period $t = 0$**: there are two types of strategies; observe other levels of two inputs a_{kl} in subsequent periods or observe other input combinations $\mathbf{a}_{ij}$ in subsequent periods. This strategy can also take at most four periods.

Remark 2. *Note that if $c(\mathcal{P}) \equiv 0$ then Ω is weakly better than all possible strategies noted above. This implies given any cost of partition $c(\mathcal{P})$ for all $\mathcal{P} \in \mathbb{P} \setminus \Omega$ and λ and for all prior beliefs μ_0 there exists a value of $\bar{c}$ such that if $c(\Omega) > \bar{c}$ optimal attention strategy takes at least two periods and if $c(\Omega) \leq \bar{c}$ optimal attention strategy takes only one period.*

Remark 3. *Under any attention strategy the coarseness of chosen partitions increases with the number of oeriods it takes to learn. Thus if the DM chooses a strategy that involves s period his expected error would be higher than a strategy that involves $t < s$*

periods for a given a value of λ. This is opposite to the usual learning models where the longer the DM learns the more informative his posterior belief would be.

This apparent counter-intuitive result is due to the fact that in this model the DM trades off between complexity of the problem with theprecision and time it takes to update his belief. A more complex strategy would give higher precision of posterior distribution but involves a higher cost of partition. If the complex strategy is broken into many parts, each part would generate lower cost of partition however the DM would sacrifice some level of precision in posterior and the learning would take longer which generates further loss in payoff.

Now we find the condition under which *selective attention* is optimal. For theorem 2 let us assume that the cost of partitions is symmetric across all input combinations, i.e, for any two partition that requires observing same number of inputs/input combinations generate the same cost of partition. This is a simplifying assumption and does not affect the results qualitatively.

Theorem 2. *For any prior belief μ_0, selective attention strategy is optimal if the following conditions hold true,*

1. *the exists $\bar{c} > \underline{c} \geq 0$ such that the ratio of the cost of partitions for the selectively attentive strategy $\mathcal{P}_{A_i}$ and the fully attentive strategy Ω is given by,*

$$\underline{c} \leq \frac{c(\mathcal{P})}{c(\Omega)} \leq \bar{c}$$

2. *the conditions for optimality of one input partition as described in result 2 holds true.*

The proof of theorem 2 is given in the Appendix. Since Ω always generates weakly more informative posterior choice distribution the lower bound is intuitve. The non-monoticity arises from the fact that the DM can choose to learn over multiple periods if $c(\Omega)$ is too high compared to $c(\mathcal{P})$.

Next we consider two other interesting strategies, namely *sequential attention* and *conditional sequential attention*. Under the former the DM observes both input in two periods but never observes two inputs together. We would call this a *breadth* strategy as the DM observes all levels of all inputs over multiple periods. On the other hand under the latter strategy the DM observes only one input in period $t = 0$ and based on his signal he observes all input combinations for a given level of input A_i in period $t = 1$. We would denote this strategy as a *depth* strategy since the DM observes in depth all combinations given a particular level of an input. The next result shows the trade-off between the two strategies.

Proposition 3. *Sequential attention is preferred over conditional sequential attention if and only if*

1. *All good and bad states are symmetric in terms of prior probability, i.e, Δ_G and Δ_B(refer equation 5) is sufficiently small.*

2. *The middle states are not symmetric,i.e., Δ_M is sufficiently high. More specifically the middle states where both inputs are payoff relevant,i.e, ω_8 and ω_9 has relatively(compared to $\mu_0(M)$) small probabilty.*

The proof of this propostition is given in the Appendix. Note that the propostition only compares the expected payoff from the two strategies but none of them need not be the optimal stratgy. However these two strategies are comparable because for both strategies require the DM to use same number of observations and takes same number of period with period $t = 0$ being the same. If we assume that the two strategies generate the same cost of partition, $c(\mathcal{P})$, then there exists a range of values of $c(\mathcal{P})$ for which for any given μ_0 one of these two strategies would be optimal.

Another feature of this model which is also found in the coarse categorization models is that the updated belief of the DM increases the variance across the blocks but reduces the variance within a block. If ω_i and ω_j belongs to the same block then the ratio of prior probabilities is same as that of posterior probbilities. But if the two ω's belong to different blocks then the ratio of the posterior belief can be higher than that under full learning with Ω. For example under selective attention the DM attributes all payoff variance to one input and never updates belief about other input.

In the next section we will discuss different markets for information. The distinction between these two strategies would be useful in that context. A conditional sequential strategy requires observing input combinations based on information obtained in the first period whereas sequential attention strategy is same irrespective of the period $t = 1$ belief.

4.4 Market for Information

In the previous section we have already shown that the DM would choose a coarser partition if the cost of observing original state space Ω is too high. As a result the DM learns only about a subset of inputs or input combination which is payoff relevant for him. Thus a market for information can increase the precision and expected payoff of the DM. In this section we will describe different types of market for information that would affect DM's decision.

In this model we interpret the cost of partition as a cognitive cost related to memory or storage of information. If another economic agent can store/report information for a lower cost, the DM may be willing to buy this information.

For example, suppose the DM does not observe input A_2 but obseves all levels of A_1 because it is too costly to observe all input combinations. Let μ_1 be the belief in period $t = 1$ after the DM updates his belief based on $\mathcal{P}_{A_1}$. We will consider two types of information provision, conditional and unconditional.

Conditional information provision means given the DM choice of attention strategy he can buy information about A_2 conditional on his belief over a_{1i}. For example, if after the period $t = 0$ learning the DM chooses a_{11}, then in period $t = 1$ he can choose to learn about A_2 conditional on a_{11} with $\mathcal{P}_{A_2|a_{11}}$ which requires observing $\mathbf{a}_{11} = (a_{11}, a_{21})$ and $\mathbf{a}_{12} = (a_{11}, a_{22})$.

Unconditional information provision only alllows the DM observes $\mathcal{P}_{A_2}$, i.e, the expected payoff from using two levels of A_2, where the expectation is defined over levels of A_1.

Continuing our assumption that the two types pf information provision generates same cost of partition we analyze which type of information provision would make the DM better off.

Given proposition 3 unconditional information provision generates higher net payoff if the prior belief is such that $\mu_0(\omega_6, \omega_7, \omega_{10}, \omega_{11}) \approx \mu_0(M)$ and the prior probabilities of all *good*(or *bad*) states are sufficiently close. If the probability of one or two *good* (or *bad*) states are higher then conditional information provision is strictly better for the DM.

Since conditional information provision depends on the period $t = 0$ learning it may lead to higher wrror probability. As $\lambda > 0$, the DM never learns perfectly about any state. This implies even if a_{12} generates a higher expected payoff there is positive probability that the DM chooses a_{11} given his information in period $t = 0$. This means if the DM makes a mistake in the period $t = 0$ then he would not learn about his mistakes in period $t = 1$ by choosing the wrong conditional information.

This generates a trade-off between choosing a conditional and the unconditional information strategies. The unconditional information strategies gives information about A_2 for all values of A_1, which reduces the probability of mistake for low probability *good*(*bad*) states compared to conditional strategies however it increases the probability of mistake for high probability *good*(*bad*) states. This is a breadth vs depth trade-off.

This trade-off is however not present in any of the earlier models. In rational inattention model the DM always weakly prefers Ω so this consideration does not exist. For coarse categorization models on the other hand the DM treats conditional information strategy strictly better since he never makes a mistake conditional on observing a partition.

Note that the conditional information only uses one level of A_1, say a_{11} and not for other level a_{12}. If λ is sufficiently low then DM is less likely to make mistake in period $t = 0$ which increases the benefit from using conditional information provision. Thus the two types of cost are complement to each other in this case. A lower level of λ enourages the DM to undertake conditional information strategies which further reduces the probability of making mistakes.

These two types of information provision is similar to two real life scenarios in agriculture. Agricultural extension program helps the farmers learn from his experience in his own land, which is similar to conditional information strategies. Whereas unconditional information provision strategies are similar to giving general information about an input to the farmer from the experience of his neighbours/results found by researchers. This model suggests that a conditional information provision is better for more able/educated farmers who has a lower λ.

However it is not unambiguously true that providing any of these two types of information would make the DM better off. Suppose the cost of partition function such that $c(\Omega)$ is not too high compared to other partitions $\mathcal{P}$. In absence of any market for information the DM optimally chooses to observe Ω.

In this scenario suppose conditional information provision is offered at a price $p < c(\mathcal{P}_{A_i|a_{-ij}})$. If p is low enough the DM would optimally choose $\mathcal{P}_{A_i}$ in period $t = 1$ and $\mathcal{P}_{A_{-i}|a_{ik}}$ in period $t = 1$. Even though it increases the net payoff of the DM the probability of making mistake would weakly increase under the new strategy.

It is often argued in the agricultural economics literature that learning by a farmer has a positive externality effect on other farmers. Thus providing information can actually reduce the welfare of the economy by encouraging individual farmers to learn less.

The reason behind this apparently paradoxical result is that the optimal strategy is not achieved where the probability of mistake is minimized but where net payoff in maximized. If cost of a coarser partition is reduced then the net gain to the DM from the reduced cost can be sufficiently high so that he switches tousing a coarser partition leading to less learning. Thus the market for information can work as a substitute rather than a complement to private learning by the DM.

However if the market for information is such that it affects the marginal cost λ then there is an unambiguous effect on welfare. For any given cost function $c(\mathcal{P})$ and prior belief μ_0 if λ is reduced by either educating the DM or giving them already processed information then the DM would always learn more. As a result both DM's individual net expected payoff would go up and through the positive externality of learning the social welfare would also be higher.

5 Literature review

This papers mainly relates to *rational inattention* and *coarse categorization* literature. In the RI literature Sims(2000), Caplin and Dean (2015), Matejka and McKay(2015) and Caplin, Dean and Leahy (2017) consider Shannon mutual entropy cost in an uni-dimensional decision problem. Given this cost function Matejka and McKay(2015) showed that in a static discreet choice environment the choice follows a weighted logistic model. Matejka, Steiner and Stewart(2017) extends the model in a dynamic choice problem and found that the choice is similar to a dynamic logistic model.

All these RI models consider that the DM is choosing between actions without any multi-dimensionality of action space. One major exception is Matejka and Tabellini (2016). They consider that the DM needs to learn about multiple attribute but the payoff from different attributes are additively separable. This implies the attention problem for each dimension can be solved independently.

Shannon cost implies in a multi-dimensional choice problem would be equivalent to a uni-dimensional choice problem as long as there is one-to-one mapping between the two state spaces which is a major departure in our model

This feature that i.e., the cost function depends on the description of the state space is similar to Woodford and Herbert (2017). In their model the DM faces a cost function where states that are "close" to each other are harder to learn about than states that are further apart. The measure of "closeness" is given by the perceptual distance between different alternatives. They show this feature of the cost function generates smoothness in the attention strategy along the sequence of possible actions.

However, in this paper we want to discuss the phenomenon of *selective attention* where the attention strategy can jump from observing one attribute to two attributes. The Fisher information matrix cost function that is used in Herbert and Woodford(2017) can not generate this type of discontinuity, i.e., along the number of dimensions to pay attention to.

Nieuwerburgh and Veldkamp(2009A, B) discusses a rational DM who faces an information processing constraint and has to choose between multiple assets for his portfolio decision which is similar to the multi-dimensional choice problem discussed here. Even though the assets can be correlated they assume that there exists a set of independent shocks in the economy that generates the payoff for all assets. So learning about that set of shocks is sufficient. They further restrict the choice of DM by assuming that he can only choose to observe independent signals for different assets/shocks.

Using Shannon cost of attention function they show that the DM would choose to learn about only one type of asset which can generate underinvestment or home bias depending on the context. They emphasize the difference between this result and the result from a pure prediction model where the DM only want to learn for prediction and does not choose any action based on his prediction. The prediction models generate more learning than their model and thus fail to explain real life scenarios.

Mondria (2010) discusses a model of rational DM who faces a portfolio choice problem similar to Nieuwerburgh and Veldkamp(2009A) but he allows the DM to choose a linear combination of two assets as a signal in addition to learning about them separately. He finds that a positive measure of DM would choose a linear combination of assets as a signal even when the assets are independent. This generates a type of bias in their information where an increase variance of one asset affects the posterior variance the other independent asset.

The formulation of the partition based state space in our model is similar to Hong and Page (2009). They describe a learning model that distinguishes between interpreted signals and generated signals and assume that the DM updates based on the interpreted signal instead of generated signal. The state space has a product space structure where each component is considered as an attribute and the DM can choose any partition to interpret his signal which is represented by a projection of the state space. Moreover there is no notion of cmplexity which implies the DM never faces any complesity-precision trade-off.

The *coarse categorization* models discusses the impact of complexity of the learning problem on the final attention choice of the DM. Mohlin (2014) describes that a rational DM in a prediction task would choose to coarsely categorize information to trade off variance within categories and bias in prediction in a category. This type of categorization is optimal for the DM as the cost of categorization is endogenously generated by the bias in prediction for a finer category.

Fryer and Jackson(2008) and Mullianathan, Schwarzstein and Schleifer(2008) assume that the DM faces an exogenous cost of categorization. In the context of a prediction problem both these papers show that if a rational DM coarsely categorize objects then he attribute the properties of one into other objects. This may result overreac-

tion/underreaction to information or systematic bias against minority group of objects that are more likely to be grouped together.

This paper is closest to Schwarzstein(2014) which we have already discussed in detail. There are two major differences between this paper and Schwarzstein(2014). First, the cost of attention in his paper does not depend on informativeness of the attention strategy which implies *selective attention* is the only type of mistake and it is payoff irrelevant,

Second, in this paper we consider the DM faces a choice problem instead of a prediction problem. As noted earlier by Nieuwerburgh and Veldkamp(2009A) if the DM's payoff depends on the prediction and does not involve any choice of action then the resulting attention would be different than the case when he chooses an action based on his prediction.

The multi-attribute nature of the choice problem discussed in this paper is similar to the context dependent choice problems as well. In the context dependent choice problems however the dimensions are additively separable. The choice of DM is biased as he pays more attention to certain dimensions even though the payoff relevance of all dimensions are same. Even though the DM exhibits selectively more attention to some attributes, attention is not chosen by the DM in these models and the attention strategy does not depend only on the payoff relevance of the dimension but the variance of the choices available for the dimension.

6 Conclusion

We construct a model of multi-attribute choice problem where the DM faces an attention constraint. In this model we analyze the accuracy-complexity trade-off faced by the DM. One way the DM can reduce complexity is by breaking the problem is several parts across multiple time periods which generates an interestinf time accuracy trade-off.

We then consider two types of attention strategies, namely breadth and depth strategy and show that a breadth strategy is preferred when the *good* (or *bad*) states have similar probability or only one of the input is payoff relevant. Then we discuss the implication of introducing a market for information on the DM's choice of attention strategy.

Next we want to test the implication of this model in a laboratory environment and analyze which type of cost function are closest to the actual decision making choices by economic agents.

Bibliography

Beaman, L., Magruder, J., and Robinson, J. (2014). Minding small change among small firms in kenya. *Journal of Development Economics*, 108:69–86.

Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., and Roberts, J. (2013). Does management matter? evidence from india. *The Quarterly Journal of Economics*, 128(1):1–51.

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

Caplin, A., Dean, M., and Leahy, J. (2017). Rationally inattentive behavior: Characterizing and generalizing shannon entropy. Technical report, National Bureau of Economic Research.

Francetich, A. and Kreps, D. M. (2016a). Choosing a good toolkit, i: Formulation, heuristics, and asymptotic properties.

Francetich, A. and Kreps, D. M. (2016b). Choosing a good toolkit, ii: Simulations and conclusions.

Hanna, R., Mullainathan, S., and Schwartzstein, J. (2014). Learning through noticing: Theory and evidence from a field experiment. *The Quarterly Journal of Economics*, 129(3):1311–1353.

Hébert, B. and Woodford, M. (2017). Rational inattention and sequential information sampling. Working Paper 23787, National Bureau of Economic Research.

Hong, L. and Page, S. (2009). Interpreted and generated signals. *Journal of Economic Theory*, 144(5):2174–2196.

Karlan, D., McConnell, M., Mullainathan, S., and Zinman, J. (2016). Getting to the top of mind: How reminders increase saving. *Management Science*, 62(12):3393–3411.

Matejka, F. and McKay, A. (2014). Rational inattention to discrete choices: A new foundation for the multinomial logit model. *The American Economic Review*, 105(1):272–298.

Matějka, F. and Tabellini, G. (2016). Electoral competition with rationally inattentive voters.

Mondria, J. (2010). Portfolio choice, attention allocation, and price comovement. *Journal of Economic Theory*, 145(5):1837–1864.

Mullainathan, S., Schwartzstein, J., and Shleifer, A. (2008). Coarse thinking and persuasion. *The Quarterly journal of economics*, 123(2):577–619.

Mullainathan, S. and Shafir, E. (2013). *Scarcity: Why having too little means so much*. Macmillan.

Schwartzstein, J. (2014). Selective attention and learning. *Journal of the European Economic Association*, 12(6):1423–1452.

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

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

Van Nieuwerburgh, S. and Veldkamp, L. (2009). Information immobility and the home bias puzzle. *The Journal of Finance*, 64(3):1187–1215.

Van Nieuwerburgh, S. and Veldkamp, L. (2010). Information acquisition and under-diversification. *The Review of Economic Studies*, 77(2):779–805.

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

A Appendix

A.1 Proof of Proposition 1

Proof. The strategy set of the DM is given by $\theta = \left\{\Pi_{t\in T}\mathbf{A}^{\mathbb{P}\times\Delta(\Omega)}\right\}_{P_t\in\mathbb{P}^t}$. Each θ_i is compact By Tychonoff's theorem and $\mathbb{P}$ and T being finite we can consider finitely many choice problems, one for each path of partition strategies.

We start by considering only one such θ_i which is compact. By assumption, π is uniformly bounded by 1 and 0. Also, K is assumed to be weakly convex which implies it will be continuous over all posterior belief distribution in the interior of $\Delta\left(\Delta\left(\Omega\right)\right)$. Also, K is assumed to be bounded below. This implies $u - K$ is bounded above and continuous in the interior of the compact set.

Continuity of the objective function in the interior of a compact set implies sup exists for the interior compact set, θ_i. This result along with the bounded above condition implies maximum exists as well. Since there are only finitely many such θ_is, the maximum i the product space θ exists as well. Hence, proved. □

A.2 Main Results: Proof

A.2.1 Proof of Lemma 1

Proof. Let us consider that the DM chooses $\mathcal{P}$ in some period t and $t+1$. WLOG let us assume that $t = 0$ and the DM chooses not learning for any period $t > 1$. It is enough to show that this strategy is not optimal for the following cost of partition function,

$$c(\mathcal{Q}) \equiv 0 \quad \forall \mathcal{Q} \in \mathbb{P}$$

Let us assume for simplicity partition $\mathcal{P}$ has only two blocks, say ω_h and ω_l and the prior belief at $t = 0$ is given by μ_0. Under strategy 1 in period $t = 0$, the DM suppose chooses two possible posterior distributions $\left\{P_0(\mathbf{a}_{ij}|\omega_h), P_0(\mathbf{a}_{ij}|\omega_l)\right\}$. Blackwell informativeness criterion implies the DM would choose only two signals h and l.

At the beginning of period $t = 1$ after the realization of period $t = 0$ signal the DM would update his belief over $\{\omega_h, \omega_l\}$. Let us these two possible beliefs as $\mu_{1|h}$ and $\mu_{1|l}$ corresponding to the h and l signal. The DM would choose the attention strategy in period $t = 1$ based on these beliefs, which would generate four possible posterior distribution of actions, i.e, $\left\{P_{1h}(\mathbf{a}_{ij}|\omega_h), P_{1h}(\mathbf{a}_{ij}|\omega_l), P_{1l}(\mathbf{a}_{ij}|\omega_h), P_{1l}(\mathbf{a}_{ij}|\omega_l)\right\}$, where the subscript $1j$ represents at the beginning of period $t = 1$ the DM has updated belief based on the realization of signal j in period $t = 0$.

Consider the following alternative strategy: the DM chooses to learn only in period $t = 0$ and uses the signal structure to obtain the following posterior distributions: $\left\{pP_{1h}(\mathbf{a}_{ij}|\omega_h) + (1-p)P_{1l}(\mathbf{a}_{ij}|\omega_h), pP_{1h}(\mathbf{a}_{ij}|\omega_l) + (1-p)P_{1l}(\mathbf{a}_{ij}|\omega_l)\right\}$ where $p = \mu_0(\omega_h)$ is the probability of state ω_h at the beginning of period $t = 0$.

For the $t = 0$ decision problem the cost of these two strategies are the same. The cost function is given by the Shannon mutual entropy which is linear in posterior probability.

So the cost from first strategy is

$$C(\mu_0,\lambda)=\lambda(\underbrace{E_{\mu_0}(H(P_0(\mathbf{a}_{ij}|\omega_s)))-H(P(\mathbf{a}_{ij}))}_{\text{cost in period } t=0}+$$

$$\underbrace{p(E_{\mu_{1|h}}H(P_{1h}(\mathbf{a}_{ij}|\omega_s))-H(P_0(\mathbf{a}_{ij}|\omega_s))+(1-p)(E_{\mu_{1|l}}H(P_{1l}(\mathbf{a}_{ij}|\omega_s))-H(P_0(\mathbf{a}_{ij}|\omega_s)))}_{\text{cost in period } t=1}$$

$$=\lambda(p(E_{\mu_{1|h}}H(P_{1h}(\mathbf{a}_{ij}|\omega_s))+(1-p)(E_{\mu_{1|l}}H(P_{1l}(\mathbf{a}_{ij}|\omega_s)))-H(P(\mathbf{a}_{ij})))$$

and cost for the second strategy is given by

$$C(\mu_0,\lambda)=\lambda(E_{\mu_1}H(P_{1s}(\mathbf{a}_{ij}|\omega))-H(P(\mathbf{a}_{ij})))$$
$$=C(\mu_0,\lambda)=\lambda(p(E_{\mu_{1|h}}H(P_{1h}(\mathbf{a}_{ij}|\omega_s))+(1-p)(E_{\mu_{1|l}}H(P_{1l}(\mathbf{a}_{ij}|\omega_s)))-H(P(\mathbf{a}_{ij})))$$

Hence the two strategies generate exact same cost of attention. If the cost of partition $\mathcal{P}$ is greater than zero then the first strategy is more costly than the second strategy because the first strategy requires the DM to observe the same partition in two periods.

The expected payoffs from these two strategies in periods $t \geq 1$ are the same since the DM is Bayesian, i.e, maximizes expected utility. In addition under the second policy the DM starts to enjoy the benefit from learning one period earlier, i.e, from $t=0$. This implies the second strategy weakly dominates the first strategy hence observing the same partition twice can not be optimal.

However, we need to ensure that the second strategy has to be dynamically consistent, i.e, if the DM updates his belief to some μ_1 at the end of period $t=0$ then he would not choose to learn using the partition in period $t=1$. This condition is satisfied by the *Locally Invariant Posterior*(LIP) property of Shannon cost function.

Under our assumption on cost of partition $c(\mathcal{Q})$ this problem is equivalent to a Shannon problem over the partitions. The LIP property ensures that the choice of posterior is invariant to changes to prior and action space. If there are two decision problem characterized by two possible priors μ_1 and μ_2 where the two priors are connected, i.,e μ_2 lies in the convex hull of posterior generated by μ_1 then the posterior distribution under the prior μ_1 would remain as optimal posterior even under μ_2 whenever the posterior is available under μ_2.

Since we do not restrict the possible information structure for a partition other than the partition consistency condition, possible posteriors available to the DM remains same as long as the partition remains same. Hence, at $t=1$, with the updated belief $\mu_{1|s}$, where $s=h,l$ the optimal posterior remains the the same, i.e., he would optimally choose not to learn.

However, for any cost function with $c(\mathcal{P})>0$ LIP does not hold true in general. If the DM chooses different partitions in different periods then he can choose different posterior distribution even when the posterior distribution from the earlier period is available because choosing a different partition implies that he can choose other posterior distributions that were not previously available.

But the locally invariant posterior property nonetheless holds true given a partition. Since the cost of partition $c(\mathcal{Q})$ does not depend on the posterior chosen optimal conditional on a partition, for $\mu_{1|s}$ the posterior chosen under μ_0 remains available. Hence the result holds true for any cost of partition function $c(\mathcal{Q})$.

Note that, this result is not only true for the cost function we have discussed in the *costly partition model* but also true for any cost of attention function that satisfies LIP and are posterior separable. First, if a cost function satisfies LIP then observing one partition only once is dynamically consistent.

Also, following Caplin, Dean and Leahy (2017) we know that if a cost of attention function is posterior separable and satisfies LIP then it is uniformly posterior separable (UPS), i.e, the cost function can be written as,

$$C(\mu_0) = E_\mu G(P(\mathbf{a}_{ij}|\omega)) - G(P(\mathbf{a}_{ij}))$$

where $G(.)$ is a strictly convex function. For UPS cost function, let us consider the two similar strategies, strategy one: choose partition $\mathcal{P}$ in both period $t = 0$ and strategy two, choose partition $\mathcal{P}$ only in $t = 0$. Suppose under strategy 2 the DM chooses the same posterior distribution in period $t = 0$ as he would do under strategy 1 in period $t = 1$. Here we assume that the same posterior distribution is available under both strategies which is trues for UPS cost functions since we do not impose any restriction on the information structures and G is independent of prior μ and the partition $\mathcal{P}$.

By linearity of the cost function and the assumption that the G(.) function does not change with change in μ the two strategies would generate same cost of learning. Similar to the earlier case the DM strictly prefers to learn using strategy 2 because he can use the information for one more period. Hence he would strictly prefer strategy 2. □

A.2.2 Proof of Lemma 2

Proof. For this proof we again assume WLOG that the DM chooses $\mathcal{P}$ in period $t = 0$ and $\mathcal{Q}$ in period $t = 1$ and no learning in any period $t > 1$. First, we assume that $\mathcal{Q}$ is finer than $\mathcal{P}$. In this case there are some posterior distribution that are available with $\mathcal{Q}$ but not under $\mathcal{P}$.

Let μ_0 denote the prior belief in period $t = 0$ and $\mu_{1|s}$ denote the belief at the beginning of period $t = 1$ if signal s is observed in period $t = 0$. Consider another strategy where the DM chooses $\mathcal{Q}$ in period $t = 0$ and no learning for any period $t > 1$.

The LIP property given a partition implies the DM will choose the same posterior belief under the first strategy in period $t = 1$ when he observes $\mathcal{Q}$ and under the second strategy in period $t = 0$. Thus the payoff from $t \geq 2$ periods are the same in both periods.

The first strategy would be better if the net benefit from breaking down the learning problem into a coarser and finer partition is higher than that of using only the finer partition under the second strategy. However, this is not the case for any cost of partition function $c(\mathcal{P})$.

Let us first assume that $c(\mathcal{Q}) \equiv 0$, i.e., the DM faces only the Shannon cost. Similar to the earlier lemma since the cost is a linear function of Shannon entropy the two strategies

would generate the same cost given by

$$C(\mu_0, \mathcal{Q}) = \lambda(E_{\mu_0}(E_{\mu_{1|s}} H(P_{1s}(\mathbf{a}_{ij}|\omega_s)) - H(P(\mathbf{a}_{ij})))$$

where $\mu_{1|s}$ is the conditional distribution of states after observing signal s in period $t = 0$ under strategy 1 and $P_{1s}(\mathbf{a}_{ij}|\omega_s)$ is the corresponding posterior distribution. Hence the DM would strictly prefer the second strategy if the cost of partition $\mathcal{P}$ is strictly positive.

Also, since the partition $\mathcal{Q}$ is coarser, by partition consistency assumption under the first strategy in period $t = 0$ the DM can only have a less informative posterior compared to period $t = 0$ with partition $\mathcal{Q}$, i.e, his net expected payoff in $t = 0$ would be lower under strategy 1.

A similar logic can be applied in the case where $\mathcal{Q}$ is coarser than $\mathcal{P}$. We compare this strategy with observing only $\mathcal{P}$ in period $t = 0$. Since under both strategies in period $t = 0$ the DM observes $\mathcal{P}$ the DM would only be better off if by observing $\mathcal{Q}$ in period $t = 1$ he can either reduce the cost of attention or increase the expected value.

By the partition consistency condition we know that all possible strategies that are available under $\mathcal{Q}$ is also available under $\mathcal{P}$ but not the other way round. Suppose under the first strategy the beginning of $t = 1$ period belief is $p_{1|s}$ and the DM chooses some posterior distribution $P_{1|s}(\mathbf{a}_{ij})$. Then under strategy 2 this posterior belief would still be available so the net expected payoff from posterior under second strategy can only be weakly higher.

However under the first strategy if $c(\mathcal{P}) > 0$, then the first strategy has an additional strictly positive cost of attention which implies that the first strategy would be strictly worse than the second strategy.

To prove the final part of lemma we note that not learning is represented by choosing a partition with only one block, i.e., the whole state space, namely $\mathcal{P}_0$. Since this partition is coarser than all possible partitions the DM would not choose the partition $\mathcal{P}_0$ before any other partition.

However the DM can choose $\mathcal{P}_0$ after any finer partition. This is true for the following two reasons, first, under $\mathcal{P}_0$ the Shannon cost is zero as prior probability is same as posterior probability. Second, the cost of this partition $c(\mathcal{P}_0) = 0$, i.e., the DM can strictly prefer to choose this partition if for all other partition $c(\mathcal{P}) > 0$. The zero cost assumption imply that when the DM chooses not to change his belief no learning, i.e, $\mathcal{P}_0$ is the best strategy. Hence, proved.

□

A.2.3 Proof of Theorem 1

Part 1:

First let us show that Ω is a weakly better strategy at any period t for any marginal cost $\lambda > 0$. First, we note that the Shannon mutual entropy satisfies the LIP property which implies the cost function does not depend on the prior belief or the state space over which the prior is defined. This implies the DM faces the same cost function for every possible partition. For a finer partition the expected entropy cost is calculated over

more states(blocks) but the additivity of the cost function in distribution of posteriors ensure the cost function does not change for the DM.

Given this observation we can rewrite any partition problem as the original problem with Ω along with the *partition consistency condition*. So any partition problem becomes a constrained original problem under this cost function. Hence the optimal solution can not be strictly better than that of Ω. So for any period t and any value of $\lambda > 0$ we get Ω to be weakly better than all possible partitions in $\mathbb{P}$.

Given this result we now use lemma 1. Lemma 1 shows the DM would never choose to observe the same partition in two periods, this implies if the DM weakly prefers to use Ω over all possible partition he would only use it in period $t = 0$ and not learn anything after $t = 0$.

Note that Ω can be indifferent to some other partition in period $t = 0$ if partition consistency condition does not bind. In that case the DM can choose the coarser partition say $\mathcal{P}$ in period $t = 0$. Since partition consistency condition does not bind, the two partitions would generate the same posterior under both partitions and hence using lemma 1 and lemma 2 we get the DM would only choose to observe $\mathcal{P}$ in period $t = 0$ and no learning thereafter.

Given the first two statement of the proof we can consider the DM's problem as a static decision problem with partition Ω where the payoff is obtained for all $t \in T$ periods. So the relevant payoff from action $\mathbf{a}_{ij}$ in state ω is $\sum_{t\in T} \pi(\mathbf{a}_{ij}|\omega)$. Thus the T period choice problem reduces to a static decision problem with the updated payoffs. We can apply the result of Matejka and McKay(2015) and show that the conditional probability of choosing action $\mathbf{a}_{ij}$ is given by

$$P(\mathbf{a}_{ij}|\omega) = \frac{exp(\sum_{t\in T} \pi(\mathbf{a}_{ij}|\omega)/\lambda)P(\mathbf{a}_{ij})}{\sum_{\mathbf{a}\in\mathbf{A}} exp(\sum_{t\in T} \pi(\mathbf{a}|\omega)/\lambda)P(\mathbf{a})}, \tag{4}$$

where $P(\mathbf{a})$ is the unconditional probability of choosing action $\mathbf{a}$ that follows Bayes rule,

$$P(\mathbf{a}) = \sum_{\omega} \mu_0(\omega)P(\mathbf{a}|\omega).$$

Part 2:

Given statement 1 in part 1 we conclude that the *selective attention* partition $\mathcal{P}_{A_1}$ is optimal only if it is indifferent to Ω. If all states except constant probability states, i.e., ω_1 and ω_{16} then there is no benefit from learning so all partitions are indifferent to each other which gives the first condition.

Next we consider the case where DM is not indifferent to all partitions but $\mathcal{P}_{A_1}$ and Ω generate the same payoff. This happens if and only if the posterior probability from the two partitions are same, i.e, the partition consistency condition does not bind. This generates some restriction on the possible posteriors.

For example, since ω_4 and ω_5 belong to the same block the posterior probability of choosing different actions has to be the same. Given equation **??** this is only possible

if either $P(\mathbf{a}_{ij}) = 0$ or $\pi(\mathbf{a}_{ij}|\omega_4) = \pi(\mathbf{a}_{ij}|\omega_4)$. This gives us either $P(\mathbf{a}_{11}) = \mathbf{a}_{12} = 0$ or $\mu_0(\omega_4) = 0 = \mu_0(\omega_5)$.

Let us assume that $\mu_0(\omega_4) = 0 = \mu_0(\omega_5)$. Similar argument shows that $\mu_0(\omega_7) = \mu_0(\omega_8) = \mu_0(\omega_9) = \mu_0(\omega_{10}) = 0$ if all combinations have positive probability of being chosen or at least two input combinations and two of the four states have zero probability.

Similarly for ω_{12} and ω_{13}, since none of the actions give same payoff either $P(\mathbf{a}_{11}) = P(\mathbf{a}_{12}) = 0$ or $\mu_0(\omega_{12}) = 0 = \mu_0(\omega_{13})$.

This gives us two sets of conditions: first, either $P(\mathbf{a}_{11}) + P(\mathbf{a}_{12}) = 1$ and $\mu_0(\omega_4) = \mu_0(\omega_5) = \mu_0(\omega_{12}) = \mu_0(\omega_{13}) = \mu_0(\omega_7) = \mu_0(\omega_8) = \mu_0(\omega_9) = \mu_0(\omega_{10}) = 0$. If $P(\mathbf{a}_{ij}) = 1$ we do not need $\mu_0(\omega_7) = \mu_0(\omega_8) = 0$and similarly for $\mathbf{a}_{12}$.

Second $P(\mathbf{a}_{21}) + P(\mathbf{a}_{22}) = 1$ and $\mu_0(\omega_2) = \mu_0(\omega_3) = \mu_0(\omega_{14}) = \mu_0(\omega_{15}) = \mu_0(\omega_6) = \mu_0(\omega_8) = \mu_0(\omega_9) = \mu_0(\omega_{11}) = 0$. This gives us the second condition in the theorem.

Note that, selective attention is not payoff relevant for the DM because he is indifferent between Ω and $\mathcal{P}_{A_1}$. If the DM was given with information about A_2 he would not update based on the information given his updated belief.

A.2.4 Proof of Result 1

Since the cost of attention is only in choosing a partition and there is non-negative benefit from learning, given any partition $\mathcal{P}$ the DM would choose to learn perfectly about the partition. This proves the first statement of the theorem.

Since given a partition the DM can learn perfectly about it, if he chooses Ω then he can get 1 for all states except ω_{16}. However if he chooses to $\mathcal{P}_{A_1}$ his expected payoff is strictly less than 1 in all the *medium* (except ω_6, ω_{11}) and *bad* states.

In case of *good* states however the DM can always choose an input combination to ensure a payoff of 1, so there are no losses from observing only input A_1.

So the expected gain from choosing Ω is to get payoff 1 in all *bad* and *medium* (except ω_6, ω_{11}) states. This gain is maximum when all states have ex-ante equal probability, i.e., DM starts with an uniform prior. Under uniform prior the expected payoff using $\mathcal{P}_{A_1}$ for all these states are 1/2. Thus the condition for choosing $\mathcal{P}_{\mathcal{A}_\infty}$ is given by,

$$\frac{1}{2}\mu_0(\{\omega_7, \omega_8, \omega_9, \omega_{10}, \omega_{12}, \omega_{13}, \omega_{14}, \omega_{15}\} < \bar{c}.$$

A.2.5 Proof of Theorem 2

Proof. We prove this proposition using the following steps: first we show that if a DM chooses Ω in period $t = 0$ then he would not choose to pay any attention for any later period $t > 0$. This directly follows from the lemma that if the DM sequentially chooses partition $\mathcal{P}$ and $\mathcal{Q}$, then $\mathcal{P}$ is neither finer not coarser than $\mathcal{Q}$. Since every other partition is coarser than Ω, the DM would never choose any other partition along with Ω under any optimal attention strategy.

This implies we only consider a static scenario where the DM chooses Ω is period $t = 0$ and we compare this with another static attention strategy choosing a selectively attentive partition $\mathcal{P}_1$ in period $t = 0$ and not paying any attention afterwards. Since the

actual payoff from choosing $\mathcal{P}_1$ in period $t=0$ followed by any other partition would be weakly higher than this static strategy with $\mathcal{P}_1$ we only need to show that this strategy is strictly better than choosing Ω in period $t=0$ given some μ_0.

WLOG we assume the selective attentive partition is $\mathcal{P}_{A_1} = \{\{\omega_1\}, \{\omega_2, \omega_3\}, \{\omega_4, \omega_5\}, \{\omega_6\}, \{\omega_7, \omega_8\}, \{\omega_9, \omega_{10}\}, \{\omega_{11}\}, \{\omega_{12}, \omega_{13}\}, \{\omega_{14}, \omega_{15}\}, \{\omega_{16}\}\}$, i.e., the DM is selectively attentive to input A_1.

Let us first solve the DM attention problem assuming $c(\Omega) = c(\mathcal{P}_1) = 0$. This is same as the Shannon model. In this case the partition problem can be written as the original problem with Ω with the additional partition consistency condition. Under $\mathcal{P}_1$ since the DM has to choose the same action for two states in the same block the two partition would generate the same expected value for the DM under the static attention choice if the DM finds it optimal to choose same $P(\mathbf{a}_{ij}|\omega)$ for all state ω in the same block under $\mathcal{P}_1$.

Theorem 1 shows there exists a set of values of μ_0 where the net expected payoff from $\mathcal{P}$ is same as Ω. The following denotes the value function with static attention strategy and without the cost of partition,

$$V(\mu_0, \mathcal{P}) = \max_{P(\mathbf{a}_{ij}|\omega)} E(\pi(\mathbf{a}_{ij}, \omega)) - C(\mu_0^p, \lambda)$$

and using this notation we get there exists values of μ_0 such that

$$\Delta V \equiv V(\mu_0, \Omega) - V(\mu_0, \mathcal{P}_1) = 0.$$

Since this difference is continuous in μ_0 we conclude that for a given value of $c(\Omega) - c(\mathcal{P}_1) = \bar{c}$, say there exists a set of prior belief μ_0 for which $\Delta V < \bar{c}$, so Ω is strictly worse option. This proves the first part of the proposition.

To prove the second part we note that when the DM chooses $\mathcal{P}_1$ over Ω then his prior belief does not satisfy the two conditions described above. Since $c(\mathcal{P}_1) > 0$ the DM would never choose to pay attention using $\mathcal{P}_1$ unless $V(\mu_0, \mathcal{P}_1) \geq c(\mathcal{P}_1)$. But under those two conditions the DM either has $P(\mathbf{a}_{ij}) = 1$ or

$$\frac{P(\mathbf{a}_{11})}{P(\mathbf{a}_{12})} = \frac{P(\mathbf{a}_{11}|\omega)}{P(\mathbf{a}_{12}|\omega)} \quad \forall \omega \in supp(\mu_0)$$

, i.e, the unconditional and conditional choice probabilities are same. This implies the DM does not learn at all. In this case he would never choose the partition finer than the inattentive partition which has a zero cost. Thus selective attention can not be optimal for this belief in this model. This implies the DM only chooses selective attention when it is payoff relevant, which implies he would always be willing to pay a positive price for the information he has not attended.

□

A.3 Proof of lemma 3

Proof. Suppose not. Let us assume the DM uses two partitions Let us consider two different strategies; strategy 1: observe $\mathcal{P}$ in period $t=0$ and $\mathcal{Q}$ in period $t=1$, strategy

2: observe $\mathcal{Q}$ in period $t = 0$ and $\mathcal{P}$ in period $t = 1$. Let us further assume WLOG that $\mathcal{Q}$ is more informative than $\mathcal{P}$, i.e, the expected posterior distribution generated using $\mathcal{Q}$ has a higher entropy than that of $\mathcal{P}$. Since the DM chooses the same two partitions under the two strategies the cost of partition for the two strategies would be same so we only consider the net payoff subject to Shannon cost.

For now let us assume that T is very large, so the loss in payoff for any delay in learning for one period is negligible. This is analytically equivalent to assuming that the DM chooses the two consequtive partition in one period so the payoff is obtained only at the end of two partitions. Since we do not allow the DM to choose two separate partitions this can not be true in the model. However if we can show that in this context strategy 1 and strategy 2 are payoff equivalent then strategy 2 that involves observing a more informative partition in earlier period would be weakly payoff improving for all $T \geq 2$.

Note that since the two partitions are orthogonal, i.e, cannot be ranked by coarseness ranking (from lemma 2) we cannot apply LIP because the period $t = 1$ probability obtained using $\mathcal{P}$ does not belong to the convex hull of any posterior distribution generated by $\mathcal{Q}$.

The only way strategy 1 can be better than strategy 2 if after observing $\mathcal{P}$ the payoff differences among input combinations increases for $\mathcal{Q}$ for every state in Ω. Since under Shannon model a higher payoff difference implies more learning strategy 1 would generate more informative posterior.

Since, $\mathcal{P}$ is a partition of the original state space the payoff difference would increase for some states and decrease for some other states. Orthogonality of $\mathcal{P}$ and $\mathcal{Q}$ ensures under $\mathcal{Q}$ there would exist at least one block in which for one state the payoff difference has increased and for another the payoff difference has decreased. The states for which the payoff difference increases or decreases depnds only on the partition and prior probability and does not depend on the order of partition.

However, the DM consider the possibility of future learning while choosing the first partition, thus combining the two partitions he faces the same set of *participation consistency* condition for both strategies. The Shannon cost of attention function does not depend on the partition.

Suppose the DM decided optimal posterior choice distribution for state ω as P^*_ω. If he follows the strategy 1 in period $t = 0$ he chooses $P^{\mathcal{P}}_\omega$ and if he follows strategy 2 he chooses $P^{\mathcal{Q}}_\omega$ in period $t = 0$.

If under strategy 1 the participation consistency condition imlplies ω_i and ω_j should have same log-odds ratio then under strategy 2 either we can separate out the two states or the two states remain in the same partition. If they remain in the same partition then the log-odds ratio of these two states under $P^{\mathcal{P}}_\omega$ would be same as P^*_ω. If with $\mathcal{Q}$ we can separte the two states then in period $t = 1$ the DM would choose a log-odds ratio which is same as P^*_ω.

Then under strategy 2 these two states would have the same log odds ratio since in period $t = 1$ the DM would choose the same log-odds as that of P^*_ω and in period $t = 1$ the log-odds ratio does not change under $\mathcal{P}$. Since the Shannon cost function does not

depend on the partition choosing the same log-odds with only one partition would be same irrespective of the order of partition.

On the other hand if two states can be separated in $\mathcal{P}$ then under $\mathcal{Q}$ either they can be separated or they cannot be separated. In the later case the log-odds ratio from $P_\omega^{\mathcal{P}}$ is same as P_ω^*. In the former case by linearity of cost function the order of the posterior distribution would not affect the cost of learning, i.e, any choice of $P_\omega^{\mathcal{P}}$ that leads to P_ω^* would cost the same.

Thus under both the two strategies the DM would choose the same posterior distribution and pay the same cost of attention. Now, if we relax the assumption of T being large, since $\mathcal{Q}$ generate more informative posterior it is weakly preferable to use $\mathcal{Q}$ before $\mathcal{P}$ since the payoff difference is bigger for earlier periods as a result of which the DM enjoys higher payoff for more periods.

Note that, the result is true for all other cost function where the cost of updating is linear in posterior probability. This is true for all other entropy cost functions are in general for all UPS cost functions where the prior belief does not affect the cost function and $G(.)$ is linear in posterior probabilities.

□

A.4 Proof of lemma 4

Proof. Let us assume for simplicity that the DM does not observe one input combination $\mathbf{a}_{21}$, i.e., chooses $\mathcal{P}_{\neg\mathbf{a}_{21}}$ in period $t = 0$, then the statement of the lemma claims that he should choose $\mathcal{P}_{\mathbf{a}_{21}}$ in period $t = 1$.

If the DM chooses to not observe more than one input combination then the cost function would determine which seqence of partitions would generate the lowest cost. For example if the DM does not observe $\mathbf{a}_{21}$ and $\mathbf{a}_{22}$ then he can choose to observe $\mathcal{P}_{A_2|a_{12}}$ in period $t = 1$ or $\mathcal{P}_{\mathbf{a}_{21}}$ and $\mathcal{P}_{\mathbf{a}_{22}}$ in period $t = 1$ and $t = 2$ respectively. But without further assumption on cost function $c(\mathcal{P})$ we cannot conclude which choice would be cheaper for the DM.

We want to show that choosing $\mathcal{P}_{\neg\mathbf{a}_{21}}$ in period $t = 0$ followed by $\mathcal{P}_{\mathbf{a}_{21}}$ in period $t = 1$ is better than choosing $\mathcal{P}_{\neg\mathbf{a}_{21}}$ and choosing any other partition under which $\mathbf{a}_{21}$ is observed but the partition is finer than $\mathcal{P}_{\mathbf{a}_{21}}$.

Suppose not. Suppose the DM chooses $\mathcal{P}_{A_2|a_{12}}$ in period $t = 1$ instead of $\mathcal{P}_{\mathbf{a}_{ij}}$. Then in period $t = 1$ the DM can choose a more informative posterior distribution in period $t = 1$ with a finer partition. But since in period $t = 0$ the DM could observe all three input combination except $\mathbf{a}_{21}$ the DM should optimally choose the log odds ratio of the posetrior probability of choosing action $\mathbf{a}_{ij}$ for all $i, j \neq (2, 1)$.

In period $t = 1$, $\mathcal{P}_{A_2|a_{12}}$ gives some information about $\mathbf{a}_{22}$ which the DM already could have obtained in period $t = 0$. If the DM chooses a different log-odds ratio for other input combination then the choice of posterior probability over input combination in perod $t = 0$ for any given state ω was not optimal, since the only restriction in period $t = 0$ was on the participation constraint imposed by $\mathcal{P}_{\neg\mathbf{a}_{21}}$ which only restricts the relative position of $\mathbf{a}_{21}$ without affecting the other input combinations.

Thus given in period $t = 0$ the DM has chosen the ratio of posterior probabilities optimally for other three inputs the optimal attention strategy would be the following: if the signal increases the probability of payoff 1 for $\mathbf{a}_{21}$ then the probability of all other input combinations decreases proportionaly so that the log-odds ratio remain the same. Similarly if the signal decreases the probability of payoff 1 for $\mathbf{a}_{ij}$ then the probability of choosing all other input combinations should go up proportionaly.

If probability of choosing all other input combination are changed proportionally, both $\mathcal{P}_{\mathbf{a}_{21}}$ and $\mathcal{P}_{A_2|a_{12}}$ would genrate the same posterior distribution in period $t = 1$ becuase both the partitions face no constraint to observe $\mathbf{a}_{21}$ so both the two partitions should generate the same choice probability of $\mathbf{a}_{21}$ in period $t = 1$. This implies the Shannon cost and gross payoff would be the same for both the two partitions. But the cost of partition $c(.)$ would be lower for the coarser partition. So the DM would be weakly better off by choosing the coarser partition. Hence, proved.

□

A.5 Characterization of Optimal Strategy

We would characterize the optimal strategy in terms of prior belief over Ω. Before that let us define certain quantities that would simplify the description of prior belief. Let $G = \{\omega_2, \omega_3, \omega_4, \omega_5\}$ denote the set of *good* states $B = \{\omega_{12}, \omega_{13}, \omega_{14}, \omega_{15}\}$ denote the set of *bad* states and $M = \{\omega_6, \omega_7, \omega_8, \omega_9, \omega_{10}, \omega_{11}\}$ be the set of *middle* states.

Let

$$\Delta_i = max_{\omega_i, \omega_j \in G} |\mu_t(\omega_i), \mu_t(\omega_j)| \tag{5}$$

for $i = G, B, M$ be the maximum difference between *good* states *bad* states and *middle* states respectively. Δ_i takes a value zero when all *good*(or *bad* or *middle*) states have same probabilities.

Let us first consider one-period strategies, which would be optimal if $c(\Omega)$ is sufficiently small relative to the cost of other partitions so that the DM prefers to observe Ω in one period than observing more than one coarser partitions over multiple periods. For any period $t \geq 0$ the following conditions characterize which type of partition would be optimal.

Result 2. *1. Observing only one input is optimal when the following two conditions are true:*

- *(a) only one of input is payoff relevant in middle states, i.e, $\mu_t(\omega_6, \omega_{11})$ OR $\mu_0(\omega_7, \omega_{10})$ are sufficiently higher than $\mu_t(\omega_8, \omega_9)$*
- *(b) there exists $\epsilon_G \geq \epsilon_B \geq 0$ such that $\Delta_G \leq \epsilon_G$ and $\Delta_B \leq \epsilon_B$*

2. Observing both inputs are optimal when the following conditions hold true:

- *(a) $\Delta_G > \epsilon_G$ and $\Delta_B > \epsilon_B$*
- *(b) The prior probability of the two middle states where both inputs are payoff relevant, i.e, $\mu_0(\omega_8, \omega_9)$ is high*

(c) there exists $\epsilon_{M\geq 0}$ such that $\Delta_M \leq \epsilon_M$

To prove the result 2 we make the following observation. For a given prior belief μ_t at time t we would say partition 1 is *better* than partition 2 for state ω, if the posterior choice distribution under partition 1 generates lower error probability than that of under partition 2, where error probability for a given state ω is the sum of choice probabilities of any input combination that gives zero in state ω. We can also define expected error as the expected value of the error probability where the expectation is over the prior probability of states in Ω.

For a given prior μ_t and state ω two partitions will generate different expected error only if they generate different *partition consisitency* restrictions. However, this does not imply that the state ω should belong to the same block in every period for both partitions.

Shannon cost of attention implies given a marginal cost λ the posterior choice probabilities depend on only the payoff difference of different actions in a given state (or block for a partition). Thus two partitions would generate same expected error for a given state if in any period the payoff difference from different input combinations are same for the blocks containing ω under the two startegies.

Thus the only way partition 1 would be better than partition 2 for state ω in a period t if under partition 1 state ω belongs to a block b_1 that generates higher payoff difference than block b_2 which contains ω under partition 2. Since the cost of attention does not change over time, learning early is always weakly better.

Note that under one-input strategy the only states that can be separated completely are ω_6, ω_{11} (using partition $\mathcal{P}_{A_1}$) and ω_7, ω_{10} (using partition $\mathcal{P}_{A_1}$)[2]. Whereas under two-input partitions these four states never belongs to a block that contains only one state. Thus for any two-input partition there exists a one-input partition which is better for states $\omega_6, \omega_7, \omega_{10}, \omega_{11}$. Thus one-input partitions are optimal when prior probability of these states are sufficiently high.

Under any one-input partitions all *good* and *bad* states are treated symmetrically, i.e, always *good* (*bad*) states are combined with *good*(*bad*) states to form a block. For example under $\mathcal{P}_{A_1}$ the four *good* and *bad* states belong to blocks of two states, ω_2 and ω_3 belong to same block and ω_4 and ω_5 belong to the same block. All *good* (*bad*) states have only one state that generates payoff 1 (or 0). So the payoff differences would be symmetric and hence the expected errors would be symmetric.

However if the DM chooses two inputs there exists partitions where a *good* state can be combined with *bad* state or *middle* states or other *good* states. Thus the payoff differences would not be symmetric. More specifically if a *good* state is combined with a *bad* state the incentives to learn would be higher compared to the case when it is combined with other *good* states. However under two input strategies the DM does not observe at least one input combination. This implies for at least one *good* (or *bad*) state the error probabilities would be higher under two-input partition than under any one-input partition that observes all levels of an input. Thus two input stratgies are optimal

[2]Here we are ignoring ω_1 and ω_{16} as no learning is always optimal for these states.

when the prior probability of one or two of the *good* (or *bad*) states are different than the other *good* (or *bad*) states.

For state ω_8 and ω_9 all one-input partitions generate no learning optimally. Hence if the prior proabablty of ω_8 and ω_9 is higher the DM would optimally choose a two-inputs partition.

A.6 Proof of theorem 2

Proof. Using Ω always generated weakly more informative posterior choice distribution. For a given μ_0 and λ let $\Delta_{\mu_0}(\mathcal{P}_{A_i})$ denote the minimum differnce in payoff between full learning and selective attention with some partition $\mathcal{P}_{A_i}$. Since the payoff function is bounded, the maximum value of payoff difference is bounded as well for any given λ. This implies we can define $\Delta(\mathcal{P}_{A_i}) = \max_{\mu_0} \Delta_{\mu_0}(\mathcal{P}_{A_i})$ as the uniform bound of the payoff difference.

If the difference in cost of partition $c(\Omega) - c(\mathcal{P}_{A_i})$ is lower than the payoff difference $\Delta(\mathcal{P}_{A_i})$ then choosing Ω is optimal. This generates the lower bound $\underline{c}$. On the other hand if $c(\Omega)$ is too high compared to $\mathcal{P}_{A_1}$ such that sequential learning which involves $\mathcal{P}_{A_i}$ and $\mathcal{P}_{A_j}$ in two consequetive periods generates lower cost than Ω, i.e, $2c(\mathcal{P}_{A_i}) < c(\Omega)$ then the DM would not choose selective attention. Instead in period $t = 1$ he would choose to observe the other selectively attentive partition. However, sequential attention also generates lower expected payoff over time compared to Ω. This generates the upper bound $\bar{c}$

The second condition guarantees that the only possible strategies are selective, sequential or full attention as observing one input is optimal. These two conditions together complete the proof. □

A.7 Proof of Proposition 3

Proof. The proof of this proposition follows directly from result 2 described in this section. For both sequential and conditional sequential attention strategy the total number of period where the DM chooses to learn are same, two periods. In period $t = 0$ the two startegies use the same partition, $\mathcal{P}_{A_i}$ hence the posterior choice distribution would be same at the end of period $t = 0$ for two strategies.

Thus to compare the two stratgies we need the states for which sequential attention is *better* than conditional sequential attention and vice versa. Given any λ if Ω_1 is the set of states for which sequential attention is *better* than conditional sequential attention then by continuity of the value function with respect to belief μ_1 there exists a threshold $\bar{\mu}$ such that if $\mu_1(\Omega_1) > \bar{\mu}$ then sequential attention would generate lower expected error than conditional sequential attention and vice versa when $\mu_1(\Omega_1) \leq \bar{\mu}$.

Under sequential attention problem the DM chooses a one-input partition in period $t = 1$ and under conditional sequential attention he chooses a two-inputs partition in period $t = 1$. Given the results in the last section sequential attention is better for all *middle* states except ω_8 and ω_9. This proves the part 1 of the statement of the proposition.

For part two we note that under conditional sequential attention the DM completely ignores two input combinations whereas under sequential attention he observes all levels of an given input. Thus the error probability would be lower under conditional sequential attention for a *good* state for which the DM observes the input combination that generates 0 but would be higher for a *good* state for which the DM does not observe the input combination that generates 0. The same logic applies to *bad* states as well.

Thus the sequential attention strategy is preferred over conditional sequential attention if the difference in probability of one or two *good*(or *bad*) states are sufficiently low so that the expected error is lower than that of under conditional sequential probability. As probability of one or two *good*(or *bad*) start to increase the expected error under conditional sequential attention strategy goes down which generates the second statement of the proposition.

□