Behavioral Economics Approach to Interpretable Deep Image Classification. Rationally Inattentive Utility Maximization Explains Deep Image Classification

Kunal Pattanayak [*1] **Vikram Krishnamurthy** [*1]

Abstract

Are deep convolutional neural networks (CNNs) for image classification consistent with utility maximization behavior with information acquisition costs? This paper demonstrates the remarkable result that a deep CNN behaves equivalently (in terms of necessary and sufficient conditions) to a rationally inattentive utility maximizer, a model extensively used in behavioral economics to explain human decision making. This implies that a deep CNN has a parsimonious representation in terms of simple intuitive human-like decision parameters, namely, a utility function and an information acquisition cost. Also the reconstructed utility function that rationalizes the decisions of the deep CNNs, yields a useful preference order amongst the image classes (hypotheses).

1. Introduction

In behavioral economics, two fundamental questions relating to human decision making are: (i) How to construct a utility function that rationalizes decisions? The area of revealed preference (and inverse reinforcement learning in machine learning) gives necessary and sufficient conditions for decisions to be consistent with utility maximization. (ii) How to model attention spans for information acquisition? The pioneering work of (Sims, 2003) constructs rationally inattentive models (in terms of Shannon mutual information) for how humans perceive information. In the last decade, necessary and sufficient conditions have been developed in the economics literature for determining if a sequence of decisions is consistent with rationally inattentive Bayesian utility maximization.

A natural question relating to interpretable deep learning is: *Are the decisions of a deep CNN in image classification, consistent with rationally inattentive Bayesian utility maximization?* In other words, *is a deep CNN equivalent to a decision maker with behavioral economics constraints?* Based on the massive CIFAR-10 dataset for image classification, and as schematically shown in Fig. 1, this paper shows the surprising result that deep CNNs for image classification are *equivalent* to rationally inattentive utility maximizers. By equivalent we mean, that the decisions of the deep learner satisfy the necessary and sufficient conditions for rationally inattentive Bayesian utility maximization.

This equivalence is remarkable for several reasons: first, it draws important parallels between human decision making and deep neural networks; namely that deep neural networks satisfy economics based rationality. Second, this result enhances interpretability of deep learning; indeed, utility functions and rational inattention cost that rationalize deep learning yield a parsimonious representation of hundreds of thousands of layer weights in terms of a few hundred parameters. Third, our Bayesian revealed preference approach introduces a useful preference ordering amongst the set of hypotheses (image labels) considered by the deep neural network; for example, how much additional priority is allocated to the classification of a cat as a cat compared to a cat as a dog. In classical deep learning, this preference ordering is not explicitly generated

The Bayesian revealed preference approach is ideally suited for interpretable deep learning because it is purely *data-driven*. It does a *post-hoc* analysis of the decisions generated by a collection of Bayesian agents, namely, it reconstructs the utility function and rational inattention model parameters. Another feature is that only the state and action of the agents need to be related to the underlying deep neural network based image classifier. The remaining latent variables in the Bayesian model are unobserved and abstracted away; yet necessary and sufficient conditions for their existence can be checked via a convex feasibility test described below.

Related Works

Since this paper focuses on interpretable deep learning via behavioral economics, we briefly discuss related works in

[*]Equal contribution [1]School of Electrical and Computer Engineering, Cornell University, Ithaca, New York, USA. Correspondence to: Kunal Pattanayak or Vikram Krishnamurthy <kp487@cornell.edu or vikramk@cornell.edu>.

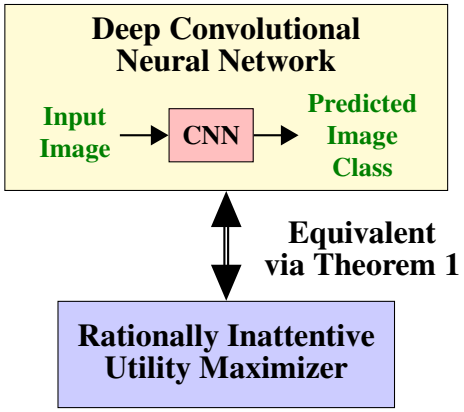

Figure 1. Schematic illustration of Bayesian revealed preference based interpretable image classification by deep CNNs. Theorem 1 establishes equivalence between the image classification behavior of a CNN and the decision-making of a rationally inattentive maximizer (see Table 2 for a summary of results). Hence, the CNN's predictive behavior can be parsimoniously represented by a utility function and an information acquisition cost.

these areas.

Bayesian revealed preference and Rational inattention. Estimating utility functions given a finite sequence of decisions and budget constraints is the central theme of revealed preference in economics. The seminal work of Afriat (1967); Diewert (1973) (see also Varian (1982)) give necessary and sufficient conditions for the existence of a utility function that rationalizes the observed decision sequence.

Bayesian utility maximization in the presence of costly information acquisition has been extensively studied under the area of "rational inattention", pioneered by Nobel Laureate Christopher Sims (Sims, 2003; 2010). The key idea is that information acquisition by human attention is costly and can be modeled in information theoretic terms as a Shannon capacity limited communication channel. In complete analogy to the classical revealed preference test by Afriat (1967), a general *data-driven* test for identifying Bayesian utility maximization was proposed by Caplin & Dean (2015) that provides necessary and sufficient conditions for the actions of a Bayesian agent to be consistent with rationally inattentive behavior. Our setup in the paper is confined to the Bayesian setting, where unlike classical revealed preference, the utility function involves discrete valued variables.

Interpretable ML. Providing interpretable models for deobfuscating 'black-box' machine learning algorithms under the area of interpretable machine learning is a subject of extensive research (Chakraborty et al., 2017; Doshi-Velez & Kim, 2017; Guidotti et al., 2018). Murdoch et al. (2019) defines interpretable machine learning as "the use of machine-learning models for the extraction of relevant knowledge about domain relationships contained in data".

Since the literature is enormous, we will only discuss a subset of works pertaining to interpretability of deep neural networks for image classification. One prominent approach, namely, saliency maps, reconstructs the most preferred or typical image pertaining to each image class the deep neural network has learned (Simonyan et al., 2013; Nguyen et al., 2016). Related work includes creating hierarchical models for determining the most important features (Hase et al., 2019). Another approach seeks to provide local approximations to the trained model, local w.r.t the input image (Lei et al., 2016; Lundberg & Lee, 2017). A third approach approximates the predictions of the deep neural networks by a linear function of simplified individual image features (Bach et al., 2015; Ribeiro et al., 2016; Shrikumar et al., 2016; Lundberg & Lee, 2017). The aim is to estimate the corresponding weights, or feature importance weights, that best explain the model predictions. Finally, deep neural networks have also been modeled by Bayesian inference frameworks using probabilistic graphical methods (Wang & Yeung, 2016).

To the best of our knowledge, a behavioral economics based approach for the post-hoc analysis of deep neural networks has not been explored in literature. However, we note that behavioral economics based interpretation models have been applied to domains outside machine learning, for example, in online finance platforms for efficient advertising (Milgrom, 1981; Huang & Liu, 2007), and more recently in YouTube to rationalize user commenting behavior (Hoiles et al., 2020). The key difference from other works in interpretable ML is that instead of asserting that a behavioral economics model explains the deep neural networks' predictions, we propose a feasibility test (necessary and sufficient conditions) which if satisfied, ensures the existence of such a rationalizing model.

2. Bayesian Revealed preference with Rational Inattention

This section describes the key ideas behind Bayesian revealed preference with rational inattention. Despite the abstract formulation below, the reader should keep in mind the deep learning context. In Sec. 3, we will use Bayesian revealed preference theory as a tool for interpretable deep learning by showing that deep CNNs are equivalent to rationally inattentive Bayesian utility maximizers.

2.1. Utility Maximization Rational Inattention (UMRI)

We start by describing the **u**tility **m**aximization model with **r**ational **i**nattention (henceforth called UMRI) for a ***single*** Bayesian decision maker/agent. Abstractly, the UMRI

model is parametrized by the tuple

$$\theta = (\mathcal{X}, \mathcal{Y}, \mathcal{A}, \pi_0, C, u, \alpha). \quad (1)$$

With respect to the abstract parametrization of the UMRI model for a single Bayesian agent, the following elements constitute the tuple θ defined in (1).

State: $\mathcal{X}$ is the finite set of ground truths with prior probability distribution π_0. With respect to our image classification context, $\mathcal{X} = \{1, 2, \dots 10\}$ is the set of image classes in the CIFAR-10 dataset and π_0 is the empirical probability distribution of the image classes in the test dataset of CIFAR-10.

Observation and Attention function: The agent chooses attention function $\alpha : \mathcal{X} \to \Delta(\mathcal{Y})$, a stochastic mapping from $\mathcal{X}$ to a finite set of observations $\mathcal{Y}$. Given state x and attention function α, the agent samples observation y with probability $\alpha(y|x)$. The agent then computes the posterior probability distribution $p(x|y)$ via Bayes formula as

$$p(x|y) = \frac{\pi_0(x)\alpha(y|x)}{\sum_{x' \in \mathcal{X}} \pi_0(x')\alpha(y|x')}. \quad (2)$$

The observation and attention function are latent variables that are abstracted away in the deep image classification context. Yet, their existence is guaranteed by the convex feasibility test in Theorem 1 below.

Action: The agent chooses action a from a finite set of actions $\mathcal{A}$ after computing the posterior probability distribution $p(x|y)$. In the image classification context, a is the image class predicted by the neural network, hence $\mathcal{A} = \mathcal{X}$. In behavioral economics, Bayesian models for decision making are widely assumed as a useful approximation (Caplin & Dean, 2015; Milgrom, 1981).

Utility function: The agent has a utility function $u(x, a) \in \mathbb{R}^+$, $x \in \mathcal{X}, a \in \mathcal{A}$ and aims to maximize its expected value, with the expectation taken wrt the random state x and random observation y. A key feature in our approach is to show that the utility function rationalizes the decisions of the deep neural networks.

Information Acquisition Cost: The information acquisition cost $C(\alpha, \pi_0) \in \mathbb{R}^+$ maps the attention function α and prior pmf π_0 to the set of positive reals. It is the cost the agent incurs for estimating the state (2). The function $C(\cdot)$ abstractly captures the cost incurred during training of the deep neural networks. In rational inattention theory from behavioral economics, a higher information acquisition cost is incurred for more accurate attention functions (equivalently, more accurate state estimates (2) given observation y). We refer the reader to the influential work of Sims (2003; 2010).

The Bayesian agent's aim is to maximize its expected utility and minimize its cost of information acquisition. Hence, the action a given observation y, and attention function α are optimally chosen by the Bayesian agent as described below.

Definition 1 (Rationally Inattentive Utility Maximizer)**.** *Consider a Bayesian agent parametrized by θ in* (1) *under the UMRI model. Then,*

(a) **Expected Utility Maximization:** *Given posterior probability distribution $p(x|y)$, the agent takes action a that maximizes its expected utility. That is, with $\mathbb{E}$ denoting mathematical expectation, the action a satisfies*

$$a \in \operatorname*{argmax}_{a' \in \mathcal{A}} \mathbb{E}_x\{u(x, a')|y\} = \sum_{x \in \mathcal{X}} p(x|y)u(x, a') \quad (3)$$

(b) **Attention function Rationality:** *The agent's chosen attention function α optimally trades off between maximizing the expected utility and minimizing the information acquisition cost.*

$$\alpha \in \operatorname*{argmax}_{\alpha'} \mathbb{E}_y\{\max_{a \in \mathcal{A}} \mathbb{E}_x\{u(x, a)|y\}\} - C(\alpha', \pi_0) \quad (4)$$

Eq. 3,4 in Definition 1 essentially comprise a bilevel optimization problem. The lower-level optimization task is to choose the the 'best' action for any observation y. The upper-level optimization task is to sample the observations optimally by choosing the 'best' attention function.

2.2. UMRI for a Collection of Agents (CUMRI)

Bayesian revealed preference theory (Caplin & Dean, 2015; Caplin & Martin, 2015) characterizes (rationalizes) optimal decision making amongst a collection of Bayesian decision makers. Hence, we need to extend the above single agent UMRI model to a multi-agent setup. This multi-agent framework is also crucial for our formulation of interpretable deep learning as will be explained below.

To this effect, we introduce the **C**ollective **UMRI** model (henceforth called CUMRI) for a ***collection*** of Bayesian agents. The model is abstractly parametrized by the tuple

$$\Theta = (\mathcal{K}, \{\theta_k, k \in \mathcal{K}\}), \quad (5)$$

where $\mathcal{K} = \{1, 2, \dots K\}$ is the collection of Bayesian agents[1]. In Θ, the tuple $\theta_k = (\mathcal{X}, \mathcal{Y}, \mathcal{A}, \pi_0, C, u_k, \alpha_k)$ is the UMRI model tuple associated with agent m. Note that the parameters $\mathcal{X}, \mathcal{Y}, \mathcal{A}, \pi_0, C$ are the same for all $\theta_k \in \Theta$, that is, the agents in $\mathcal{K}$ only differ in their utility function and the chosen attention function.

In complete analogy to Definition 1 we can define a collection of rationally inattentive utility maximizers (CUMRI).

Definition 2 (Collection of Rationally Inattentive Utility Maximizers)**.** *Consider a collection of Bayesian agents*

[1]Equivalently, the CUMRI tuple Θ (5) can be defined as $\Theta = (\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{K}, \pi_0, C, \{u_k, \alpha_k, k \in \mathcal{K}\})$.

parametrized by Θ (5) *under the CUMRI model.*
(a) **Expected Utility Maximization:** *Given posterior probability distribution* $p(x|y)$*, every agent* $k \in \mathcal{K}$ *chooses action* a *that maximizes its expected utility.*

$$a \in \underset{a' \in \mathcal{A}}{\operatorname{argmax}} \ \mathbb{E}_x\{u_k(x, a')|y\} = \sum_{x \in \mathcal{X}} p(x|y) u_k(x, a')$$

(b) **Attention function Rationality:** *For every agent* k*, the attention function* $\alpha_k \in \theta_k$ *optimally trades off between maximizing its expected utility and minimizing its information acquisition cost.*

$$\alpha_k \in \underset{\alpha'}{\operatorname{argmax}} \mathbb{E}_y\{ \max_{a \in \mathcal{A}} \mathbb{E}_x\{u_k|y\}\} - C(\alpha', \pi_0)$$

The CUMRI model is an aggregation of multiple Bayesian agents following the UMRI model of decision-making.

2.3. Relevance of CUMRI to Interpretable Deep Image Classification

Let us now explain the relevance of CUMRI in interpretable deep learning. The motivation for a *collection* of agents (discussed above) is clear: We require the decisions of a collection of $K \geq 2$ trained CNNs in order to interpret their decision-making behavior. (If $K = 1$, explaining the CNN's decisions becomes trivial since any information acquisition cost rationalizes its decisions; see discussion following Theorem 1.) So let us start with the simplest case involving two trained deep neural networks N_1 and N_2; so $\mathcal{K} = \{1, 2\}$ in the above setup. Suppose N_1 was observed to make highly accurate predictions on a sufficiently rich input dataset while N_2 makes relatively less accurate predictions on the same dataset.

Analogous to the CUMRI model, we first abstract the precision of the feature representations of the input image data learned by N_1 and N_2 via two attention functions α_1 and α_2 in (4). Second, we assume the existence of an information acquisition cost function $C(\cdot)$ that abstracts the computational resources expended for learning the representations. That is, the training cost incurred by N_1 and N_2 are $C(\alpha_1)$ and $C(\alpha_2)$ respectively. The existence of such a cost function is justified since learning a more accurate representation (more accurate attention function) entails higher computation resources which implies a higher cost of information acquisition.

Given this abstracted setup, if it is known that the predictions of N_1 and N_2 can be explained by the CUMRI model (and Theorem 1 below will give necessary and sufficient conditions for this), then there exist two utility functions u_1 and u_2 for N_1 and N_2, that satisfy the following inequality:

$$\mathbb{E}_{\alpha_i}\{u_i\} - C(\alpha_i) \geq \mathbb{E}_{\alpha_j}\{u_i\} - C(\alpha_j), \ i, j \in \{1, 2\}.$$

The above inequality says that CNNs N_1 and N_2 would be worse off (in an expected utility sense) if they make predictions based on swapping each other's learned representations. Put differently, N_1 and N_2 are always better off performing image classification using their own learned representations. At a deeper (and more subtle) level, since the information acquisition cost is a latent variable (recall model parametrization in Sec. 2.1), one can construct a cost function $C(\cdot)$ so that N_1 and N_2 will not benefit from choosing *any* other representation for making predictions; this can be viewed as a form of global optimality. To summarize, a CNN modeled by CUMRI learns the best representation given the cost it incurs for learning the representation.

2.4. Bayesian Revealed Preference (BRP) Test for Rationally Inattentive Utility Maximization

Thus far, we have described the CUMRI model of a collection of rationally inattentive utility maximizers. We are now ready to state our key result: Theorem 1 below says that the decisions of a collection of Bayesian agents is rationalized by a CUMRI tuple Θ if and only if a set of convex inequalities have a feasible solution. These inequalities comprise our **B**ayesian **R**evealed **P**reference (henceforth called BRP) test for rationally inattentive utility maximization.

Before stating the theorem, we specify one necessary requirement for BRP Test in Theorem 1. Theorem 1 assumes knowledge of the following dataset obtained from the Bayesian agents in $\mathcal{K}$.

$$\mathbb{D} = \{\pi_0, p_k(a|x), x, a \in \mathcal{X}, k \in \mathcal{K}\}. \tag{6}$$

In (6), $\pi_0 \in \Delta^{|\mathcal{X}|-1}$ denotes the prior pmf over the set of states $\mathcal{X}$ in Θ (5). The variable $p_k(a|x)$ is the conditional probability that agent $k \in \mathcal{K}$ takes action a given state x. $\mathbb{D}$ characterizes the input-output behavior of the collection of Bayesian agents and serves as the input for BRP feasibility test described below.

Theorem 1 (BRP Test for Rationally Inattentive Utility Maximization (Caplin & Dean, 2015))**.** *Given the input-output dataset* $\mathbb{D}$ (6) *obtained from a collection of Bayesian agents* $\mathcal{K}$*. Then,*
1. *Existence: There exists a CUMRI tuple* Θ (5) *that rationalizes dataset* $\mathbb{D}$ *if and only if there exists a feasible solution that satisfies the set of convex inequalities*

$$\textit{BRP}(\mathbb{D}) \leq \mathbf{0}. \tag{7}$$

In (7)*,* $\textit{BRP}(\cdot)$ *corresponds to a set of convex inequalities stated in Algorithm 1 below.*
2. *Reconstruction: The set-valued estimate of* Θ *that rationalizes* $\mathbb{D}$ *is the set of all feasible solutions to* (7)*.*

The proof of Theorem 1 and additional insight is given in the supplementary material. It is important to stress the

"iff" in Theorem 1. If the inequalities in (7) are not feasible, then the Bayesian agents that generate the dataset $\mathbb{D}$ are not rationally inattentive utility maximizers. If (7) does have a feasible solution, then there exists a reconstructable family of utility functions and information acquisition costs that rationalize $\mathbb{D}$. Finally, a key feature of Theorem 1 is that the estimated utilities are set-valued; every utility function in the set of viable utility functions explains $\mathbb{D}$ equally well. In terms of interpretable deep learning, of all parameters in the CUMRI tuple Θ, we are only interested in the utility functions u_k of the agents and the cost of information acquisition C.

Algorithm 1 BRP Convex Feasibility Test of Theorem 1

Require: Dataset $\mathbb{D} = \{\pi_0, p_k(a|x), x, a \in \mathcal{X}, k \in \mathcal{K}\}$ from a collection of Bayesian agents $\mathcal{K}$.

Find: Positive reals $c_k,\ u_k(x,a) \in (0,1)$ for all $x \in \mathcal{X}$, $a \in \mathcal{A},\ k \in \mathcal{K}$ that satisfy the following inequalities:

$$\underline{\textbf{NIAS}}: \sum_x p_k(x|a)\ (u_k(x,b) - u_k(x,a)) \leq 0, \qquad (8)$$
$$\forall a, b \in \mathcal{A},\ k \in \mathcal{K},$$

$$\underline{\textbf{NIAC}}: \sum_a \left(\max_b \sum_x p_j(x,a) u_k(x,b) \right) - c_j \qquad (9)$$
$$- \sum_{x,a} p_k(x,a) u_k(x,a) + c_k \leq 0,\ \forall j, k \in \mathcal{K},$$

where $p_k(x,a) = \pi_0(x) p_k(a|x)$, $p_k(x|a) = \frac{p_k(x,a)}{\sum_{x'} p_k(x',a)}$.

Return: Set of feasible utility functions $u_k = \{u_k(x,a)\}$ and information costs c_k for all agents $k \in \mathcal{K}$.

Discussion.

(i) *Intuition.* The BRP convex feasibility test in Theorem 1 comprises two sets of inequalities, namely, the *NIAS* (No-Improving-Action-Switches) (8) and *NIAC* (No-Improving-Action-Cycles) (9) inequalities (Caplin & Dean, 2015). NIAS ensures that the agent takes the best action given a posterior pmf. NIAC ensures that every agent chooses the best attention function. BRP test checks if there exist K utility functions and K positive reals that, together with $\mathbb{D}$, satisfy the NIAS and NIAC inequalities.

(ii) *Identifiability.* The BRP feasibility test requires the dataset $\mathbb{D}$ to be generated from $K > 2$ Bayesian agents. If $K = 1$, then (7) holds trivially since any information acquisition cost satisfies the convex inequalities of BRP.

(iii) *Parsimonious Representation.* In the deep image classification context, another important consequence is that due to the CUMRI model's parsimonious parametrization in (5), the predictions of K CNNs can be rationalized by just K utility functions and an information acquisition cost function, thus bypassing the need of several million parameters to describe the deep CNNs. As illustrated in Fig. 1, this serves as our motivation for behavioral economics based interpretable deep image classification.

(iv) *Computational Aspects of BRP Test.* Suppose the dataset $\mathbb{D}$ is obtained from K Bayesian agents. Then, BRP($\mathbb{D}$) comprises a feasibility test with $K\ (|\mathcal{X}||\mathcal{A}| + 1)$ free variables and $K^2 + K\ (|\mathcal{A}|^2 - |\mathcal{A}| - 1)$ convex inequalities. Thus, the number of free variables and inequalities in the BRP feasibility test scale linearly and quadratically, respectively, with the number of observed Bayesian agents.

2.5. Summary

Since this section formulated a behavioral economics model that may not be familiar to a machine learning reader, we briefly summarize the main ideas. We introduced two rationally inattentive utility maximization models, UMRI for a single Bayesian agent and CUMRI for a collection of agents. Building on the CUMRI model, our key result is Theorem 1 which outlines a decision test BRP for rationally inattentive utility maximization given the decisions from a collection of agents. This test comprises a set of convex feasibility inequalities that have a feasible solution *if and only if* the collection of agents are rationally inattentive utility maximizers. The set of feasible utility functions and information acquisition costs thus parsimoniously represent the input-output relationship of the agent decisions.

In the rest of the paper, we present our experimental findings on the CIFAR-10 image dataset. We will focus on using Theorem 1 to construct interpretable models for deep CNNs.

3. Bayesian Revealed Preference explains CIFAR-10 Image Classification by CNNs

Thus far we have proposed a BRP test (Theorem 1) that is both necessary and sufficient for a dataset to be consistent with rationally inattentive Bayesian utility maximization. This section provides experimental evidence that image classification using deep *convolutional neural networks* (CNNs) is consistent with rationally inattentive utility maximization behavior, namely, the decisions of the CNNs satisfy both necessity and sufficiency of Theorem 1. Our experimental findings are summarized in Table 2 and schematically shown in Fig. 1. Our experiments are performed using the widely used benchmark image dataset CIFAR-10 (Krizhevsky et al., 2009) for the training and testing of the deep CNNs.

For the reader's convenience, our main results are summarized as follows:

1. The predictions of the trained convolution neural networks (CNNs) on the test image dataset pass the BRP feasibility test (7) Theorem 1. Table 2 summarizes our test

Index	CNN Architecture Description (sequentially, excluding input layer and output softmax layer)
1	**1 hidden layer:** conv2D $(32, (3 \times 3))$
2	**3 hidden layers:** conv2D $(32, (3 \times 3))$, conv2D $(32, (3 \times 3))$, Dropout 0.2, conv2D $(64, (3 \times 3))$, maxpool2D $(2, 2)$
3	**4 hidden layers:** conv2D $(32, (3 \times 3))$, conv2D $(32, (3 \times 3))$, maxpool2D $(2, 2)$, Dropout 0.2, conv2D $(128, (3 \times 3))$, conv2D $(128, (3 \times 3))$, maxpool2D $(2, 2)$, Dropout 0.4
4	**6 hidden layers:** conv2D $(32, (3 \times 3))$, conv2D $(32, (3 \times 3))$, maxpool2D $(2, 2)$, Dropout 0.2, conv2D $(64, (3 \times 3))$, conv2D $(64, (3 \times 3))$, maxpool2D $(2, 2)$, Dropout 0.3, conv2D $(128, (3 \times 3))$, conv2D $(128, (3 \times 3))$, maxpool2D $(2, 2)$, Dropout 0.4
5	**8 hidden layers:** conv2D $(32, (3 \times 3))$, conv2D $(32, (3 \times 3))$, maxpool2D $(2, 2)$, Dropout 0.2, conv2D $(64, (3 \times 3))$, conv2D $(64, (3 \times 3))$, maxpool2D $(2, 2)$, Dropout 0.3, conv2D $(64, (3 \times 3))$, conv2D $(64, (3 \times 3))$, maxpool2D $(2, 2)$, Dropout 0.3, conv2D $(128, (3 \times 3))$, conv2D $(128, (3 \times 3))$, maxpool2D $(2, 2)$, Dropout 0.4

Table 1. Parameter description of the CNNs constructed for rational inattention testing. All convolutional layers are followed by an ELU activation unit. Further, the weights of all convolutional layers are l2 regularized with weight 0.01. All neural networks were trained online on Google Colab using the GPU setting. The maximum RAM used was 0.91GB. Highest per epoch execution time was recorded to be 432 seconds for the deepest neural network (CNN # 5).
Notation: conv2D $(n, (a \times a))$ means the layer has n number of $(a \times a)$ convolutional filters; Dropout x means x fraction of the neuron links are randomly dropped in the layer; maxpool2D (a, a) means a max-pooling operation is performed at that using an (a, a) window.

results. Therefore we conclude that there exists a CUMRI tuple Θ which rationalizes the predictions of the CNNs and shows the CNNs are equivalent to rationally inattentive utility maximizers. Put differently, this shows that CNNs behave identically to a collection of behavioral economics-constrained rational decision makers.

2. Second, the BRP feasibility test generates a set-valued estimate of viable utility functions for the CNNs that explain (rationalize) their image classification behavior. These utility functions map the true image label and the predicted image label to the set of positive reals thereby introducing a useful preference ordering among the set of true and predicted image labels for each CNN. Such an ordering cannot be obtained otherwise by just viewing the decisions of the CNNs.

3. Finally, we explore robustness aspects of how well our proposed rational inattention models fit the decisions of the CNNs. We show that the fit is highly robust; namely, the decisions pass the BRP test by a large margin.

3.1. Experimental Setup : CNN Architecture, Training Specification and Construction of Dataset

CNNs have achieved state-of-the-art results on real-world image classification applications (Ciregan et al., 2012; Russakovsky et al., 2015; He et al., 2016). In this paper, we confine our experimental analysis only to CNNs. However, our methodology in Sec. 2 is not limited to CNNs and can be applied to any neural network architecture.

CNN Architecture. Our experimental setup comprises 5 convolutional neural networks. The CNN architectures and computation details are provided in Table 1. The smallest neural network has 1 convolutional layer, and the largest neural network has 8 convolutional layers. All convolutional layers are followed by an ELU (Exponential Linear Unit) (Clevert et al., 2015) activation unit. The weights of the convolutional layers are uniformly L2 regularized with regularization parameter 0.01 for all neural networks.

Image Dataset. All 5 CNNs were trained on the training images and validated on the test images available from the CIFAR-10 dataset. The CIFAR-10 dataset (Krizhevsky et al., 2009) is a public dataset that consists of 60000 32x32 colour images in 10 distinct classes (for example, airplane, automobile, ship, cat, dog etc.), with 6000 images per class. There are 50000 training images and 10000 test images. In the rest of the paper, we will use the terms image classes and image labels interchangeably.

Training Parameters. The training of the 5 CNNs via data augmentation[2] and testing on the CIFAR-10 test image dataset was conducted in **three** separate phases. The phases differ in the parameters used for training the CNNs as described below. Hence, to differentiate between the CNNs trained in different phases, we denote by $\mathcal{K}_1$, $\mathcal{K}_2$ and $\mathcal{K}_3$ the collection of trained CNNs in Phases 1, 2 and 3, respectively. That is, with respect to the 5 CNNs described in Table 1, and index i taking values in $\{1, 2, 3\}$,

$$\mathcal{K}_i := \text{collection of 5 CNNs trained in Phase } i \qquad (10)$$

All 3 phases use the RMSprop optimizer (Hinton et al., 2012) with the decay parameter set to 10^{-6}.

Phase 1. Train for 125 epochs using adaptive learning rate of 0.001 till 75 epochs followed by 0.005 for 25 epochs and 0.003 for the remaining 25 epochs.

[2]the details of the data augmentation process at the start of each epoch is provided in the supplementary material.

Phase 2. Train for 60 epochs using adaptive learning rate of 0.001 till 30 epochs followed by 0.005 for 15 epochs and 0.003 for the remaining 15 epochs.

Phase 3. Train for 30 epochs using fixed rate of 0.001.

Testing Procedure and Construction of Dataset for BRP Test. The image class predictions over all images in the CIFAR-10 test image dataset generated by all 5 CNNs was aggregated, separately for Phase 1,2 and 3, to construct three datasets $\mathbb{D}_1, \mathbb{D}_2$ and $\mathbb{D}_3$, in complete analogy to the dataset $\mathbb{D}$ used to test for rationally inattentive utility maximization (BRP feasibility test) in Theorem 1. The motivation and construction of the datasets is discussed below.

Suppose we view the CNNs in our experimental setup as Bayesian agents making decisions. The output predictions of the CNNs are analogous to the decisions of the agents. Relating the CUMRI model to the deep image classification context (recall Sec. 2.2), the true image class of the input test image to the CNNs is analogous to the state x of the Bayesian agents. Also, the predicted image class of the CNNs is analogous to the action a of the Bayesian agents. Hence, it is natural to check if the recorded predictions of the CNNs are consistent with rationally inattentive utility maximization. Keeping the above analogy in mind, we now construct dataset $\mathbb{D}_1$ corresponding to the collection of CNNs $\mathcal{K}_1$ to perform the BRP convex feasibility test.

Consider the collection of CNNs $\mathcal{K}_1$ trained in Phase 1. Let s_i denote the image class of the k^{th} test image in the CIFAR-10 test image dataset. Also, let $f_{i,k} \in \Delta^9$ denote the softmax output vector of the k^{th} CNN, $k \in \mathcal{K}_1$ for the i^{th} test image. The components of $f_{i,k}$ sum up to 1, and the j^{th} component is the confidence with which the CNN classifies the test image into image class j. Then, the corresponding dataset $\mathbb{D}_1$ is defined as

$$\begin{aligned}\mathbb{D}_1 =&\{\pi_0, p_k(a|x), x, a \in \mathcal{X}, k \in \mathcal{K}_1\}, \text{ where}\\ &\pi_0(x) = \sum_{i=1}^{N} \frac{\mathbb{1}\{s_i = x\}}{N},\\ &p_k(a|x) = \frac{\sum_{i=1}^{N} \mathbb{1}\{s_i = x\} f_{i,k}(a)}{\sum_{i=1}^{N} \mathbb{1}\{s_i = x\}},\\ &N = 10000,\ \mathcal{X} = \mathcal{A} = \{1, 2, \ldots 10\}. \qquad (11)\end{aligned}$$

The datasets $\mathbb{D}_2, \mathbb{D}_3$ for the collection of CNNs $\mathcal{K}_2, \mathcal{K}_3$ are constructed in an identical manner as (11). Let us briefly explain the above computations. $\pi_0(x)$ is the empirical probability that the image class of a test image in the CIFAR-10 test dataset is x. Since the output of the CNN is a probability vector, we compute $p_k(a|x)$ for the k^{th} CNN by averaging the a^{th} component of its output vector over all test images belonging to image class x. Finally, N is the number of test images in the CIFAR-10 test dataset, and since the set of true and predicted image classes are the same, $\mathcal{X} = \mathcal{A}$ which justifies the last equation in (11).

3.2. BRP Test Results for CNN datasets $\mathbb{D}_1, \mathbb{D}_2, \mathbb{D}_3$

In Sec. 3.1, we related the output predictions of the collections of trained CNNs $\mathcal{K}_1, \mathcal{K}_2$ and $\mathcal{K}_3$ (10) to decisions taken by collections of Bayesian agents, and constructed their corresponding datasets $\mathbb{D}_1, \mathbb{D}_2$ and $\mathbb{D}_3$ (11) for rational inattention testing. Recall from Theorem 1 that if the datasets pass the BRP feasibility test, it proves the existence of a human decision-making model tuple that explains the predictions of the CNNs. In this section, we show that indeed, the datasets pass the BRP feasibility test and the feasible utility functions comprise our interpretable deep image classification model for CNNs. The test results are described below and summarized in Table 2.

We applied the BRP feasibility test to each of the datasets $\mathbb{D}_1, \mathbb{D}_2, \mathbb{D}_3$ and $\mathbb{D}_1 \cup \mathbb{D}_2 \cup \mathbb{D}_3$. The datasets $\mathbb{D}_1$-$\mathbb{D}_3$ can be viewed as the predictions of 3 distinct collections of trained CNNs. Similarly, the dataset $\mathbb{D}_1 \cup \mathbb{D}_2 \cup \mathbb{D}_3$ can be viewed as the predictions of $5 \times 3 = 15$ distinct trained CNNs.
We observed that, remarkably, all 4 datasets pass the BRP test. This shows that the predictions of all 4 collections of trained CNNs $\mathcal{K}_1, \mathcal{K}_2, \mathcal{K}_3$ and $\mathcal{K}_1 \cup \mathcal{K}_2 \cup \mathcal{K}_3$ can be rationalized by the set-valued utility functions and information costs estimated by the feasibility test. Table 2 shows the computational details of the feasibility test results. Indeed, since $\mathbb{D}_1 \cup \mathbb{D}_2 \cup \mathbb{D}_3$ passed the BRP test, its subsets $\mathbb{D}_1 \cup \mathbb{D}_2$, $\mathbb{D}_1 \cup \mathbb{D}_3$ and $\mathbb{D}_2 \cup \mathbb{D}_3$ are guaranteed to pass the feasibility test.

3.3. Sparsity Enhanced and Robustness Optimized BRP Test for CNN Prediction Datasets

The previous section showed that the input-output datasets of the CNNs pass the BRP feasibility test. It is then natural to address the following questions wrt the set-valued utility function estimated by the test.
1. Sparsity: *How to represent the set of feasible utility functions with the fewest parameters, i.e., what is the sparsest solution to the BRP test?*
2. Robustness: *Which point in the feasible set passes the feasibility test with the largest margin, i.e., what is the most robust solution to the BRP test?*

Theorem 2 below provides answers to the above questions.

Theorem 2. *Given the input-output dataset $\mathbb{D}$ in (11) obtained from a collection of K rationally inattentive utility maximizers (BRP$(\mathbb{D}) \leq \mathbf{0}$ has a feasible solution). Then,*

1. *The sparsest solution is obtained by minimizing the sum of row-wise $\mathcal{L}_1$ norm of the feasible utility func-*

Dataset $\mathbb{D}$ (6) of decisions generated by CNNs	$\mathbb{D}_1$	$\mathbb{D}_2$	$\mathbb{D}_3$	$\mathbb{D}_1 \cup \mathbb{D}_2 \cup \mathbb{D}_3$
Number of CNNs that generate $\mathbb{D}$	5	5	5	15
BRP Feasibility Test Execution Time	20.63s	20.07s	17.56s	723.61s
Do CNN responses pass the BRP Feasibility Test?	**Yes**	**Yes**	**Yes**	**Yes**
Robustness metric $\mathcal{R}(\mathbb{D})$ (13)	0.0787	0.0256	0.0506	NA
Computation time to obtain sparsest solution	218.24s	210.37s	352.72s	144.5 min
Computation time to obtain most robust solution	62 min	68 min	67 min	>24 hours

Table 2. Summary of Experimental Results showing that Bayesian revealed preference test (BRP) of Theorem 1 rationalizes the image classification of a collection of deep CNNs. Remarkably, it is observed that all datasets $\mathbb{D}_1, \mathbb{D}_2$ and $\mathbb{D}_3$ generated by the CNNs pass the BRP test. All tests were performed in MATLAB using the general purpose numerical optimization algorithm `fmincon`. This uses the interior reflective Newton method (Coleman & Li, 1996) for large scale problems and sequential quadratic programming (Bertsekas, 1997) for medium scale problems. The tests were run on a quad core Intel i7-8650U processor using 252 MB of RAM, and running Windows 10 64-bit operating system.

tions $u_1, u_2, \ldots, u_K$ of the K agents that generate $\mathbb{D}$.

$$\min_{u_{1:K} \in \mathbb{R}_+^{|\mathcal{X}| \times |\mathcal{A}|}} \sum_{k=1}^{K} \|u_k\|_1, \ \textit{s.t. BRP}(\mathbb{D}) \leq \mathbf{0}, \quad (12)$$

where $\|\cdot\|_1$ denotes the row-wise $\mathcal{L}_1$ norm.

2. *The most robust solution is obtained by maximizing the margin with which the dataset $\mathbb{D}$ passes the BRP test. That is, in addition to decision variables $u_{1:K}, c_{1:K}$ in Algorithm 1, find $\varepsilon > 0$ such that*

$$\varepsilon = \max_{\varepsilon' > 0} \varepsilon', \ \textit{s.t. BRP}(\mathbb{D}) \leq -\varepsilon' \, [1 \, 1 \ldots \, 1]^T. \quad (13)$$

The optimal value of ε normalized wrt the row-wise $\mathcal{L}_2$-norm of the utility functions is the robustness metric of the dataset denoted by $\mathcal{R}(\mathbb{D})$.

Theorem 2 provides a useful procedure to compute two point estimates from the set of feasible solutions of the BRP test, namely, the sparsest and most robust solutions. The variables $u_{1:K}$ in Theorem 2 are the decision variables in the BRP feasibility test in Algorithm 1.

For a dataset that is apriori known to pass the BRP feasibility test of Theorem 1, the sparsest solution contains the minimum number of non-zero variables in the feasible solution set and minimizes the $\mathcal{L}_1$-norm of the vector of decision variables (12). The advantage of the sparsity enhanced solution is *memory-centric*; it represents the interpretable deep image classification model of the CNNs using the least number of parameters.

The motivation behind the robustness optimized solution in (13) in Theorem 2 stems from the minimum perturbation tests proposed by Varian et al. (1991) in the area of classical revealed preference theory (Afriat, 1967). Intuitively, the most robust solution needs the largest perturbation of all points in the feasible set of solutions, to fail the BRP feasibility test, which is precisely what (13) achieves. The advantage of the most robust solution is *experiment-centric*; this solution is most robust to measurement errors while constructing the dataset $\mathbb{D}$ (11) for the BRP feasibility test.

Summary. The main outcome of our behavioral economics based post-hoc analysis is that, remarkably, image classification using deep neural networks can be explained by rationally inattentive Bayesian utility maximization models. Table 2 summarizes our test results. The set-valued utility function estimated from BRPtest in Theorem 1 forms our interpretable deep image classification model. We also provided a procedure to compute two point estimates of the resulting set-valued utility function. One achieves the most sparsest representation of the feasible set of utilities, the other achieves the most robust (to errors in decision dataset computation) solution in the feasible set of utilities.

4. Conclusions and Extensions

This paper has developed behavioral economics models for interpretable deep learning. We showed that the decisions of a collection of deep CNNs performing image classification can be explained (rationalized) by rationally inattentive Bayesian utility maximization. Our methodology comprised two steps: First, using the theory of Bayesian revealed preference, Theorem 1 gave a necessary and sufficient condition for the actions of a collection of decision makers to be consistent with rationally inattentive Bayesian utility maximization. Second we showed that deep CNNs operating on the CIFAR-10 dataset satisfy these necessary and sufficient conditions. Moreover, we studied the robustness margin by which the deep CNNs satisfy Theorem 1; we found that the margins were sufficiently large implying robustness of the results. Indeed using Theorem 2 we reconstructed a sparse utility function and also maximized the robustness of the fit. At a more conceptual level, our results suggest that deep CNNs for image classification are equivalent to a behavioral-economics constrained Bayesian decision system (which is used in behavioral economics to approximate human decision making).

An immediate extension of this work is to replace the state, namely, the input class label, by appropriately designed image features. This would result in a richer descriptive model of the deep neural networks due to a larger number of parameters in the utility function Another extension of this work is to evaluate if other types of deep learning architectures are equivalent to rationally inattentive Bayesian utility maximization. Finally, it is also worthwhile compute the robustness of the Bayesian revealed preference tests using Shapley values (Lundberg & Lee, 2017) from coalitional game theory.

Supplental python and MATLAB codes required to completely reconstruct the results in this paper are available upon request.

A. Proof of Theorem 1

Proof of necessity of NIAS and NIAC:

1. NIAS (8): For agent $k \in \mathcal{K}$, define the subset $\mathcal{Y}_a \subseteq \mathcal{Y}$ so that for any observation $y \in \mathcal{Y}_a$, given posterior pmf $p_k(x|y)$, the optimal choice of action is a (3). We define the revealed posterior pmf given action a as $p_k(x|a)$. The revealed posterior pmf is a stochastically garbled version of the actual posterior pmf $p_k(x|y)$, that is,

$$p_k(x|a) = \sum_{y\in\mathcal{Y}} \frac{p_k(x,y,a)}{p_k(a)} = \sum_{y\in\mathcal{Y}} p_k(y|a)p_k(x|y) \quad (14)$$

   Since the optimal action is a for all $y \in \mathcal{Y}_a$, it follows from (3) that

$$\begin{aligned}
&\sum_{x\in\mathcal{X}} p_k(x|y)(u_k(x,b) - u_k(x,a)) \leq 0 \\
\implies &\sum_{y\in\mathcal{Y}_a} p_k(y|a) \sum_{x\in\mathcal{X}} p_k(x|y)(u_k(x,b) - u_k(x,a)) \leq 0 \\
\implies &\sum_{y\in\mathcal{Y}} p_k(y|a) \sum_{x\in\mathcal{X}} p_k(x|y)(u_k(x,b) - u_k(x,a)) \leq 0 \\
&\text{(since } p_k(y|a) = 0,\ \forall y \in \mathcal{Y}\backslash\mathcal{Y}_a) \\
\implies &\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p_k(y|a)p_k(x|y)(u_k(x,b) - u_k(x,a)) \leq 0 \\
\implies &\sum_{x\in\mathcal{X}} p_k(x|a)(u_k(x,b) - u_k(x,a)) \leq 0 \text{ (from (14))}
\end{aligned}$$

   This is precisely the NIAS inequality. Hence, the true utility functions of the collection of agents $\mathcal{K}$ comprise a feasible solution to the set of NIAS inequalities.
2. NIAC (9): Define scalars $c_k = C(\alpha_k)$, where $C(\cdot)$ denotes the information acquisition cost of the collection of agents $\mathcal{K}$. By definition, $c_k > 0$. Also, let $J(\alpha_k, u_k)$ denote the expected utility of the k^{th} agent when using attention function α_k (first term in RHS of (4)). Here, the expectation is taken wrt both the state x and observation y. Finally, for the k^{th} agent, we define the revealed attention function α'_k over the set of actions $\mathcal{A}$ as

$$\alpha'_k(a|x) = p_k(a|x),\ \forall a \in \mathcal{A},$$

   where the variable $p_k(a|x)$ is obtained from the dataset $\mathbb{D}$. Clearly, the revealed attention function is a stochastically garbled version of the true attention function since

$$\alpha'_k(a|x) = p_k(a|x) = \sum_{y\in\mathcal{Y}} p_k(a|y)\alpha_k(y|x) \quad (15)$$

   As an intermediate step to show the necessity of NIAC for rationally inattentive utility maximization to

hold, we obtain a relationship between $J(\alpha_k, u_j)$ and $J(\alpha'_k, u_j)$ as follows.

$$\begin{aligned}
&J(\alpha'_k, u_j)\\
=&\sum_a \max_{b\in\mathcal{A}} \sum_x p_k(x,a)u_j(x,b)\\
=&\sum_a \max_{b\in\mathcal{A}} \sum_{y,x} p_k(a|y)\alpha_k(y|x)\pi_0(x)u_j(x,b)\\
\leq&\sum_a \sum_y p_k(a|y) \sum_x \max_{b\in\mathcal{A}} \alpha_k(y|x)\pi_0(x)u_j(x,b)\\
=&\sum_y \left(\sum_a p_k(a|y)\right) \sum_x \max_{b\in\mathcal{A}} \alpha_k(y|x)\pi_0(x)u_j(x,b)\\
=&J(\alpha_k, u_j). \qquad (16)
\end{aligned}$$

The second equality follows from (15), and the inequality follows from Jensen's inequality. Hence, we establish the relation that $J(\alpha_k, u_j) \geq J(\alpha'_k, u_j)$, where equality holds if $j = k$ (this condition is due to the NIAS inequality (8)).

Using condition (4) for optimal choice of attention function by the agents in $\mathcal{K}$, and the above inequalities (16), we are now ready to show the necessity of the NIAC inequality (9) for rationally inattentive utility maximization. That is, for any agent $k \in \mathcal{K}$,

$$\begin{aligned}
J(\alpha_k, u_k) - c_k =& J(\alpha'_k, u_k) - c_k\\
\geq\ & J(\alpha_j, u_k) - c_j\\
\geq\ & J(\alpha'_j, u_k) - c_j,\ \forall\, j \in \mathcal{K}, j \neq k.
\end{aligned}$$

This is precisely the NIAC inequality, hence the true utility functions and associated information acquisition costs incurred by the agents comprise a feasible solution to the set of NIAC inequalities.

Proof for sufficiency of NIAS and NIAC: To prove sufficiency, we provide a constructive map from the dataset $\mathbb{D}$ and any point in the feasible set of solutions generated by BRP test to a viable CUMRI model tuple Θ that rationalizes $\mathbb{D}$. Specifically, we show that the constructed tuple Θ satisfies (3) and (4) for all agents that generate $\mathbb{D}$.

Construction of Θ*:* Given dataset $\mathbb{D}$ and a point in the feasible set of solutions generated by BRP test. The parameters $\mathcal{K}, \mathcal{X}, \mathcal{A}$ in the tuple Θ are obvious from the dataset $\mathbb{D}$. Let the utility functions of the agents be the feasible utility values $\{u_k, k \in \mathcal{K}\}$ in the feasible point. Define the set of observations $\mathcal{Y} = \{y_1, y_2, \ldots, y_{|\mathcal{A}|}\}$, where for each agent $k \in \mathcal{K}$, there exists a bijective map r_k from the constructed $\mathcal{Y}$ to the set of observed actions in $\mathbb{D}$ such that $p_k(r_k(y)|y) = 1$. Define the k^{th} agent's attention function $\alpha_k(y|x) = p_k(r_k(y)|x)$. The feasible point also comprises positive reals $\{c_k, k \in \mathcal{K}\}$ that we will use to define the information acquisition cost below.

$$C(\alpha) = \max_{k\in\mathcal{K}}\ c_k + J(\alpha, u_k) - J(\alpha_k, u_k) \qquad (17)$$

In (17), $C(\cdot)$ is a piece-wise linear concave function of the attention function α. The variable $J(\alpha, u)$ denotes the expected value of utility function u obtained by using attention function α (first term in the RHS of (4)), where the expectation is taken wrt both state x and action a. Further, since the NIAC inequality is satisfied, (17) implies $C(\alpha_k) = c_k$.

Having constructed the CUMRI tuple Θ, it only remains to show that inequalities (3) and (4) in Definition 1 are satisfied for all agents in $\mathcal{K}$.

1. *NIAS implies* (8) *holds.* From NIAS (8), we know that for any agent $k \in \mathcal{K}$, actions $a, b \in \mathcal{A}, a \neq b$, the following inequalities hold.

$$\begin{aligned}
&\sum_x p_k(x|a)(u_k(x,b) - u_k(x,a)) \leq 0\\
\implies &\sum_x p_k(x|y = r_k^{-1}(a))(u_k(x,b) - u_k(x,a)) \leq 0\\
\implies &a \in \operatorname*{argmax}_{b\in\mathcal{A}} \sum_x p_k(x|y)u_k(x,b) \implies (3).
\end{aligned}$$

2. *Information Acquisition Cost* (17) *implies* (9) *holds.* Then, for any attention function α, the following inequalities result.

$$\begin{aligned}
&C(\alpha) = \max_{k\in\mathcal{K}}\ c_k + J(\alpha, u_k) - J(\alpha_k, u_k)\\
\implies &C(\alpha) \geq c_k + J(\alpha, u_k) - J(\alpha_k, u_k),\ \forall k \in \mathcal{K}\\
\implies &J(\alpha_k, u_k) - c_k \geq J(\alpha_k, u_k) - C(\alpha),\ \forall k \in \mathcal{K}\\
\implies &\alpha_k \in \operatorname*{argmax}_{\alpha'} J(\alpha', u_k) - C(\alpha') \implies (4).
\end{aligned}$$

This completes the proof of Theorem 1.

References

Afriat, S. N. The construction of utility functions from expenditure data. *International economic review*, 8(1):67–77, 1967.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Bertsekas, D. P. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.

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

Caplin, A. and Martin, D. A testable theory of imperfect perception. *The Economic Journal*, 125(582):184–202, 2015.

Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M., Cerutti, F., Srivastava, M., Preece, A., Julier, S., Rao, R. M., et al. Interpretability of deep learning models: a survey of results. In *2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI)*, pp. 1–6. IEEE, 2017.

Ciregan, D., Meier, U., and Schmidhuber, J. Multi-column deep neural networks for image classification. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3642–3649. IEEE, 2012.

Clevert, D.-A., Unterthiner, T., and Hochreiter, S. Fast and accurate deep network learning by exponential linear units (ELUs). *arXiv preprint arXiv:1511.07289*, 2015.

Coleman, T. F. and Li, Y. An interior trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on optimization*, 6(2):418–445, 1996.

Diewert, W. E. Afriat and revealed preference theory. *The Review of Economic Studies*, 40(3):419–425, 1973.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5):1–42, 2018.

Hase, P., Chen, C., Li, O., and Rudin, C. Interpretable image recognition with hierarchical prototypes. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7-1, pp. 32–40, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hinton, G., Srivastava, N., and Swersky, K. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8), 2012.

Hoiles, W., Krishnamurthy, V., and Pattanayak, K. Rationally Inattentive Inverse Reinforcement Learning Explains YouTube commenting behavior. *The Journal of Machine Learning Research*, 21(170):1–39, 2020.

Huang, L. and Liu, H. Rational inattention and portfolio selection. *The Journal of Finance*, 62(4):1999–2040, 2007.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Lei, T., Barzilay, R., and Jaakkola, T. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*, 2016.

Lundberg, S. and Lee, S.-I. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

Milgrom, P. Good news and bad news: Representation theorems and applications. *Bell Journal of Economics*, 12 (2):380–391, 1981.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*, 2019.

Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., and Clune, J. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *arXiv preprint arXiv:1605.09304*, 2016.

Ribeiro, M. T., Singh, S., and Guestrin, C. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252, 2015.

Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

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

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

Varian, H. R. The nonparametric approach to demand analysis. *Econometrica: Journal of the Econometric Society*, pp. 945–973, 1982.

Varian, H. R. et al. *Goodness-of-fit for revealed preference tests*. Department of Economics, University of Michigan Ann Arbor, 1991.

Wang, H. and Yeung, D.-Y. Towards bayesian deep learning: A framework and some existing methods. *IEEE Transactions on Knowledge and Data Engineering*, 28 (12):3395–3408, 2016.