

## A Deep Choice Model

**Makoto Otsuka**

CREST, JST  
4-1-8 Honcho, Kawaguchi,  
Saitama 332-0012, Japan  
motsuka@ucla.edu

**Takayuki Osogami**

IBM Research - Tokyo  
19-21 Hakozaeki, Chuo-ku,  
Tokyo 103-8510, Japan  
osogami@jp.ibm.com

### Abstract

Human choice is complex in two ways. First, human choice often shows complex dependency on available alternatives. Second, human choice is often made after examining complex items such as images. The recently proposed choice model based on the restricted Boltzmann machine (RBM choice model) has been proved to represent three typical phenomena of human choice, which addresses the first complexity. We extend the RBM choice model to a deep choice model (DCM) to deal with the features of items, which are ignored in the RBM choice model. We then use deep learning to extract latent features from images and plug those latent features as input to the DCM. Our experiments show that the DCM adequately learns the choice that involves both of the two complexities in human choice.

### Introduction

The models of choice have been extensively studied in artificial intelligence (Faradani, Hartmann, and Ipeirotis 2011; Pfeiffer et al. 2012; Zhen et al. 2015) and other areas (Farias, Jagabathula, and Shah 2013). The research has revealed that people show rather irrational but systematic bias in their choices, and our choices are influenced by available alternatives in a complex manner (Rieskamp, Bussemeyer, and Mellers 2006). For example, the probability of choosing an item, A, can be increased by adding an item, D, into the choice set, when D is inferior to A in all aspects under consideration (i.e., D acts as a decoy). This phenomenon is known as the attraction effect (Rieskamp, Bussemeyer, and Mellers 2006). In addition, we make choices after examining complex information. For example, we often choose or purchase items after looking at photos or reading textual information about products. However, these natural stimuli have not been the scope of research on choice due to the difficulty of handling their high-dimensional features. In fact, it has been confirmed in psychological experiments that phenomena such as the attraction effect appear when people make choices from images of human faces (Ariely 2010).

Existing choice models either cannot represent some of the biases in human choice or cannot effectively handle natural stimuli. A choice model using a restricted Boltzmann machine (RBM), called an RBM choice model, was

recently proposed (Osogami and Otsuka 2014). The RBM choice model is the only choice model that has been proved to represent three typical phenomena of human choice (the similarity effect, the compromise effect, and the attraction effect) and can learn those phenomena from choice data. The RBM choice model, however, cannot handle features of items. In this model, two items are either identical or distinct, and there is no notion of similarity between items. Conventional choice models, including the standard multinomial logit model (MLM), take into account the features of items, but there has been no work that deals with natural stimuli with choice models in the literature.

Deep learning (Hinton and Salakhutdinov 2006; Vincent et al. 2010; Salakhutdinov and Hinton 2009a) extracts distributed representations of features that can be used for additional learning. The performance of classification and regression is greatly improved using deep learning in various fields including image recognition (Krizhevsky, Sutskever, and Hinton 2012) and audio classification (Lee et al. 2009). To date, deep learning has not been applied to choice models, and it is non-trivial how deep learning should be applied, particularly to the RBM choice model, in a way that represents the typical phenomena of human choice.

We propose a deep choice model (DCM) by extending the RBM choice model to incorporate features extracted through deep learning, which is the first contribution of this paper. We show that the DCM generalizes the RBM choice model, so that the DCM is guaranteed to represent all of the typical choice phenomena in the sense of Osogami and Otsuka (2014) when it corresponds to the RBM choice model. We will also see that the DCM generalizes the MLM in such a way that it represents the typical choice phenomena. In addition, the DCM can be trained by following the exact log-likelihood gradient without approximation.

We validate the effectiveness of the DCM with experiments that capture the essence of human choice from a set of natural stimuli, which constitute our second contribution. Specifically, our choice set consists of images, and the choice exhibits the attraction effect. Our experiments show that we can train the DCM in such a way that the trained DCM can adequately predict the choice probabilities for unseen images or unseen combinations of images.

## Related work

Much of the research on choice models has been on the MLM and its variants. The MLM can be derived from the axiom of independence from irrelevant alternatives (IIA), which essentially states that the ratio between choice probabilities of two items should not depend on what other items are in the choice set (Luce 1959; McFadden 1974). As the human choice often violates the IIA axiom (Rieskamp, Busemeyer, and Mellers 2006), many variants of the MLM have been studied in the literature. These include hierarchical MLM (Chapelle and Harchaoui 2005), multinomial probit model (Hausman and Wise 1978), and nested logit model (Ben-Akiva 1973). More generally, these variants of the MLM fall into the class of random utility models (McFadden 1981). Although these random utility models can violate the IIA axiom, they inherently cannot represent the attraction effect (Rieskamp, Busemeyer, and Mellers 2006).

Sequential sampling models (Busemeyer and Townsend 1993; Usher and McClelland 2004; Otter et al. 2008), on the other hand, mimic the cognitive process of the human making a choice and have been shown to represent the typical phenomena of human choice, including the attraction effect, for some specific cases. However, no algorithms have been proposed to train a sequential sampling model in a way that the trained model exhibits the typical choice phenomena.

Research on choice models has recently focused on learning and representing the typical phenomena of human choice (Osogami and Katsuki 2014; Osogami and Otsuka 2014; Shenoy and Yu 2013; Takahashi and Morimura 2015). The RBM choice model has the unique characteristic that it can be proved to represent all three typical phenomena of human choice (Osogami and Otsuka 2014).

Unlike other choice models, however, the RBM choice model cannot deal with the features of items, which has significantly limited its applicability. More precisely, Osogami and Otsuka (2014) briefly suggest a way to incorporate the features of items into the RBM choice model but does not discuss (and it is nontrivial) how to train the RBM choice model when the features are incorporated in the suggested approach. The DCM incorporates the features differently from what was suggested in Osogami and Otsuka (2014) by using the ideas from replicated soft max (Salakhutdinov and Hinton 2009b; Srivastava, Salakhutdinov, and Hinton 2013).

## Deep choice model

The DCM gives the probability of selecting each item, or each subset of items, in a given choice set. In this section, we introduce the DCM and show that the DCM generalizes the MLM and the RBM choice model. We then present a training algorithm for the DCM.

### Model architecture

A choice set consisting of  $D$  items is represented by a set of  $D$  vectors,  $\mathcal{X} \equiv \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(D)}\}$ . For  $d \in [1, D]$ , a  $K$ -dimensional binary vector,  $\mathbf{x}^{(d)} \in \{0, 1\}^K$ , represents the features of an item. The DCM gives the probability of selecting an item or a subset of items from a given choice set. The set of selected items (the selected set) consisting of  $C$  items

is represented by  $\mathcal{Y} \equiv \{\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(C)}\}$ , where  $\mathbf{y}^{(c)} \in \mathcal{X}$  for  $c \in [1, C]$ . We do not a priori fix the size of the choice set or the selected set, allowing  $C$  and  $D$  to vary.

The DCM is a restricted Boltzmann machine (RBM; Hinton (2002)) that has the particular structure shown in Fig. 1a. An RBM is composed of a layer of visible nodes (i.e., visible layer) and a layer of hidden nodes (i.e., hidden layer). While a node in one layer can be connected to a node in the other layer, there are no intra-layer connections. In the DCM, the visible layer (bottom part of Fig. 1a) is separated into two parts: input layers and output layers.

The input layers (the left part of the visible layer) represent the choice set,  $\mathcal{X}$ . For  $d \in [1, D]$ , the  $d$ -th input layer represents  $\mathbf{x}^{(d)}$ , the  $K$ -dimensional binary feature of the  $d$ -th item in  $\mathcal{X}$ . The output layers (the right part of the visible layer) represent the selected set,  $\mathcal{Y}$ . For  $c \in [1, C]$ , the  $c$ -th output layer represents  $\mathbf{y}^{(c)}$ , the feature of the  $c$ -th item in  $\mathcal{Y}$ . The number of input layers and output layers is variable, because  $C$  and  $D$  are variable.

To absorb the impact of the varying  $C$  and  $D$  as well as the imbalance between  $C$  and  $D$ , we replicate the input layer  $C$  times and the output layer  $D$  times. This results in  $C D$  input layers and  $C D$  output layers. Let  $x_{c,d,i}$  denote the value of the  $i$ -th node in the  $d$ -th layer of the  $c$ -th replica of the input:  $x_{c,d,i} = (\mathbf{x}^{(d)})_i$  for  $i \in [1, K]$ ,  $d \in [1, D]$ , and  $c \in [1, C]$ . Let  $y_{d,c,j}$  denote the value of the  $j$ -th node in the  $c$ -th layer of the  $d$ -th replica of the output:  $y_{d,c,j} = (\mathbf{y}^{(c)})_j$  for  $j \in [1, K]$ ,  $c \in [1, C]$ , and  $d \in [1, D]$ . The hidden layer is not replicated and is composed of  $L$  nodes, taking the binary value  $h_\ell \in \{0, 1\}$  for  $\ell \in [1, L]$ . Let  $\mathbf{h} \equiv (h_\ell)_{\ell \in [1, L]}$ .

The DCM is parameterized by bias and weight. Bias is associated with each node, and weight is associated with each connection of a pair of nodes. We let some of these parameters share common values to keep the probability of selecting  $\mathcal{Y}$  from  $\mathcal{X}$  invariant to the order of the items in  $\mathcal{Y}$  or  $\mathcal{X}$ .

For reasons analogous to the RBM choice model (Osogami and Otsuka 2014), the bias on the input nodes would have no effect and are not considered here. For  $\ell \in [1, L]$ , let  $b_\ell^{\text{hid}}$  denote the bias associated with the  $\ell$ -th hidden node. Let  $\mathbf{b}^{\text{hid}} \equiv (b_\ell^{\text{hid}})_{\ell \in [1, L]} \in \mathbb{R}^L$  denote the vector of the bias for the hidden nodes. Let  $b_{d,c,j}^{\text{out}}$  denote the bias associated with the  $j$ -th node in the  $c$ -th layer of the  $d$ -th replica of the output. The bias of the output nodes shares common values and is normalized, depending on the size of the choice set and the selected set, as follows:

$$b_{d,c,j}^{\text{out}} = \frac{1}{C D} b_j^{\text{out}} \quad (1)$$

for  $j \in [1, K]$ . Let  $\mathbf{b}^{\text{out}} \equiv (b_j^{\text{out}})_{j \in [1, K]} \in \mathbb{R}^K$  denote the bias shared among the output nodes.

The weight is associated with each connection between a hidden node and one of the input or output nodes. Let  $W_{c,d,i,\ell}$  denote the weight for the connection between the  $\ell$ -th hidden node and the  $i$ -th node of the  $d$ -th layer of the  $c$ -th replica of the input. Let  $U_{d,c,j,\ell}$  denote the weight for the connection between the  $\ell$ -th hidden node and the  $j$ -th node of the  $c$ -th layer of the  $d$ -th replica of the output. The weight

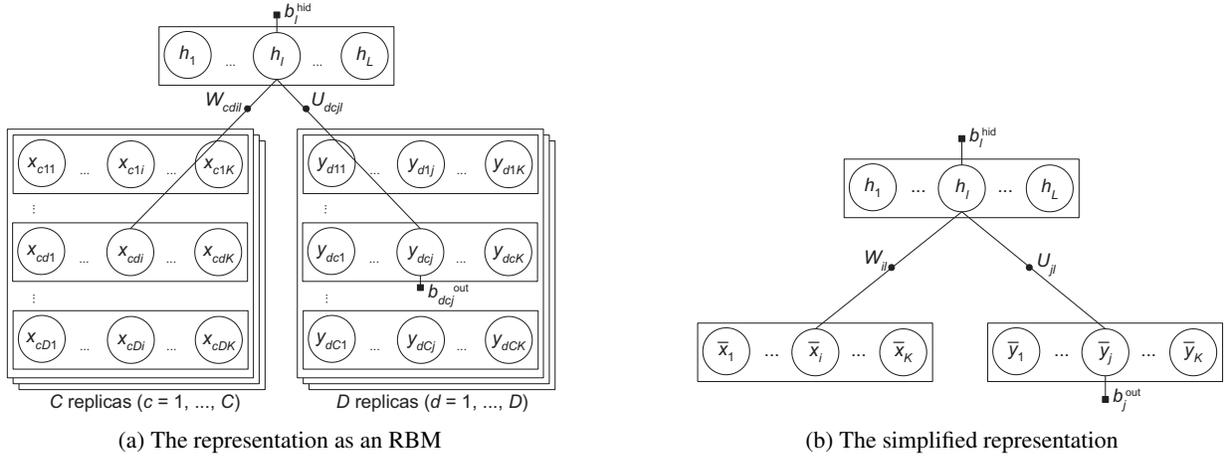


Figure 1: Two representations of the DCM, where the input (bottom left part) is connected to the output (bottom right part) via a hidden layer (top part). (a) Here, each node takes a binary value. (b) Here, input and output nodes take real values, while hidden nodes take binary values.

shares common values and is normalized as follows:

$$W_{c,d,i,\ell} = \frac{1}{CD} W_{i,\ell}, \quad U_{d,c,j,\ell} = \frac{1}{CD} U_{j,\ell} \quad (2)$$

for  $i, j \in [1, K]$  and  $\ell \in [1, L]$ . To denote the weight shared among the connections, let  $\mathbf{W} \equiv (W_{i,\ell})_{i \in [1, K], \ell \in [1, L]}$  and  $\mathbf{U} \equiv (U_{j,\ell})_{j \in [1, K], \ell \in [1, L]}$ , where  $\mathbf{W}, \mathbf{U} \in \mathbb{R}^{K \times L}$ .

The parameters of the DCM are collectively denoted by  $\theta \equiv \{\mathbf{W}, \mathbf{U}, \mathbf{b}^{\text{out}}, \mathbf{b}^{\text{hid}}\}$ . The values of  $(\mathcal{X}, \mathcal{Y}, \mathbf{h})$  and  $\theta$  determine the energy of the DCM,  $E(\mathcal{X}, \mathcal{Y}, \mathbf{h}; \theta)$ , which in turn defines the probability distribution over the values of the nodes of the DCM.

In accordance with the standard properties of the RBM, the energy of the DCM is represented as follows:

$$\begin{aligned} E(\mathcal{X}, \mathcal{Y}, \mathbf{h}; \theta) = & - \sum_{\ell=1}^L h_{\ell} b_{\ell}^{\text{hid}} - \sum_{d=1}^D \sum_{c=1}^C \sum_{j=1}^K y_{d,c,j} b_{d,c,j}^{\text{out}} \\ & - \sum_{d=1}^D \sum_{c=1}^C \sum_{i=1}^K \sum_{\ell=1}^L x_{c,d,i} W_{c,d,i,\ell} h_{\ell} \\ & - \sum_{d=1}^D \sum_{c=1}^C \sum_{j=1}^K \sum_{\ell=1}^L y_{d,c,j} U_{d,c,j,\ell} h_{\ell}. \end{aligned} \quad (3)$$

Plugging the shared values (1)-(2) into (3), we obtain

$$E(\mathcal{X}, \mathcal{Y}, \mathbf{h}; \theta) = -\mathbf{h}^{\top} \mathbf{b}^{\text{hid}} - \bar{\mathbf{y}}^{\top} \mathbf{b}^{\text{out}} - \bar{\mathbf{x}}^{\top} \mathbf{W} \mathbf{h} - \bar{\mathbf{y}}^{\top} \mathbf{U} \mathbf{h}, \quad (4)$$

where

$$\bar{\mathbf{x}} \equiv (\bar{x}_i)_{i \in [1, K]} \equiv \frac{1}{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{X}} \mathbf{x} \quad (5)$$

$$\bar{\mathbf{y}} \equiv (\bar{y}_j)_{j \in [1, K]} \equiv \frac{1}{|\mathcal{Y}|} \sum_{\mathbf{y} \in \mathcal{Y}} \mathbf{y}. \quad (6)$$

The simplified expression (4) provides an alternative perspective of the DCM, as illustrated in Fig. 1b. Similar to Fig. 1a, the DCM in Fig. 1b is represented with a hidden

layer and a visible layer. Here, the visible layer is split into a single input layer and a single output layer, as opposed to the  $CD$  layers in Fig. 1a. The visible nodes in Fig. 1b take real values,  $\bar{\mathbf{x}}$  and  $\bar{\mathbf{y}}$ , so that this representation is not a proper RBM. The following argument is based on the representation in Fig. 1a, where the nodes take binary values.

To define the choice probability with the DCM, let  $F(\mathcal{X}, \mathcal{Y}; \theta)$  be the equilibrium free energy given  $(\mathcal{X}, \mathcal{Y})$ :

$$F(\mathcal{X}, \mathcal{Y}; \theta) \equiv \sum_{\tilde{\mathbf{h}}} E(\mathcal{X}, \mathcal{Y}, \tilde{\mathbf{h}}; \theta), \quad (7)$$

where the summation with respect to  $\tilde{\mathbf{h}}$  is over all of the possible binary vectors of  $L$  dimensions. Then the probability of selecting  $\mathcal{Y}$  given  $\mathcal{X}$  is defined as follows:

$$p(\mathcal{Y}|\mathcal{X}; \theta) = \frac{\exp(-F(\mathcal{X}, \mathcal{Y}; \theta))}{\sum_{\tilde{\mathcal{Y}} \in \mathcal{Z}} \exp(-F(\mathcal{X}, \tilde{\mathcal{Y}}; \theta))}, \quad (8)$$

where  $\mathcal{Z} \subset 2^{\mathcal{X}}$  denotes the candidates of the selected set.

Using the energy function of the DCM defined in Eq. (4), we express the free energy given  $\mathcal{X}$  and  $\mathcal{Y}$  as follows:

$$F(\mathcal{X}, \mathcal{Y}; \theta) = -\bar{\mathbf{x}}^{\top} \mathbf{W} \hat{\mathbf{h}} - \bar{\mathbf{y}}^{\top} \mathbf{U} \hat{\mathbf{h}} - \bar{\mathbf{y}}^{\top} \mathbf{b}^{\text{out}} - \hat{\mathbf{h}}^{\top} \mathbf{b}^{\text{hid}}, \quad (9)$$

where, letting  $\sigma(z) \equiv (1 + \exp(-z))^{-1}$  be the logistic sigmoid function, we define  $\hat{\mathbf{h}} \equiv (\hat{h}_{\ell})_{\ell \in [1, L]}$  such that

$$\hat{h}_{\ell} = \sigma\left(b_{\ell} + \sum_{i=1}^K \bar{x}_i W_{i,\ell} + \sum_{j=1}^K \bar{y}_j U_{j,\ell}\right) \quad (10)$$

is the probability that the hidden node  $\ell \in [1, L]$  takes the value of 1 given  $\bar{\mathbf{x}}$  and  $\bar{\mathbf{y}}$ . We will discuss ways to calculate (8) when we introduce a training algorithm in the following.

The DCM extends replicated softmax (Salakhutdinov and Hinton 2009b; Srivastava, Salakhutdinov, and Hinton 2013)

to the model with input and output. Replicated softmax has been used to absorb the impact of the varying size of the document, where a visible softmax layer representing a word is replicated to the number of total words. Each softmax layer has a shared bias and is connected to the hidden layer with shared weights. In replicated softmax, the bias for the hidden layer is multiplied by the total number,  $B$ , of words, which we interpret as making  $B$  copies of RBMs sharing a common hidden activation to deal with the case where both input and output have varying size. Unlike replicated softmax, the layers of the DCM are not softmax but binary.

### Relationship to other choice models

The DCM is reduced to the MLM, when  $C = 1$ ,  $\mathbf{W} = \mathbf{0}$ ,  $\mathbf{U} = \mathbf{0}$ , and  $\mathbf{b}^{\text{hid}} = \mathbf{0}$ . In this case, the choice probability of  $\mathbf{y}$  from  $\mathcal{X}$  is expressed as follows:

$$p(\mathbf{y}|\mathcal{X}; \mathbf{b}^{\text{out}}) = \frac{\exp(\mathbf{y}^\top \mathbf{b}^{\text{out}})}{\sum_{\tilde{\mathbf{y}} \in \mathcal{X}} \exp(\tilde{\mathbf{y}}^\top \mathbf{b}^{\text{out}})}, \quad (11)$$

which is exactly the choice probability given by the MLM. The MLM can also be seen as the DCM with no hidden nodes ( $L = 0$ ) when exactly one item is selected ( $C = 1$ ).

The DCM is reduced to the RBM choice model when the binary features of the items are the standard basis of the  $K$ -dimensional Euclidean space. Here, there are  $K$  distinct items that can be included in a choice set. For  $1 \leq k \leq K$ , the feature of the  $k$ -th such item is the unit vector whose  $k$ -th element is 1 and the other elements are 0. This means that the  $i$ -th element of the feature vector denotes whether the item is the  $i$ -th item in  $\mathcal{I}$ . Then  $\bar{x}_i = 1/C$  if  $\mathbf{x}^{(i)} \in \mathcal{X}$ , and  $\bar{x}_i = 0$  otherwise. Likewise,  $\bar{y}_j = 1/D$  if  $\mathbf{y}^{(j)} \in \mathcal{Y}$ , and  $\bar{y}_j = 0$  otherwise. Multiplying  $\mathbf{x}$  by  $C$  and  $\mathbf{y}$  by  $D$ , then dividing  $\mathbf{W}$  by  $C$  and  $\mathbf{U}$  by  $D$ , we keep the energy of the DCM unchanged and obtain the energy of the RBM choice model shown in Osogami and Otsuka (2014).

### Training

We train the DCM, parameterized by  $\theta$ , by following the gradient of conditional log-likelihood:

$$\nabla_\theta \log \prod_{(\mathcal{X}, \mathcal{Y}) \in \mathcal{D}} p(\mathcal{Y}|\mathcal{X}; \theta) = \sum_{(\mathcal{X}, \mathcal{Y}) \in \mathcal{D}} \nabla_\theta \log p(\mathcal{Y}|\mathcal{X}; \theta), \quad (12)$$

where  $\mathcal{D}$  is the set of training data, consisting of pairs of a choice set,  $\mathcal{X}$ , and a selected set,  $\mathcal{Y}$ . By (8) and (9), we have

$$\begin{aligned} & \nabla_\theta \ln p(\mathcal{Y}|\mathcal{X}; \theta) \\ &= -\nabla_\theta F(\mathcal{X}, \mathcal{Y}; \theta) + \sum_{\tilde{\mathcal{Y}} \in \mathcal{Z}} p(\tilde{\mathcal{Y}}|\mathcal{X}; \theta) \nabla_\theta F(\mathcal{X}, \tilde{\mathcal{Y}}; \theta). \end{aligned} \quad (13)$$

The gradient of the free energy in (13) with respect to each parameter is given as follows:

$$\begin{aligned} \nabla_{\mathbf{W}} F(\mathcal{X}, \mathcal{Y}; \theta) &= -\bar{\mathbf{x}} \hat{\mathbf{h}}^\top, & \nabla_{\mathbf{U}} F(\mathcal{X}, \mathcal{Y}; \theta) &= -\bar{\mathbf{y}} \hat{\mathbf{h}}^\top, \\ \nabla_{\mathbf{b}^{\text{out}}} F(\mathcal{X}, \mathcal{Y}; \theta) &= -\bar{\mathbf{y}}, & \nabla_{\mathbf{b}^{\text{hid}}} F(\mathcal{X}, \mathcal{Y}; \theta) &= -\hat{\mathbf{h}}. \end{aligned} \quad (14)$$

Eq. (9) and Eq. (14) suggest that  $F(\mathcal{X}, \mathcal{Y}; \theta)$  can be evaluated in  $O(KL)$  time. With Eq. (8), we can then evaluate Eq. (13) in  $O(|\mathcal{Z}|KL)$  time, where recall that  $\mathcal{Z}$  denotes the candidates of selected sets. When the size of  $\mathcal{Z}$  is small, Eq. (13) can be evaluated exactly. For example, if we consider only the cases where exactly one item is selected, we have  $|\mathcal{Z}| = |\mathcal{X}|$ . When  $|\mathcal{Z}|$  is prohibitively large, we can estimate the second term in Eq. (13) using Gibbs sampling.

## Numerical experiments

We now validate the effectiveness of the combination of the DCM and deep learning through numerical experiments. Because data of choosing from natural stimuli is unavailable, we will create a dataset from publicly available images.

### Digit choice task

We consider the setting where an agent chooses an image from a set of images. The goal of our task, which we refer to as the digit choice task, is to learn the probability distribution that the agent follows when the agent chooses images. Training data consist of the pairs of  $\mathcal{X}$ , a choice set of images, and  $\mathcal{Y}$ , the selected image (here,  $C = |\mathcal{Y}| = 1$ ). Specifically, we use the gray-scale images of handwritten digits from the MNIST dataset<sup>1</sup>. Due to the size ( $28 \times 28$ ) of the images in the MNIST dataset, the item shown to the agent is represented as a 784-dimensional vector  $[0, 1]^{784}$ , where the pixel intensity is scaled between 0 and 1.

To create the training dataset, we use 100 images for each of the three digits: 0, 1, and 9. These 300 images are randomly selected from the MNIST dataset without duplication. Our choice set,  $\mathcal{X}$ , consists of two or three images of the three digits:  $\{0, 1\}$ ,  $\{0, 1, 9\}$ , or  $\{1, 9\}$ . Fig. 2a shows the probability distribution that the agent follows. Specifically, when the choice set consists of images of 0 and 1, the agent chooses the image of 0 with probability 0.9 and the image of 1 otherwise. When the choice set consists of images of 0, 1, and 9, the choice probabilities are 0.1, 0.8, and 0.1, respectively. When the choice set consists of images of 1 and 9, the choice probabilities are 0.1 and 0.9, respectively.

Notice that the choice probabilities in Fig. 2a cause the attraction effect. Specifically, the images of 0 are more attractive (selected more frequently) than the images of 1, when only images of 0 and 1 are in the choice set. However, the images of 1 are more attractive than the images of 0, when an image of 9 is in the choice set together with images of 0 and 1. That is, an image of 9 acts as a decoy and increases the relative attractiveness of an image of 1.

For each of the three combinations of the three digits ( $\{0, 1\}$ ,  $\{0, 1, 9\}$ , or  $\{1, 9\}$ ), we create 1,000 pairs of  $\mathcal{X}$  and  $\mathcal{Y}$  in such a way that the empirical distribution of the choices exactly matches the target distribution in Fig. 2a. For example, for the combination of  $\{0, 1\}$ , we randomly choose 1,000 pairs of an image of 0 and an image of 1 from the 100 images prepared for training. These constitute the 1,000 choice sets. For each choice set of two images,  $\mathcal{X}$ , one of the two images is chosen as the single element of the selected set,  $\mathcal{Y}$ . Specifically, an image of 1 is selected as the element

<sup>1</sup><http://yann.lecun.com/exdb/mnist/index.html>.

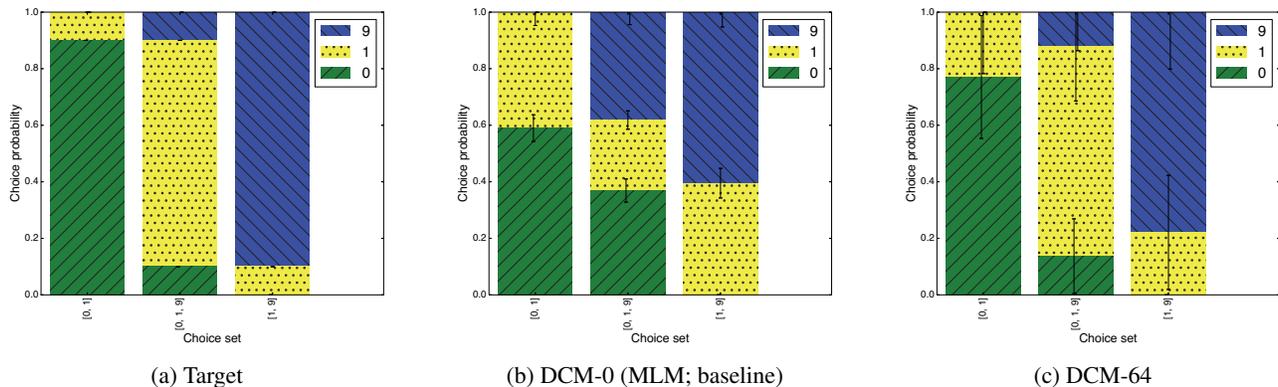


Figure 2: The digit choice task. (a) The target distribution, of selecting each digit given a choice set, that is used to generate training data. (b) The distribution learned by the MLM. (c) The distribution learned by the DCM with 64 hidden nodes.

of  $\mathcal{Y}$  for 900 choice sets, and an image of 0 is selected for the remaining 100 choice sets.

When we test the quality of trained choice models, we use another set of 300 images (100 images for each of the three digits) that have not been used for training. The choice set,  $\mathcal{X}$ , for testing is created analogously to that for training. The goal of the digit choice task is to train a choice model, parameterized with  $\theta$ , such that the conditional probability distribution,  $p(\{y\}|\mathcal{X};\theta)$  for  $y \in \mathcal{X}$ , well approximates the corresponding conditional probability distribution that the agent follows, for each  $\mathcal{X}$  of the choice sets for testing.

### Deep learning

To handle images with the DCM, we first create distributed binary representation for each image of handwritten digits in an unsupervised fashion using deep learning. The aim of deep learning is to extract binary features reflecting latent structures embedded in images without using class labels. With the MNIST dataset, the latent structure is expected to reflect the digit associated with the image. We use 50,000 images from the MNIST dataset for deep learning. These 50,000 images do not overlap with any of the 300 images used to train choice models or with the 300 images used to test the choice models.

We employ the stacked denoising autoencoder (SDA; Vincent et al. (2010)) for deep learning. We prepare 500 binary nodes with the hyperbolic tangent nonlinearity for the first hidden layer, which is fully connected to the nodes in the input layer with 784 binary nodes. We prepare another 500 binary nodes with the hyperbolic tangent nonlinearity for the second hidden layer, which is fully connected to the first hidden layer. We use the activation of the second hidden layer as the latent features of the corresponding input images. More specifically, the values of the second hidden layer are scaled to  $[0, 1]^{500}$  and then rounded to  $\{0, 1\}^{500}$ .

The SDA is trained by the use of Pylearn2 (Goodfellow et al. 2013). Before training the SDA, all of the weights and biases associated with the two hidden layers of the SDA are initialized by random samples from the uniform distribution over  $(-0.05, 0.05)$ . We first train the weights and biases for the first hidden layer of the SDA for 1,000 epochs. In

each epoch, the SDA is trained with the 50,000 images with stochastic gradient descent using the minibatches of 100 images. Here, the learning rate is set to a small constant value of 0.001. The corruption level for the input layer, which is the probability of randomly setting the value of nodes to 0, is set to 0.2. After training the first layer of hidden units, we fix the parameters associated with the first hidden layer and train the parameters associated with the second hidden layer with exactly the same settings used for the first hidden layer, except for the corruption level, which is now set to 0.3.

### Results

We train the DCM using the training data, which consist of the 3,000 pairs of the choice set of images and the selected image ( $\mathbf{X}$  and  $\mathbf{Y}$ ). We use the binarized latent features extracted from an image via deep learning as the feature vector of that image. Hence, the feature vector has  $K = 500$  dimensions. The number of hidden nodes,  $L$ , is varied among 0, 4, 16, 64, 256, and 1024. Weights  $\mathbf{W}$  and  $\mathbf{U}$  are initialized by samples from the uniform distribution over  $[-a, a]$ , where  $a \equiv m/\sqrt{\max\{K, L\}}$ ; we use  $m = 10$  for  $\mathbf{W}$  and  $m = 0.1$  for  $\mathbf{U}$ . Biases  $\mathbf{b}_h$  and  $\mathbf{b}_y$  are initialized to 0. The DCM is trained for 20 epochs, where all of the 3,000 pairs of  $(\mathcal{X}, \mathcal{Y})$  are used as training data in each epoch, and the parameters of the DCM are updated with stochastic gradient descent (Eq. 13) using the minibatches of 10 images. The learning rate is set to 0.001.

Fig. 2b shows the choice probabilities that are predicted by the trained DCM with no hidden nodes, which corresponds to the MLM, as is shown in Eq. (11). Here, the choice probabilities are evaluated for the 1,000 choice sets of images for each of the three combinations of digits:  $\{0, 1\}$ ,  $\{0, 1, 9\}$ , and  $\{1, 9\}$ . For each of the three combinations of digits, the predicted choice probabilities are averaged over the 1,000 choice sets and shown as bars in the figure. Error bars show corresponding standard deviations.

As is evident in Fig. 2b, the target distribution in Fig. 2a cannot be represented by the MLM. In the target distribution, an image of 9 acts as the decoy and causes the attraction effect in which the relative probability of choosing an

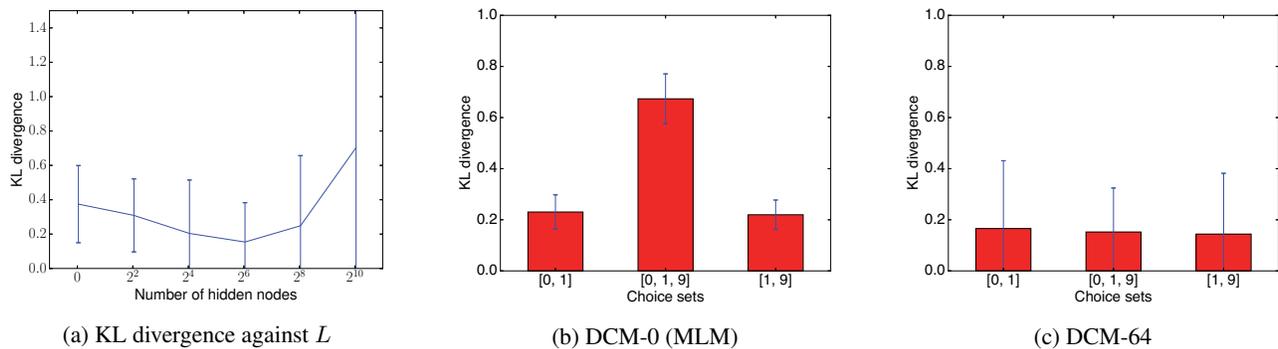


Figure 3: Kullback-Leibler (KL) divergence between the target distribution and the distribution predicted by the trained DCM. (a) The average KL divergence is shown as a function of the number of hidden nodes in the DCM. (b-c) The average KL divergence is shown for each choice set. The error bar shows the corresponding standard deviation.

image of 1 over an image of 0 increases. The trained MLM shows that the ratio between the probabilities of choosing an image of 0 or 1 is essentially independent of the existence of an image of 9 in the choice set.

Fig. 2c shows the choice probabilities that are predicted by the DCM with 64 hidden nodes (DCM-64), which has been trained analogously to the MLM. Contrary to the MLM, the choice probabilities predicted by the DCM-64 exhibits the attraction effect that appears in the target distribution. Although prior work shows that the RBM choice model can also be trained to represent the attraction effect (Osogami and Otsuka 2014), the RBM choice model cannot give any useful prediction for the digit choice task. Recall that the images used for training are never used in testing the choice model, so that the prediction given by the RBM choice model will be essentially random.

We use the results in Fig. 3 to quantitatively evaluate the predicted distributions. As a measure of the quality of prediction, we use the Kullback-Leibler divergence,

$$\text{KL}_{\mathcal{X}}(q||p) \equiv \sum_{\mathcal{Y}} q(\mathcal{Y}|\mathcal{X}) \log \frac{q(\mathcal{Y}|\mathcal{X})}{p(\mathcal{Y}|\mathcal{X})}, \quad (15)$$

between the target distribution,  $q$ , and the predicted distribution,  $p$ . In Fig. 3a, the KL divergence is plotted as a function of  $L$ , the number of hidden nodes used in the DCM. Here,  $\text{KL}_{\mathcal{X}}(q||p)$  is averaged over the 3,000 choice sets prepared for testing. In this particular setting, the average KL divergence is 0.374 for the DCM-0 (or equivalently MLM), which is reduced by increasing  $L$  and reaches the minimum value of 0.153 with the DCM-64. The predictive error is thus reduced by 58.8 %. Adding further hidden nodes deteriorates the predictive performance in this case, but the optimal number of hidden nodes depends on the particular settings such as the amount of training data. In Fig. 3b-3c, we show the KL divergence for each of the three combinations of digits. Here, the KL divergence is averaged over the 1,000 choice sets for each case. The DCM-64 has a smaller average KL divergence than the MLM for each of the three cases, and the improvement is most significant for the cases when the choice set consists of three images of 0, 1, and 9.

We ran the experiments on a Windows workstation having

16 cores of Intel Xeon CPU E5-2670 2.6 GHz and 64 GB memory. Training the SDA took 406 minutes, including the time to store intermediate results and other overhead that constitute about 25 % of the total execution time. Training the deep choice model took 268 seconds, and training the multinomial logit model took 190 seconds. Notice that we did not optimize the hyperparameters such as the learning rate for acceleration, which is not a major focus of this paper.

## Conclusion

The study of choice has a long history (Train 2009; Greene 2011), but this paper addresses the problem of learning choices from natural stimuli, or images in particular, for the first time in the literature. The combination of the MLM and deep learning, which is used as the baseline in our experiments, is in fact new. The key idea of the DCM when combined with deep learning is in creating the average vector of the features of items in the choice set and using that average vector as input to the choice model. This idea can potentially be applied to other choice models to be combined with deep learning, opening up a new direction of research on the use of deep learning for modeling and learning human choice.

While our digit choice task focuses on the attraction effect, we show in the supplementary material (Otsuka and Osogami 2016) that the DCM can represent three typical phenomena of human choice (the similarity effect, the compromise effect, and the attraction effect) in the settings of (Shenoy and Yu 2013). These three typical phenomena are considered to be robust and significant in the human choice (Rieskamp, Bussemeyer, and Mellers 2006; Otter et al. 2008).

## Acknowledgements

This research was supported by JST, CREST.

## References

Ariely, D. 2010. *Predictably Irrational: The Hidden Forces That Shape Our Decisions*. Harper Perennial, revised and expanded edition.

- Ben-Akiva, M. 1973. *The structure of passenger travel demand models*. Ph.D. Dissertation.
- Busemeyer, J. R., and Townsend, J. T. 1993. Decision field theory: A dynamic cognition approach to decision making. *Psychological Review* 100:432–459.
- Chapelle, O., and Harchaoui, Z. 2005. A machine learning approach to conjoint analysis. In Saul, L. K.; Weiss, Y.; and Bottou, L., eds., *Advances in Neural Information Processing Systems 17*. 257–264.
- Faradani, S.; Hartmann, B.; and Ipeirotis, P. G. 2011. What's the right price? Pricing tasks for finishing on time. In *Human Computation, volume WS-11-11 of AAAI Workshops*, 26–31. AAAI.
- Farias, V. F.; Jagabathula, S.; and Shah, D. 2013. A non-parametric approach to modeling choice with limited data. *Management Science* 59(2):305–322.
- Goodfellow, I. J.; Warde-Farley, D.; Lamblin, P.; Dumoulin, V.; Mirza, M.; Pascanu, R.; Bergstra, J.; Bastien, F.; and Bengio, Y. 2013. Pylearn2: A machine learning research library. *arXiv preprint arXiv:1308.4214*.
- Greene, W. H. 2011. *Econometric Analysis*. Pearson, 7th edition.
- Hausman, J. A., and Wise, D. A. 1978. A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica* 46(2):403–426.
- Hinton, G. E., and Salakhutdinov, R. 2006. Reducing the dimensionality of data with neural networks. *Science* 313(5786):504–507.
- Hinton, G. E. 2002. Training products of experts by minimizing contrastive divergence. *Neural computation* 14(8):1771–1800.
- Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*. 1106–1114.
- Lee, H.; Pham, P. T.; Largman, Y.; and Ng, A. Y. 2009. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in Neural Information Processing Systems 22*. 1096–1104.
- Luce, R. D. 1959. *Individual choice behavior: A theoretical analysis*. New York, NY: John Wiley and Sons.
- McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In Zarembka, P., ed., *Frontiers in Econometrics*. Academic Press. 104–142.
- McFadden, D. 1981. Econometric models of probabilistic choice. In Manski, C. F., and McFadden, D., eds., *Structural Analysis of Discrete Data with Economic Applications*. MIT Press. 198272.
- Osogami, T., and Katsuki, T. 2014. A Bayesian hierarchical choice model with visibility. In *Proc. 22nd International Conference on Pattern Recognition*, 3618–3623.
- Osogami, T., and Otsuka, M. 2014. Restricted Boltzmann machines modeling human choice. In *Advances in Neural Information Processing Systems 27*. 73–81.
- Otsuka, M., and Osogami, T. 2016. Supplementary material for a deep choice model. Technical Report RT0970, IBM Research.
- Otter, T.; Johnson, J.; Rieskamp, J.; Allenby, G. M.; Brazell, J. D.; Diederich, A.; Hutchinson, J. W.; MacEachern, S.; Ruan, S.; and Townsend, J. 2008. Sequential sampling models of choice: Some recent advances. *Marketing Letters* 19(3-4):255–267.
- Pfeiffer, T.; Gao, X. A.; Chen, Y.; Mao, A.; and Rand, D. G. 2012. Adaptive polling for information aggregation. In *Proceedings of AAAI-12*, 122–128.
- Rieskamp, J.; Busemeyer, J. R.; and Mellers, B. A. 2006. Extending the bounds of rationality: Evidence and theories of preferential choice. *Journal of Economic Literature* 44:631–661.
- Salakhutdinov, R., and Hinton, G. E. 2009a. Deep Boltzmann machines. In *Proc. 12th International Conference on Artificial Intelligence and Statistics*, 448–455.
- Salakhutdinov, R., and Hinton, G. E. 2009b. Replicated softmax: An undirected topic model. In *Advances in Neural Information Processing Systems 22*. 1607–1614.
- Shenoy, P., and Yu, A. J. 2013. Rational preference shifts in multi-attribute choice: What is fair? In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci 2013)*, 1300–1305.
- Srivastava, N.; Salakhutdinov, R.; and Hinton, G. E. 2013. Modeling documents with deep Boltzmann machines. In *Proc. 29th Conference on Uncertainty in Artificial Intelligence*.
- Takahashi, R., and Morimura, T. 2015. Predicting preference reversals via Gaussian process uncertainty aversion. In *Proc. 18th International Conference on Artificial Intelligence and Statistics*.
- Train, K. 2009. *Discrete Choice Methods with Simulation*. Cambridge University Press, second edition.
- Usher, M., and McClelland, J. L. 2004. Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological Review* 111(3):757–769.
- Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; and Manzagol, P.-A. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research* 11:3371–3408.
- Zhen, Y.; Rai, P.; Zha, H.; and Carin, L. 2015. Cross-modal similarity learning via pairs, preferences, and active supervision. In *Proceedings of AAAI-15*, 3203–3209.