Softmax and McFadden's Discrete Choice under Interval (and Other) Uncertainty*

Bartlomiej Jacek Kubica[1], Laxman Bokati[2], Olga Kosheleva[2][0000-0003-2587-4209], and Vladik Kreinovich[2][0000-0002-1244-1650]

[1] University of Life Sciences, Warsaw, Poland
bartlomiej.jacek.kubica@gmail.com
[2] University of Texas at El Paso, El Paso TX 79968, USA
lbokati@miners.utep.edu, olgak@utep.edu, vladik@utep.edu

Abstract. One of the important steps in deep learning is softmax, when we select one of the alternatives with a probability depending on its expected gain. A similar formula describes human decision making: somewhat surprisingly, when presented with several choices with different expected equivalent monetary gain, we do not just select the alternative with the largest gain; instead, we make a random choice, with probability decreasing with the gain – so that it is possible that we will select second highest and even third highest value. Both formulas assume that we know the exact value of the expected gain for each alternative. In practice, we usually know this gain only with some uncertainty. For example, often, we only know the lower bound $\underline{f}$ and the upper bound $\overline{f}$ on the expected gain, i.e., we only know that the actual gain f is somewhere in the interval $[\underline{f}, \overline{f}]$. In this paper, we show how to extend softmax and discrete choice formulas to such cases of interval uncertainty.

Keywords: Deep learning · Softmax · Discrete choice · Interval uncertainty.

1 Formulation of the Problem

Deep learning: a brief reminder. At present, the most efficient machine learning technique is *deep learning* (see, e.g., [2, 7]), in particular, *reinforcement deep learning* [12], where, in addition to processing available information, the system also (if needed) automatically decides which additional information to request – and if an experimental setup is automated, to produce.

For selecting the appropriate piece of information, the system estimates, for each possible alternative, how much information this particular alternative will bring.

* This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science) and HRD-1242122 (Cyber-ShARE Center of Excellence).

It is important to add randomness. And here comes an interesting part. A reader who is not familiar with details of deep learning algorithms may expect that the system selects the alternative with the largest estimate of expected information gain. This idea was indeed tried – but it did not work well: instead of looking for the best possible model, i.e., the model for which the averaged difference between the predictions and observations is the smallest, the corresponding system would often get stuck in a local minimum of the corresponding objective function.

In numerical analysis, a usual way to get out of a local minimum is to perform some random fluctuation. This is, e.g., the main idea behind simulated annealing. Crudely speaking, it means that we do not always follow the smallest – or the largest – value of the corresponding objective function, we can follow the next largest (smallest), next next largest, etc. – with some probability.

Softmax: how randomness is currently added. Of course, the actual maximum should be selected with the highest probability, the next value with lower probability, etc. In other words, if we want to maximize some objective function $f(a)$, and we have alternatives $a_1, \ldots, a_n$ for which this function has values $f_1 \stackrel{\text{def}}{=} f(a_1), \ldots, f_n \stackrel{\text{def}}{=} f(a_n)$, then the probability p_i of selecting the i-th alternative should be increasing with f_i, i.e., we should have $p_i \sim F(f_i)$ for some increasing function $F(z)$, i.e., $p_i = c \cdot F(f_i)$, for some constant c.

We should always select one of the alternatives, so these probabilities should add up to 1: $\sum\limits_{j=1}^n p_j = 1$. From this condition, we conclude that $c \cdot \sum\limits_{j=1}^n F(f_j) = 1$. Thus,

$$c = \frac{1}{\sum\limits_{j=1}^n F(f_j)} \tag{1}$$

and so,

$$p_i = \frac{F(f_i)}{\sum\limits_{j=1}^n F(f_j)}. \tag{2}$$

Which function $F(z)$ should we choose? In deep learning – a technique that requires so many computations that it cannot exist without high performance computing – computation speed is a must. Thus, the function $F(z)$ should be fast to compute – which means, in practice, that it should be one of the basic functions for which we have already gained an experience of how to compute it fast. There are a few such functions: arithmetic functions, the power function, trigonometric functions, logarithm, exponential function, etc.

The selected function should be increasing, and it should always return non-negative results – otherwise, we will end up with meaningless negative probability. Among basic functions, only one function has this property – the exponential function $F(z) = \exp(k \cdot z)$ for some $k > 0$. For this function, the probability p_i

takes the form

$$p_i = \frac{\exp(k \cdot f_i)}{\sum\limits_{j=1}^{n} \exp(k \cdot f_j)}. \tag{3}$$

This expression is known as *softmax*.

It is desirable to further improve deep learning. Deep learning has lead to many interesting results, but it is not a panacea. There are still many challenging problems where the existing deep learning algorithms has not yet led to fully successful learning. It is therefore desirable to look at all the stages of deep leaning and see if we can modify them so as to improve the overall learning performance.

Need to generalize softmax to the case of interval uncertainty. One of the stages in which there is a potential for improvement is softmax. Indeed, when we apply the softmax formula, we only take into account the corresponding estimates $f_1, \ldots, f_n$. However, in practice, we do not just have these estimates, we often have some idea of how accurate is each estimate. Some estimates may be more accurate, some may be less accurate. It is desirable to take this information about accuracy into account.

For example, as a measure of such accuracy, we may know the upper bound Δ_i on the absolute value $|f_i - f_i^{\rm act}|$ of the difference between the estimate f_i and the (unknown) actual value $f_i^{\rm act}$ of the objective function. In this case, the only information that we have about the actual value $f_i^{\rm act}$ is that this value is located somewhere in the interval $[f_i - \Delta_i, f_i + \Delta_i]$.

How to take this interval information into account when computing the corresponding probabilities p_i? This is the problem that we study in this paper – and for which we provide a reasonable solution.

Another important case where a softmax-type formula is used. There is another application area where a similar formula is used: the analysis of human choice.

If a person needs to select between several alternatives $a_1, \ldots, a_n$, and this person knows the exact monetary value $f_1, \ldots, f_n$ associated with each alternative, then we expect this person to always select the alternative with the largest possible monetary value – actual or equivalent. We also expect that if we present the person with the exact same choice several times in a row, this person will always make the same decision – of selecting the best alternative.

Interestingly, this is *not* how most people make decisions. It turns out that we make decisions probabilistically: instead of always selecting the best alternative, we select each alternative a_i with probability p_i described exactly by the softmax-like formula

$$p_i = \frac{\exp(k \cdot f_i)}{\sum\limits_{j=1}^{n} \exp(k \cdot f_j)}. \tag{4}$$

for some $k > 0$.

In other words, in most cases, we usually indeed select the alternative with the higher monetary value, but with some probability, we will also select the next highest, with some smaller probability, the next next highest, etc.

This fact was discovered by an economist D. McFadden – who received a Nobel Prize in Economics for this discovery; see, e.g., [10, 11, 13].

But why? At first glance, such a probabilistic behavior sounds irrational – why not select the alternative with the largest possible monetary value?

A probabilistic choice would indeed be irrational if this was a stand-alone choice. In reality, however, no choice is stand-alone, it is a part of a sequence of choices, some of which involve conflict – and it is known that in conflict situations, a probabilistic choice makes sense. Let us give a simple example explaining why it is so.

Suppose that a person own two stores, one bigger, one smaller, but has just enough money to hire one all-night-long security person. A seemingly rational idea is to use this person to protect the bigger store – since in case of theft, the corresponding losses will be larger. However, if the owner does this, the thieves will know that the smallest store is not protected at all and rob it that very night. A much better strategy is, every night, to assign the security person to one the stores with some probability. We need to send this person to the larger store with a higher probability – i.e., in a larger portion of nights – but we still need to maintain some probability of providing security for the smaller store instead. Then, the thieves will weigh the risk of being caught (and sent to jail) vs. possible gain and thus, hopefully, refrain from an attack.

In practice, we usually only know gain with some uncertainty. McFadden's formula describes people's behavior in an idealized situation when the decision maker knows the exact momentary consequences f_i of each alternative a_i. In practice, this is rarely the case. At best, we know a lower bound $\underline{f}_i$ and an upper bound $\overline{f}_i$ of the actual (unknown) value f_i. In such situations, all we know is that the unknown value f_i is somewhere within the interval $[\underline{f}_i, \overline{f}_i]$.

How can we extend McFadden formula to this case of interval uncertainty? This is also what we will do in this paper.

2 Formulating the Problem in Precise Terms

Discussion. Let $\mathcal{A}$ denote the class of all possible alternatives. We would like, given any finite set of alternative $A \subseteq \mathcal{A}$ and a specific alternative $a \in A$, to describe the probability $p(a \mid A)$ that out of all the alternatives from the set A, the alternative a will be selected.

Once we know these probabilities, we can then compute, for each set $B \subseteq A$, the probability $p(B \mid A)$ that one of the alternatives from the set B will be selected as $p(B \mid A) = \sum\limits_{b \in B} p(b \mid A)$. In particular, we have $p(a \mid A) = p(\{a\} \mid A)$.

A natural requirement related to these conditional probabilities is that if we have $A \subseteq B \subseteq C$, then we can view the selection of A from C as either a direct selection, or as first selecting B, and then selecting A out of B. The resulting probability should be the same, so we must have $p(A \mid C) = p(A \mid B) \cdot p(B \mid C)$. Thus, we arrive at the following definition.

Definition 1. *Let $\mathcal{A}$ be a set. Its elements will be called* alternatives. *By a* choice function, *we mean a function $p(a \mid A)$ that assigns to each pair $\langle A, a\rangle$ of a finite set $A \subseteq \mathcal{A}$ and an element $a \in A$ a number from the interval $(0,1]$ in such a way that that the following two conditions are satisfied:*

- *for every set A, we have $\sum\limits_{a\in A} p(a \mid A) = 1$, and*
- *whenever $A \subseteq B \subseteq C$, we have $p(A \mid C) = p(A \mid B) \cdot p(B \mid C)$, where $p(B \mid A) \stackrel{\text{def}}{=} \sum\limits_{b\in B} p(b \mid A)$.*

Proposition 1. *For each set $\mathcal{A}$, the following two conditions are equivalent to each other:*

- *the function $p(a \mid A)$ is a choice function, and*
- *there exists a function $v : \mathcal{A} \to \mathbb{R}^+$ that assigns a positive number to each alternative $a \in \mathcal{A}$ such that*

$$p(a \mid A) = \frac{v(a)}{\sum\limits_{b\in A} v(b)}. \tag{5}$$

Proof. It is easy to check that for every function v, the expression (5) indeed defines a choice function. So, to complete the proof, it is sufficient to prove that every choice function has the form (5).

Indeed, let $p(a \mid A)$ be a choice function. Let us pick any $a_0 \in \mathcal{A}$, and let us define a function v as

$$v(a) \stackrel{\text{def}}{=} \frac{p(a \mid \{a, a_0\})}{p(a_0 \mid \{a, a_0\})}. \tag{6}$$

In particular, for $a = a_0$, both probabilities $p(a \mid \{a, a_0\})$ and $p(a_0 \mid \{a, a_0\})$ are equal to 1, so the ratio $v(a_0)$ is also equal to 1. Let us show that the choice function has the form (5) for this function v.

By definition of $v(a)$, for each a, we have $p(a \mid \{a, a_0\}) = v(a) \cdot p(a_0 \mid \{a, a_0\})$.

By definition of a choice function, for each set A containing a_0, we have $p(a \mid A) = p(a \mid \{a, a_0\}) \cdot p(\{a, a_0\} \mid A)$ and $p(a_0 \mid A) = p(a_0 \mid \{a, a_0\}) \cdot p(\{a, a_0\} \mid A)$. Dividing the first equality by the second one, we get

$$\frac{p(a \mid A)}{p(a_0 \mid A)} = \frac{p(a \mid \{a, a_0\})}{p(a_0 \mid \{a, a_0\})}. \tag{6}$$

By definition of $v(a)$, this means that

$$\frac{p(a \mid A)}{p(a_0 \mid A)} = v(a). \tag{7}$$

Similarly, for each $b \in A$, we have

$$\frac{p(b \mid A)}{p(a_0 \mid A)} = v(b). \tag{8}$$

Dividing (7) by (8), we conclude that for each set A containing a_0, we have

$$\frac{p(a \mid A)}{p(b \mid A)} = \frac{v(a)}{v(b)}. \tag{9}$$

Let us now consider a set B that contains a and b but that does not necessarily contain a_0. Then, by definition of a choice function, we have

$$p(a \mid B) = p(a \mid \{a, b\}) \cdot p(\{a, b\} \mid B) \tag{10}$$

and

$$p(b \mid B) = p(b \mid \{a, b\}) \cdot p(\{a, b\} \mid B). \tag{11}$$

Dividing (10) by (11), we conclude that

$$\frac{p(a \mid B)}{p(b \mid B)} = \frac{p(a \mid \{a, b\})}{p(b \mid \{a, b\})}. \tag{12}$$

The right-hand side of this equality does not depend on the set B. So the left-hand side, i.e., the ratio

$$\frac{p(a \mid B)}{p(b \mid B)} \tag{13}$$

also does not depend on the set B. In particular, for the sets B that contain a_0, this ratio – according to the formula (9) – is equal to $v(a)/v(b)$. Thus, the same equality (9) holds for all sets A – not necessarily containing the element a_0.

From the formula (9), we conclude that

$$\frac{p(a \mid A)}{v(a)} = \frac{p(b \mid A)}{v(b)}. \tag{14}$$

In other words, for all elements $a \in A$, the ratio

$$\frac{p(a \mid A)}{v(a)} \tag{15}$$

is the same. Let us denote this ratio by c_A; then, for each $a \in A$, we have

$$p(a \mid A) = c_A \cdot v(a). \tag{16}$$

From the condition that $\sum\limits_{b \in A} p(b \mid A) = 1$, we can now conclude that $c_A \cdot \sum\limits_{b \in A} v(b) = 1$, thus

$$c_A = \frac{1}{\sum\limits_{b \in A} v(b)}. \tag{17}$$

Substituting this expression (17) into the formula (16), we get the desired expression (5).

The proposition is proven.

Comment. This proof is similar to the proofs from [4, 8].

Discussion. As we have mentioned earlier, a choice is rarely a stand-alone event. Usually, we make several choices – and often, at the same time. Let us consider a simple situation. Suppose that we need to make two independent choices:

- in the first choice, we must select one the alternatives $a_1, \ldots, a_n$, and
- in the second choice, we must select one of the alternatives $b_1, \ldots, b_m$.

We can view this as two separate selection processes. In this case, in the first process, we select each alternative a_i with probability

$$\frac{v(a_i)}{\sum\limits_{k=1}^{n} v(a_k)}, \tag{18}$$

and, in the second process, we select each alternative b_j with probability

$$\frac{v(b_j)}{\sum\limits_{\ell=1}^{m} v(b_\ell)}. \tag{19}$$

Since the two processes are independent, for each pair $\langle a_i, b_j \rangle$, the probability of selecting this pair is equal to the product of the corresponding probabilities:

$$\frac{v(a_i)}{\sum\limits_{k=1}^{n} v(a_k)} \cdot \frac{v(b_j)}{\sum\limits_{\ell=1}^{m} v(b_\ell)}. \tag{20}$$

Alternatively, we can view the whole two-stage selection as a single selection process, in which we select a pair $\langle a_i, b_j \rangle$ of alternatives out of all $n \cdot m$ possible pairs. In this case, the probability of selecting a pair is equal to

$$\frac{v(\langle a_i, b_j \rangle)}{\sum\limits_{k=1}^{n} \sum\limits_{\ell=1}^{m} v(\langle a_k, b_\ell \rangle)}. \tag{21}$$

The probability of selecting a pair should be the same in both cases, so the values (20) and (21) must be equal to each other. This equality limits possible functions $v(a)$.

Indeed, if all we know about each alternative a is the interval $[\underline{f}(a), \overline{f}(a)]$ of possible values of the equivalent monetary gain, then the value v should depend only on this information, i.e., we should have $v(a) = V(\underline{f}(a), \overline{f}(a))$ for some function $V(x, y)$. Which functions $V(x, y)$ guarantee the above equality?

To answer this question, let us analyze how the gain corresponding to selecting a pair $\langle a_i, b_j \rangle$ is related to the gains corresponding to individual selections of a_i and b_j. Suppose that for the alternative a_i, our gain can take any value from the interval $[\underline{f}(a_i), \overline{f}(a_i)]$, and for the alternative b_j, our gain can take any value from the interval $[\underline{f}(b_j), \overline{f}(b_j)]$. These selections are assumed to be independent. This means that we can have all possible combinations of values $f(a_i) \in [\underline{f}(a_i), \overline{f}(a_i)]$ and $f(b_j) \in [\underline{f}(b_j), \overline{f}(b_j)]$.

The smallest possible value of the overall gain $f(a_i) + f(b_j)$ is when both gains are the smallest. In this case, the overall gain is $\underline{f}(a_i) + \underline{f}(b_j)$. The largest possible value of the overall gain $f(a_i) + f(b_j)$ is when both gains are the largest. In this case, the overall gain is $\overline{f}(a_i) + \overline{f}(b_j)$. Thus, the interval of possible values of the overall gain is

$$[\underline{f}(\langle a_i, b_j, \rangle), \overline{f}(\langle a_i, b_j, \rangle)] = \underline{f}(a_i) + \underline{f}(b_j), \overline{f}(a_i) + \overline{f}(b_j)]. \tag{22}$$

In these terms, the requirement that the expressions (20) and (21) coincide takes the following form, where we denoted $f_i \stackrel{\text{def}}{=} f(a_i)$, $g_j \stackrel{\text{def}}{=} g(b_j)$:

Definition 2. *We say that a function $V : \mathbb{IR} \times \mathbb{IR} \to \mathbb{IR}^+$ is* consistent *if for every two sequences of intervals $[\underline{f}_1, \overline{f}_1], \ldots, [\underline{f}_n, \overline{f}_n]$, and $[\underline{g}_1, \overline{g}_1], \ldots, [\underline{g}_m, \overline{g}_m]$, for every i and j, we have*

$$\frac{V(\underline{f}_i, \overline{f}_i)}{\sum\limits_{k=1}^{n} V(\underline{f}_k, \overline{f}_k)} \cdot \frac{V(\underline{g}_j, \overline{g}_j)}{\sum\limits_{\ell=1}^{m} V(\underline{g}_\ell, \overline{g}_\ell)} = \frac{V(\underline{f}_i + \underline{g}_j, \overline{f}_i + \overline{g}_j)}{\sum\limits_{k=1}^{n} \sum\limits_{\ell=1}^{m} V(\underline{f}_k + \underline{g}_\ell, \overline{f}_k + \overline{g}_\ell)}. \tag{23}$$

Monotonicity. Another reasonable requirement is that the larger the expected gain, the more probable that the corresponding alternative is selected.

The notion of larger is easy when gains are exact, but for intervals, we can provide the following definition.

Definition 3. *We say that an interval A is* smaller than or equal to *an interval B (and denote it by $A \le B$) if the following two conditions hold:*

- *for every element $a \in A$, there is an element $b \in B$ for which $a \le b$, and*
- *for every element $b \in B$, there is an element $a \in A$ for which $a \le b$.*

Proposition 2.

$$[\underline{a}, \overline{a}] \le [\underline{b}, \overline{b}] \Leftrightarrow (\underline{a} \le \underline{b} \,\&\, \overline{a} \le \overline{b}). \tag{24}$$

Proof is straightforward.

Definition 4. *We say that a function $V : \mathbb{IR} \times \mathbb{IR} \to \mathbb{IR}^+$ is* monotonic *if for every two intervals $[\underline{a}, \overline{a}]$ and $[\underline{b}, \overline{b}]$, if $[\underline{a}, \overline{a}] \le [\underline{b}, \overline{b}]$ then $V(\underline{a}, \overline{a}) \le V(\underline{b}, \overline{b})$.*

Proposition 3. *For each function $V : \mathbb{IR} \times \mathbb{IR} \to \mathbb{IR}^+$, the following two conditions are equivalent to each other:*

- *the function V is consistent and monotonic;*

– *the function $V(\underline{f}, \overline{f})$ has the form*

$$V(\underline{f}, \overline{f}) = C \cdot \exp(k \cdot (\alpha_H \cdot \overline{f} + (1 - \alpha_H) \cdot \underline{f})), \tag{25}$$

for some values $C > 0$, $k > 0$, and $\alpha_H \in [0, 1]$.

Conclusion. Thus, if we have n alternatives $a_1, \ldots, a_n$, and for each alternative a_i, we know the interval $[\underline{f}_i, \overline{f}_i]$ of possible values of the gain, we should select each alternative i with the probability

$$p_i = \frac{\exp(k \cdot (\alpha_H \cdot \overline{f}_i + (1 - \alpha_H) \cdot \underline{f}_i))}{\sum\limits_{j=1}^{n} \exp(k \cdot (\alpha_H \cdot \overline{f}_j + (1 - \alpha_H) \cdot \underline{f}_j))}. \tag{26}$$

Thus, *we have found the* desired *extension of softmax and McFadden's discrete choice to the case of interval uncertainty.*

Comment 1. The formula (26) is the original McFadden's formula in which, instead of the exact gain f_i, we use the expression $\alpha_H \cdot \overline{f}_i + (1 - \alpha_H) \cdot \underline{f}_i$ for some $\alpha_H \in [0, 1]$. The formula (3) was first proposed by a future Nobelist Leo Hurwicz and is thus known as Hurwicz optimism-pessimism criterion [3, 5, 6, 9].

Comment 2. For the case when we know the exact values of the gain, i.e., when we have a degenerate interval $[f, f]$, we get a *new justification for the original McFadden's formula.*

Comment 3. Similar ideas can be used to extend softmax and McFadden's formula to other types of uncertainty. As one can see from the proof, by taking logarithm of V, we reduce the problem to additivity, and additive functions are known; see, e.g., [6]. For example, for probabilities, the equivalent gain is the expected value – since, due to large numbers theorem, the sum of many independent copies of a random variable is practically a deterministic number. Similarly, a class of probability distributions is equivalent to the interval of mean values corresponding to different distributions, and specific formulas are known for the fuzzy case.

Proof of Proposition 2. It is easy to check that the function (26) is consistent and monotonic. So, to complete the proof, it is sufficient to prove that every consistent monotonic function has the form (25).

Indeed, let us assume that the function V is consistent and monotonic. Then, due to consistency, it satisfies the formula (23). Taking logarithm of both sides of the formula (23), we conclude that for the auxiliary function $u(\underline{a}, \overline{a}) \stackrel{\text{def}}{=} \ln(V(\underline{a}, \overline{a}))$, for every two intervals $[\underline{a}, \overline{a}]$ and $[\underline{b}, \overline{b}]$, we have

$$u(\underline{a}, \overline{a}) + u(\underline{b}, \overline{b}) = u(\underline{a} + \underline{b}, \overline{a} + \overline{b}) + c \tag{27}$$

for an appropriate constant c. Thus, for $U(\underline{a}, \overline{a}) \stackrel{\text{def}}{=} u(\underline{a}, \overline{a}) - c$, substituting $u(\underline{a}, \overline{a}) = U(\underline{a}, \overline{a}) + c$ into the formula (27), we conclude that

$$U(\underline{a}, \overline{a}) + U(\underline{b}, \overline{b}) = U(\underline{a} + \underline{b}, \overline{a} + \overline{b}), \tag{28}$$

i.e., that the function U is additive. Similarly to [6], we can use the general classification of additive locally bounded functions (and every monotonic function is locally bounded) from [1] to conclude that $U(\underline{a},\overline{a}) = k_1 \cdot \overline{a} + k_2 \cdot \underline{a}$. Monotonicity with respect to changes in $\underline{a}$ and $\overline{a}$ imply that $k_1 \geq 0$ and $k_2 \geq 0$. Thus, for

$$V(\underline{a},\overline{a}) = \exp(u(\underline{a},\overline{a})) = \exp(U(\underline{a},\overline{a}) + c) = \exp(c) \cdot \exp(U(\underline{a},\overline{a})),$$

we get the desired formula, with $C = \exp(c)$, $k = k_1 + k_2$ and $\alpha_H = k_1/(k_1 + k_2)$.

The proposition is proven.

References

1. Aczél, J., Dhombres, J.: Functional equations in several variables, Camridge University Press, Cambridge, UK (2008)
2. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning, MIT Press, Cambridge, Massachusetts (2016)
3. Hurwicz, L.: Optimality Criteria for Decision Making Under Ignorance, Cowles Commission Discussion Paper, Statistics, No. 370 (1951)
4. Kosheleva, O., Kreinovich, V., Sriboonchitta, S.: Econometric models of probabilistic choice: beyond McFadden's formulas, In: Kreinovich, V., Sriboonchitta, S., Huynh V. N. (eds.): Robustness in econometrics, pp. 79–88, Springer Verlag, Cham, Switzerland (2017)
5. Kreinovich, V.: Decision making under interval uncertainty (and beyond), In: Guo, P., Pedrycz W. (eds.): Human-centric decision-making models for social sciences, pp. 163–193, Springer Verlag (2014)
6. Kreinovich, V.: Decision making under interval (and more general) uncertainty: monetary vs. utility approaches, Journal of Computational Technologies **22**(2), 37–49 (2017)
7. Kreinovich, V.: From traditional neural networks to deep learning: towards mathematical foundations of empirical successes, In: Shahbazova, S. N. et al. (eds.): Proceedings of the World Conference on Soft Computing, Baku, Azerbaijan, May 29–31, 2018 (2018)
8. Luce, D.: Inividual choice behavior: a theoretical analysis. Dover, New York (2005)
9. Luce, R. D., Raiffa, R.: Games and decisions: introduction and critical survey, Dover, New York (1989)
10. McFadden, D.: Conditional logit analysis of qualitative choice behavior, In: Zarembka, P. (ed.), Frontiers in econometrics, pp. 105–142. Academic Press, New York (1974)
11. McFadden, D.: Economic choices, American Economic Review **91**, 351–378 (2001)
12. Sutton, R. S., Barto, A. G.: Reinforcement learning. An introduction, 2nd edition, MIT Press, Cambridge, Massachusetts (2018)
13. Train, K.: Discrete choice methods with simulation, Cambridge University Press, Cambridge, Massachusetts (2003)