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

Bartlomiej Jacek Kubica[1], Laxman Bokati[2]
Olga Kosheleva[2], and Vladik Kreinovich[2]

[1]Department of Applied Informatics
Warsaw University of Life Sciences
ul. Nowoursynowska 159
02-776 Warsaw, Poland
bartlomiej.jacek.kubica@gmail.com

[2]University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
olgak@utep.edu, vladik@utep.edu

1. Deep Learning: A Brief Reminder

- At present, the most efficient machine learning technique is *deep learning*.
- An important particular case is *reinforcement deep learning* where:
  - in addition to processing available information,
  - the system also (if needed) automatically decides which additional information to request,
  - (and if an experiment setup is automated, which information to produce).
- For selecting the appropriate piece of information, the system estimates:
  - for each possible alternative
  - how much information this particular alternative will bring.

2. It Is Important to Add Randomness

- One 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 finding the model that best fits the training data,
  - the algorithm would sometimes 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 change.
- 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 smallest (largest), next next smallest, etc. – with some probability.

3. 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.; so:
  - 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.:
$$p_i = \frac{F(f_i)}{\sum\limits_{j=1}^{n} F(f_j)}.$$
- Which function $F(z)$ should we choose?
- Deep learning requires so many computations that it cannot exist without high performance computing.

4. Softmax (cont-d)

- Thus, in deep learning, computation speed is a must.
- Thus, the function $F(z)$ should be fast to compute.
- This 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 return non-negative results for all real values z (positive or negative).
- 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, $p_i = \dfrac{\exp(k \cdot f_i)}{\sum\limits_{j=1}^{n} \exp(k \cdot f_j)}$.
- This expression is known as the *softmax* formula.

5. Need to Generalize Softmax to Interval Uncertainty

- hen 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 uncertainty into account.
- For example, we may know the upper bound Δ_i on $|f_i - f_i^{\rm act}|$, where $f_i^{\rm act}$ is the (unknown) actual value 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 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?

6. Another Important Use of a Softmax-Type Formula

- 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 values $f_1, \ldots, f_n$ 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 set of alternatives 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.

7. Human Choice (cont-d)

- 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, for some $k > 0$.
- In other words, in most cases, we usually indeed select the alternative with the higher monetary value.
- However, 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.

8. 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.

9. We Usually Only Know Gain with Some Certainty

- McFadden's formula describes people's behavior in an idealized situation when we know the exact monetary 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 $\left[\underline{f}_i, \overline{f}_i\right]$.
- It is therefore desirable to extend McFadden's formula to the case of interval uncertainty.

10. Formulating the Problem in Precise Terms

- Let $\mathcal{A}$ denote the class of all possible alternatives; we would like:
  - given any finite set of alternatives $A \subseteq \mathcal{A}$ and an 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.
- We can then compute, for each set $B \subseteq A$, the probability $p(B \mid A)$ that one of the alternatives from B will be selected: $p(B \mid A) = \sum\limits_{b \in B} p(b \mid A)$.
- 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:
  - either as a direct selection,
  - or as first selecting B, and then selecting A out of B.
- The resulting probability should be the same:

$$p(A \mid C) = p(A \mid B) \cdot p(B \mid C).$$

- Thus, we arrive at the following definition.

11. Definitions and the First Result

- Let $\mathcal{A}$ be a set. Its elements will be called *alternatives*.
- By a *choice function*, we mean a function $p(a \,|\, A)$ that assigns to each pair $\langle A, a\rangle$ of a finite set $A \subseteq \mathcal{A}$ and $a \in A$ a number $p \in (0, 1]$ so that:
  - for every set A, we have $\sum\limits_{a \in A} p(a \,|\, A) = 1$, and
  - whenever $A \subseteq B \subseteq C$, we have $p(A \,|\, C) = p(A \,|\, B) \cdot p(B \,|\, C)$, where $p(B \,|\, A) \stackrel{\text{def}}{=} \sum\limits_{b \in B} p(b \,|\, A)$.
- **Proposition 1.** *The following two conditions are equivalent to each other:*
  - *the function $p(a \,|\, A)$ 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*

$$p(a \,|\, A) = \frac{v(a)}{\sum\limits_{b \in A} v(b)}.$$

12. 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.
- 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 the first process, we select each alternative a_i with probability $\dfrac{v(a_i)}{\sum\limits_{k=1}^{n} v(a_k)}$.
- In the 2nd process, we select each b_j with probability $\dfrac{v(b_j)}{\sum\limits_{\ell=1}^{m} v(b_\ell)}$.

13. Discussion (cont-d)

- Since the two processes are independent, the probability of selecting this pair is equal to the product:

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

- Alternatively, we can view the whole two-stage selection as a single selection process, in which we select a pair $\langle a_i, b_j\rangle$ with probability

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

- The probability of selecting a pair should be the same in both cases.
- This equality limits possible functions $v(a)$.

14. Case of Interval Uncertainty

- We consider the case when 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 $v(a) = V\left(\underline{f}(a), \overline{f}(a)\right)$ 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 $\langle a_i, b_j\rangle$.
- For the alternative a_i, our gain $f_i \stackrel{\text{def}}{=} f(a_i)$ can take any value from the interval $\left[\underline{f}_i, \overline{f}_i\right] \stackrel{\text{def}}{=} [\underline{f}(a_i), \overline{f}(a_i)]$.
- For the alternative b_j, our gain $g_j \stackrel{\text{def}}{=} f(b_j)$ can take any value from the interval $\left[\underline{g}_j, \overline{g}_j\right] \stackrel{\text{def}}{=} [\underline{f}(b_j), \overline{f}(b_j)]$.

15. Case of Interval Uncertainty (cont-d)

- These selections are assumed to be independent.
- This means that we can have all possible combinations of values $f_i \in \left[\underline{f}_i, \overline{f}_i\right]$ and $g_j \in \left[\underline{g}_j, \overline{g}_j\right]$.
- The smallest possible value of the overall gain $f_i + g_j$ is when both gains are the smallest: $\underline{f}_i + \underline{g}_j$.
- The largest possible value of the overall gain $f_i + g_j$ is when both gains are the largest: $\overline{f}_i + \overline{g}_j$.
- Thus, the interval of possible values of the overall gain is

$$[\underline{f}(\langle a_i, b_j\rangle), \overline{f}(\langle a_i, b_j\rangle)] = \left[\underline{f}_i + \underline{g}_j, \overline{f}_i + \overline{g}_j\right].$$

- In these terms, the requirement that the two expressions coincide takes the following form.

16. Definitions

- We say that a function $V : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ is *consistent* if we always have

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

- 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.
- 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 $a \le b$;
  - for every element $b \in B$, there is an element $a \in A$ for which $a \le b$.
- One can easily check that $[\underline{a}, \overline{a}] \le [\underline{b}, \overline{b}] \Leftrightarrow (\underline{a} \le \underline{b} \,\&\, \overline{a} \le \overline{b})$.

17. Second Result

- We say that a function $V : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ is *monotonic* if for every two intervals $[\underline{a}, \overline{a}]$ and $\left[\underline{b}, \overline{b}\right]$, if $[\underline{a}, \overline{a}] \leq \left[\underline{b}, \overline{b}\right]$ then $V(\underline{a}, \overline{a}) \leq V\left(\underline{b}, \overline{b}\right)$.
- **Proposition 2.** *For each function $V : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$, the following two conditions are equivalent to each other:*

  – *the function V is consistent and monotonic;*

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

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

18. Relation to Hurwicz Criterion

- Thus, we should select each alternative i with the probability

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

- So, *we have extended the softmax/McFadden's discrete choice formula to the case of interval uncertainty.*
- It should be mentioned that the above formula coincides with what we would have obtained from the original McFadden's formula if:
  - instead of the exact gain f_i, we substitute into this original formula,
  - the expression $f_i = \alpha_H \cdot \overline{f}_i + (1-\alpha_H) \cdot \underline{f}_i$ for some $\alpha_H \in [0,1]$.
- This expression was first proposed by a future Nobelist Leo Hurwicz.
- It is thus known as Hurwicz optimism-pessimism criterion.
- For the case $\underline{f} = \overline{f}$ when we know the exact values of the gain, we get a *new justification for the original McFadden's formula.*

19. Extending to Other Types of Uncertainty

- 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 consistency condition to additivity.
- Good news is that all additive functions are known.
- For example, for probabilities, the equivalent gain is the expected value.
- Indeed, due to large numbers theorem, the sum of many independent copies of a random variable is deterministic.
- Similarly, a class of probability distributions is equivalent to the interval of their mean values.
- Specific formulas are known for the fuzzy case.

20. Conclusion

- Currently, one of the most promising Artificial Intelligence techniques is deep learning.
- The successes of using deep learning are spectacular:
  - from winning over human champions in Go (a very complex game that until recently resisted computer efforts)
  - to efficient algorithms for self-driving cars.
- All these successes require a large amount of computations on high performance computers.
- While deep learning has been very successful, there is a lot of room for improvement.
- For example, the existing deep learning algorithms implicitly assume that all the input data are exact.
- In reality, data comes from measurements and measurement are never absolutely accurate.

21. Conclusion (cont-d)

- The simplest situation is when we know the upper bound Δ on the measurement error.
- In this case, based on the measurement result $\widetilde{x}$, the only thing that we can conclude about the actual value x is that x is in the interval

$$[\widetilde{x} - \Delta, \widetilde{x} + \Delta] .$$

- One of the important steps in deep learning algorithms is computing softmax.
- We have shown how computing softmax can be naturally extended to the case of such interval uncertainty.
- The resulting formulas are almost as simple as the original ones.
- So their implementation will take about the same time on the same high performance computers.

22. Acknowledgments

This work was supported in part by the US National Science Foundation grant HRD-1242122 (Cyber-ShARE Center of Excellence).

23. Proof of Proposition 1

- It is easy to check that for every function v, the expression $p(a\,|\,A)=\dfrac{v(a)}{\sum\limits_{b\in A} v(b)}$ indeed defines a choice function.
- So, to complete the proof, it is sufficient to prove that every choice function has this form.
- Indeed, let $p(a\,|\,A)$ be a choice function.
- Let us pick any $a_0\in\mathcal{A}$, and let us define a function v as
$$v(a)\stackrel{\text{def}}{=}\frac{p(a\,|\,\{a,a_0\})}{p(a_0\,|\,\{a,a_0\})}.$$
- For $a=a_0$, both probabilities $p(a\,|\,\{a,a_0\})$ and $p(a_0\,|\,\{a,a_0\})$ are equal to 1, so the ratio $v(a_0)$ is also equal to 1.
- Let us show that the choice function has the desired form for this v.
- By definition of $v(a)$, we have $p(a\,|\,\{a,a_0\})=v(a)\cdot p(a_0\,|\,\{a,a_0\})$.

24. Proof of Proposition 1 (cont-d)

- By definition of a choice function, for each set A containing a_0, we have:

$$p(a\,|\,A) = p(a\,|\,\{a, a_0\}) \cdot p(\{a, a_0\}\,|\,A) \text{ and}$$

$$p(a_0\,|\,A) = p(a_0\,|\,\{a, a_0\}) \cdot p(\{a, a_0\}\,|\,A).$$

- Dividing the first equality by the second one, we get

$$\frac{p(a\,|\,A)}{p(a_0\,|\,A)} = \frac{p(a\,|\,\{a, a_0\})}{p(a_0\,|\,\{a, a_0\})}.$$

- By definition of $v(a)$, this means that $\dfrac{p(a\,|\,A)}{p(a_0\,|\,A)} = v(a)$.
- Similarly, for each $b \in A$, we have $\dfrac{p(b\,|\,A)}{p(a_0\,|\,A)} = v(b)$.
- Dividing these two equalities, we conclude that for each set A containing a_0, we have $\dfrac{p(a\,|\,A)}{p(b\,|\,A)} = \dfrac{v(a)}{v(b)}$.

25. Proof of Proposition 1 (cont-d)

- 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

$p(a\,|\,B) = p(a\,|\,\{a,b\})\cdot p(\{a,b\}\,|\,B), p(b\,|\,B) = p(b\,|\,\{a,b\})\cdot p(\{a,b\}\,|\,B)$.

- Dividing these equalities by each other, we conclude that

$$\frac{p(a\,|\,B)}{p(b\,|\,B)} = \frac{p(a\,|\,\{a,b\})}{p(b\,|\,\{a,b\})}.$$

- The right-hand side of this equality does not depend on the set B.
- So the left-hand side, i.e., the ratio $\dfrac{p(a\,|\,B)}{p(b\,|\,B)}$, also does not depend on the set B.
- In particular, for the sets B that contain a_0, this ratio – according to a previous formula – is equal to $v(a)/v(b)$.

26. Proof of Proposition 1 (cont-d)

- Thus, the same equality holds for all sets A – not necessarily containing the element a_0.
- From this equality, we conclude that $\dfrac{p(a\,|\,A)}{v(a)} = \dfrac{p(b\,|\,A)}{v(b)}$.
- In other words, for all elements $a \in A$, the ratio $\dfrac{p(a\,|\,A)}{v(a)}$ is the same.
- Let us denote this ratio by c_A; then, for each $a \in A$, we have:
$$p(a\,|\,A) = c_A \cdot v(a).$$
- From $\sum\limits_{b\in A} p(b\,|\,A) = 1$, we can now conclude that: $c_A \cdot \sum\limits_{b\in A} v(b) = 1$, thus $c_A = \dfrac{1}{\sum\limits_{b\in A} v(b)}$.
- Substituting this expression into the formula for $p(a\,|\,A)$, we get the desired expression.
- The proposition is proven.

27. Proof of Proposition 2

- It is easy to check that the above function is consistent and monotonic.
- So, to complete the proof, it is sufficient to prove that every consistent monotonic function has the desired form.
- Indeed, let us assume that the function V is consistent and monotonic.
- Then, due to consistency, it satisfies the corresponding formula.
- Taking logarithm of both sides of this formula, we conclude that for the auxiliary function $u(\underline{a},\overline{a}) \stackrel{\text{def}}{=} \ln(V(\underline{a},\overline{a}))$, we have:

$$u(\underline{a},\overline{a})+u\left(\underline{b},\overline{b}\right)=u\left(\underline{a}+\underline{b},\overline{a}+\overline{b}\right)+c \text{ for some } c.$$

- 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 this formula, we conclude that

$$U(\underline{a},\overline{a})+U\left(\underline{b},\overline{b}\right)=U\left(\underline{a}+\underline{b},\overline{a}+\overline{b}\right).$$

- So, the function U is additive.

28. Proof of Proposition 2 (cont-d)

- Every monotonic function is locally bounded.
- We can use the general classification of additive locally bounded functions 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, we get the desired formula for

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

- Here, $C = \exp(c)$, $k = k_1 + k_2$ and $\alpha_H = \dfrac{k_1}{k_1 + k_2}$.
- The proposition is proven.