Logit Discrete Choice Model: A New Distribution-Free Justification

Ruey L. Cheu[1,2], Hung T. Nguyen[4]
Tanja Magoc[3], and Vladik Kreinovich[2,3]
[1]Department of Civil Engineering
[2]Center for Transportation Infrastructure Systems
[3]Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968, USA
[4]Department of Mathematical Sciences
New Mexico State University
Las Cruces, NM 88003, USA
contact email vladik@utep.edu

Abstract

According to decision making theory, if we know the user's utility $U_i = U(s_i)$ of all possible alternatives s_i, then we can uniquely predict the user's preferences. In practice, we often only know approximate values $V_i \approx U_i$ of the user's utilities. Based on these approximate values, we can only make probabilistic predictions of the user's preferences. It is empirically known that in many real-life situations, the corresponding probabilities are described by a *logit* model, in which the probability p_i of selecting the alternative s_i is equal to $p_i = \dfrac{e^{\beta \cdot V_i}}{\sum\limits_{j=1}^{n} e^{\beta \cdot V_j}}$. There exist many theoretical explanations of this empirical formula, some of these explanations led to a 2001 Nobel prize. However, it is known that the logit formula is empirically valid even when the assumptions behind the existing justifications do not hold. To cover such empirical situations, it is therefore desirable to provide a new distribution-free justification of the logit formula. Such a justification is provided in this paper.

1 Formulation of the Problem

Traditional approach to decision making. In decision making theory, it is proven that under certain reasonable assumption, a person's preferences are defined by his or her *utility function* $U(x)$ which assigns to each possible outcome

x a real number $U(x)$ called *utility*; see, e.g., Keeney and Raiffa [5], Raiffa [10]. In many real-life situations, a person's choice s does not determine the outcome uniquely, we may have different outcomes $x_1, \ldots, x_n$ with probabilities, correspondingly, $p_1, \ldots, p_n$.

For example, drivers usually select the path with the shortest travel time. However, when a driver selects a path s, the travel time is often not uniquely determined: we may have different travel times $x_1, \ldots, x_n$ with corresponding probabilities $p_1, \ldots, p_n$.

For such a choice, we can describe the utility $U(s)$ associated with this choice as the expected value of the utility of outcomes: $U(s) = E[U(x)] = p_1 \cdot U(x_1) + \ldots + p_n \cdot U(x_n)$. Among several possible choices, a user selects the one for which the utility is the largest: a possible choice s is preferred to a possible choice s' (denoted $s > s'$) if and only if $U(s) > U(s')$.

It is important to mention that the utility function is not uniquely determined by the preference relation. Namely, for every two real numbers $a > 0$ and b, if we replace the original utility function $U(x)$ with the new one $U'(x) \stackrel{\text{def}}{=} a \cdot U(x) + b$, then for each choice s, we will have

$$U'(s) = E[a \cdot U(x) + b] = a \cdot E[U(x)] + b = a \cdot U(s) + b$$

and thus, $U'(s) > U'(s')$ if and only if $U(s) > U(s')$.

Situations in which we can only predict probabilities of different decision. One important application of decision making theory is predicting the user decisions. If we know the exact values $U(s)$ of the utilities, then we can predict the exact choice. For example, if the user has to choose between alternatives s and s', then the user chooses s if $U(s) \geq U(s')$ and s' if $U(s) \leq U(s')$.

In practice, we do not know the exact values $U(s)$ of the user's utility, we only know the approximate values $V(s) \approx U(s)$. Due to the difference between the observed (approximate) values $V(s)$ and the actual (unknown) values $U(s)$, we are no longer able to uniquely predict the user's behavior: e.g., even when $V(s) > V(s')$, we may still have $U(s) < U(s')$, and thus, it is possible that the user will prefer s.

If the differences $V(s) - U(s)$ and $V(s') - U(s')$ are small, then for $V(s) \gg V(s')$, we can be reasonably sure that $U(s) > U(s')$ and thus, that the user will select s. Similarly, if $V(s) \ll V(s')$, we can be reasonably sure that $U(s) < U(s')$ and thus, that the user will select s'. However, when the values $V(s)$ and $V(s')$ are close, then there is a certain probability that $U(s) > U(s')$ and thus, that the user will select s, and there is also a certain probability that $U(s) < U(s')$ and thus, that the user will select s'.

In this situation, based on the (approximate) utility values $V(s)$ and $V(s')$, we cannot exactly predict whether the user will prefer s or s' – because for the same values of $V(s)$ and $V(s')$, the user can prefer s and the user can also prefer s'. The best we can do in this situation is to predict the *probability* $P(s > s')$ of selecting s over s'.

Discrete choice: a formal description of the problem. Let us formulate the problem in precise terms. We have n different alternatives $s_1, \ldots, s_n$. For each of these alternative s_i, we know the (approximate) utility value $V_i \stackrel{\text{def}}{=} V(s_i)$. Based on these utility values $V(s_1), \ldots, V(s_n)$, we would like to predict the probability p_i that a user will select the alternative s_i.

Models used for such prediction are usually called *discrete choice models* [11].

Invariance requirements in discrete choice models. As we have mentioned, the utility function is not uniquely determined by the preference relation. Namely, whenever the original utility function $U(s)$ describes the user's preference, then, for every $a > 0$ and b, the new function $U'(s) = a \cdot U(s) + b$ also describes the same preference. In other words, we can shift all the values of the utility function $u(s) \to U(s) + b$, and we can re-scale all the values $U(s) \to a \cdot u(s)$, and the resulting utility function will still describe the same preferences.

It is therefore reasonable to assume that if we shift the values of the approximate utility function, i.e., if we replace the original values $V(s_i)$ with the new values $V'(s_i) = V(s_i) + b$, then we should get the same preference probabilities:

$$p_i(V(s_1), V(s_2) \ldots, V(s_n)) = p_i(V(s_1) + b, V(s_2) + b, \ldots, V(s_n) + b).$$

In particular, if we take $b = -V(s_1)$, then we conclude that

$$p_i(V(s_1), V(s_2) \ldots, V(s_n)) = p_i(0, V(s_2) - V(s_1), \ldots, V(s_n) - V(s_1)),$$

i.e., that the probabilities depend only on the *differences* between the utility values – but not on the values themselves.

At first glance, it may seem reasonable to similarly require that the probability not change under re-scaling. However, in this case, re-scaling does not make intuitive sense, because we have a natural scale. For example, as a unit for such a scale, we can choose a standard deviation of the difference $U(s) - V(s)$ between the (unknown) actual utility $U(s)$ and the (known) approximate value of this utility $V(s)$.

In line with this analysis, in discrete choice models, it is usually assumed that the probabilities do not change with shift but it is *not* assumed that these probabilities are scale-invariant.

Logit: the most widely used discrete choice model. The most widely used discrete choice model is a *logit* model in which

$$p_i(V_1, \ldots, V_n) = \frac{e^{\beta \cdot V_i}}{\sum\limits_{j=1}^{n} e^{\beta \cdot V_j}} \tag{1}$$

for some parameter β. This model was first proposes in [6].

Logit: original justification. In Luce [6], this model was justified based on the assumption of *independence of irrelevant alternatives*, according to which the relative probability of selecting s_1 or s_2 should not change if we add a third alternative s_3. In formal terms, this means that the probability of selecting s_1 out of two alternatives s_1 and s_2 should be equal to the conditional probability of selecting s_1 from three alternatives s_1, s_2, and s_3 under the condition that either s_1 or s_2 are selected.

It can be proven that under this assumption, the ratio p_i/p_j of the probabilities p_i and p_j should only depend on V_i and V_j; moreover, that we must have $p_i/p_j = f(V_i)/f(V_j)$ for some function $f(z)$. The requirement that this ratio be shift-invariant then leads to the conclusion that $f(z) = e^{\beta \cdot z}$ for some β – and thus, to the logit model.

Limitations of the original justification. At first glance, the above independence assumption sounds reasonable (and it is often reasonable). However, there are reasonable situations where this assumption is counter-intuitive; see, e.g., Chipman [2], Debreu [3], Train [11].

For example, assume that in some cities, all the buses were originally blue. To get from point A to point B, a user can choose between taking a taxi (s_1) and taking a blue bus (s_2). A taxi is somewhat better to this user, so he selects a taxi with probability $p_1 = 0.6$ and a blue bus with the remaining probability $p_2 = 1 - 0.6 = 0.4$. In this case, the ratio p_1/p_2 is equal to 1.5.

Suppose now that the city decided to buy some new buses, and to paint them red. Let us also suppose that the comfort of the travel did not change, the buses are exactly the same. From the common sense viewpoint, it does not matter to the user whether buses are blue or red, so he should still select a taxi with probability $p_1 = 0.6$ and buses with probability 0.4. However, from the purely mathematical viewpoint, we now have *three* options: taking a taxi (s_1), taking a blue bus (s_2), and taking a red bus (s_3). Here, the probability of taking a bus is now $p_2 + p_3 = 0.4$. Hence, $p_2 < 0.4$ and so, the ratio p_1/p_2 is different from what we had before – contrary to the above independence assumption.

Current justification. An alternative justification for logit started with the unpublished result of Marley first cited in Luce and Suppes [7]. Marley has shown that if we assume that the approximation errors $\varepsilon(s) \stackrel{\text{def}}{=} U(s) - V(s)$ are independent and identically distributed, and if this distribution is the Gumbel distrubution, then the probability of selecting s_i indeed follows the logit formula.

Gumbel distribution can be characterized by the cumulative distribution function $F(\varepsilon) = e^{-e^{-\varepsilon}}$; it is a known distribution of extreme values.

In 1974, McFadden [8] showed that, vice versa, if we assumed that the approximation errors $\varepsilon(s)$ are independent and identically distributed, and the choice probabilities are described by the logit formula, then the errors $\varepsilon(s)$ must follow the extreme value (Gumbel) distribution.

This justification was one of the main achievements for which D. McFadden received a Nobel prize in 2001 [9].

Limitations of the current justification. The problem with this justification is that the logit model is known to work well even in the cases when different approximation errors are differently distributed; see, e.g., Train [11].

For such situations, the only known alternative explanation is Luce's original one. The main limitation of this explanation was that it is based on the independence assumption. This is not so critical if we have three or more alternatives. Indeed, in this case, the empirical logit formula (that we are trying to explain) satisfies this assumption, so making this assumption in the situations when the logit formula holds makes sense.

This limitation, however, becomes crucial if we only consider the case of two alternatives. In this case, the independence assumption cannot even be formulated and therefore, Luce's justification does not apply. So, we arrive at the following problem.

Formulation of the problem. We need to come up with a new distribution-free justification for the logit formula, i.e., with a justification that does not depend on the assumption that approximation errors are independent and identically distributed. Such a justification is provided in this paper.

2 Preliminary Analysis

In accordance with the above formulation of the problem, we are interested in the case of $n = 2$ alternatives s_1 and s_2. We know the approximate utility values V_1 and V_2, and we know that the probability p_1 of selecting the first alternative p_1 should only depend on the difference $V_1 - V_2$: $p_1 = F(V_1 - V_2)$ for some function $F(z)$. Our objective is to find this function $F(z)$. Let us first describe reasonable properties of this function $F(z)$.

When s_2 is fixed (hence V_2 is fixed) but the alternative s_1 is improving (i.e., V_1 is increasing), then the probability of selecting s_1 can only increase (or at least remain the same – e.g., if that probability was already equal to 1, it cannot further increase). In other words, as the difference $V_1 - V_2$ increases, the probability $p_1 = F(V_1 - V_2)$ should also increase (or at least remain the same). Thus, it is reasonable to require that the function $F(z)$ should be (non-strictly) increasing.

When s_2 and V_2 are fixed and s_1 becomes better and better, i.e., $V_1 \to +\infty$, then we should select s_1 with probability tending to 1. So, we must have $F(z) \to 1$ as $z \to +\infty$.

Similarly, s_2 and V_2 are fixed, and s_1 becomes worse and worse, i.e., $V_1 - V_2 \to -\infty$, then we should prefer s_2. So, we must have $F(z) \to 0$ as $z \to -\infty$.

Since we only have two alternatives, the probability $p_1 = F(V_1 - V_2)$ and the probability $p_2 = F(V_2 - V_1)$ must always add up to 1. Thus, we must have $F(z) + F(-z) = 1$ for all z.

So, we arrive at the following definition.

Definition 1 *By a* choice function*, we mean a function $F : R \to [0,1]$ which is (non-strictly) increasing, and for which $F(z) \to 1$ as $z \to +\infty$, $F(z) \to 0$ as $z \to -\infty$, and $F(z) + F(-z) = 1$ for all z.*

3 Main Idea

Our main idea is as follows. Up to now, we have discussed how to *describe* the user's behavior, but often, the ultimate objective is how to *modify* this behavior. For example, in transportation problems, the goal is often to use public transportation to relieve traffic congestion and related pollution. In this case, the problem is not just to estimate the probability of people using public transportation, but to find out how to increase this probability.

One way to increase this probability is to provide incentives. If we want to encourage people to prefer alternative s_1, then we can provide those who select this alternative with an additional benefit of value v_0. In this case, for alternatives $s_i \neq s_1$, the corresponding utility V_i remains the same, but for the alternative s_1, we have a new value of utility $V_1' = V_1 + v_0$.

After this addition, the original probability

$$p_1 = F(V_1 - V_2) \tag{2}$$

of selecting the alternative s_1 changes to a new value

$$p_1' = F(V_1' - V_2) = F(V_1 + v_0 - V_2). \tag{3}$$

These formulas can be simplified if we denote the difference $V_1 - V_2$ between the approximate utility values by ΔV. In these new notations, the original probability

$$p_1 = F(\Delta V) \tag{4}$$

is replaced by the new probability

$$p_1' = F(\Delta V + v_0). \tag{5}$$

This change of probability can be described in general terms: we receive new information – that there are now incentives. Based on this new information, we update our original probabilities p_i of selecting different alternatives s_i.

From the statistical viewpoint (see, e.g., [4, 12]), when we receive new information, the correct way of updating probabilities is by using the Bayes formula. Specifically, if we have n incompatible hypotheses $H_1, \ldots, H_n$ with initial probabilities

$$P_0(H_1), \ldots, P_0(H_n), \tag{6}$$

then, after observations E, we update the initial probabilities to the new values:

$$P(H_i \,|\, E) = \frac{P(E \,|\, H_i) \cdot P_0(H_i)}{P(E \,|\, H_1) \cdot P_0(H_1) + \ldots + P(E \,|\, H_n) \cdot P_0(H_n)}. \tag{7}$$

Thus, we should require that the function $F(z)$ be such for which the transition from the old probability (4) to the new probability (5) can be described by the (fractionally linear) Bayes formula (7).

4 From the Main Idea to the Exact Formulas

Let us formalize the above requirement. In the case of two alternatives s_1 and s_2, we have two hypotheses: the hypothesis H_1 that the user will prefer s_1 and the opposite hypothesis H_2 that the user will prefer s_2. Initially, we did now know about any incentives, we only knew the approximate utility V_1 of the first alternative and the approximate utility V_2 of the second alternative. Based on the information that we initially had, we concluded that the probability of the hypothesis H_1 is equal to $p_1 = p(H_1) = F(\Delta V)$ (where $\Delta V = V_1 - V_2$), and the probability of the opposite hypothesis H_2 is equal to $p_2 = p(H_2) = 1 - p_1$.

Now, suppose that learn that there was no incentive to select alternative s_2 and an incentive of size v_0 to select alternative s_1. This new information E changes the probabilities of our hypotheses H_1 and H_2. Namely, according to Bayes formula, after the new information E, the probability p_1 should be updated to the following new value $p_1' = P(H_1 \,|\, E)$:

$$p_1' = \frac{P(E \,|\, H_1) \cdot P(H_1)}{P(E \,|\, H_1) \cdot p_1 + P(E \,|\, H_2) \cdot P(H_2)}. \tag{8}$$

The probability $P(E \,|\, H_1)$ is the conditional probability with which we can conclude that there was an incentive of size v_0 based on the fact that the user actually selected the alternative s_1. This conditional probability is, in general, different for different values v_0. To take this dependence into account, we will denote this conditional probability $P(E \,|\, H_1)$ by $A(v_0)$.

Similarly, the probability $P(E \,|\, H_2)$ is the conditional probability with which we can conclude that there was an incentive of size v_0 for alternative s_1 based on the fact that the user actually selected the alternative s_2. This conditional probability is also, in general, different for different values v_0. To take this dependence into account, we will denote this conditional probability $P(E \,|\, H_2)$ by $B(v_0)$.

If we substitute the expressions $P(E \,|\, H_1) = A(v_0)$, $P(E \,|\, H_2) = B(v_0)$, $P(H_1) = F(\Delta V)$, and $P(H_2) = 1 - P(H_1) = 1 - F(\Delta V)$ into the above formula (8), then we conclude that

$$p_1' = \frac{A(v_0) \cdot F(\Delta V)}{A(v_0) \cdot F(\Delta V) + B(v_0) \cdot (1 - F(\Delta V))}. \tag{9}$$

On the other hand, once we know that there was an incentive v_0 to select the alternative s_1 and no incentive for the alternative s_2, then we have a better idea of the resulting utilities of the user: namely, the new value of the approximate utility is $V_1 + v_0$ for alternative s_1 and V_2 for the alternative s_2. In accordance with our expression for the choice probability based on the approximate utility values, the new probability of selecting s_1 should be equal to $F((V_1 + v_0) - V_2)$, i.e., to $F(\Delta V + v_0)$ (expression (4)).

If the probability update was done correctly, in full accordance with the Bayes formula, then this new value (4) must be equal to the value (9) that comes from using the Bayes formula. So, we arrive at the following definition:

Definition 2 *A choice function $F(z)$ is called* Bayes correct *if, for every v_0, there exist values $A(v_0)$ and $B(v_0)$ for which*

$$F(\Delta V + v_0) = \frac{A(v_0) \cdot F(\Delta V)}{A(v_0) \cdot F(\Delta V) + B(v_0) \cdot (1 - F(\Delta V))} \tag{10}$$

for all ΔV.

Comment. In other words, we require that the 2-parametric family of functions $F = \left\{ \dfrac{A \cdot F(\Delta V)}{A \cdot F(\Delta V) + B} \right\}$ corresponding to Bayesian updates be *shift-invariant* under a shift $\Delta V \to \Delta V + v_0$.

5 Main Result

Theorem 1 *Every Bayes correct choice function $F(z)$ has the form*

$$F(\Delta V) = \frac{1}{1 + e^{-\beta \cdot \Delta V}} \tag{11}$$

for some real number β.

If we substitute $\Delta V = V_1 - V_2$ into this formula, and multiply the numerator and the denominator of the resulting formula by $e^{\beta \cdot V_1}$, then we conclude that for every Bayes correct choice function $F(z)$, we have

$$p_1 = F(V_1 - V_2) = \frac{e^{\beta \cdot V_1}}{e^{\beta \cdot V_1} + e^{\beta \cdot V_2}}. \tag{12}$$

Thus, for the desired case of two alternatives, we indeed provide a new distribution-free justification of the logit formula.

6 Proof

It is known that many formulas in probability theory can be simplified if instead of the probability p, we consider the corresponding odds

$$O = \frac{p}{1 - p}. \tag{13}$$

(If we know the odds O, then we can reconstruct the probability p as $p = O/(1 + O)$.) The right-hand side of the formula (10) can be represented in terms of odds $O(\Delta V)$, if we divide both the numerator and the denominators by $1 - F(\Delta V)$. As a result, we get the following formula:

$$F(\Delta V + v_0) = \frac{A(v_0) \cdot O(\Delta V)}{A(v_0) \cdot O(\Delta V) + B(v_0)}. \tag{14}$$

Based on this formula, we can compute the corresponding odds $O(\Delta V + v_0)$: first, we compute the value

$$1 - F(\Delta V + v_0) = \frac{B(v_0)}{A(v_0) \cdot O(\Delta V) + B(v_0)}, \tag{15}$$

and then divide (14) by (15), resulting in:

$$O(\Delta V + v_0) = c(v_0) \cdot O(\Delta V), \tag{16}$$

where we denoted $c(v_0) \stackrel{\text{def}}{=} A(v_0)/B(v_0)$. It is known (see, e.g., [1]) that all monotonic solutions of the functional equation (16) are of the form $O(\Delta V) = C \cdot e^{\beta \cdot \Delta V}$. Therefore, we can reconstruct the probability $F(\Delta V)$ as

$$F(\Delta V) = \frac{O(\Delta V)}{O(\Delta V) + 1} = \frac{C \cdot e^{\beta \cdot \Delta V}}{C \cdot e^{\beta \cdot \Delta V} + 1}. \tag{17}$$

The condition $F(z) + F(-z) = 1$ leads to $C = 1$. Dividing both the numerator and the denominator of the right-hand side by $e^{\beta \cdot \Delta V}$, we get the desired formula (11). Q.E.D.

Acknowledgments. This work was supported in part by NSF grant EAR-0225670, by Texas Department of Transportation grant No. 0-5453, and by the Japan Advanced Institute of Science and Technology (JAIST) International Joint Research Grant 2006-08.

References

[1] J. Aczel, *Lectures on functional equations and their applications*, Dover, New York, 2006.

[2] J. Chipman, "The foundations of utility", *Econometrica*, 1960, Vol. 28, pp. 193–224.

[3] G. Debreu, "Review of R. D. Luce, 'Individual Choice Behavior' ", *American Economic Review*, 1960, Vol. 50, pp. 186–188.

[4] E. T. Jaynes and G. L. Bretthorst (ed.), *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.

[5] R. L. Keeney and H. Raiffa, *Decisions with Multiple Objectives*, John Wiley and Sons, New York, 1976.

[6] D. Luce, *Individual Choice Behavior*, John Wiley and Sons, New York, 1959.

[7] D. Ruce and P. Suppes, "Preference, utility, and subjective probability", In: D. Luce, R. Bush, and E. Galanter, *Handbook on Mathematical Psychology*, John Wiley and Sons, New York, 1965, pp. 249–410.

[8] D. McFadden, "Conditional logit analysis of qualitative choice behavior", In: P. Zarembka (ed.), *Frontiers in Econometrics*, Academic Press, New York, 1974, pp. 105–142.

[9] D. McFadden, "Economic choices", *American Economic Review*, 2001, Vol. 91, pp. 351–378.

[10] H. Raiffa, *Decision Analysis*, Addison-Wesley, Reading, Massachusetts, 1970.

[11] K. Train, *Discrete Choice Methods with Simulation*, Cambridge University Press, Cambridge, Massachusetts, 2003.

[12] H. M. Wadsworth (ed.), *Handbook of statistical methods for engineers and scientists*, McGraw-Hill Publishing Co., New York, 1990.