Answers to Problem Set 1

John Rust, Professor of Economics Georgetown University Spring 2022

I. This problem set leads you through the steps to derive the multinomial logit model (MNL) as a random utility model with additive random utilities that have independent Type 1 extreme value distributions. Recall from the notes that the MNL model was initially derived by the mathematical psychologist Duncan Luce based on the axiom of *independence from irrelevant alternatives* (IIA), leading to this formula for the conditional choice probability

$$P(d|x) = \frac{\exp\{u(x,d)\}}{\sum_{d'\in D(x)} \exp\{u(x,d')\}} \tag{1}$$

where $u(x,d)$ is some function of x and the discrete choice d, and for each x, $D(x)$ is a finite choice set. McFadden derived the MNL model as a *random utility model* (RUM). That is, Mcfadden (following a lead of Thurstone who worked on similar models in the 1930s) assumed that the choice of an alternative $d \in D(x)$ is not only affected by x but also by a *random utility component* $\varepsilon(d)$ that reflects other factors and "states" of the individual making the choice that the econometrician does not observe. So the individual's choice is governed by a *decision rule* $d(x,\varepsilon)$ that depends on a vector of variables that the econometrician can observe (both states of the individual and potentially characteristics or "attributes" of the items the individual is choosing between) given by

$$d(x,\varepsilon) = \underset{d\in D(x)}{argmax}[u(x,d)+\varepsilon(d)] \tag{2}$$

Note that McFadden's formulation of the random utility model invokes an implicit *additive separability* (AS) assumption: the utility, which might be written in general as $u(x,\varepsilon,d)$ takes the specific additively separable form $u(x,d)+\varepsilon(d)$ where we make an assumption about the *probability distribution* of the unobserved components of the utility function, $\varepsilon \equiv \{\varepsilon(d)|d \in D(x)\}$. If ε is a multivariate continuous random vector with full support over $R^{|D(x)|}$ (where $|D(x)|$ is the number of elements in the choice set), with CDF $F(\varepsilon|x)$ then it is not hard to show that the implied *conditional choice probability* $P(d|x)$ given by

$$P(d|x) = \int_{\varepsilon} I\{d(x,\varepsilon) = d\}\, dF(\varepsilon|x) \tag{3}$$

satisfies $P(d|x) > 0$ for each $d \in D(x)$. That is, there is positive probability of observing the individual choosing any alternative $d \in D(x)$. McFadden showed that if ε has a multivariate Type 1 extreme value distribution (also called a Gumbel distribution), with CDF given by

$$F(\varepsilon) = \prod_{d\in D(x)} \exp\{-\exp\{-(\varepsilon(d)-\mu(d))/\sigma\}\} \tag{4}$$

where $\mu(d) \in (-\infty,\infty)$ is the *location parameter* of the continuous random variable $\varepsilon(d)$ and σ is the *scale parameter.*

A. Show that $F(\varepsilon)$ is a valid multivariate CDF and shows its support is all of $R^{|D(x)|}$. Are the random variables $\{\varepsilon(d)|d \in D(x)\}$ *IID*? Are they independently distributed?

Answer: This is fairly easy to show. First it is easy to show that for any $\varepsilon \in R^{|D(x)|}$ we have $0 < F(\varepsilon) < 1$, $F(\varepsilon)$ is strictly monotonically increasing in ε and

$$\begin{aligned} \lim_{\varepsilon \to +\infty} F(\varepsilon) &= 1 \\ \lim_{\varepsilon \to -\infty} F(\varepsilon) &= 0. \end{aligned} \tag{5}$$

So we conclude $F(\varepsilon)$ is indeed a multivariate CDF. It has "full support" on the entire Euclidean space $R^{|D(x)|}$ because the CDF is strictly increasing and thus has a positive density, $\frac{\partial}{\partial \varepsilon} F(\varepsilon) = f(\varepsilon) > 0$ for any $\varepsilon \in R^{|D(x)|}$. Further it is easy to see from equation (4) that $F(\varepsilon)$ is a product of its marginal distributions since the marginal (univariate CDF) $F_d(\varepsilon(d))$ is just the limit given by

$$F_d(\varepsilon(d)) = \lim_{\varepsilon(d_\sim) \to \infty} F(\varepsilon(d), \varepsilon(d_\sim)) \tag{6}$$

where $\varepsilon(d_\sim)$ are all components of the vector ε except for component d, i.e.

$$\varepsilon(d_\sim) = (\varepsilon(1), \ldots, \varepsilon(d-1), \varepsilon(d+1), \ldots, \varepsilon(|D(x)|)). \tag{7}$$

It follows that $F(\varepsilon)$ is a product of its marginals

$$F(\varepsilon) = \prod_{d \in D(x)} F_d(\varepsilon(d)) = \prod_{d \in D(x)} \exp\{-\exp\{-(\varepsilon(d) - \mu(d))/\sigma\}\}, \tag{8}$$

so by definition the components of the random vector $\tilde{\varepsilon} = (\tilde{\varepsilon}(1), \ldots, \tilde{\varepsilon}(|D(x)|))$ are independently distributed random variables. These random variables are not *IID* (i.e. independent and identically distributed) unless all of them have the same location parameters $\mu(d)$, i.e. only if $\mu(d) = \mu$ for some $\mu \in R$ for all $d \in D(x)$.

B. Show that the collection of extreme value distriubuted random variables $\{\varepsilon(d) | d \in D(x)\}$ is *max-stable* i.e. show that $\eta = \max\{\varepsilon(d) | d \in D(x)\}$ is also a Type 1 extreme value distribution and calculate its location parameter μ and scale parameter σ.

Answer: Remember how to derive the probability distribution of the maximum of a vector of random variables, i.e. the CDF of the scalar random variable $\tilde{Z}$ given by $\tilde{Z} = \max[\tilde{X}_1, \ldots, \tilde{X}_N]$ where $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_N)$ is a random vector where the component random variables $\tilde{X}_i$, $i = 1, \ldots, N$ may or may not be independently distributed. Let $F_X(x_1, \ldots, x_N)$ be the multivariate CDF of the random vector $\tilde{X}$. We have

$$\begin{aligned} F_Z(z) &= Pr\{\tilde{Z} \le z\} \qquad (9) \\ &= Pr\{\max(\tilde{X}_1, \ldots, \tilde{X}_N) \le z\} \\ &= F_X(z, z, \ldots, z, z) \qquad (10) \end{aligned}$$

since it is easy to see that $\{\omega \in \Omega | Z(\omega) \le z\} = \{\omega \in \Omega | (X_1(\omega) \le z, \ldots, X_N(\omega) \le z\}$, where Ω is the measure space over which the random variables are defined. In words, we have $\tilde{Z} \le z$ if and only if $\{\tilde{X}_1 \le z, \ldots, \tilde{X}_N \le z\}$.

Now apply this formula for the max to the extreme value distribution. We have

$$\begin{aligned} Pr\{\max(\varepsilon) \leq z\} &= F(z,\dots,z) \\ &= \prod_{d \in D(x)} F_d(z) \\ &= \prod_{d \in D(x)} \exp\{-\exp\{-(z-\mu(d))/\sigma\}\}, \\ &= \exp\{-\exp\{-(z-I)/\sigma\}\} \end{aligned} \tag{11}$$

where I is called the *inclusive value* and is given by

$$I = \sigma \log\left(\sum_{d \in D(x)} \exp\{\mu(d)/\sigma\} \right). \tag{12}$$

We can see from equation (12) that the distribution of $\max(\varepsilon)$ is a univariate Type 1 extreme value distribution with location parameter I given by the inclusive value in equation (12) and scale parameter σ. So we have shown that the maximum of independent Type 1 extreme value random variables is itself a Type 1 extreme value. That justifies the term "max-stable" — i.e. the Type 1 extreme value family is closed under the operation of taking the maximum just like the class of *stable random variables* is closed under the operation of summation, $+$: i.e. the sum of stable random variables is also a member of the class of stable random variables. Note that Gaussian random variables are an example of stable random variables: a sum of normally distributed random variables is normally distributed. But the normal family is not max-stable: the maximum of a collection of normally distributed random variables is not normally distributed. On the other hand the Type 1 extreme value family is not a subclass of stable random variables: sums of Type 1 extreme value random variables are generally not Type 1 extreme value. Note that there is a class of *generalized extreme value* random variables that are not necessarily independently distributed but which also has the max stable property.

C. Using the result in part B and the fact that if η is a Type 1 extreme value distribution with location parameter μ and scale parameter σ its mean is given by

$$E\{\eta\} = \mu + \gamma\sigma \tag{13}$$

where $\gamma \simeq 0.577\dots$ is *Euler's constant* and

$$\text{var}(\eta) = \sigma^2 \frac{\pi^2}{6} \tag{14}$$

write a formula for the *expected maximum utility*

$$E\left\{ \max_{d \in D(x)} [u(x,d) + \varepsilon(d)] \right\} \tag{15}$$

which McFadden called the *social surplus function.* Why would he call it that? Is there any analogy you can draw between the social surplus function and the *indirect utility function* of consumer theory?

answer Suppose we normalize the location parameter μ of the extreme value random variables $\varepsilon(d)$ entering the utility maximization problem (15) to zero, $\mu = 0$. However the utility component $u(x,d)$ is now equivalent to the location parameter of the extreme variable random variable $\eta(d) = u(x,d) + \varepsilon(d)$. The family $\eta = \{\eta(d)|d \in D(x)\}$ is a family of independent but non-identically distributed Type 1 Extreme value random variables with common scale parameter $\sigma > 0$. So, using the answer to part B above, it follows from the max-stability property of Type 1 extreme random variables that $\tilde{Z} = \max(\eta(1), \ldots, \eta(|D(x)|))$ is also a Type 1 extreme value random variable with location parameter I given by

$$I(\{u(x,d)|d \in D(x)\}) = \sigma \log\left(\sum_{d \in D(x)} \exp\{u(x,d)/\sigma\}\right), \tag{16}$$

and scale parameter σ. It follows that $E\{\tilde{Z}\} = E\{\max_{d \in D(x)}[u(x,d) + \varepsilon(d)]\} = \gamma\sigma + I(\{u(x,d)|d \in D(x)\})$ and the variance of $\tilde{Z}$ is $\text{var}(\tilde{Z}) = \sigma^2 \frac{\pi^2}{6}$ by equation (14).

Why did McFadden call the inclusive value $I(\{u(x,d)|d \in D(x)\})$ the "social surplus function"? Possibly because of its interpretation as the expectation of maximum utility. Imagine instead of a single consumer the ε vector is an index of individual consumer heterogeneity, so we can imagine an economy with a continuum of different consumers and each consumer's preferences are deterministic and indexed by the ε parameter. Each of these consumers makes a choice of an alternative $d \in D(x)$ so $\max_{d \in D(x)}[u(x,d) + \varepsilon(d)]$ is the maximized utility of a consumer of type ε. Let $f(\varepsilon)$ be the joint density of the ε in the population of consumers. Then we have

$$I(\{u(x,d)|d \in D(x)\}) = \int \left[\max_{d \in D(x)}[u(x,d) + \varepsilon(d)]\right] f(\varepsilon)d\varepsilon \tag{17}$$

so we can interpret the inclusive value as the expected or average social welfare over this hypothetical population of heterogeneous consumers. McFadden called it the "social surplus function" but we could also call it a "social welfare function" or we can call it the "expected maximum utility" function, or EMAX function. It can also be called the *smoothed max function* as we will show in the answer to part D below.

What is the relation to the indirect utility function? We discuss this in more detail in the answer to part D below, but recall what an indirect utility function is: it is the maximized value of utility subject to a consumer's budget constraint. The indirect utility function is defined in terms of the classical static model of continuous choice of consumption vectors subject to a budget constraint, whereas here we are considering discrete choice (choice of a specific alternative d from a finite choice set $D(x)$) and while in this formulation there are not necessarily any prices or income explicitly in the model, we can treat the state dependent choice set $D(x)$ as the analog of the "budget set" in static consumer theory. So the expected maximum utility, i.e. the inclusive value or the social surplus function or whatever you want to call it, has some rough analogy to the indirect utility function. We will further draw out this analogy in the answer to part D below.

D. Now forget about the Type 1 extreme value distribution for a moment, and suppose the unobserved additively separable components of utility $\varepsilon = \{\varepsilon(d)|d \in D(x)\}$ could be any continuous multivariate distribution with CDF $F(\varepsilon|x)$ such as a multivariate normal distribution. Show under as much generality as you can that the *Williams-Daly-Zachary Theorem* holds:

$$P(d|x) = \frac{\partial}{\partial u(x,d)} E\left\{\max_{d \in D(x)}[u(x,d) + \varepsilon(d)]\right\}. \tag{18}$$

HINT: Use the *Lebesgue Dominated Convergence Theorem* to show you can interchange the partial derivative and expectation operators in equation (18) to get equation (3). Can you draw a further analogy between equation (18) and Roy's Identity?

answer How does the Lebesgue Dominated Convergence Theorem play a role here? Well, it can be used to show that we can change the order of differentiation and integration in equation (18). Let's suppose that we have shown this first part already and thus we have justified switching the partial differentiation and integration operations in equation (18). Doing this we get

$$\begin{aligned} & \frac{\partial}{\partial u(x,d)} E\left\{ \max_{d \in D(x)} [u(x,d) + \varepsilon(d)] \right\} \\ = \; & E\left\{ \frac{\partial}{\partial u(x,d)} \max_{d \in D(x)} [u(x,d) + \varepsilon(d)] \right\} \\ = \; & E\left\{ I\left\{ d \in \underset{d' \in D(x)}{argmax} [u(x,d') + \varepsilon(d')] \right\} \right\} \\ = \; & \int_\varepsilon I\left\{ u(x,d) + \varepsilon(d) = \max_{d' \in D(x)} [u(x,d') + \varepsilon(d')] \right\} f(\varepsilon) d\varepsilon \\ = \; & P(d|x). \end{aligned} \tag{19}$$

To see why the first equation in (19) holds, note that the partial derivative $\frac{\partial}{\partial u(x,d)}$ of the maximum utility $\max_{d' \in D(x)}[u(x,d') + \varepsilon(d')]$ tells us what happens to this maximum utility when we increase $u(x,d)$ a little bit. If d is the utility maximizing alternative, then increasing $u(x,d)$ a little bit increases $\max_{d' \in D(x)}[u(x,d') + \varepsilon(d')]$ by the same amount, so the partial derivative equals 1 in this case. If d is not the utility maximizing alternative, then increasing $u(x,d)$ by a sufficiently small amount will not succeed in making alternative d the utility maximizing choice, so the partial derivative will be zero in this case. So the partial derivative equals 1 for values of (x,ε) where d is the utility maximizing choice, and 0 for values of (x,ε) where d is not the utility maximizing choice, so this implies that the partial derivative with respect to $u(x,d)$ is just the *indicator function* of the event that alternative d is the utility maximizing choice. This is the third equation in (19). When we integrate this indicator function over all ε in $R^{|D(x)|}$ holding x fixed (i.e. conditioning on the observed state x), we just get the conditional probability of chosing alternative $d \in D(x)$ (fourth and fifth equations in (19)), $P(d|x)$. In other words, the choice probability $P(d|x)$ is just the integral over the region in ε-space where it is optimal for the decision maker to choose alternative d: this is the content of the last equation in (19).

Note that the argument that the partial derivative of $\max_{d \in D(x)}[u(x,d) + \varepsilon(d)]$ equals either 0 or 1 actually holds for *almost all vectors* $\varepsilon \in R^{|D(x)|}$. There are values of ε that result in "ties" where more than 1 alternative $d \in D(x)$ is the utility maximizing choice. For the cases where there are ties, there can be kinks in the max function at these points and the partial derivative may not exist. However the set of points $\varepsilon \in R^{|D(x)|}$ where ties occur and the maximum utility is not differentiable is a set of *measure zero* with respect to Lebesgue measure on $R^{|D(x)|}$. Since the CDF of a Type 1 extreme value distribution is *absolutely continuous* with respect to Lebesgue measure (this means that any subset of the space that has Lebesgue measure zero will also have probability zero under the multivariate Type 1 extreme CDF on $R^{|D(x)|}$), we can conclude that equation (19) holds *almost everywhere* and thus with probability 1 with respect to the probability distribution for ε.

Now this is where the Lebesgue Dominated Convergence Theorem comes in. This theorem tells us that if we have a sequence of functions $\{g_n(\varepsilon)\}$ that converge pointwise to a function $g(\varepsilon)$ for almost all ε, and if the $g_n(\varepsilon)$ are dominated almost everywhere by some integrable function $h(\varepsilon)$ in the sense that $|g_n(\varepsilon)| \leq h(\varepsilon)$, for all n and almost all ε, and we have

$$\int h(\varepsilon)d\varepsilon < \infty, \tag{20}$$

then we have

$$\lim_{n\to\infty} \int g_n(\varepsilon)d\varepsilon = \int g(\varepsilon)d\varepsilon, \tag{21}$$

where we can conclude that all of the function $\{g_n(\varepsilon)\}$ and the limiting function $g(\varepsilon)$ are all integrable and dominated by the integral of the dominating function h

$$\int g_n(\varepsilon)d\varepsilon \leq \int h(\varepsilon)d\varepsilon < \infty \tag{22}$$

and

$$\int g(\varepsilon)d\varepsilon \leq \int h(\varepsilon)d\varepsilon < \infty. \tag{23}$$

Now how does this relate to the ability to interchange the operations of partial differentiation and integration in equation (19) above? Well, look at equation (21). It tells us that we can interchange the operations of "taking a limit" and integration. But a partial derivative like any ordinary derivative is itself a limit: a limit of secant approximations to the slope of a function at a given point. So with a bit of extra thought, you can express the partial derivative of the integral in the first equation in (19) as a limit of these secant approximations to the partial derivative, but you should be able to show that these secant approximations have slopes that are never greater than 1. So the "dominating function" $h(\varepsilon)$ just equals $1 \times f(\varepsilon)$ which is integrable since $f(\varepsilon)$ is the density of a vector of Type 1 extreme value random variables and this integrates to 1 which is finite. Thus with some thought you can see that the interchange of the order ot taking limits and integration as in the original statement of the Lebesgue Dominated Convergence Theorem in equation (21) implies also that the order of taking the partial differentiation operator and the integral operator does not matter and we get the same result regardless of which order we take these operations. Finally we see that taking the partial derivative inside the expectation operator we see again as shown in equation (19) that the partial derivative of the expected maximum function with respect to $u(x,d)$ equals the conditional choice probability $P(d|x)$.

Bonus question: If the Lebesgue Dominated Convergence Theorem justifies the interchange of the derivative and expectation operators, would it also jusity the interchange of the expectation and maximization operators, say, in the defintion of expected maximum utility, so expected maximum utility equals the maximum of expected utilities, e.g. does this equation hold?

$$E\left\{\max_{d\in D(x)} [u(x,d)+\varepsilon(d)]\right\} = \max_{d\in D(x)} E\{u(x,d)+\varepsilon(d)\}? \tag{24}$$

Now returning to the answer of question D, I asked you to draw an analogy between the Williams-Daly-Zachary result and Roy's Identity in static consumer theory. In what way is equation (18) similar to Roy's identity? Recall that Roy's Identity tells us that if $V(p,y)$ is an indirect utility function, i.e.

$$V(p,y) = \max_{\{x|px \le y\}} u(x) \tag{25}$$

then we have

$$x(p,y) = -\frac{\frac{\partial}{\partial p}V(p,y)}{\frac{\partial}{\partial y}V(p,y)} \tag{26}$$

where $x(p,y)$ is the "Marhalliand demand function" and $u(x)$ is the "direct utility function." So Roy's Identity tells us that demand, $x(p,y)$, is what we get by measuring the effect of raising prices p a little bit on indirect utility. Raising prices without raising income will reduce indirect utility, so that is why we need to put a negative sign in equation (26) and divide by the marginal utility of income, $\frac{\partial}{\partial y}V(p,y)$ to put things in units of quantities x rather than in units of "utils" which is what $V(p,y)$ is measured in. But the analogy is that we can recover demand from the indirect utility function, but in a discrete choice problem the expected maximum utility $\max_{d' \in D(x)}[u(x,d') + \varepsilon(d')]$ is similar to the indirect utility function $V(p,y)$. Similarly the analog of the Marshallian demand function $x(p,y)$ in the discrete choice context is the conditional choice probability $P(d|x)$. But rather than looking at the effect of changing prices on demand in the continuous choice setting for Roy's identity, the Willams-Daly-Zachary Theorem just tells us to look at the effect of making a single alternative d a little bit more attractive. The effect of doing this on "demand" is precisely just the conditional choice probability $P(d|x)$. Maybe it is a stretch to analogize the Williams-Daly-Zachary Theorem to Roy's Identity, but later we will also study the "Hotz-Miller Inversion Theorem" and this is very much in the spirit of duality theory in continuous consumer choice where it is shown that given a Marshallian demand function, one can recover a direct utility function that yields it via "integration" of the partial differential equation for $x(p,y)$ given in Roy's identity, and then further inverting the indirect utility function $V(p,y)$ to obtain the direct utility function $u(x)$. The Hotz-Miller Inversion Theorem states, roughly, that we can recover utility differences $u(x,d) - u(x,0)$ where $0 \in D(x)$ is some "normalizing choice" from the conditional choice probabilities $\{P(d|x)\}$.

E. Now using the result you derived in part C where you (hopefully) were able to derive a closed form expression for the social surplus function, use the Williams-Daly-Zachary Theorem (18) to show McFadden's result, namely that Type 1 extreme value distributed preference shocks (random utilities) results in the classic multinomial logit model formula (1), except that the utilities $u(x,d)$ have to be divided by the scale parameter σ.

answer Once we have proven the Williams-Daly-Zachary Theorem then deriving the multinomial choice probabilities is super easy, using only a few lines of calculus and the expression for the inclusive value in equation (16).

$$\begin{aligned} P(d|x) &= \frac{\partial}{\partial u(x,d)} E\left\{\max_{d' \in D(x)}[u(x,d') + \varepsilon(d')]\right\} \\ &= \frac{\partial}{\partial u(x,d)} I\left(\{u(x,d')|d' \in D(x)\}\right) \\ &= \frac{\exp\{u(x,d)/\sigma\}}{\sum_{d' \in D(x)} \exp\{u(x,d')/\sigma\}}. \end{aligned} \tag{27}$$

F. Show that the MNL model is the same as what people in the machine learning literature refer to as the *soft-max function* and prove this

$$\lim_{\sigma \downarrow 0} P(d|x) = \begin{cases} 1/n & \text{if } u(x,d) \geq u(x,d') \quad \forall d' \in D(x) \\ 0 & \text{otherwise} \end{cases} \tag{28}$$

where n is the number of alternatives in $D(x)$ that achieve the maximal utility. Similarly, we might call the social surplus function for the Type 1 extreme value distribution the *smoothed max function* since you should also prove that

$$\lim_{\sigma \downarrow 0} E\left\{max_{d \in D(x)}[u(x,d) + \varepsilon(d)]\right\} = \max_{d \in D(x)}[u(x,d)]. \tag{29}$$

answer Since the scaling parameter σ is proportional to the standard deviation of a Type 1 extreme value random variable, then clearly as $\sigma \to 0$ there are no unobserved random factors affecting choice. We can express the scaling parameter σ as a scaled version of a "standardized" Type 1 Extreme value distribution with a scale value of $\sigma = 1$ just the same way we can express any normally distributied random variable $\tilde{X} \sim N(\mu, \sigma)$ in terms of a standard normal random variable $\tilde{Z} \sim N(0,1)$ by writing $\tilde{X} = \mu + \sigma\tilde{Z}$. So now let $\{\varepsilon(d)|d \in D(x)\}$ denote a family of IID Type 1 extreme value random variables each with $\mu = 0$ and $\sigma = 1$. Then we can write the expected maximum utiity when the scale of the extreme value shocks are $\sigma \neq 1$ as follows $E\{\max_{d \in D(x)}[u(x,d) + \sigma\varepsilon(d)]\}$. Then it is quite obvious that we have

$$\lim_{\sigma \downarrow 0} E\left\{max_{d \in D(x)}[u(x,d) + \sigma\varepsilon(d)]\right\} = \max_{d \in D(x)}[u(x,d)]. \tag{30}$$

For this reason we can call $E\{\max_{d \in D(x)}[u(x,d) + \sigma\varepsilon(d)]\}$ the *smoothed max function* because by the addition of small idiosyncratic shocks $\sigma\varepsilon(d)$ we are getting something very close to the max function, except that the idiosyncratic shocks smooth out the kinks in the deterministic max function, $\max_{d \in D(x)}[u(x,d)]$ treated now as a function of x. The kinks occur at the points x where the consumer is indifferent between two or more alternatives, so there will ordinarily be a kink in the maximum utility treated as a function of x. But using the Williams-Daly-Zachary Theorem it is not hard to show that the smooth max function will be continuously differentiable in x if each of the $u(x,d)$ functions is continuously differentiable in x. In fact, using the Williams-Daly-Zachary Theorem you can show that

$$\frac{\partial}{\partial x} E\left\{\max_{d \in D(x)}[u(x,d) + \sigma\varepsilon(d)]\right\} = \sum_{d \in D(x)} P(d|x) \frac{\partial}{\partial x} u(x,d). \tag{31}$$

Now consider the MNL model with the MNL probabilities given in equation (28). Let $\overline{u}(x) = \max_{d \in D(x)} u(x,d)$. Then it is easy to see that we can rewrite the MNL choice probabilities in a mathematically equivalent but much better way for doing calculations on the computer, since no matter how large the utilities $u(x,d)$, this latter mathematically equivalent expression will only experience problems of *underflow* on the computer which is far less dangerous of a problem than *overflow* when there are values of $u(x,d)/\sigma$ that are too large to be accurately calculated as $\exp\{u(x,d)/\sigma\}$ on digital computers.

$$\begin{aligned} P(d|x) &= \frac{\exp\{u(x,d)/\sigma\}}{\sum_{d' \in D(x)} \exp\{u(x,d')/\sigma\}} \\ &= \frac{\exp\{[u(x,d) - \overline{u}(x)]/\sigma\}}{\sum_{d' \in D(x)} \exp\{[u(x,d') - \overline{u}(x)]/\sigma\}}. \end{aligned} \tag{32}$$

Notice that in the numerically stable evaluation of the MNL choice probability in the last equation of (32), the expressions entering the exponentials are always by construction *negative numbers.* A problem of underflow occurs when you ask a computer to evaluate the exponential of a large enough negative number such as $\exp\{-1000000000000\}$. The computer cannot represent such a small number using only 64 bits of precision, and thus the computer will treat this as *machine zero* since it is smaller than the smallest positive real number that can be represented as a binary real number with 64 bits.

Now let's go back to the original question. We can now see, using (32) why the limiting value of the softmax function (or what we call the logit choice probabilities) equals $1/n$ where n is the number of alternatives that equal the highest utility $\overline{u}(x)$ as $\sigma \downarrow 0$. The reason is that from equation (32) if d is any such alternative that results in the maximum utility $\overline{u}(x) = \max_{d\in D(x)} u(x,d)$, the numerator will be 0, but the denominator will equal n, where n is the total number of alternatives $d' \in D(x)$ that equal the maximized utility. If there is only 1 such alternative, then $n = 1$ and then the limit of the MNL choice probability as $\sigma \downarrow 0$ is equal to 1. The interpretation is clear: when there is one utility maximizing choice, the choice proability converges to 1 if d equals that utiity maximizing choice and 0 otherwise when $\sigma \downarrow 0$. So the softmax function could also be called the *smoothed argmax function* but that is not the terminology that people in the machine learning literature use. For some strange reason they refer to the logit formula as the "softmax". Go figure.

G. Write down the Axiom of Independence from Irrelevant Alternatives and show that the MNL model satisfies this axiom. Provide an example of a random utility model that does not satisfy the IIA axiom. Is the IIA axiom "reasonable" and consistent with "rational choice" or does it imply unrealistic restrictions on choice probabilities?

answer This axiom was introduced by the mathematical psychologist R Duncan Luce in his 1959 book, *Individual Choice Behavior: A Theoretical Analysis.* There are different ways to state the axiom but the point of departure is *probabilistic choice theory* that assumes for various pyschological reasons individuals choosing from some finite choice set $D(x)$ (which may depend on observed covariates x) behave probabilistically, so the same person may make different choices from the same choice set at different points in time due to unobservable pscyhological factors that affect their choice. Thus, the mathematical psychologists take choice probabilities $P(d|D(x))$, which denotes the probability a subject chooses alternative d from a finite choice set $D(x)$, as "givens". Luce then made a key assumption or axiom that he expected choice probabilities to behave in order to get deeper insight into the "representation" of these choice probabilities. Luce's axiom of *Independence from Irrelevant Alternatives* (IIA) states roughly that the odds of choosing one alternative $d \in D(x)$ over another alternative $d' \in D(x)$ is independent of the other items in the choice set $D(x)$. That is, the odds of choosing d' over d is independent of how many or the characteristics of any other alternative $\delta \in D(x)$. In particular, if $E(x)$ is some other expanded choice set that contains $D(x)$, so we have $D(x) \subset E(x)$, Luce's IIA axiom states that if alternatives d and d' are elements of $D(x)$ (and therefore elements of $E(x)$) we have

$$\frac{P(d'|D(x))}{P(d|D(x))} = \frac{P(d'|E(x))}{P(d|E(x))}. \tag{33}$$

An alternative way to state Luce's Axiom is to let $P(D(x)|E(x))$ denote the probability that an individual chooses some alternative d in the choice set $D(x)$ given that the overall choice set id

$E(x)$. The alternative way to state IIA is this

$$P(d|E(x)) = P(d|D(x))P(D(x)|E(x)), \tag{34}$$

and it is easy to see that the version of the IIA axiom in equation (34) implies the version in equation (33). Luce's Theorem 3 establishes that if the IIA axiom holds there are what pyschologists call *ratio scales* $u(d,x)$ (or what economists call "utilities") such that the choice probability has the following form

$$P(d|E(x)) = \frac{\exp\{u(d,x)\}}{\sum_{d' \in D(x)} \exp\{u(d',x)\}}, \tag{35}$$

which you will recognize as the multinomial logit model.

This is the end of the "official answers" for this problem set. Below I go a bit further in the answers for your benefit, but of course I did not expect you to provide such a detailed answer! But to give you more perspective into the subject I strongly recommend reading further below.

Luce explained the motivation for his IIA axiom as follows

> When a person chooses among alternatives, very often their responses appear to be governed by probabilities that are conditioned on the choice set. But ordinary probability theory with its standard definition of conditional probability does not seem to be quite what is needed. An example illustrates the difficulty. When deciding how to travel from home to another city, your choice may be by airplane (a), bus (b), or car (c). Let A,B,C denote the uncertain states of nature associated with each form of travel. Note that if one elects c all of the uncertainties of A and B remain because planes fly and busses run whether or not you are on them. However, if you elect either a or b , then your car remains in the garage and the set C is radically altered from when the car is driven. So there really is no universal event underlying the sources of uncertainty. The choice axiom of chapter 1 was introduced as a first attempt to construct a probability-like theory of choice that by-passed the fixed, universal sample space assumption.

McFadden derived the MNL model via a different strategy, i.e. as a *random utility model* where a utility maximizing individual chooses an alternative d from a finite choice set $D(x)$ where the choice of any individual depends on variables x which the econometrician can observe, but also other factors that the econometrician cannot observe. The inability to observe these other factors can make the individual's choice appear probabilistic from the standpoint of the econometrician even if from an individual's perspective their choices are perfectly deterministic and not *individually probabilistic* as mathematical psychologists such as Luce presumed. In his pioneering article "Conditional Logit Analysis of Qualitative Choice Behavior" (P. Zarembka, ed., *Frontiers in Econometrics* (New York: Academic Press, 1973), 105–42) McFadden noted that

> A fundamental concern of economics is understanding human choice behavior. Models or hypotheses are formed on the nature of decision processes, and are evaluated in light of observed behavior. This task is complicated because the econometrician cannot observe or control all the factors influencing behavior, and because the process of observation itself influences acts of the decision maker through the vehicle of experience. It becomes necessary to make inferences on a model of *individual* choice behavior by sampling from a *population* of individuals (or sampling from a population of 'experience levels' for a single individual). When the model

> of choice behavior under examination depends on unobserved characteristics in the population, the testable implications of the individual choice model are obscured. However it is possible deduce from the individual choice model properties of population choice behavior which can be subjected to empirical test.

So the key contributions of McFadden's 1973 paper are twofold: First, he showed how the tractable MNL could be derived as a random utility model under the assumptions of additively separable unobservable shocks affecting choices (AS), and the assumption that these unobservable shocks ε have a multivariate *IID* extreme value distribution (EV). The statistical independence of the components $\varepsilon(d)$ of the random vector ε imply the IIA property that Luce identified. The second big contribution of this article was to show how MNL models could be estimated econometrically by the method of maximum likeihood. Though back in 1973 it was hard to do the nonlinear style of maximum likelihood estimation since computers were not very powerful then, the method eventually took off and lead to huge numbers of empirical applications of the logit model of discrete choice. As of January 2022, McFadden's 1973 article has over 22,000 citations according to Google Scholar.

However while the IIA property (and the indpendent Type 1 extreme value utility shocks) result in a very nice close-form expression for choice probabilties, this axiom (and the predictions for choice probabilities it implies) has some undesirable side effects as Gerard Debreu noted in his 1960 review of Luce's book in the *American Economic Review.* Though Debreu's original example concerned a hypothetical choice "thought experiment" involving the choice of listening to 3 different pieces of recorded music (Debussy versus Beethoven's 8th symphony, but the latter played by two different orchestras for which the listener was presumed to be indifferent), Debreu's example has since become known as the *red bus/blue bus paradox* that McFadden provided on page 113 of his article. McFadden states the paradox this way

> The primary limitation of the model is that the independence of irrelevant alternatives axiom is implausible for alternative sets containing choices that are close substitutes. An example illustrates this point. Suppose a population faces the alternatives of transportation by auto and by bus, and two thirds choose to use auto. Now suppose a second 'brand' of bus travel is introduced that is in all essential respects the same as the first. Intuitively, two thirds of the population will still choose auto, and the remainder will be split between the bus alternatives. However if the selection probabilities satisfy the Independence from Irrelevant Alternatives Axiom, only half the population will choose auto when the second bus is introduced. The reason this is counter-intuitive is that we expect individuals to lump the two bus alternatives together in making the auto-bus choice. This example suggests that the application of the model should be limited to situations where the alternatives can plausibly be assumed to be distinct and weighed independently in the eyes of the decision maker.

Make sure you understand why the introduction of the 2nd bus alternative results in the counter-intuitive counterfactual prediction fom the MNL model. Note in the first case, where there are only 2 alternatives bus and auto with utilities v_b and v_a, respectively, if the IIA axiom holds, the choice probability of $2/3$ for the auto implies that the utility difference $v_b - v_a$ is given by $\log(1/2)$ (show this by inverting the logit probability for the choice of auto, which McFadden assumed equals $2/3$, and calculate the utility difference $v_b - v_a$ that implies a $2/3$ choice probability and show it equals $v_b - v_a = \log(1/2)$). Now for the counterfactual of introducing a 2nd bus alternative that is in all *observable* respects identical to the existing bus, this means the utility of the new bus alternative, call it v_{nb} equals the existing bus utility, so $v_{nb} = v_b$. This implies a counterfactual choice probability for choosing auto relative to the two bus alternatives will be

$$P(a|\{a,b,nb\}) = \frac{\exp\{v_a\}}{\exp\{v_a\}+\exp\{v_b\}+\exp\{v_{nb}\}} = \frac{1}{1+2\exp\{v_b-v_a\}} = \frac{1}{2} \tag{36}$$

as McFadden claimed in his example. Here the solution to the paradox comes from the interpretation of the random utility model. Even though the *observable* characteristics of the two bus alternatives is the same (which implies $v_b = v_{nb}$), the independence in the error terms in the random utility model imply that there is still a significant *idiosyncratic component* affecting choice of alternatives that is not identical for the two bus alternatives and this implies the counterintuitive counterfactual prediction of the MNL model.

One solution to this problem is to look for other multivariate distributions for unobservables that imply that when *observable characteristics of two alternatives are identical, so also will be the unobservable components of random utility as well.* One way this can be done is via *random coefficients.* Suppose now that x is a vector of the observed characteristics of different alternatives. For example in the case of the auto versus bus example, x contains observed covariates measuring travel time of car versus bus, the waiting time to use a car versus waiting for a bus, the comfort and privacy of a car versus bus, the travel cost of car versus bus, and so forth. Then $x(d)$ can denote the characteristics of choice d in the choice set $D(x)$. Let $u(x,d) = x(d)\tilde{\beta}$ be the utility a consumer gets from choosing alternative $d \in D(x)$, where $\tilde{\beta}$ are *random coefficients* i.e. random weights that different consumers put on the various attributed in $x(d)$ when determining how much they like or dislike different alternatives. For example a poor person might put a large weight on travel cost, whereas a rich person might put more weight on convenience and comfort in their transportation experience. Now let $\beta = E\{\tilde{\beta}\}$ denote the population mean of the random coefficents and let $\Omega = E\{(\tilde{\beta}-\beta)(\tilde{\beta}-\beta)'\}$ be the variance-covariance matrix fo the random coefficients. A popular distribution for the random coefficients is multivariate normal, so $\tilde{\beta} \sim N(\beta,\Omega)$. In any event, we can represent the random coefficients discrete choice model as an additive random utility model by defining the unobserved utility components $\varepsilon(d)$ as follows

$$u(x,d)+\varepsilon(d) \equiv x(d)\beta+\varepsilon(d) = x(d)\beta + x(d)(\tilde{\beta}-\beta). \tag{37}$$

Will this model satisfy the IIA property? The answer is no, because the random utility model implies that the unobservable terms $\varepsilon(d)$ and $\varepsilon(d')$ will remain independently distributed even if the observable characteristics of the two alternatives are identical, whereas the random coefficients model implies that the error terms will be perfectly correlated with each other if the observed characteristics are identical, $x(d) = x(d')$. So this implies that the consumer is indifferent between the two bus alternatives d and d' if $x(d) = x(d')$ regardless of what random weights $\tilde{\beta}$ they put on the attributes of these two bus alternatives. Thus the random coefficients model does not exhibit the IIA property and does not suffer from the counterintuitive counterfactual predictions of McFadden's "red-bus/blue-bus" example.

However a big drawback of random coefficient models is that they generally do not have the simple closed-form expressions for choice probabilities that MNL models enjoy. McFadden did not rest by just observing the limitations of the MNL model. He defined a new larger class of multivariate extreme value random variables that includes *IID* Type 1 extreme value random variables as a subclass, for which simple closed-form expressions for choice probabilities continue to hold, but which relax the restrictive independence property of the MNL model. This class is called *Generalized Extreme Value* (GEV) and I will briefly define the class below and show a prominent subclass known as *nested multinomial logit models* (NMNL).

We assume that the multivariate distrubtion of ε is generalized extreme value family and thus has joint CDF of the form

$$F(\varepsilon_1,\ldots,\varepsilon_n) = \exp\{-G(e^{-\varepsilon_1},\ldots,e^{-\varepsilon_n})\} \tag{38}$$

where the function G maps the positive orthant of R^n into the positive real line, and satisfies the properties that guarantee F is a CDF and ensure the logical consistency of the corresponding random utility model in McFadden 1978, including

1. **homogeneity of degree** $1/\sigma$ for some $\sigma > 0$, i.e. for any constant $\lambda > 0$ and any $y = (y_1,\dots,y_n)$ satisfying $y \geq 0$ we have
$$G(\lambda y_1,\dots,\lambda y_n) = \lambda^{\frac{1}{\sigma}} G(y_1,\dots,y_n) \tag{39}$$

2. **alternating sign property** for any distinct indices $(i_1,\dots,i_k)$ where each index $i_j \in \{1,\dots,n\}$, we have
$$\begin{aligned} \frac{\partial^k G}{\partial y_{i_1},\dots,\partial y_{i_k}} &\geq 0 \qquad \text{if k is odd} \\ \frac{\partial^k G}{\partial y_{i_1},\dots,\partial y_{i_k}} &\leq 0 \qquad \text{if k is even.} \end{aligned} \tag{40}$$

3. **unboundedness** For each $i \in \{1,\dots,n\}$ we have
$$\lim_{y_i \to \infty} G(y_1,\dots,y_{i-1},y_i,y_{i+1},\dots,y_n) = \infty \tag{41}$$

Theorem *The function given in equation (38) defines a valid multivariate extreme value CDF if the generating function G satisfies properties 1, 2 and 3 above.*

It is not extremely hard to prove this result. The alternating sign property is required to ensure the montonicity of the CDF in all of its arguments. The alternating sign property implies in particular that G is non-decreasing in each of its arguments, and for higher order mixed partial derivatives, the alternating sign property is sufficient to establish that all mixed partial derivatives of $F(\varepsilon_1,\dots,\varepsilon_n)$ with respect its arguments is non-negative, which any valid CDF must satisfy. The unboundedness property implies that $\lim_{\varepsilon_i \to -\infty} F(\varepsilon_1,\dots,\varepsilon_{i-1},\varepsilon_i,\varepsilon_{i+1},\dots,\varepsilon_n) = 0$ which is also a requirement to be a valid CDF. Finally, the homogeneity property implies that $G(0,\dots,0) = 0$, which implies the other key property of a CDF, namely
$$\lim_{\varepsilon \to \infty} F(\varepsilon) = 1. \tag{42}$$

Thus, McFadden introduced a new class of multivariate CDFs F that include the standard multivariate Type 1 extreme value distribution as a special case. Appropriately, he called this the *Generalized Extreme Value family* (GEV) of distributions though unfortunately *Wikipedia* refers to the GEV family as a different class of *univariate* extreme value distributions, so to distinguish, the class of multivariate CDFs that McFadden introduced are also called *Multivariate Generalized Extreme Value family* or (MGEV). Note that the CDF of a collection of *IID* Type 1 extreme value random variables is a special case of equation (38) when $G(y_1,\dots,y_n) = \sum_{i=1}^n y_i$. This case corresponds to the CDF in equation (4) in the case where the scale parameter $\sigma = 1$ and all the location parameters $\mu + i = 0$. If G is given by $G(y_1,\dots,y_n) = \sum_{i=1}^n [y_i]^{\frac{1}{\sigma}}$ we get exactly the CDF in equation (4) when we allow non-zero location parameters for the marginal distributions of $F(\varepsilon)$, i.e. the CDF defined by
$$F(\varepsilon_1,\dots,\varepsilon_n) = \exp\{-G(e^{-(\varepsilon_1-\mu_1)},\dots,e^{-(\varepsilon_n-\mu_n)})\} \tag{43}$$

coindcides with the CDF (4).

McFadden's objective in introducing the MGEV class of distributions was that by specifying different G functions he could capture forms of dependence in the components ε_d vector that violate the IIA property while still preserving the nice closed-form solution properties of the Type 1 Extreme value function for the choice probabilities. To see this, note that the components of the random variable $\tilde{\varepsilon}$ will not be mutually independent of each other unless $G(y_1,\ldots,y_n)$ is an additive function with a representation such as

$$G(y_1,\ldots,y_n) = \sum_{i=1}^{n} f_i(y_i) \tag{44}$$

such as the case discussed above where $f_i(y_i) = [y_i]^{\frac{1}{\sigma}}$. Below we will discuss examples of G functions that are not additive and thus exhibit dependence in the components of $\tilde{\varepsilon}$, so the IIA property no longer strictly holds.

But before we do that, let's show that we still retain the nice closed-form expression properties for the choice probabilities implied by random utility models using the MGEV family of distributions. Recall that we did this for the "base-case" of independent Type 1 extreme valued random variables in equation (4) by showing this family is "max-stable" as we did in the answer to part B above. We can repeat this same argument to show that the MGEV family of distributions is max-stable: i.e. the maximum of a random vector that has an MGEV distribution has a Type 1 extreme value distribution. So to test your understanding, consider the random utility model where the decision rule is

$$d(x,\varepsilon) = \max_{d\in D(x)} [u(x,d)+\varepsilon(d)] \tag{45}$$

but now $\varepsilon = \{\varepsilon(d)|d \in D(x)\}$ has a MGEV distribution for some generating function G, where we assume that the location parameters of each of the component extreme value random variables $\varepsilon(d)$ are normalized so that $E\{\varepsilon(d)\} = 0$. For notational simplicity, let's renumber the elements of $D(x)$ so it is equivalent tot he set of integers, $D(x) = \{1,\ldots,n\}$. It follows that $u(x,d)+\varepsilon(d)$ has an extreme value distribution with location parameter $u(x,d)$ and by the max-stability property the random variable $\tilde{Z} = \max_{d\in D(x)}[u(x,d)+\varepsilon(d)]$ also has a Type 1 extreme value distribution given by

$$\begin{aligned}
F_{\tilde{Z}}(z) &= Pr\left\{\tilde{Z} \leq z\right\} \\
&= F(z,z,\ldots,z) \\
&= \exp\left\{-G\left(e^{-(z-u(x,1))},\ldots,e^{-(z-u(x,n))}\right)\right\} \\
&= \exp\{-\exp\{-(z-I)/\sigma\}\}
\end{aligned} \tag{46}$$

where the "inclusive value" I in this case is given by

$$I\left(\{u(x,d)|d \in D(x)\}\right) = \sigma \log\left(G(e^{u(x,1)},\ldots,e^{u(x,n)})\right). \tag{47}$$

It follows that the maximized utility $\tilde{Z}$ is a univariate Type 1 extreme value random variable with expected value equal to

$$E\left\{\max_{d\in D(x)}[u(x,d+\varepsilon(d)]\right\} = \sigma \log\left(G(e^{u(x,1)},\ldots,e^{u(x,n)})\right) + \gamma\sigma. \tag{48}$$

Now to derive the formula for the conditional choice probabilities, $P(d|x)$ we apply the Williams-Daly-Zachary Theorem to get

$$\begin{aligned} P(d|x) &= \frac{\partial}{\partial u(x,d)} E\left\{ \max_{d \in D(x)} [u(x,d+\varepsilon(d)]\right\} \\ &= \frac{\partial}{\partial u(x,d)} \left[\sigma \log\left(G(e^{u(x,1)},\dots,e^{u(x,n)})\right) + \gamma\sigma\right] \\ &= \frac{\sigma e^{u(x,d)} G_d\left(e^{u(x,1)},\dots,e^{u(x,n)}\right)}{G\left(e^{u(x,1)},\dots,e^{u(x,n)}\right)} \end{aligned} \tag{49}$$

where $G_r(y_1,\dots,y_n)$ is a short hand for the partial derivative of the generating function G with respect to y_r

$$G_r(y_1,\dots,y_n) \equiv \frac{\partial}{\partial y_r} G(y_1,\dots,y_n). \tag{50}$$

To test your understanding, calculate the choice probability in the case where $G(y_1,\dots,y_n) = \sum_{i=1}^n y_i^{\frac{1}{\sigma}}$ and show that the resulting choice probability reduces to the standard MNL formula but extended to the case where the scale parameter σ is present, as in formula (28) in the answer to part E above. So in particular, is the IIA property satisfied in this case?

To further test your understanding, find the formula for the choice probabilities for the MGEV distribution with this generating function: $G(y_1,\dots,y_n) = \left[\sum_{i=1}^n y_i^{\frac{1}{\sigma}}\right]^\sigma$. Are the components of the vector $\varepsilon = (\varepsilon_1,\dots,\varepsilon_n)$ still independently distributed in this case? If so, or if not, does the IIA property continue to hold? Derive the choice probabilities and compare them to the case above. Compare and contrast the two cases in terms of the implied choice behavior. In the first case, in the limit as $\sigma \downarrow 0$ what happens to the variance of maximized utility, i.e. what is $\lim_{\sigma\downarrow 0} \mathrm{var}(\tilde{Z})$ where $\tilde{Z} = \max_{d\in D(x)}[u(x,d)+\varepsilon(d)]$ where $D(x) = \{1,\dots,n\}$? Now consider the second case where the generating function G is given by $G(y_1,\dots,y_n) = \left[\sum_{i=1}^n y_i^{\frac{1}{\sigma}}\right]^\sigma$. Calculate the limiting variance of maximized utility $\tilde{Z}$ in this case as $\sigma \downarrow 0$. Is it zero or positive? Note the difference in the two cases in the G functions: in the first case G is homogeneous of degree $\frac{1}{\sigma}$ and this goes to infinity as $\sigma \downarrow 0$ whereas in the second case G is homogenous of degree 1 for all $\sigma \geq 0$ (including in the limit when $\sigma = 0$ where you can show that $G(y_1,\dots,y_n) = \max(y_1,\dots,y_n)$.

So with some thought, you can see the differences in these two cases: in the first case where G is additively separable and is homogeneous of degree $\frac{1}{\sigma}$ we have independence in the components of the $\varepsilon = (\varepsilon_1,\dots,\varepsilon_n)$ shocks, they are all Type 1 exreme value with scale σ and IIA holds, but as $\sigma \downarrow 0$ the idiosynractic shocks disappear and the choice probability converges to a degenrate probability that puts probability $1/K$ on the K alternatives that have the highest "strict utilities" $u(x,d)$ (thus allowing for potential ties) but if there is only 1 alternative with the highest strict utility $u(x,d)$ then the limiting choice probability is a unit mass on the alternative that generates the highess strict utility. So $\tilde{Z}$ converges to a constant in this case, i.e. $\lim_{\sigma\downarrow 0} \tilde{Z} = \max_{d\in D(x)}[u(x,d)]$ with probability 1.

But in the second case G is linearly homogeneous and we no longer have independence in the components of $\varepsilon = (\varepsilon_1,\dots,\varepsilon_n)$: there is some sort of "common shock" affecting all of them. For this reason, the disribution of maximum utility $\tilde{Z}$ is *non-degenerate* (and thus has positive limiting variance) and in fact the variance of $\tilde{Z}$ equals $\frac{\pi^2}{6}$ for any $\sigma \geq 0$ including $\sigma = 0$. On the other hand there appear to be "idiosyncratic components" in $\varepsilon = (\varepsilon_1,\dots,\varepsilon_n)$, and the variance of the idiosyncratic components of ε_i appeaers to be going to zero as $\sigma \downarrow 0$ in the sense that the agent's choices are given by exactly the *same MNL choice probability* as in the IIA case. So what is going on here?

Some insight comes from considering the possibility that the components of $\varepsilon = (\varepsilon_1,\dots,\varepsilon_n)$ have a *variance components structure* i.e. the possibility that $\varepsilon_i = \tilde{z} + \sigma\eta_i$ where $\eta = (\eta_1,\dots,\eta_n)$ are *IID standardized* Type 1 extreme value random variables (i.e. the scale parameter is standardized to 1, and the common component $\tilde{z}$ has some distribution, though not yet clear what it might be (or if it even exists). Recall we said that extreme value random variables are *max-stable* but not *stable* — i.e. the maximum of a collection of extreme value random variables is another extreme value random variable but sums of extreme value random variables (even if they are independently distributed) are generally *not* extreme value distributed random variables. So if this component $\tilde{z}$ does exist, it would have to have some probability distribution *different* from the extreme value distribution. In fact Cardell in his 1997 paper "Variance Components Structures for the Extreme-Value and Logistic Distributions with Application to Models of Heterogeneity" (*Econometric Theory*), proved that there is a random variable $\tilde{z}$ with a distribution he calls the $C(\sigma)$ distribution such that the following result holds:

Theorem 2.1 (Cardell) *If* $\tilde{\eta}$ *is distributed as a Type 1 Extreme value random variable with scale parameter* 1*, there is a random variable* $\tilde{z}$ *that has a distribution* $C(\sigma)$ *to be defined below that is independently distributed from* η *such that the random variable* $\tilde{\varepsilon}$ *given by*

$$\tilde{\varepsilon} = \tilde{z} + \sigma\tilde{\eta} \tag{51}$$

is also distributed Type 1 extreme value with scale parameter 1*. The probability density function for* $\tilde{z}$ *is given by*

$$f_\sigma(z) = \left[\frac{1}{\sigma}\right]\sum_{i=0}^{\infty}\left[\frac{(-1)^i e^{-iz}}{i!\Gamma(-\sigma i)}\right]. \tag{52}$$

Note that by definition a $C(0)$ random variable is an ordinary Type 1 extreme value random variable.

How is Theorem 2.1 by Cardell relevant to thinking about the two cases above, one where the G function is homogeneous of degree $\frac{1}{\sigma}$ and the other where it is homogeneous of degree 1? Well, we showed in the latter case there is variability among the components of $\varepsilon = (\varepsilon_1,\dots,\varepsilon_n)$ that does not disappear as $\sigma \downarrow 0$ whereas in the former case, all uncertainty disappears as $\sigma \downarrow 0$ and each component ε_i converges in distribution to a degenerate random variable with zero variance (i.e. a constant), equal to the location parameter μ or 0 if the location parameter is zero. This must mean that in the case where G is linearly homogeneous, the shocks ε_i have the following variance-components representation

$$\varepsilon_i = \tilde{z} + \sigma\eta_i, \quad i = 1,\dots,n \tag{53}$$

where η_i is a standardized (i.e. has scale normalized to 1) Type extreme value distribution and the common component (common utility shock) $\tilde{z}$ has a $C(\sigma)$ distribution. Given this, consider the maximized utility in this case

$$\tilde{Z} = \max_{d\in D(x)}[u(x,d)+\varepsilon_d] = \max_{d\in D(x)}[u(x,d)+\tilde{z}+\sigma\eta_d] = \tilde{z} + \max_{d\in D(x)}[u(x,d)+\sigma\eta_d]. \tag{54}$$

Now we can see that the "common shock" $\tilde{z}$ comes out of the maximization (since it is common to all alternatives) and this common shock can survive even as $\sigma \downarrow 0$. Thus the *idiosyncratic shocks* specific to each alternative $\sigma\eta_i$ converge in distribution to 0 as $\sigma \downarrow 0$ but the common component $\tilde{z}$ survives and converges in distribution to a $C(0)$ random variable, which happens to have a standardized Type 1 extreme value distribution. This is formalized in the following Theorem of Cardell (1997)

Theorem 4.1 *For any $\sigma \in (0,1)$ let $\tilde{z}$ be distributed as a $C(\sigma)$ random variable and independent of the random vector $\eta = (\eta_1,\dots,\eta_n)$ which are IID type 1 extreme value random variables with location standardized to 0 and scale parameter* 1. *Let $\varepsilon_i = \tilde{z} + \sigma\eta_i$, $i = 1,\dots,n$. Then the CDF of $\varepsilon = (\varepsilon_1,\dots,\varepsilon_n)$ is given by*

$$F_\sigma(\varepsilon_1,\dots,\varepsilon_n) = \exp\{-G_\sigma(e^{-\varepsilon_1},\dots,e^{-\varepsilon_n})\}, \tag{55}$$

where the generating function G_σ is given by

$$G_\sigma(y_1,\dots,y_n) = \left[\sum_{i=1}^{n}[y_i]^{\frac{1}{\sigma}}\right]^{\sigma}. \tag{56}$$

While something of a detour, Cardell's "variance components" representation for a subclass of MGEV random variables that result in *nested logit* models (NMNL) provides very helpful intuition for the type of correlation in the unobservable components $\varepsilon = (\varepsilon_1,\dots,\varepsilon_n)$ the nested logit model specification implies. We will describe NMNL models in more detail below but first, let's return to the derivation of choice probabilities in the general MGEV case.

Note that we can further simplify the expression for the choice probability $P(d|x)$ implied by the MGEV distribution of ε in equation (49) but appealing to Euler's Theorem for homogeneous functions

$$\sum_{r=1}^{n} y_r G_r(y_1,\dots,y_n) = \frac{1}{\sigma}G(y_1,\dots,y_n), \tag{57}$$

so using equation (57) we can re-write the last equation of (49) as follows

$$\begin{aligned} P(d|x) &= \frac{\sigma e^{u(x,d)} G_r\left(e^{u(x,1)},\dots,e^{u(x,n)}\right)}{G\left(e^{u(x,1)},\dots,e^{u(x,n)}\right)} \\ &= \frac{e^{u(x,d)} G_d\left(e^{u(x,1)},\dots,e^{u(x,n)}\right)}{\sum_{r=1}^{n} e^{u(x,r)} G_r\left(e^{u(x,1)},\dots,e^{u(x,n)}\right)} \\ &= \frac{\exp\left\{u(x,d) + \log\left(G_d\left(e^{u(x,1)},\dots,e^{u(x,n)}\right)\right)\right\}}{\sum_{r=1}^{n} \exp\left\{u(x,r) + \log\left(G_r\left(e^{u(x,1)},\dots,e^{u(x,n)}\right)\right)\right\}}, \end{aligned} \tag{58}$$

which looks very similar to the familiar MNL formula (1). To test your understanding, verify that when $G(y_1,\dots,y_n) = \sum_{i=1}^{n}[y_i]^{\frac{1}{\sigma}}$ the alternative logit-like formula for $P(d|x)$ under the general MGEV distribution reduces to the σ=scaled version of the MNL choice probability formula in equation (28).

The class of *nested logit models* is a subclass of the MGEV family for particular generating functions G that have a hierarchical structure that can be explained via a graphical *choice tree.* Let's take a particular example where large choice trees have a natural interpretation: the choice of cars. Here we follow the paper of Iskhakov, Gillingham, Munk-Nielsen, Rust and Scherning (2022) "Equilibrium Trade in Automobiles" (forthcoming *Journal of Political Economy*). Nested logit models have been commonly used by other authors for modeling the automobile market: see the citations to the work of Penelopi Goldberg and James Berkovec in the paper above. In the case of the Iskhakov *et. al.* paper the model is one of *dynamic discrete choice* whereas previous auto choice models have been framed in a static context. In the Iskhakov *et. al.* paper the vector x denotes the *car ownership state* of the household and they use the notation $x = (i,a)$ to denote a household who owns a car of type i (e.g. i could index a particular make/model such as Volvo XC90, etc) and a denotes the age of the car. The choice set $D(x)$ includes the options to 1) "purge" the current car and enter the no-car state (or "outside good" choice), denoted by $d = \emptyset$, or 2) keep the current car, denoted by $d = \kappa$, or 3) trade the car (i,a) for some other car (j,d) where i is the type and d is the age of the car that the consumer chooses to trade their current car for. (Sorry for the overloaded notation here: up to now d indexed a decision made by the individual, but now d indexes the age of the car the individual chooses. So in what follows below, now consider d to index only part of the household's decision, i.e. the decision on what age car to trade the current car aged a for).

Even though in the dynamic version of the auto problem dynamic programming must be used to describe the strategy of households on whether to purge, keep or trade their car in every time period, Iskhakov *et. al.* show that under certain assumptions on the pattern of temporal dependence in the uobservable shocks that this DP problem, namely that the vector of unobservable MGEV shocks at time t, ε_t, is independent of the vector of shocks in any other time period $s \neq t$, ε_s, the DP problem results in *choice specific value functions* $v(x,d)$ that superficially look very similar to static utility functions, $u(x,d)$. Here we have reverted to the notation where x is the state of the household and d indexes the decision the household makes, but in the extended notation of Ishkakov *et. al.* if a household owns a car $x = (i,a)$ and chooses to trade for a car (j,d) the value or discounted utility for this choice is given by $v(i,a,j,d)$. Similarly if the household chooses to keep the car the value of this is $v(i,a,\kappa)$ and if the household chooses to purge the car (sell it but not buy another one, so the household enters the no-car state) the value of this decision is $v(i,a,\phi)$.

The actual choice of a household who is in state $x = (i,a)$ also depends on idiosyncratic shocks (one for each possible alternative choice) just as in a static discrete choice model, so we can write the optimal dynamic choice rule (or trading strategy) for the household as a function $\delta(x,\varepsilon) = \delta(i,a,\varepsilon)$ given by

$$\delta(i,a,\varepsilon) = argmax \left[v(i,a,\phi) + \varepsilon_{\phi}, v(i,a,\kappa) + \varepsilon_{\kappa}, \underset{j,d}{argmax}[v(i,a,j,d) + \varepsilon_{j,d}] \right] \tag{59}$$

where ε_{ϕ} is an idiosyncratic shock associated with choosing the "outside good" (i.e. to purge the current car and have no car next period), ε_{κ} is the unobserved component of utility associated with choosing to keep the current car (i,a), and finally the last set of choices involving trading the current car for some other car of type j and age d so $\varepsilon_{j,d}$ is the unobserved utility shock associated with that choice. Recall that "*argmax*" is shorthand for "argument that maximizes" so $\delta(i,a,\varepsilon)$ is just the index of the household's optimal decision on the current car: whether to purge it, keep it, or trade if for some other make or model.

Thus at each time t the household has an observable car ownership state and also experiences idiosyncratic shocks ε_t that affect the household's choices. In the auto problem, we index the arguments of the joint CDF of this vector of idiosyncratic shocs as $\varepsilon \equiv (\varepsilon_{\kappa}, \varepsilon_{\phi}, \{\varepsilon_{j,d}\})$. Thus, $F(\varepsilon_{\phi}, \varepsilon_{\kappa}, \{\varepsilon_{d,j}\})$ is the CDF of these preference shocks and we assume it has the representation as an MGEV distribution

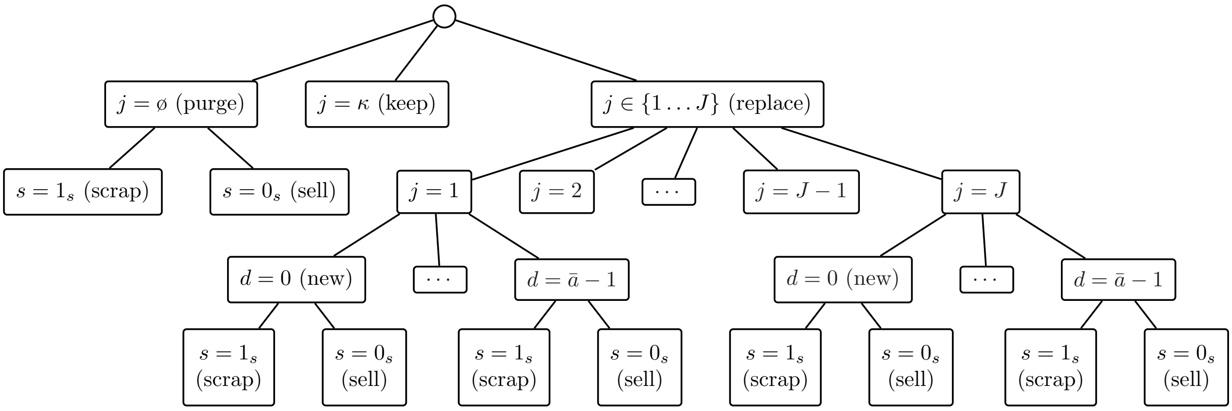
$$F(\varepsilon_{\phi}, \varepsilon_{\kappa}, \{\varepsilon_{j,d}\}) = \exp\{-G(e^{-\varepsilon_{\phi}}, e^{-\varepsilon_{\kappa}}, \{e^{-\varepsilon_{j,d}}\})\} \tag{60}$$

where the function $G(y_{\phi}, y_{\kappa}, \{y_{j,d}\})$ is given by

$$G(y_{\phi}, y_{\kappa}, \{y_{j,d}\}) = y_{\phi}^{\frac{1}{\sigma}} + y_{\kappa}{}^{\frac{1}{\sigma}} + \left(\sum_{j=1}^{J} \left(\sum_{d=0}^{\bar{a}_j - 1} y_{j,d}{}^{\frac{1}{\sigma_j}} \right)^{\frac{\sigma_j}{\sigma_{j>0}}} \right)^{\frac{\sigma_{j>0}}{\sigma}} . \tag{61}$$

where $\bar{a} - 1$ is the oldest age of car that can be held or purchased in the market (so all cars of age $\bar{a}$ or older must be scrapped) and where σ, $\sigma_{j>0}$ and $\{\sigma_j\}$, $j = 1, \ldots, J$ are $J+2$ "scale/similarity parameters" of the GEV distribution that implies the nesting structure illustrated in Figure 1. Note that in this tree there a total

Figure 1: Choice tree for nested logit specification of the GEV distribution for ε in the auto trading problem



of 4 levels but the G function notation above has omitted a choice at the lowest level of this tree, namely, for any choice involving trading the current car (i,a) for another car (j,d), there is also a binary decision s about whether to sell the car (which we denote in the figure by the choice $s=0_s$) or scrap it (which we denote by $s=1_s$). Rather than further complicate the already complicated formula for the nested logit G function above, we leave it as an exercise for you to adapt the formula above to acommodate this "lowest level" scrap versus sell decision.

Note that the parameter σ controls the level of variability of the preference shocks in the top level of the choice tree which is a choice of whether to purge, keep or trade for another car. As $\sigma\to 0$ the overall scale of the preference shocks tend to zero due to the inequalities

$$\sigma \geq \sigma_{j>0} \geq \sigma_j \geq 0 \quad j=1,\dots,J \tag{62}$$

that must be satisfied for F to be a valid CDF. If we impose the restriction that $\sigma=\sigma_{j>0}$, then the three level NMNL choice tree collapses to the two level choice tree structure similar to Figure 1 where the scale parameters preference shocks for the top level choices are the same, and we only allow different scale parameters σ_j to reflect correlation/similarity in different ages of type j of car. If we further restrict that $\sigma=\sigma_{j>0}=\sigma_j$ for $j=1,\dots,J$, then the NMNL collapses to a standard MNL model, where there is a common scale parameter σ for all preference shocks, but these shocks are *IID*. Then the IIA property holds.

But otherwise we can use the variance components interpretation from the theorems of Cardell to get intuition into the pattern of dependence among the components implied by the nested logit choice tree in figure 1. Consider the 3rd level of the tree (which is actually the bottom of the tree if we were to ignore the scrap vs sell decision, and our equation for the CDF of ε in equation (61) does ignore the scrap/sell decision), then for each car type, when choosing over the car age there is a variance components representation for the preference shocks $\varepsilon_{j,d}$ given by

$$\varepsilon_{j,d} = \tilde{z}_j + \sigma_j \eta_{j,d} \tag{63}$$

where $\{\eta_{j,d}\}$ are *IID* standardized extreme value random variables capturing the idiosyncratic features of different cars of different ages d but all of them type j cars, whereas $\tilde{z}_j$ has a $C(\sigma)$ distribution (independently distributed from $\{\eta_{j,d}\}$) and it can be thought of as the "brand effect for car type j" and this is common to all ages of cars of type j.

But now in the next level up, where the customer chooses over different car types, j, this "brand effect" $\tilde{z}_j$ plays a role in the brand or type of car the household chooses. But conditional on the choice to replace the current car, you should be able to show that the brand shocks z_j and $z_{j'}$ are independently distributed for two diffent types of cars $j \neq j'$. This will imply that the choice of a particular type of car j will be governed by a MNL model, but where the covariates entering this model are *inclusive values capturing the expected maximum utility of choosing the optimal age of car of each type j.*

Now consider the top level of the tree: it represents a choice over three alternatives: 1) purge the current car, 2) keep the current car, or 3) scrap the car. Again, using the variance component interpretation, we can consider this to be given by a trinomial logit model where there are shocks specific to each of these three choices. To better appreciate this, the max-stability property implies that for the choice to trade, the maximum utilty of the choice of trading is the maximum of all of the choices in the subtree descending below the node of the trade marked "$j \in \{1,\dots,J\}$ replace" in figure 1. Since the maximum of extreme value random variables is also extreme value distributed, it implies that the random variable $\tilde{Z} = \max_{j,d}[v(i,a,j,d)+\varepsilon_{j,d}]$ has an extreme value distribution, and similarly for the choices to purge or to keep the car. So the top level of the choice tree is equivalent to a trinomial choice model and the specification of the G in equation (61) implies that these three extreme value random variables corresponding to a choice at the top level of the tree are indpendently distributed (and hence satisfy the IIA property) whereas for choices further down in the tree (such as choice of car type or car age given car type) there is correlation in the $\varepsilon_{j,d}$ shocks at those choice levels. You should verify by looking at the G function in equation (61) that this is indeed the case.

It is very important to note that the choice tree in figure 1 does *NOT* imply a sequential choice process where the household first chooses whether to scrap, keep or trade their car, and then, for example if they chose to trade their car, then to choose the optimal type j and age d to replace their car with. Instead the choices households make are *fully simultaneous and not sequential* and are based on full simultaneous knowledge of all the value functions $\{v(i,a,\emptyset), v(i,a,\kappa), \{v(i,a,j,d)|j=1,\dots,J, d=0,\dots,\bar{a}-1\}\}$ and the unobserved preference shocks $\varepsilon = \left\{\varepsilon_\emptyset, \varepsilon_\kappa, \{\varepsilon_{j,d}|j=1,\dots,J, d=0,\dots,\bar{a}-1\}\right\}$. However despite this, below I derive the choice probabilities and the choice probabilities can be factored into products of "transition probabilities" that have the interpretation of a sequential decision process when in fact the decision at any given time period is made simultaneously over all feasible alernatives in the household's choice set $D(i,a)$.

We haven't fully explained the household's dynamic programming problem and you can see the paper by Gillingham, Iskhakov, Munk-Nielsen, Rust and Scherning (2022) for those details but here I briefly sketch them so as to keep these answers self-contained. In dynamic programming, there is a *value function* that is key to the solution. In this case, the value function for a household who owns a car (i,a) is given by $V(i,a,\varepsilon)$ given by

$$V(i,a,\varepsilon) = \max\left[v(i,a,\emptyset)+\varepsilon_\emptyset, v(i,a,\kappa)+\varepsilon_\kappa, \max_{j,d}[v(i,a,j,d)+\varepsilon_{j,d}]\right] \tag{64}$$

which is a parallel equation to equation (59) defining the optimal decision rule $\delta(i,a,\varepsilon)$ except that we replace the *argmax* operators by max operators. So $V(i,a,\varepsilon)$ is the expected discounted utility for a household who owns a car (i,a) and faces a vector of idiosyncratic shocks ε at the start of the period when they are making their car trading decision.

Now consider the *decision specific* value functions $v(i,a,\emptyset)$, $v(i,a,\kappa)$ and $\{v(i,a,j,d)|j=1,\dots,J, a=$

$0,\dots,\bar{a}-\}$. These reflect the current part of discounted utility associated with particular decisions. For example if the household chooses to purge their current car, i.e. a decision of $d = \o$, the value of this decision is

$$v(i,a,\o) = u(\o) + \beta EV(\o) \tag{65}$$

where $\beta \in (0,1)$ is the household's discount factor and $EV(\o)$ is the expected value (at the start of next period) of not owning a car, i.e.

$$EV(\o) = \int V(\o,\varepsilon) f(\varepsilon|\o) \tag{66}$$

where $f(\varepsilon|\o)$ is the MGEV probability density for the vector of ε shocks in the no car state. In the no car state there is no possible choice $d = \kappa$ corresponding to "keeping the current car" so the value function $V(\o,\varepsilon)$ is given by

$$V(\o,\varepsilon) = \max\left[v(\o,\o) + \varepsilon_{\o}, \max_{j,d}[v(\o,j,d) + \varepsilon_{j,d}]\right], \tag{67}$$

which is very similar to the equation for the value of having a car (i,a) in equation (64) above except we removed the option of keeping the current car (since you cannot keep a car you don't have!). Once again, if the ε shocks have a distribution in the MGEV family, since this family is max-stable, then $V(\o,\varepsilon)$ is a Type 1 extreme value distributed random variable (when we treat ε as an unobserved random variable) so when we take expectations with respect to the ε random variable we get the usual log-sum closed form expectation for the expectation

$$EV(\o) = \sigma \log\left(\exp\{v(\o,\o)/\sigma\} + \exp\{I_{j>0}(\o)/\sigma\}\right) \tag{68}$$

where $I_{j>0}(\o)$ is the *inclusive value* or *ex ante* expected maximized utility corresponding to the choice of moving out of the no car state and buying some car (d,j) that provides the consumer the *ex post* highest discounted utility after observing the preference shocks for each (d,j) alternative, and is given by

$$I_{j>0}(\o) = \sigma_{j>0} \log\left(\sum_{j=1}^{J} \exp\{I_j(\o)/\sigma_{J>0}\}\right). \tag{69}$$

where $I_j(\o)$ is the inclusive value or expected maximal discounted utility associated with the choice of a particular car type j by a household that does not own a car, and is given by

$$I_j(\o) = \sigma_j \log\left(\sum_{d=0}^{\bar{a}_j-1} \exp\{v(\o,j,d)/\sigma_j\}\right). \tag{70}$$

Now consider the choice of keeping the current car. We said its decision-specific value is $v(i,a,\kappa)$ and it is defined by

$$v(i,a,\kappa) = u(i,a) + \beta\left[(1-\alpha(i,a)]EV(i,a+1) + \alpha(i,a)EV(i,\bar{a})\right], \tag{71}$$

where $\alpha(i,a)$ is the probability that the car the household keeps is involved in a collision that totally destroys (or makes it not worth repairing) so it must be scrappage, a transition that we represent as the next period age of the car being equal to the scrap age $\bar{a}$. If the car is not involved in an accident that results in its being scrapped, which occurs with probability $1-\alpha(i,a)$, then the car is one year older next year and the expected value of having such a car is $EV(i,a+1)$.

It follows that we also need to have formulas for $EV(i,\bar{a})$ (i.e. the expected value of a car of type i that needs to be scrapped) and the expected value of state $(i,a+1)$ is $EV(i,a+1)$. We can write the generic formula for $EV(i,a)$ as

$$EV(i,a) = \int V(i,a,\varepsilon)f(\varepsilon|i,a), \tag{72}$$

where again $f(\varepsilon|i,a)$ is the probability density of the MGEV-distributed random variable ε. Via the max-stability property $V(i,a,\varepsilon)$ has a Type 1 extreme value distribution, so using the value function definition in equation (64) we can write a closed-form expression for $EV(i,a)$ as follows

$$EV(i,a) = \begin{cases} \sigma\log\left(\exp\{v(i,a,\emptyset)/\sigma\} + \exp\{I_{j>0}(i,a)/\sigma\}\right) & \text{if } a = \bar{a}_j \\ \sigma\log\left(\exp\{v(i,a,\emptyset)/\sigma\} + \exp\{v(i,a,\kappa)/\sigma\} + \exp\{I_{j>0}(i,a)/\sigma\}\right) & \text{otherwise} \end{cases} \tag{73}$$

where the first equation for $EV(i,a)$ in (73) is for the case where the current car is at the scrappage age threshold, $a = \bar{a}$, so keeping this car is no longer an option, and the second formula is for the case where $a < \bar{a}$ so the consumer has the additional option to choose to keep the current car (i,a) rather than trade it, which has the decision-specific value $v(i,a,\kappa)$. The inclusive value for the decision to trade for another car given that the household currently owns a car (i,a) is denoted by $I_{j>0}(i,a)$ and is given by

$$I_{j>0}(i,a) = \sigma_{j>0}\log\left(\sum_{j'=1}^{J}\exp\{I_{j'}(i,a)/\sigma_{j>0}\}\right) \tag{74}$$

where $I_{j'}(i,a)$ is the inclusive value associated with the decision to trade the current car (i,a) for some car (j',d') of car type j' given by

$$I_{j'}(i,a) = \sigma_{j'}\log\left(\sum_{d'=0}^{\bar{a}-1}\exp\{v(i,a,j',d')/\sigma_{j'}\}\right). \tag{75}$$

The final decision-specific value we need to define is $v(i,a,j,d)$, the value of trading the current car (i,a) for another car (j,d). Note the latter car will not necessarily be a new car: it will only be a new car if $d = 0$. The value of this decision is

$$v(i,a,j,d) = u(j,d) + \mu[P_{jd} - P_{ia} + T_s(i,a) + T_b(j,d)] + \beta\left[(1-\alpha(j,d))EV(j,d+1) + \alpha(j,d)EV(j,\bar{a})\right], \tag{76}$$

where P_{ja} is the price of the car (j,d) the household buys, P_{ia} is the resale price of the existing car the household sells, $T_b(j,d)$ are buyer-side transactions and search costs to find the car (j,d) the household purchases, and $T_s(i,a)$ are the seller-side transactions costs associated with selling the existing car (i,a). Here, the parameter μ denotes the "marginal utility of money" i.e. we are assuming a *quasi-linear utility function* where the parameter μ constitutes the utility value of a 1 dollar of expenditure. For notational simplicity, we suppressed the dependence of the value functions on the set of car prices. However, to describe equilibrium is necessary note that all the value functions implicitly depend on the vector of prices P but we are not going to consider the issue of defining and computing equilibrium here. See the Gillingham *et. al.* paper for that.

Now given these decision-specific value functions, let's derive the conditional choice probabilities for the consumer implied by the nested logit formulation. The beauty of a nested logit model is that implies a decomposition in the overall choice probability into a product of "transition proabilities" corresponding to the different subtrees of the overall decision tree in figure 1. First, the overall choice probability for trading the current car (i,a) for a car (j,d) can be denoted by $\Pi(j,d|i,a)$ and it is a conditional probability since we have to condition on the consumer's current *state* which involves owning a car $x=(i,a)$. The choice probability for a consumer who has no car is given by $\Pi(j,d|\emptyset)$ and this will generally be different than the choice probability of someone who owns a car. Now the nested logit structure of the MGEV distribution for the ε shocks enables us to further decompose the choice probability for a household who owns a car as follows

$$\Pi(j,d|i,a) = \Pi(d|j,i,a)\Pi(j|j>0,i,a)\Pi(j>0|i,a) \tag{77}$$

where $\Pi(d|j,i,a)$ is the conditional probability that the household buys a car of age d given they chose to trade their curent car (i,a) for a car of type j, and $\Pi(j|j>0,i,a)$ is the conditional probability that a household chooses to trade their current car for a car of type j given the overall decision to trade their current car (which we denote by $j>0$ with $j=0$ being equivalent to a decision not to trade the current car, i.e. the decision κ), and $\Pi(j>0|i,a)$ is the probability of choosing to trade the current car (i,a) for some other age and type of car traded on the market.

Now look at the top three branches of the choice tree in figure 1. There are three choices here: 1) purge the car, which has probability $\Pi(\emptyset|i,a)$, 2) keep the current car, which has probability $\Pi(\kappa|i,a)$, and finally 3) trade for some other car, which has probability $\Pi(j>0|i,a)$. The probabilities for these three choices are given by

$$\Pi(\emptyset|i,a) = \begin{cases} \frac{\exp\{v(i,a,\emptyset)/\sigma\}}{\exp\{v(i,a,\emptyset)/\sigma\}+\exp\{v(i,a,\kappa)/\sigma\}+\exp\{I_{j>0}(i,a)/\sigma\}} & \text{if } a<\bar{a} \\ \frac{\exp\{v(i,a,\emptyset)/\sigma\}}{\exp\{v(i,a,\emptyset)/\sigma\}+\exp\{I_{j>0}(i,a)/\sigma\}} & \text{if } a=\bar{a} \end{cases} \tag{78}$$

which reflects the constraint that for any car type i a consumer who holds the oldest possible age $a=\bar{a}$ is forced to scrap it, so we have

$$\Pi(\kappa|i,a) = \begin{cases} \frac{\exp\{v(i,a,\kappa)/\sigma\}}{\exp\{v(i,a,\emptyset)/\sigma\}+\exp\{v(i,a,\kappa)/\sigma\}+\exp\{I_{j>0}(i,a)/\sigma\}} & \text{if } a<\bar{a} \\ 0 & \text{if } a=\bar{a} \end{cases} \tag{79}$$

Finally we have

$$\Pi(j>0|i,a) = \begin{cases} \frac{\exp\{I_{j>0}(i,a)\sigma\}}{\exp\{v(i,a,\emptyset)/\sigma\}+\exp\{v(i,a,\kappa)/\sigma\}+\exp\{I_{j>0}(i,a)/\sigma\}} & \text{if } a<\bar{a} \\ \frac{exp\{I_{j>0}(i,a)\sigma\}}{\exp\{v(i,a,\emptyset)/\sigma\}+\exp\{I_{j>0}(i,a)/\sigma\}} & \text{if } a\geq\bar{a} \end{cases} \tag{80}$$

Now consider the probability that the household who chooses to trade their car for another $(j>0)$ will purchase a car of type j

$$\Pi(j|j>0,i,a) = \frac{\exp\{I_j(i,a)/\sigma_{j>0}\}}{\sum_{j'=1}^{J}\exp\{I_{j'}(i,a)/\sigma_{j>0}\}}. \tag{81}$$

Finally, let $\Pi(d|j,i,a)$ be the conditional probability that the consumer chooses a car of age d given that the consumer has chosen a car type j

$$\Pi(d|j,i,a) = \frac{\exp\{v(i,a,d,j)/\sigma_j\}}{\sum_{d'=0}^{\bar{a}}\exp\{v(i,a,j,d')/\sigma_j\}}. \tag{82}$$

Finally we derive the corresponding choice probability for a person who is in the no car state, $x=\emptyset$. Again the nested logit structure implies the following decomposition of the conditional choice probability

$$\Pi(j,d|\emptyset) = \Pi(d|\emptyset)\Pi(j|j>0,\emptyset)\Pi(j>0|\emptyset). \tag{83}$$

In the case of a consumer whose has no car, it is easy to see that the choice tree in figure 1 has only two branches at the top of the tree corresponding to the choices to either stay in the no car state, $d = \text{ø}$, or to purchase some type of car, $j > 0$. The probability of remaining in the no car state is $\Pi(\text{ø}|\text{ø})$ given by

$$\Pi(\text{ø}|\text{ø}) = \frac{\exp\{v(\text{ø},\text{ø})/\sigma\}}{\exp\{v(\text{ø},\text{ø})/\sigma\} + \exp\{I_{j>0}(\text{ø})/\sigma\}} \tag{84}$$

and $\Pi(j > 0|\text{ø})$ is the probability that the consumer chooses some type of car,

$$\Pi(j > 0|\text{ø}) = 1 - \Pi(\text{ø}|\text{ø}) = \frac{\exp\{I_{j>0}(\text{ø})/\sigma\}}{\exp\{v(\text{ø},\text{ø})/\sigma\} + \exp\{I_{j>0}(\text{ø})/\sigma\}}, \tag{85}$$

and $\Pi(j|j > 0,\text{ø})$ is the probability the consumer chooses car type j given the decision to choose some type of car,

$$\Pi(j|j > 0,\text{ø}) = \frac{\exp\{I_j(\text{ø})/\sigma_{j>0}\}}{\sum_{j'=1}^{J} \exp\{I_{j'}(\text{ø})/\sigma_{j>0}\}} \tag{86}$$

and $\Pi(d|j,\text{ø})$ is the probability the consumer chooses a car of age d given that they chose a car of type j

$$\Pi(d|j,\text{ø}) = \frac{\exp\{v(d,j,\text{ø})/\sigma_j\}}{\sum_{d'=0}^{\bar{a}} \exp\{v(d',j,\text{ø})/\sigma_j\}}. \tag{87}$$