Chapter 18
Continuous Approximations of Discrete Choice Models Using Point Process Theory

Hannes Malmberg and Ola Hössjer

Abstract We analyze continuous approximations of discrete choice models with a large number of options. We start with a discrete choice model where agents choose between different options, and where each option is defined by a characteristic vector and a utility level. For each option, the characteristic vector and the utility level are random and jointly dependent. We analyze the optimal choice, which we define as the characteristic vector of the option with the highest utility level. This optimal choice is a random variable. The continuous approximation of the discrete choice model is the distributional limit of this random variable as the number of offers tends to infinity. We use point process theory and extreme value theory to derive an analytic expression for the continuous approximation, and show that this can be done for a range of distributional assumptions. We illustrate the theory by applying it to commuting data. We also extend the initial results by showing how the theory works when characteristics belong to an infinite-dimensional space, and by proposing a setup which allows us to further relax our distributional assumptions.

Keywords Discrete choice · Random utility · Extreme value theory · Random fields · Point processes · Concomitant of order statistics

H. Malmberg (✉)
Stanford Institute for Economic Policy Research, 366 Galvez Street, Stanford, CA 94305-6015, USA
email: malmberg@stanford.edu

H. Malmberg
Department of Economics, University of Minnesota, 4-101 Hanson Hall 1925 4th Street S, Minneapolis, MN 55455, USA
email: pmalmber@umn.edu

O. Hössjer
Department of Mathematics, Stockholm University, 106 91 Stockholm, Sweden
e-mail: ola@math.su.se

S. Silvestrov et al. (eds.), *Stochastic Processes and Applications*, Springer Proceedings in Mathematics & Statistics 271,
https://doi.org/10.1007/978-3-030-02825-1_18

18.1 Introduction

There is a long tradition in economics to use random utility theory to study discrete choices such as the choice of mode of transportation. Early contributions are Luce [12] and Mcfadden [15]. Over time, random utility theory has been extended to encompass more functional forms, distributional assumptions, and applications (Ben-Akiva and Lerman [2], Anderson et al. [1], Train [20]). The theory posits that agents maximize utility, but that utility is random from the econometrician's point of view. Utility is expressed as a random variable

$$U_i = f(X_i) + \varepsilon_i \quad i = 1, \dots, n,$$

where U_i is the utility of option i, X_i are random variables that describe the characteristics of option i, $f(X_i)$ is the deterministic component of utility, and ε_i are independently and identically distributed random variables. The agent chooses the option with the highest utility.

Insofar each option has distinct characteristics, we can equivalently view this as a choice over the characteristics X_i. We write $X_{[n:n]}$ for the X_i corresponding to the largest U_i. This is a random variable taking values in the set $\{X_1, \dots, X_n\} \subseteq \Omega$, where Ω is a general characteristics space.

We are interested in a continuous approximation to the discrete choice problem when the number of options is large, and we define the approximation as the distributional limit of the law of $X_{[n:n]}$ as $n \to \infty$. A continuous approximation takes an offer distribution density Λ, a deterministic utility component $f(\cdot)$, and the distribution of the random utility component ε_i, as inputs. The output is a probability distribution of choices over Ω.

The theory is relevant in situations where agents face discrete choices and a large number of options. For example, the choice of residential location in a city is a discrete choice as agents only buy one residence. This makes a random utility approach natural. On the other hand, the number of potential residential locations is large. In this case, it can be useful to approximate the discrete choice process with a continuous probability distribution over space.

We approach the problem by interpreting the collection of characteristics-utility pairs $(X_1, U_1), \dots, (X_n, U_n)$ as the realizations of a point process ξ_n on the Cartesian product $\Omega \times \mathbb{R}$ of the characteristics space and the utility space. With this interpretation, the best choice $X_{[n:n]}$ is a function of ξ_n. More details on point process theory can be found, for instance, in Cox and Isham [4] and Jacobsen [9]. More specifically, we can build on the results in point process theory presented in Resnick [17] to derive sharp results on the limiting behavior of $X_{[n:n]}$. In particular, we show that a monotone transformation of the underlying point process ξ_n converges to a Poisson process on $\Omega \times \mathbb{R}$ and we derive the limiting behavior of $X_{[n:n]}$ using continuity properties of the mapping from ξ_n to $X_{[n:n]}$. We show that there is a tractable continuous approximation for a range of distributional assumptions.

After our theoretical result, we illustrate our theory with an empirical example taken from Burke and Brown [3] who analyze commuter walking distances. We show that our theory predicts that walking distances are gamma-distributed and verify that this prediction is confirmed by the data. In the discussion section, we also propose an extension which would allow us to analyze the asymptotic behavior under an even wider range of distributional assumptions.

In Sect. 18.2 we outline the model environment. In Sect. 18.3, we provide the necessary theoretical background on point processes. Section 18.4 derives the limiting behavior of our point processes and use this to derive the limiting behavior of choice probabilities. Section 18.5 outlines the empirical application and other applications, whereas Sect. 18.6 proposes an extension to encompass a wider range of distributional assumptions. Section 18.7 concludes.

The paper is similar in aim to Malmberg and Hössjer [14]. However, they used asymptotic properties of deterministic point processes in order to analyze random utilities by methods developed in the literature on random sup measures (see O'Brien et al. [16], Resnick and Roy [18], and Stoev and Taqqu [19]). The novel approach in this paper is to analyze the problem using random point process theory instead, and this method allows for a mathematically simpler formulation than the one used in Malmberg and Hössjer [14]. Since we analyze the values X associated with the maximum U, the paper also connects to the theory of concomitants of extremes (see Ledford and Tawn [11]). The theory proposed in the extension section also relates to conditional extreme value theory, which is discussed in Heffernan and Tawn [8].

In this paper, we illustrate our theory using commuting patterns. Earlier work on random choice models with an infinite number of options has also been used to model distance dependence in international trade (Kapiarz et al. [10]). Even though the motivation for our setup comes from random choice theory, the theory has also been used in machine learning by Maddison et al. [13], who use methods in Malmberg and Hössjer [14] to derive a new way of sampling from a posterior distribution in problems of Bayesian statistics.

18.2 Model Environment

18.2.1 Model Setup and Assumptions

Consider a sequence of independent and identically distributed pairs of random variables $\{(X_i, U_i)\}_{i=1}^{\infty}$, where $X_i \in \Omega$ and $U_i \in \mathbb{R}$, and where Ω is a complete, separable, metric space.

Here X_i gives the characteristics of the choice i and Ω is the characteristic space. In case of residential choice, we might have $\Omega \subseteq \mathbb{R}^2$, where X_i gives the location of choice i. In industrial organization, $\Omega \subset \mathbb{R}^n$ might denote a multidimensional product characteristics vector, and X_i is the characteristics of a particular good. It aids intuition to think of Ω as a subset of Euclidean space, but the analysis is done for the general case of a complete, separable, metric space. This means that the setup can be used to analyze cases where choice options are functions, for example choices of continuous consumption paths over finite intervals when the valuation is random from the econometrician's perspective.

We define $U_{n:i}$ as the ith order statistic increasing order of $\{U_1, \ldots, U_n\}$. For each n, we define the characteristic $X_{[n:i]}$ to be the X-value (the concomitant) associated with $U_{n:i}$ for a sample of size n.

We are interested in the limiting probability distribution of the characteristics of the optimal choice $X_{[n:n]}$, and to this end, we study the asymptotic behavior of the sequence of probability measures

$$C_n(\cdot) = P\left(X_{[n:n]} \in \cdot\right). \tag{18.1}$$

The distribution of (X, U) is

$$P((X, U) \in A \times B) = \int_A \mu(x; B) d\Lambda(x),$$

where $F_X = \Lambda$ is the marginal distribution of X over Ω, and $\mu(x; \cdot)$ is the regular conditional probability measure of U_i given $X_i = x$.[1] The interpretation here is that the characteristics of offers are distributed according to Λ. For example, Λ gives the distribution of potential dwellings over space in the case of residential choice, or the distribution of products over the characteristic space in case of industrial organization applications. For each offer, there is a distribution of utility $\mu(x; \cdot)$ depending on the characteristics x. We make the following assumption on μ:

Assumption 1 For the collection $\mu = \{\mu(x; \cdot); x \in \Omega\}$, there exists a function

$$p : \Omega \to (0, \infty), \tag{18.2}$$

and sequences $a_n > 0$, b_n, independent of x, and a distribution function G_α with $\alpha \in \mathbb{R}$, such that

$$\mu(x; (-\infty, a_n u + b_n])^n \to G_\alpha(u)^{p(x)} \tag{18.3}$$

as $n \to \infty$, where G_α is a distribution function of one of the following three forms:

$$G_\alpha(u) = \begin{cases} \exp(-(-u)^{-\alpha} I(u < 0)), & \alpha < 0, \\ \exp(-\exp(-u)), & \alpha = 0, \\ I(u > 0)\exp(-u^{-\alpha}), & \alpha > 0, \end{cases}$$

and $I(\cdot)$ is the indicator function.

The assumption above essentially asserts that all $\mu(x; \cdot)$ belong to the domain of attraction of the same extreme value distribution, indexed by α, and that their limiting

[1] In terms of the example in the introduction, Λ corresponds to the law of the random variables X_i, and $\mu(x; \cdot)$ corresponds to the law of the random variable $f(X_i) + \varepsilon_i | X_i = x$.

relative sizes can be described by the one dimensional parameter $p(x)$. The function $p(x)$ captures the deterministic "quality" inherent in characteristics x, which determines the limiting behavior of offer quality. The following example makes it clear in what sense $p(x)$ captures a deterministic component of utility.

Example 18.1 Assume there is a function $h(x)$ such that $\mu(x;\cdot)$ is given by an exponential distribution shifted $h(x)$ to the right. Formally, let $\mu(x;\cdot)$ be the law of a random variable $h(x)+\varepsilon$ where $\varepsilon \sim Exp(1)$.

This collection of distributions satisfies Assumption 1 when $p(x)=e^{h(x)}$, $a_n=1$, $b_n=\log(n)$, and $\alpha=0$. Equation (18.3) follows from

$$\begin{aligned}\mu(x;(-\infty,b_n+a_n u])^n &= (1-\exp(-u-h(x)-\log n))^n\\ &\to \exp(-\exp(-u)p(x)),\\ &= G_0(u)^{p(x)}.\end{aligned} \tag{18.4}$$

The distributional assumption is not vacuous. Below is a class of distributions which does not satisfy Assumption 1.

Example 18.2 Suppose there exists a non-constant function $h(x)$ such that $\mu(x;\cdot)$ is the law of a normal distribution with mean $h(x)$ and variance 1.

Let $F \sim N(0,1)$ and assume without loss of generality that there exists an $x_0 \in \Omega$ such that $h(x_0)=0$, and find a_n, b_n such that $F^n\,(a_n y+b_n)$ converges to a non-degenerate distribution function $G(y)$. Extreme value theory means that the normalization constants for a normal distribution satisfies $a_n \to 0$ and $b_n \to \infty$, and we know that the limiting distribution function $G(\cdot)$ is of Gumbel type $\alpha=0$ (Resnick [17]).

But this means that $F^n(a_n y+b_n-h(x))$ converges to 0 if $h(x)>0$, as

$$\begin{aligned}\lim_{n\to\infty} F^n(a_n y+b_n-h(x)) &= \lim_{n\to\infty} F^n\left[a_n\left(y-\frac{h(x)}{a_n}\right)+b_n)\right]\\ &\le \lim_{n\to\infty} F^n\left[a_n\left(y-\frac{h(x)}{a_N}\right)+b_n)\right]\\ &= G\left(y-\frac{h(x)}{a_N}\right)\end{aligned}$$

for any sufficiently large N. Let $N\to\infty$ and we obtain the conclusion. As the limit is 0, we need $p(x)=\infty$ which violates that $p(x)<\infty$. On the other hand, we can use an analogous reasoning to conclude that $F^n(a_n y+b_n-h(x))$ converges to 1 if $h(x)<0$, so that $p(x)=0$. This violates $p(x)>0$.

We conclude that a non-constant function $h(x)$ is not consistent with Assumption 1.

It limits the theory that the traditional normal regression structure does not satisfy Assumption 1. The reason is that the normal distribution is too thin-tailed. Formally, the condition for when the linear regression formulation works is whether the limit

$$\lim_{u\to\infty} \frac{P(U + h(x_1) > u)}{P(U + h(x_2) > u)}$$

exists and is not 0 or ∞ when $h(x_1) \neq h(x_2)$. This condition holds when U is exponentially distributed but not when U is normally distributed. When U is normally distributed, the limit is ∞ for $h(x_1) > h(x_2)$ and 0 for $h(x_1) < h(x_2)$. In Sect. 18.6, we propose an extension which would allow us to analyze normal regression functions.

18.2.2 Point Process Formulation and Strategy

The sequence $\{(X_i, U_i)\}_{i=1}^n$ can be viewed as a random collection of points in $\Omega \times \mathbb{R}$, and can be described as a sequence of point processes ξ_n. We will show that after a suitable transformation, this sequence of point processes ξ_n converges to a Poisson point process ξ in a sense which will be formalized later. As

$$C_n(A) = P(X_{[n:n]} \in A) = P\left(\sup_{i:X_i\in A} U_i > \sup_{i:X_i\notin A} U_i\right)$$

is a functional on our point process ξ_n, the problem of finding $\lim_{n\to\infty} C_n$ reduces to determine whether this functional is continuous. In this case, we can use the limiting point process ξ to calculate our results.

We will start with an introduction to point processes – in particular sufficient conditions for convergence. After this, we will apply the point process machinery to our setup, and characterize the limit of our point process. Once this is done, we will define random fields taking point processes as inputs, and derive the asymptotic behavior of C_n from continuity properties of these random fields.

18.3 Background on Point Processes and Convergence Results

This section contains background results and a notational machinery for point processes. See Chapter 3 of Resnick [17] for a more detailed treatment.

Throughout this discussion, the generic point process will take values in a locally compact set E, with an associated σ-algebra $\mathcal{E}$. For the purpose of our discussion, E will be a subset of $\Omega \times \mathbb{R}$, and we assume that $\mathcal{E} = \mathcal{B}(E)$ is the Borel σ-algebra. A *point mass* is a set function, defined by

$$\delta_z(F) = \begin{cases} 1, & \text{if } z \in F, \\ 0, & \text{if } z \notin F, \end{cases}$$

where $F \subseteq E$, $F \in \mathcal{E}$. A *point measure* is a measure $m(\cdot)$ on E such that there exists a countable collections of points $\{z_k\} \subseteq E$ and numbers $\{w_k\} \geq 0$, such that

$$m(\cdot) = \sum_{z_k} w_k \delta_{z_k}(\cdot).$$

We will confine our attention to the case $w_k \equiv 1$. Let $\mathcal{M}_P(E)$ be the set of point measures on E, and let it have the minimal σ-algebra which makes

$$\{m \in \mathcal{M}_P(E) : m(F) \in B\}$$

measurable for all $F \in \mathcal{E}$, $B \in \mathcal{B}(\mathbb{R})$ where $m(F)$ is the point measure m evaluated at the set F and $\mathcal{B}(\mathbb{R})$ is the Borel σ-algebra on $\mathbb{R}$. We define a *point process* to be a random element of $\mathcal{M}_P(E)$.

If N is an arbitrary point process, we define the Laplace transform ψ associated with N as

$$\begin{aligned} \psi_N(f) &= E\left(\exp\left\{-\int_E f(x)N(dx)\right\}\right) \\ &= \int_{M_P(E)} \exp\left\{-\int_E f(x)m(dx)\right\} P^N(dm). \end{aligned} \tag{18.5}$$

Here P^N is a probability measure over the set $\mathcal{M}_P(E)$ which corresponds to the distribution of N. Moreover, the class of functions f for which we are interested in ψ_N is usually the continuous non-negative functions on E with a compact support. We write $C_K^+(E)$ to denote this set.

Definition 18.1 A sequence of point processes N_n, $n \geq 0$, converges in a point process sense to N_0, written $N_n \Rightarrow_p N_0$, if

$$\psi_{N_n}(f) \to \psi_{N_0}(f)$$

for all $f \in C_K^+(E)$.

We use the notation $\Longrightarrow$ for weak convergence of vector valued random variables in Euclidean space or on Ω, in contrast to $\Rightarrow_p$ for point process convergence.

Definition 18.2 Let E be a metric space. We call $F \subseteq E$ relatively compact if its closure $\bar{F}$ in E is compact.

Definition 18.3 Let μ be a measure on a metric space $\mathcal{X}$. We say that a sequence of measures μ_n converges *vaguely* to μ, written

$$\mu_n \Rightarrow_v \mu,$$

if

$$\mu_n(F) \to \mu(F)$$

for all relatively compact F with $\mu(\partial F) = 0$, where ∂F is the boundary of the set F.

Definition 18.4 For a point process N, the Laplace functional associated with N is defined by

$$\Psi_N(f) = E\left[\exp(-N(f))\right]$$

where $N(f) = \sum_{x \in N} f(x)$.

It is known from point process theory that the Laplace functional uniquely defines a point process. Thus, the Laplace functional can be used to define a Poisson process and derive its properties (see, for example, Resnick [17], p. 130).

Definition 18.5 A Poisson process with intensity measure μ is a point process defined by the Laplace functional

$$\Psi_N(f) = e^{-\int_E (1-e^{-f(x)})d\mu(x)}.$$

Proposition 18.1 *For any $F \in \mathcal{E}$, and any non-negative integer k, a Poisson process satisfies*

$$P(N(F) = k) = \begin{cases} e^{-\mu(F)}(\mu(F))^k/k!, & \text{if } \mu(F) < \infty, \\ 0, & \text{if } \mu(F) = \infty, \end{cases}$$

and that for any $k \geq 1$, if $F_1, \dots, F_k$ are mutually disjoint sets in $\mathcal{E}$, then $\{N(F_i)\}$ are independent random variables.

Our main theorem will also depend on the following proposition which is a modification of a result presented in the proof of a more extensive Proposition 3.21 in Resnick [17].

Proposition 18.2 *For each n, suppose $\{Z_{n,j} : 1 \leq j \leq n\}$ are independent and identically distributed (i.i.d.) random variables on E and that*

$$nP(Z_{n,1} \in \cdot) \Rightarrow_v \mu.$$

where μ is a measure on E. Then

$$N_n = \sum_{j=1}^{n} \delta_{Z_{n,j}} \Rightarrow_p N$$

where N is a Poisson random measure on E with intensity μ.

Proof This proof is essentially equivalent to the first half of the proof of Proposition 3.21 in Resnick [17]. We use that convergence in point measures is equivalent to convergence in Laplace functionals. Indeed, pick an arbitrary $f \in C_K^+(E)$, with a compact support $F \subseteq E$. Then:

$$
\begin{aligned}
\psi_{N_n}(f) &= E \exp\{-N_n(f)\} \\
&= E \exp\left\{-\sum_{j=1}^{n} f(Z_{n,j})\right\} \\
&= \left(E \exp\{-f(Z_{n,1})\}\right)^n \\
&= \left(1 - \frac{\int_E \left(1-e^{-f(z)}\right) nP\left[Z_{n,1} \in dz\right]}{n}\right)^n \\
&= \left(1 - \frac{\int_F \left(1-e^{-f(z)}\right) nP\left[Z_{n,1} \in dz\right]}{n}\right)^n \\
&\to e^{-\int_F (1-e^{-f(z)}) d\mu(z)} \\
&= e^{-\int_E (1-e^{-f(z)}) d\mu(z)} \\
&= \psi_N(f),
\end{aligned} \tag{18.6}
$$

where the convergence step is obtained from the vague convergence of $nP[Z_{n,1} \in \cdot]$. Indeed, vague convergence is equivalent to weak convergence on every compact subspace. As $1 - e^{-f(z)}$ continuous and bounded, and weak convergence means that the integral of every continuous and bounded function converges, we get the desired result. Thus,

$$N_n \Rightarrow_p N$$

as required. □

Before giving the full proof of Theorem 18.1, we state and prove the following lemma:

Lemma 18.1 *If* $(\mathcal{X} \times \mathcal{U}, \Lambda \times \nu)$ *is a product measure space, where* $\mathcal{X}$ *and* $\mathcal{U}$ *are two complete, separable metric spaces, and if* $F \subseteq \mathcal{X} \times \mathcal{U}$ *satisfies*

$$(\Lambda \times \nu)(\partial F) = 0,$$

then

$$\nu(\partial F_x) = 0 \quad \Lambda - a.e.$$

where $F_x = \{u \in \mathcal{U} : (x, u) \in F\}$ *is the cross-section of* F *at the point* x*, and a.e. refers to convergence almost everywhere (or almost surely).*

Proof We note that if we write

$$B = \{(x, u) \in \mathcal{X} \times \mathcal{U} : u \in \partial F_x\},$$

we have

$$B \subseteq \partial F$$

(as each ball around a point $(x, u) \in B$ contains both a point within and outside F). Thus, as

$$(\Lambda \times \nu)(B) = \int_{\mathcal{X}} \nu(\partial F_x) d\Lambda(x) \leq (\Lambda \times \nu)(\partial F) = 0,$$

we get that $\nu(\partial F_x) = 0$ Λ-almost everywhere. □

18.4 Limiting Behavior of Choice Probabilities

In this section, we use point process theory to derive the limit of the choice probabilities $C_n(\cdot) = \mathbb{P}(X_{[n:n]} \in \cdot)$. We first show that the point process generated by the collection $\{(X_i, U_i)\}$ converges to a Poisson process after a suitable transformation. We then use this fact to calculate the limit of C_n.

18.4.1 Convergence of Point Process

We consider a sequence of transformations

$$g_n(u) = (u - b_n)/a_n$$

where a_n, b_n are chosen to ensure extreme value convergence for all $x \in \Omega$ as in (18.3) of Assumption 1.

Let $\delta_{(x,u)}$ denote a one point distribution at (x, u) and define the extremal marked point process (cf. Resnick [17])

$$\xi_n = \sum_{i=1}^{n} \delta_{(X_i, g_n(U_i))} \tag{18.7}$$

for a sample of size n. This is a point process on $(\Omega \times \mathbb{R}, \mathcal{B}(\Omega \times \mathbb{R}))$.

We are now ready to formulate our first main result. It states that ξ_n converges to a Poisson process with a product intensity measure which multiplies the initial measure Λ on Ω with $p(x)$. Before stating this result, we first introduce a few concepts.

Definition 18.6 A random variable X stochastically dominates a random variable Y if

$$\mathbb{P}(X \geq x) \geq P(Y \geq x) \quad \forall x \in \mathbb{R}.$$

We also say that a measure μ_X on the real numbers dominates μ_Y if they are laws of random variables X and Y and X stochastically dominates Y.

Theorem 18.1 *Let G_α and p be as in Assumption 1. Suppose that for each compact subset $A \subseteq \Omega$, the function $p : \Omega \to (0, \infty)$ is bounded on that subset, and that $\mu(x_0, \cdot)$, for some $x_0 \in \Omega$ is an upper bound for all $\{\mu(x; \cdot);\ x \in A\}$ in the sense of stochastic dominance. Then*

$$\xi_n \Rightarrow_p \xi,$$

as $n \to \infty$ where ξ_n is given by (18.7), and ξ is a Poisson random measure on $(\Omega \times \mathbb{R}, \mathcal{B}(\Omega \times \mathbb{R}))$ with mean intensity $\Lambda_p \times \nu_\alpha$, where

$$\Lambda_p(A) = \int_A p(x)\Lambda(dx)$$

for all relatively compact $A \in \mathcal{B}(\Omega)$ *and*

$$\nu_\alpha([u,\infty)) = -\log(G_\alpha(u)) = \begin{cases} I(u < 0)(-u)^{-\alpha}, & \text{if } \alpha < 0 \text{ and } u < 0\,, \\ \exp(-u), & \text{if } \alpha = 0, \\ u^{-\alpha}, & \text{if } \alpha > 0 \text{ and } u > 0. \end{cases}$$

Proof Note that we have $G_\alpha(u) = 0$ for $\alpha > 0$ and $u \le 0$. Whenever $\alpha > 0$, it is therefore implicit in the proof that $u > 0$. Using the proof of Proposition 18.2, it suffices to show that

$$nP((X_1, g_n(U_1)) \in \cdot) \Rightarrow_v \Lambda_p \times \nu_\alpha,$$

i.e. that

$$nP((X_1, g_n(U_1)) \in F) \to (\Lambda_p \times \nu_\alpha)(F),$$

for all $F \subseteq \Omega \times \mathbb{R}$ which are relatively compact sets with respect to $\mathcal{B}(\Omega \times \mathbb{R})$ and satisfy

$$(\Lambda_p \times \nu_\alpha)(\partial F) = 0.$$

Henceforth, let F be an arbitrary set with these properties. Now, we note that

$$nP((X_1, g_n(U_1)) \in F) = \int_\Omega nP(g_n(U_1) \in F_x | X_1 = x) d\Lambda(x),$$

where F_x is the x-cross section of F. Thus, our task is to show that

$$\int_\Omega nP(g_n(U_1) \in F_x | X_1 = x) d\Lambda(x) \to \int_\Omega p(x)\nu_\alpha(F_x) d\Lambda(x).$$

We do this first by showing that the integrand converges almost everywhere to the desired quantity, and then we show that the sequence of integrands satisfy regularity conditions allowing us to infer convergence of integrals from pointwise convergence.

We observe that for every x,

$$nP(g_n(U_1) \in \cdot | X_1 = x) \Rightarrow_v p(x)\nu_\alpha(\cdot). \tag{18.8}$$

Indeed, it is true that for any sequence x_n such that

$$(x_n)^n \to a, \tag{18.9}$$

we have

$$n(1 - x_n) \to -\log(a). \tag{18.10}$$

Thus, letting $x_n = P(g_n(U_1) < u|X_1 = x)$ and using Assumption 1, we obtain

$$nP(g_n(U_1) \geq u|X_1 = x) \to -p(x)\log(G_\alpha(u)) = p(x)\nu_\alpha\left([u,\infty)\right). \quad (18.11)$$

In order to deduce (18.8) from (18.11), we can note that if we have a measure γ with

$$\gamma([u,\infty)) < +\infty$$

for some u, then vague convergence of γ_n to γ is equivalent to

$$\gamma_n([u,\infty)) \to \gamma([u,\infty)), \quad (18.12)$$

for all u such that $\gamma(\{u\}) = 0$. This can be seen by noting that if (18.12) is true, then the sequence $P_{nu}(\cdot) = \gamma_n(\cdot \cap [u,\infty))/\gamma_n([u,\infty))$ of probability measures converges weakly for all continuity points u of $\gamma([u,\infty))$ to $P_u(\cdot) = \gamma(\cdot \cap [u,\infty))/\gamma[u,\infty))$, and hence $P_{nu}(F) \to P_u(F)$ for all such u, from which (18.8) follows.

Now, using the previous lemma, we know that

$$\nu_\alpha(\partial F_x) = 0 \quad \Lambda_p - a.e.,$$

which means that

$$p(x)\nu_\alpha(\partial F_x) = 0 \quad \Lambda_p - a.e.$$

as $p(x) > 0$ implies that $p(x)\nu_\alpha$ and ν_α are equivalent for all $x \in \Omega$. Thus, we can use (18.8) to conclude that

$$nP(g_n(U_1) \in F_x|X_1 = x) \to p(x)\nu_\alpha(F_x) \quad \Lambda_p - a.e.$$

Therefore, we have established pointwise convergence of the integrand almost everywhere.

Now, we seek to show that $nP(g_n(U_1) \in F_x|X_1 = x)$ is uniformly bounded over n and Ω to ensure that pointwise convergence almost everywhere implies convergence in integrals. To do so, we want to define a maximal random variable which dominates $nP(g_n(U_1) \in F_x|X_1 = x)$ for all n and x.

We write

$$\pi_\Omega : (x,u) \mapsto x$$

and

$$\pi_\mathbb{R} : (x,u) \mapsto u$$

for the projection on Ω and $\mathbb{R}$ respectively. In this case, we know that $\pi_\Omega(F)$ and $\pi_\mathbb{R}(F)$ are relatively compact sets of Ω and $\mathbb{R}$ respectively. By the assumptions in the theorem, there is an $x_0(F) \in \Omega$ that maximizes p on $\pi_\Omega(F)$. This means that a random variable $\bar{U}(F)$ with measure $\mu(x_0(F);\cdot)$ dominates $U_1|X_1 = x$ stochastically for all $x \in \pi_\Omega(F)$. Write $\bar{p}(F) = p(x_0(F))$ for the corresponding p-value of p. Furthermore, we can define $\underline{\mathrm{u}}$ as the smallest u-value attained on the whole set $\pi_{\mathbb{R}}(F)$, which again is finite by the assumption of F being relatively compact. Combining these two definitions gives us

$$\begin{aligned} nP(g_n(U_1) \in F_x | X_1 = x) &\leq nP(g_n(U_1) \geq \underline{\mathrm{u}} | X_1 = x) \\ &\leq nP(g_n(\bar{U}(F)) \geq \underline{\mathrm{u}} | X_1 = x) \\ &= nP(g_n(\bar{U}(F)) \geq \underline{\mathrm{u}}) \\ &\to \bar{p}(F)\nu_\alpha([\underline{\mathrm{u}}, \infty)) \\ &< +\infty, \end{aligned}$$

which means that $nP(g_n(U_1) \in F_x | X_1 = x)$ is uniformly bounded. Using the bounded convergence theorem, we get

$$\begin{aligned} nP((X_1, g_n(U_1)) \in F) &= \int_\Omega nP(g_n(U_1) \in F_x | X_1 = x) d\Lambda(x) \\ &\to \int_\Omega \nu_\alpha(F_x) p(x) d\Lambda(x) \\ &= (\Lambda_p \times \nu_\alpha)(F), \end{aligned}$$

which completes the proof. □

This theorem is similar to Proposition 3.21 in Resnick [17]. There are two differences. First, in [17], the author considers $\xi_n = \sum_{j=1} \delta_{(jn^{-1}, g_n(U_j))}$ where $\{U_j\}$ is a sequence of independent and identically distributed random variables. Thus, the difference is that we model the first coordinate as a random variable, and let the distribution of the second coordinate depend on this first coordinate. Furthermore, we let X take values in a general separable metric space. The differences add some technicalities to the proof, but they turn out not to affect the main result.

We can also note that the distributional assumptions ensure that the optimal choice and the maximum value are independent in the limit, which means that we can write the product measure as a direct product of measures on the two spaces. See Fosgerau et al. [6] for a general discussion of probability distributions having this invariance property.

The assumption that p is bounded on compact sets is for example satisfied whenever p is continuous. The assumption that we can construct a stochastically dominating random variable for each compact set is a technical assumption required to apply the bounded convergence theorem. As a counterexample when the theorem fails, consider the model of Example 18.1, with Λ having a uniform distribution on $\Omega = [0, 1)$, $h(x) = -\log(x)$ and $p(x) = x^{-1}$ for $x \neq 0$, whereas $h(0) = 0$ and $p(0) = 1$. In order to have convergence $\xi_n(F) \Longrightarrow \xi(F)$ for relatively compact sets $F \in \mathcal{E}$ with $\mu(\partial F) = 0$, for which the closure of the projection of F onto Ω does not contain 0, we take $a_n = 1$ and $b_n = \log(n)$. On the other hand, if $F = [0, \delta] \times [-K, K]$, it can be seen that $\xi_n(F)$ tends to infinity with probability 1 as $n \to \infty$, for any values of $0 < \delta < 1$ and $K > 0$.

18.4.2 Convergence of Choice Probability Distribution

Recall that our task is to study the limiting behavior of C_n as defined in (18.1). The key to connect this limit to point processes is the observation that because g_n is strictly increasing for all n: $C_n(A) = P(X_{[n:n]} \in A) = P(M_{\xi_n}(A) > M_{\xi_n}(A^c))$ for all $A \in \mathcal{B}(\Omega)$, with M_{ξ_n} a random field defined as $M_{\xi_n}(A) = \max_{\substack{X_i \in A \\ 1 \le i \le n}} g_n(U_i)$, $A \in \mathcal{B}(\Omega)$, where $\mathcal{B}(\Omega)$ is the Borel sigma algebra over Ω, and ξ_n is the point process from (18.7). This formulation of the argmax-measure C_n in terms of random fields defined over point processes allows us to generalize the notion of argmax to the limiting case where the number of offers goes to infinity. We will study the limiting behavior of finite-dimensional distributions of M_{ξ_n}. This will allow us to calculate the limit of C_n.

The mean intensity $\Lambda_p \times \nu_\alpha$ in Theorem 18.1 is a non-finite measure defined over $\Omega \times \mathbb{R}$. However, if $\Lambda(\Omega) < \infty$, it is possible to write $\Omega \times \mathbb{R}$ as a countable union of sets with finite measures $\Lambda_p \times \nu_\alpha$. Hence, the realization of the point process ξ has countably infinite many points almost surely. If we write $\{X_i^\infty, U_i^\infty\}_{i=1}^\infty$ for the sequence of random variables giving the locations of these points, we can define, $M_\xi(A) = \max_{i; X_i^\infty \in A} U_i^\infty$ as a random field giving the highest variable attained for a given set $A \subseteq \Omega$, and $C(A) = P(M_\xi(A) > M_\xi(A^c))$ for the probability that A will contain the largest U-element.

Proposition 18.3 *If $\Lambda_p(\Omega) < \infty$, we have $C(A) = \Lambda_p(A)/\Lambda_p(\Omega)$.*

Proof Suppose first that $\Lambda_p(A^c) = 0$ or $\Lambda_p(A) = 0$. In this case, it is clear that we have $C(A) = 1$ or $C(A) = 0$ respectively as required by the formula for $A \in \mathcal{B}(\Omega)$. Indeed, using the convention that the supremum of an empty set is minus infinity, if $\Lambda_p(A) = 0$, then $M_\xi(A) = -\infty$ almost surely. As $M_\xi(A^c) > -\infty$ almost surely, we will get $C(A) = 0$. A similar reasoning applies to A^c.

Furthermore, since ξ is a Poisson random measure with mean measure $\Lambda_p \times \nu_\alpha$, we note that if $\Lambda_p(\Omega) < \infty$ we have that $M_\xi(A)$ and $M_\xi(A^c)$ are two independent, proper random variables with

$$P(M_\xi(A) \le y) = P(\xi(A \times (y, \infty)) = 0) = e^{-\Lambda_p(A)\nu_\alpha((y,\infty))} \qquad (18.13)$$

$$P(M_\xi(A^c) \le y) = P(\xi(A^c \times (y, \infty)) = 0) = e^{-\Lambda_p(A^c)\nu_\alpha((y,\infty))}. \qquad (18.14)$$

Standard calculations yield

$$P(M_\xi(A) > M_\xi(A^c)) = \frac{\Lambda_p(A)}{\Lambda_p(A) + \Lambda_p(A^c)} = \Lambda_p(A)/\Lambda_p(\Omega)$$

and the proof is complete. □

From this result, we automatically get that C is a probability measure as it is a normalized version of Λ_p which is a finite measure.

In order to prove that C_n converges weakly, we need some additional results. We use that

$$\nu_1 \ll \mu_1 \text{ and } \nu_2 \ll \mu_2 \Rightarrow \nu_1 \times \nu_2 \ll \mu_1 \times \mu_2, \tag{18.15}$$

where $\ll$ means "absolutely continuous with respect to".

We will also use that if ξ_n are point processes, ξ is a Poisson process, and

$$\xi_n \Rightarrow_p \xi,$$

then

$$P(\xi_n(F) = 0) \to P(\xi(F) = 0) \tag{18.16}$$

for all $F \in \mathcal{E}$ with $\mu(\partial F) = 0$, where μ is the intensity measure of ξ.

After these preliminaries, we are ready to state our second main result:

Theorem 18.2 *If $\Lambda_p(\Omega) < \infty$, we have*

$$C_n(\cdot) \Rightarrow C(\cdot) = \frac{\Lambda_p(\cdot)}{\Lambda_p(\Omega)}. \tag{18.17}$$

Proof Assume we have A with $C(\partial A) = 0$. We aim to prove that $C_n(A) \to C(A)$. By Proposition 18.3, C and Λ_p are equivalent, and we have $\Lambda_p(\partial A) = 0$. Noting that the result is clearly true whenever $\Lambda_p(A) = 0$ or $\Lambda_p(A^c) = 0$, we can assume that both are different from 0. By (18.13) and (18.14), this means that $(M_\xi(A), M_\xi(A^c))$ is a proper random variable on $\mathbb{R}^2$, and we will show that $(M_{\xi_n}(A), M_{\xi_n}(A^c))$ jointly converge weakly to this random variable. Indeed, consider

$$\begin{aligned} P(M_{\xi_n}(A) \le x_1, M_{\xi_n}(A^c) \le x_2) &= P\left(\xi_n(A \times (x_1, \infty) \cup A^c \times (x_2, \infty)) = 0\right) \\ &\to P\left(\xi(A \times (x_1, \infty) \cup A^c \times (x_2, \infty)) = 0\right) \\ &= P(M_\xi(A) \le x_1, M_\xi(A^c) \le x_2) \\ &= F_{M_\xi(A), M_\xi(A^c)}(x_1, x_2). \end{aligned}$$

The convergence step uses (18.16) and that

$$\partial\left(A \times (x_1, \infty) \cup A^c \times (x_2, \infty)\right) \subset \partial A \times (\min(x_1, x_2), \infty) \cup A \times (\{x_1\} \cup \{x_2\})$$

and we have $(\Lambda_p \times \nu_\alpha)(\partial A \times (\min(x_1, x_2), \infty) \cup A \times (\{x_1\} \cup \{x_2\})) = 0$ since $\Lambda_p(\partial A) = 0$ and $\nu_\alpha(\{x_1\} \cup \{x_2\}) = 0$, where $\Lambda_p \times \nu_\alpha$ is the intensity measure of ξ. Hence

$$(M_{\xi_n}(A), M_{\xi_n}(A^c)) \Rightarrow (M_\xi(A), M_\xi(A^c)).$$

Defining

$$D = \{(a, b) \in \mathbb{R}^2 : a > b\}$$

and using (18.15), with $\nu_1 \sim M_\xi(A)$, $\nu_2 \sim M_\xi(A^c)$, and μ_1, μ_2 Lebesgue measure in $\mathbb{R}$, to conclude that

$$P((M_\xi(A), M_\xi(A^c)) \in \partial D) = 0$$

we get

$$\begin{aligned} C_n(A) &= P(M_{\xi_n}(A) > M_{\xi_n}(A^c)) \\ &= P((M_{\xi_n}(A), M_{\xi_n}(A^c)) \in D) \\ &\to P((M_\xi(A), M_\xi(A^c)) \in D) \\ &= C(A) \end{aligned}$$

and the proof is complete. □

18.5 Examples

Here we provide a few examples to illustrate our theory.

Example 18.3 (*Exponential and mixture models*) This example extends Example 18.1, and calculates the argmax distribution associated with that example. Consider a family of models where the regular conditional probability measure $\mu(x; \cdot)$ is indexed by α, and where for each $A \in \mathcal{B}(\mathbb{R})$ we have

$$\mu_\alpha(x; A) = \begin{cases} P\left((2 \times 1_{\{V_1 < p(x)\}} - 1)(1 - V_2^{-1/\alpha}) \in A\right), & \alpha < 0, \\ P\left(\log(p(x)/V_1) \in A\right), & \alpha = 0, \\ P\left((2 \times 1_{\{V_1 < p(x)\}} - 1)V_2^{-1/\alpha} \in A\right), & \alpha > 0, \end{cases}$$

where $V_1, V_2 \sim U(0, 1)$ are two independent and uniformly distributed random variables on $(0, 1)$. A bit less formal, we may write

$$\mu_\alpha(x) \sim \begin{cases} -(1 - p(x))\mathrm{Beta}(1, -\alpha) + p(x)\mathrm{Beta}(1, -\alpha), & \alpha < 0, \\ \mathrm{Exp}(\log(p(x)), 1), & \alpha = 0, \\ -(1 - p(x))\mathrm{Pareto}(\alpha, 1) + p(x)\mathrm{Pareto}(\alpha, 1), & \alpha > 0, \end{cases}$$

where $\mathrm{Beta}(a, b)$ refers to a beta distribution with density $Cx^{a-1}(1 - x)^{b-1}$ on $(0, 1)$, $\mathrm{Exp}(a, b)$ is a shifted exponential distribution with location parameter a and scale parameter b, having distribution function $1 - e^{-(x-a)/b}$ for $x \geq a$, $\mathrm{Pareto}(\alpha, b)$ is a Pareto distribution with shape parameter α and scale parameter b, corresponding to a distribution function $1 - (x/b)^{-\alpha}$ for $x \geq b$. We let $x_0 \in \Omega$ be an arbitrary point for which $p(x_0) = 1$.

We have chosen the parameter α for $\mu_\alpha(x, \cdot)$ in a way so that (18.3) holds, with $a_n = n^{1/\alpha}$, $b_n = 1$ when $\alpha < 0$, $a_n = 1$, $b_n = \log(n)$ when $\alpha = 0$, and $a_n = n^{1/\alpha}$, $b_n = 0$ when $\alpha > 0$. When $\alpha = 0$, this follows from tail properties of the exponential distribution, as shown in Example 18.1. For $\alpha \neq 0$, we have that

$$\begin{aligned}\mu_\alpha(x; (-\infty, b_n + a_n u])^n &= \{1 - p(x)(1 - \mu_\alpha(x_0; (-\infty, b_n + a_n u]))\}^n \\ &\to G_\alpha(u)^{p(x)}.\end{aligned}$$

In the last step we used that $\mu_\alpha(x_0, (-\infty, b_n + a_n u])^n \to G_\alpha(u)$. This is a well known fact of univariate extreme value theory (see for instance Fisher and Tippett [5], Gnedenko [7], and Chapter 1 in Resnick [17]), and it follows from tail properties of the beta and Pareto distributions.

This means that for all these three families of distribution, the choice probabilities will give us a tilted distribution $p \times \Lambda$ which modifies the underlying Λ-distribution with p. This effect captures that areas with a high deterministic utility component p are relatively more likely to get chosen. For the case $\alpha = 0$ this effect means that if utility is given by $U_i = h(x_i) + \varepsilon_i$, where $\varepsilon_i \sim Exp(1)$, then the choice distribution is an exponential tilt $e^{h(x)}\Lambda(dx)$ of the original distribution.

Example 18.4 (*An example from the commuting literature*) If we focus on $\alpha = 0$ in the previous example, we have an interesting special case. Suppose that a person has received a new job, and potential residencies are distributed uniformly on $B(0, R)$, a disk in $\mathbb{R}^2$. There is a linear cost $c||x||$ associated with travelling to a location $x \in B(0, R)$, and there is an exponentially distributed random component associated with each residence. This means that utility is given by $U|X = x \sim \text{Exp}(-c||x||, 1)$, where $||x||$ is the Euclidean distance from the origin. This gives a very simple model to think about commuting choices. In this case, Λ has a uniform distribution on $B(0, R)$, and $p(x) = \text{Exp}(-c||x||)$. Thus, we get $C(A) = \dfrac{\int_A e^{-c||x||}dx}{\int_{B(0,R)} e^{-c||x||}dx}$. The particular direction of commuting is often not as interesting as the distribution of distances. The probability that we commute less than r is given by

$$C(\{x : ||x|| \le r\}) = \frac{\int_0^r se^{-cs}ds}{\int_0^R se^{-cs}ds},$$

which we recognize as a truncated Gamma$(2, 1/c)$-distribution.

There is suggestive evidence that travel patterns follow a gamma distribution over short distances. One good source is Burke and Brown [3], which documents the distances people walk for transport purposes to different destinations. The data was collected from a survey in Brisbane. Even though the investigators not only measured the time walked to work, the situation is somewhat analogous to the example above in that walking is a roughly linear cost. They found that the distance walked for one-leg trips is very close to a gamma distribution with shape parameter α and scale parameter β, and the same for the total distance walked from train stations to end destinations (see Figs. 18.1 and 18.2).

We see that the estimated parameters $(\hat{\alpha}, \hat{\beta})$ are $(1.42, 0.66)$ and $(2.13, 0.37)$ respectively. The estimated shape parameter is close to but not exactly 2 as would be

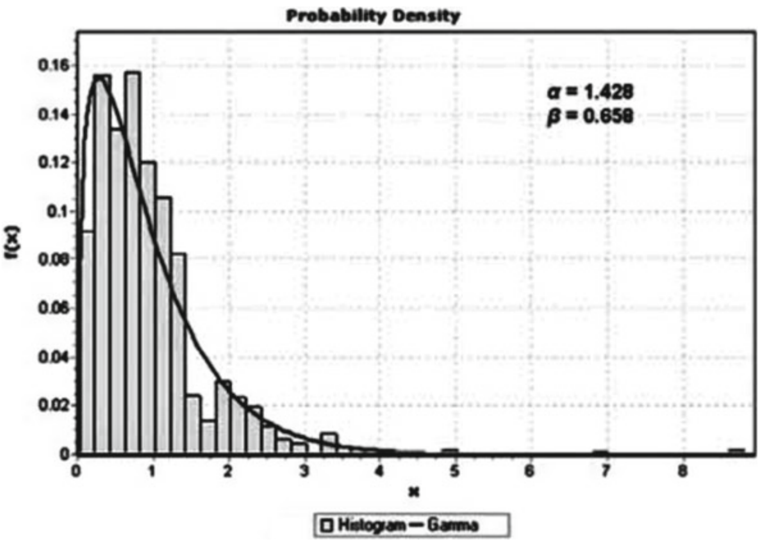

Fig. 18.1 Histogram over walking distances to final destination and fitted Gamma(α, β)-distribution (Burke and Brown [3])

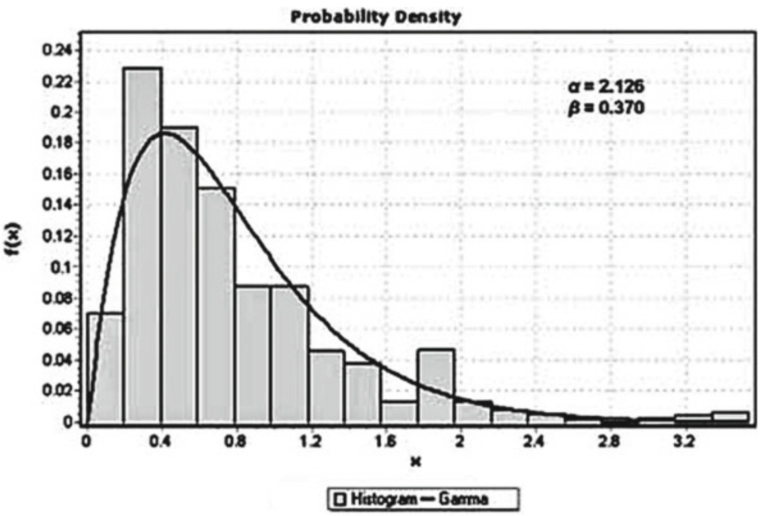

Fig. 18.2 Histogram over walking distances from train station to final destination and fitted Gamma(α, β)-distribution (Burke and Brown, [3])

predicted by the theory. The focus in the paper is to test the distributional assumption rather than to find the exact parameters, and the authors report an Anderson-Darling test but no standard errors on the parameter estimates. Hence, we do not know if α is significantly different from 2.

Example 18.5 (*The logit model: a special case*) Let Λ be a uniform distribution on the finite support $\{x_1, \ldots, x_{n_0}\}$. As in Example 18.1, let utilities be given by

$$U_j | X_j = x \sim \mathrm{Exp}(h(x), 1). \tag{18.18}$$

This corresponds to $p(x_i) = e^{h(x_i)}$ and we get

$$C(\{x_i\}) = \frac{e^{h(x_i)}}{\sum_{j=1}^{n_0} e^{h(x_j)}}.$$

This corresponds to the famous logit model from the random choice literature (McFadden [15]).

The following is an example where we let Ω be a functional space. This shows that the methodology can be applied to more general spaces than subsets of Euclidean space, and motivates the more general space definition we introduced in Sect. 18.2.

Example 18.6 Let Ω be the space of bounded functions on [0, 1], metrized by the sup-norm. A function $x \in \Omega$ describes a continuum of choice characteristics. An agent values a function $x \in \Omega$ by sampling k points of [0, 1] according to a density function g, and valuing them according to their sum and an exponentially distributed noise term on each observation. In this case X is a random variable taking values in Ω with law Λ. Algebraically,

$$\begin{aligned} U | X = x &= \frac{\sum_{j=1}^{k} x(T_j)}{k} + \frac{\sum_{j=1}^{k} \varepsilon_j}{k} \\ &= h_k(x) + \varepsilon \end{aligned}$$

where T_j are i.i.d. distributed on [0, 1] with density function g, and $\varepsilon_j \sim \mathrm{Exp}(1)$ independently. We want to find the argmax distribution on Ω. We will treat a sequence of approximations as equalities, and verify ex post that such a treatment is justified.

The random variable ε has a Gamma$(k, 1/k)$ distribution, which means that

$$\bar{F}_\varepsilon(z) \equiv 1 - F_\varepsilon(z) = \sum_{m=0}^{k-1} e^{-kz} \frac{(kz)^m}{m!} \sim \frac{e^{-kz}(kz)^{k-1}}{(k-1)!},$$

where the ratio of the last two expressions tends to 1 when z gets large. Now, we use that x is bounded to get $\underline{y}(x) \leq \inf_{t \in [0,1]} x(t) \leq \sup_{t \in [0,1]} x(t) \leq \bar{y}(x)$. We write $\mu(x; \cdot)$ for the law of $U | X = x$, and approximate the upper tail when u is large:

$$
\begin{aligned}
1-\mu\,(x;(-\infty,u]) &= 1-\int_{\underline{y}(x)}^{\bar{y}(x)} F_\varepsilon(u-y)dF_{h_k(x)}(y)\\
&= \int_{\underline{y}(x)}^{\bar{y}(x)} \bar{F}_\varepsilon(u-y)dF_{h_k(x)}(y) \sim \int_{\underline{y}(x)}^{\bar{y}(x)} \frac{(k(u-y))^{k-1}}{(k-1)!}e^{-k(u-y)}dF_{h_k(x)}(y)\\
&\sim \frac{(ku)^{k-1}}{(k-1)!}\int_{\underline{y}(x)}^{\bar{y}(x)} e^{-k(u-y)}dF_{h_k(x)}(y) \sim \bar{F}_\varepsilon(u)p_k(x),
\end{aligned}
$$

where $p_k(x)$ is the moment generating function of $F_{h_k(x)}$ with argument k. We write $\eta_k(u) = \bar{F}_\varepsilon(u)p_k(x) - (1-\mu\,(x;(-\infty,u]))$ for the approximation error. Now define $a_n = 1/k$ and $b_n = \bar{F}_\varepsilon^{-1}(1/n)$, which gives us

$$
\begin{aligned}
\mu(x;(-\infty,a_n u+b_n])^n &= \left(1-\bar{F}_\varepsilon\left(\tfrac{u}{k}+b_n\right)p_k(x)+\eta_k\left(\tfrac{u}{k}+b_n\right)\right)^n\\
&\sim \left(1-\bar{F}_\varepsilon\left(\tfrac{u}{k}+b_n\right)p_k(x)\right)^n\\
&\sim \left(1-\frac{e^{-k\left(\frac{u}{k}+b_n\right)}k^{k-1}\left(\frac{u}{k}+b_n\right)^{k-1}}{(k-1)!}p_k(x)\right)^n\\
&\sim \left(1-p_k(x)e^{-u}\bar{F}_\varepsilon(b_n)\right)^n\\
&= \left(1-\frac{p_k(x)e^{-u}}{n}\right)^n\\
&\to (e^{-e^{-u}})^{p_k(x)} = G_0(u)^{p_k(x)}
\end{aligned}
$$

as $n\to\infty$. It follows that (18.3) holds with $\alpha=0$ and $p(x)$ replaced by $p_k(x)$, provided our approximations are justified. In particular, we need $\lim_{n\to\infty} n\bar{F}_\varepsilon(u/k+b_n) = e^{-u}$ and $\lim_{n\to\infty} n\times\eta_k\left(\frac{u}{k}+b_n\right)=0$. The first of these two equations follows from

$$
\begin{aligned}
\lim_{n\to\infty} n\bar{F}_\varepsilon(u/k+b_n) &= \lim_{n\to\infty} n\bar{F}_\varepsilon(b_n)\frac{\sum_{m=0}^{k-1}e^{-kb_n}k^m(u/k+b_n)^m/m!}{\sum_{m=0}^{k-1}e^{-kb_n}k^m b_n^m/m!}e^{-u}\\
&= e^{-u}
\end{aligned}
$$

as $\bar{F}_\varepsilon(b_n)=1/n$, $b_n\to\infty$ and u/k is bounded, and for the second equation we use that

$$
\begin{aligned}
\lim_{n\to\infty} n\eta_k\,(u/k+b_n) &= \lim_{n\to\infty}\left|n\left\{\bar{F}_\varepsilon(u/k+b_n)p_k(x)-(1-\mu\,(x;(-\infty,u/k+b_n]))\right\}\right|\\
&= \lim_{n\to\infty}\left|n(\bar{F}_\varepsilon(u/k+b_n)\right|\left|\frac{\bar{F}_\varepsilon(u/k+b_n)p_k(x)-\int_{\underline{y}}^{\bar{y}}\bar{F}_\varepsilon(u/k+b_n-y)dF_{h_k(x)}(y)}{\bar{F}_\varepsilon(u/k+b_n)}\right|\\
&= 0.
\end{aligned}
$$

The first term on the second line is bounded and it can be checked that the second term converges to zero, and our result follows.

Given that Assumption 1 holds with $p(x)$ replaced by $p_k(x)$, we get the argmax measure

$$
C(A;k) = \frac{\int_A p_k(x)d\Lambda(x)}{\int_\Omega p_k(x)d\Lambda(x)}.
$$

This measure has a nice consistency property when $k \to \infty$. Indeed, by the Law of Large Numbers, as $k \to \infty$ the probability distribution of $h_k(x)$ converges to a point mass at $h(x) = E_g(x) = \int_0^1 g(s)x(s)ds$, so $p_k(x) \sim e^{kh(x)}$. We define the maximum value that h attains as

$$\bar{h} = \sup_{h'} \left\{ h' : \Lambda(x : h(x) \leq h') < 1 \right\}.$$

This definition ensures that $A_\delta = \{x \in \Omega : h(x) > \bar{h} - \delta\}$ has non-zero Λ-measure for every δ.

Now, this means that $\lim_{k\to\infty} \dfrac{C(A_\delta; k)}{C(\Omega - A_{2\delta}; k)} \geq \lim_{k\to\infty} \dfrac{e^{k(\bar{h}-\delta)}\Lambda(A_\delta)}{e^{k(\bar{h}-2\delta)}\Lambda(\Omega - A_{2\delta})} = \infty,$

Hence $\lim_{k\to\infty} C(A_\delta; k) = 1$ for all $\delta > 0$. We can interpret this as when k grows, the choice becomes less random from the point of view of the statistician and the agent will choose the option x with the highest expected value $h(x)$ with probability 1.

18.6 Extension

We have derived a way to calculate the asymptotic behavior of the best choice $C_n = X_{[n:n]}$, and have done so for a number of assumptions on the joint distribution of (X_i, U_i) of characteristics and values. However, in order to extend our results to a wider range of distributional assumptions, we must relax the requirement that $X_{[n:n]}$ should converge to a non-degenerate distribution. For example, when X and U are distributed bivariate normally with positive correlation, $|X_{[n:n]}| \to \infty$ almost surely, whereas for other models, $X_{[n:n]}$ converges to a one-point distribution.

In these cases, it can nevertheless be possible to find a sequence of functions h_n such that $h_n(X_{[n:n]}) \Rightarrow C$ for a non-degenerate random variable C. In this case, we would have $X_{[n:n]} \overset{d}{\approx} h_n^{-1}(C)$ for large n, where $\overset{d}{\approx}$ means that the two random variables have approximately the same distribution.

We have done some exploratory studies on this extension, and there are indications that for a much larger class of distributions than studied in the present paper, it is possible to find sequences h_n and g_n such that $\sum_{i=1}^{n} \delta_{(h_n(X_i), g_n(U_i))} \Rightarrow_p \xi$ for some non-degenerate Poisson process ξ with intensity measure μ on $E = \Omega \times \mathbb{R}$. The asymptotic argmax distribution of $h_n(X_{[n:n]})$ is then $C(A) = \int_{\mathbb{R}} \dfrac{\mu(A, dx)}{\mu(\Omega, dx)} F_U(dx)$, for all $A \in \mathcal{B}(\Omega)$, where $U = M_\xi(\Omega)$ is the maximum utility of ξ. In particular, if $\mu = \Lambda_p \times \nu$, this argmax distribution coincides with the one in Theorem 18.2. We also conjecture that this extension can be connected to the theory of conditional extreme values, as discussed in Heffernan and Tawn [8].

18.7 Conclusion

We have shown that point process theory can be used to derive continuous approximations of discrete choice problems with a large number of options. When the random component of utility is exponentially distributed, or a convex linear combination of beta or Pareto distributions, we have derived analytical solutions to the approximation problem. Potential applications involve commuting choices, and we have provided suggestive evidence that some observed commuting flows distributions can be justified within our framework.

However, there is still a need to generalize the theory to allow for more flexible distributional assumptions. Essentially, functional forms outside our assumed domain might lead to all choices asymptotically diverging, or asymptotically collapsing on one point. For example, if the tail of utility is too thin, the distribution of choices will converge to the set of values with the highest deterministic utility value. In other cases, $C_n(A) \to 0$ for any compact A, and the choice probabilities will drift to infinity. In Sect. 18.6, we have outlined a potential extension of that would allow for a more flexible set of assumptions on the distribution of utilities. The idea is to renormalize the characteristics space to analyze the rate at which choice probabilities converge or diverge as the number of points $n \to \infty$.

Acknowledgements The authors wish to thank Dmitrii Silvestrov for valuable comments on the manuscript. Ola Hössjer's work was supported by the Swedish Research Council, contract nr. 621-2008-4946, and the Gustafsson Foundation for Research in Natural Sciences and Medicine.

References

1. Anderson, S.P., De Palma, A., Thisse, J.-F.: Discrete Choice Theory of Product Differentiation. MIT Press, Cambridge (1992)
2. Ben-Akiva, M., Leman, S.: Discrete Choice Analysis: Theory and Applications to Travel Demand, vol. 9. MIT Press, Cambridge (1985)
3. Burke, M., Brown, A.L., et al.: Distances people walk for transport (2007)
4. Cox, D.R., Isham, V.: Point Processes, vol. 24. Chapman and Hall/CRC, London (1980)
5. Fisher, R.A., Tippett, L.H.C.: Limiting forms of the frequency distribution of the largest or smallest member of a sample. Mathematical Proceedings of the Cambridge Philosophical Society, vol. 24, pp. 180–190. Cambridge University Press, Cambridge (1928)
6. Fosgerau, M., Lindberg, P.O., Mattsson, L.-G., Weibull, J.: Invariance of the distribution of the maximum. Technical report, MPRA Paper No. 63529, University Library of Münich, Germany (2015)
7. Gnedenko, B.V.: Limit theorems for the maximum term of variational series. Dokl. Akad. Nank SSSR (1941)
8. Heffernan, J.E., Tawn, J.A.: A conditional approach for multivariate extreme values (with discussion). J. R. Stat. Soc. Ser. B **66**(3), 497–546 (2004)
9. Jacobsen, M.: Point Process Theory and Applications: Marked Point and Piecewise Deterministic Processes. Birkhäuser, Boston (2005)
10. Karpiarz, M., Fronczak, P., Fronczak, A.: International trade network: fractal properties and globalization puzzle. Phys. Rev. Lett. **113**(24), 248701 (2014)

11. Ledford, A.W., Tawn, J.A.: Concomitant tail behaviour for extremes. Adv. Appl. Probab. **30**(1), 197–215 (1998)
12. Luce, R.D.: Individual Choice Behavior a Theoretical Analysis. Wiley, New York (1959)
13. Maddison, C.J., Tarlow, D., Minka, T.: A^* sampling. Adv. Neural Inf. Process. Syst. **27**, 3086–3094 (2014)
14. Malmberg, H., Hössjer, O.: Probabilistic choice with an infinite set of options: an approach based on random sup measures. In: Silvestrov, D., Martin-Löf, A. (eds.) Modern Problems in Insurance Mathematics, pp. 291–312. Springer, Berlin (2014)
15. McFadden, D.I.: Econometric models for probabilistic choice among products. J. Bus. **53**(3), 13–29 (1980)
16. O'Brien, G.L., Torfs, P.J.J.F., Vervaat, W.: Stationary self-similar extremal processes. Probab. Theory Relat. Fields **87**(1), 97–119 (1990)
17. Resnick, S.I.: Extreme Values, Regular Variation, and Point Processes. Springer, New York (2007)
18. Resnick, S.I., Roy, R.: Random USC functions, max-stable processes and continuous choice. Ann. Appl. Probab. **1**(2), 267–292 (1991)
19. Stoev, S.A., Taqqu, M.S.: Extremal stochastic integrals: a parallel between max-stable processes and α-stable processes. Extremes **8**(4), 237–266 (2005)
20. Train, K.E.: Discrete Choice Methods with Simulation, 2nd edn. Cambridge University Press, Cambridge (2009)