The probability is simulated by drawing from the distribution, calculating the function for each draw, and averaging the results. We give two examples below, but researchers will inevitably develop others that meet the needs of their particular projects, such as Bhat's (1999) use of mixed ordered logit.

7.6.1 Mixed Nested Logit

The mixed logit model does not exhibit the independence from irrelevant alteratives property of logit, and can approximate any substitution pattern by appropriate specification of variables and mixing distribution. This fact has lead some people to feel that there is no further need for nested logit models. A mixed logit can be estimated that provides an analogous correlation/substitution patterns as a nested logit. For example, consider a nested logit with two nests of alternatives labeled A and B. Provided the logsum coefficients are between 0 and 1, substitution within each nest is greater than substitution across nests. This substitution pattern can be represented in a mixed logit model by specifying a dummy variable for each nest and allowing the coefficients on the dummies to be random (constraining, for identification purposes, the means to be zero if a full set of alternative-specific constants are included and the two variances to be the same.)

While a mixed logit can be specified in this way, doing so misses the point of simulation. As discussed in Chapter 1, simulation is used as a way to approximate integrals when a closed form does not exist. Analytic integration is always more accurate than simulation and should be used whenever feasible, unless there is a compelling reason to the contrary. Using a mixed logit to represent the substitution patterns of a nested logit, while feasible, replaces the closed-form integral of the nested logit with an integral that needs to be simulated. From a numerical perspective, this replacement can only reduce accuracy. The only possible advantages of mixed logit in this context are that (1) it might be easier for the researcher to test numerous nesting structures, including overlapping nests, within a mixed logit than a nested logit, and (2) the researcher might specify other coefficients to be random such that a mixed logit is already being used.

The second reason suggests a mixed nested logit. Suppose the researcher believes that some of the coefficients in the model are random and also that, conditional on these coefficients, the unobserved factors are correlated over alternatives in a way that can be represented by a nested logit. A mixed nested logit model can be specified to represent this situation. Conditional on the coefficients that enter utility, the choice probabilities are nested logit, which is a close-form and can be calculated exactly. The unconditional probability is the nested logit formula integrated over the distribution of the the random coefficients. Software for mixed logit can be modified by simply locating the logit formula within the code and changing it to the appropriate nested logit formula. Experience indicates that maximizing the likelihood function for unmixed nested logits is often difficult numerically, and mixing the model will compound this difficulty. Hierarchical Bayes estimation (Chapter 12) could prove particularly useful in this situation, since it does not involve maximizing the likelihood function.

7.6.2 Mixed Probit

A constraint of probit models, and in fact their defining characteristic, is all random terms enter utility linearly and are randomly distributed such that utility itself is normally distributed. This constraint can be removed by specifying a mixed probit. Suppose that some random terms enter non-linearly or are not randomly distributed, but that *conditional* on these, utility is normally distributed. For example, a price coefficient might be lognormal to assure that it is negative for all people, and yet all other coefficients are either fixed or normal and the final error terms are jointly normal. A mixed probit model is appropriate for this specification. Conditional on the price coefficient, the choice probabilities follow the standard probit formula. The unconditional probabilities are the integral of this probit formula over the distribution of the price coefficient. Two layers of simulation are used to approximate the probabilities: (1) a draw of the price coefficient is taken, and (2) for this draw, the GHK or other probit simulator is used to approximate the conditional choice probability. This process is repeated many times and the results are averaged.

Long run times can be expected for the mixed probit model since the GHK simulator is calculated for each draw of the price coefficient. However, the number of draws in the GHK simulator can be reduced, since the averaging over draws of the price coefficient reduces the variance generated by the GHK simulator. In principle, the GHK simulator can be based on only one draw for each draw of the price coefficient. In practice, it might be advisable to use more than one draw but far fewer than would used in a unmixed probit.

The mixed probit model provides a way for the researcher to avoid some of the practical difficulties that can arise with a mixed logit model. For example, to represent pure heteroskedasticity (*i.e.,* a different variance for each alternative's utility) or a fixed correlation pattern among alternatives (*i.e.*, a covariance matrix that does not depend on the variables), it can often be easier to estimate a probit instead of specifying numerous error components within a mixed logit. As emphasized by Ben-Akiva et al. (2001), specification of covariance and heteroskedasticity can be more complex in a mixed logit model than a probit, due to the fact that iid extreme value terms are necessarily added to whatever other random elements the researcher specifies. Probit is a more natural specification in these situations. However, if the researcher wants to include some non-normal random terms, an unmixed probit cannot be used. Mixing the probit allows the researcher to include non-normal terms while still maintaining the simplicity of probit's representation of fixed covariance for additive errors. Conceptually, the specification and estimation procedure are straightforward. The cost comes only in extra computation time, which becomes less relevant as computers get faster.

7.7 Dynamic optimization

In previous chapters we examined certain types of dynamics, by which choices in one period affect choices in another period. For example, we described how a lagged dependent variable can be included to capture inertia or variety-seeking behavior. These discussions suggest a much wider realm of dynamics than we had actually considered. In particular: if past choices affect current choices, then current choice affect future choices, and a decision-maker who is aware of this fact will take these future impacts into consideration. A link from the past to the present necessarily implies a link from the present to the future.

In many situations, the choices that a person makes at one point in his life have profound impact on the options that are available to him in the future. Going to college, while expensive and sometimes irritating, enhances future job possibilities. Saving money now allows a person to buy things later than he otherwise would not be able to afford. Going to the gym today means that we can skip going tomorrow. Most of us take future impacts like these into consideration when choosing among current alternatives.

The question is: how can behavior such as this be represented in discrete choice models? In general the situation can be described as follows. A person makes a series of choices over time. The alternative that is chosen in one period affects the attributes and availability of alternatives in the future. Sometimes the future impacts are not fully known, or depend on factors that have not yet transpired (such as the future state of the economy). However, the person knows that he will, in the future, maximize utility among the alternatives that are available at that time under the conditions that prevail at that time. This knowledge enables him to choose the alternative in the current period that maximizes his expected utility over the current and future periods. The researcher recognizes that the decision-maker acts in this way, but does not observe everything that the decision-maker considers in the current and future periods. As usual, the choice probability is an integral of the decision-maker's behavior over all possible values of the factors that the researcher does not observe.

In this section we specify models in which the future consequences of current decisions are incorporated. For these models, we will assume that the decision-maker is fully rational in the sense that he optimizes perfectly in each time period given the information that is available to him at that point in time and given that he knows he will act optimally in the future when future information is revealed. The procedures for modeling these decisions were first developed for various applications by, for example,Wolpin (1984) on women's fertility, Pakes (1986) on patent options, Wolpin (1987) on job search, Rust (1987) on engine replacement, Berkovec and Stern (1991) on retirement, and others. Eckstein and Wolpin (1989) provide an excellent survey of these early contributions. The thrust of more recent work has primarily between toward solving some of the computational difficulties that can arise in these models, as discussed below.

Before embarking on this endeavor, it is important to keep the concept of rationality in perspective. A model of rational decision-making over time does not necessarily represent behavior more accurately than a model of myopic behavior, where the decision-maker ignores future consequences. In fact, the truth in a given situation might lie between these two extremes: decision-makers might be acting in ways that are neither completely myopic nor completely rational. As we will see, the truly optimizing behavior is very complex. People might engage in behavior that is only approximately optimal simply because they (we) can't figure out the truly optimal way to proceed. Viewed in another light, one could argue that people always optimize when the realm of optimization is broadened sufficiently. For example, rules of thumb or other behavior that seem only to approximate optimality might actually be optimal when the costs of optimization are considered.

The concepts and procedures that are developed to examine optimizing behavior carry over, in modified form, to other types of behavior that recognize future impacts from current choices. Furthermore, the researcher can often test alternative behavioral representations. Myopic behavior nearly always appears as a testable restriction on a fully rational model, namely, a zero coefficient for the variable that captures future impacts. Sometimes, the standard rational model is a restriction on a supposedly non-rational one. For example, O'Donoghue and Rabin (1999), among others, argue that people are time inconsistent: when it is Monday, we weigh the benefits and costs that will come on, say, Wednesday only marginally more than those that will arrive on Thursday, and yet when Wednesday actually arrives, we weigh Wednesday's (today's) benefits and costs far more than Thursday's. Essentially, we have a bias for the present. The standard rational model, where the same discount rate is used between any two periods independent of whether the person is in one of the periods, constitutes a restriction on the time-inconsistent model.

The concepts in this area of analysis are more straightforward than the notation. To develop the concepts with a minimum of notation, we will start with a two period model in which the decision-maker knows the exact impact of first-period choices on the second-period alternatives and utilities. We will then expand the model to more periods and to situations where the decision-maker faces uncertainty about future impacts.

7.7.1 Two-periods, no uncertainty about future impacts

To make the explication concrete, consider a high school student's choice of whether or not to go to college. The choice can be examined in the context of two periods: the college years and the post-college years. In the first period, the student either goes to college or not. Even though these are called the college years, the student need not go to college but can take a job instead right out of high school. In the second period the student chooses among the jobs that are available to him at that time. Going to college during the "college years" means less income during that period but better job options in the post-college years. U_{1C} is the utility that the student obtains in period 1 from going to college, and U_{1W} is the utility he obtains in the first period if he works in the first period instead of going to college. If the student were myopic, he would choose college only if $U_{1C} > U_{1W}$. However, we assume that he is not myopic. For the second period, let J denote the set of all possible jobs. The utility of job j in period 2 is U_{2j}^C if the student went to college and U_{2j}^W if he worked in the first period. The utility from a job depends on the wage that the person is paid as well as other factors. For many jobs, people with a college degree are paid higher wages and granted greater autonomy and responsibility. For these jobs, $U_{2j}^C > U_{2j}^W$. However, working in the first period provides on-the-job experience that commands higher wages/responsibility than a college degree for some jobs; for these, $U_{2j}^W > U_{2j}^C$. A job not being available is represented as having a utility of negative infinity. For example, if job j is available only to college graduates, then $U_{2j}^W = -\infty$.

How will the high school student decide whether to go to college? We assume for now that the student knows U_{2j}^C and U_{2j}^W for all $j \in J$ when deciding whether to go to college in the first period. That is, the student has perfect knowledge of his future options under whatever choice he makes in the first period. We will later consider how the decision-process changes when the student is uncertain about these future utilities. The student knows that when the second period arrives he will choose the job that provides the greatest utility. That is, he knows in the first period that the utility that he will obtain in the second period if he chooses college in the first period is the maximum of U_{2j}^C over all possible jobs. We label this utility as $U_2^C = max_j(U_{2j}^C)$. The student therefore realizes that, if he chooses college in the first period, his total utility over both periods will be:

$$\begin{aligned} TU_C &= U_{1C} + \lambda U_2^C \\ &= U_{1C} + \lambda max_j(U_{2j}^C) \end{aligned}$$

where λ reflects the relative weighting of the two periods' utilities in the student's decision-process. Given the way we have defined time periods, λ incorporates the relative time spans of each period as well as the traditional discounting of future utility relative to current utility. Thus, λ can exceed one, even with discounting, if the second period represents say forty years while the first period is four years. Myopic behavior is represented as $\lambda = 0$.

The same logic is applied to the option of working in the first period instead of going to school. The student knows that he will choose the job that offers the greatest utility such that $U_2^W = max_j(U_{2j}^W)$ and the total utility over both period from choosing to work in the first period is

$$\begin{aligned} TU_W &= U_{1W} + \lambda U_2^W \\ &= U_{1W} + \lambda max_j(U_{2j}^W). \end{aligned}$$

The student chooses college if $TU_C > TU_W$ and otherwise chooses to work in the first period.

This completes the description of the decision-maker's behavior. We now turn to the researcher. As always, the researcher observes only some of the factors that affect the student's utility. Each utility in each period is decomposed into an observed and unobserved component:

$$\begin{aligned} U_{1C} &= V_{1C} + \varepsilon_{1C} \\ U_{1W} &= V_{1W} + \varepsilon_{1W} \end{aligned}$$

and

$$\begin{aligned} U_{2j}^C &= V_{2j}^C + \varepsilon_{2j}^C \\ U_{2j}^W &= V_{2j}^W + \varepsilon_{2j}^W \end{aligned}$$

for all $j \in J$. Collect the unobserved components into vector $\varepsilon = \langle \varepsilon_{1C}, \varepsilon_{1W}, \varepsilon_{2j}^C, \varepsilon_{2j}^W, \ \forall j \rangle$, and denote the density of these terms as $f(\varepsilon)$. The probability of the student choosing college is

$$\begin{aligned} P_C &= Prob(TU_C > TU_W) \\ &= Prob[U_{1C} + max_j(U_{2j}^C) > U_{1W} + max_j(U_{2j}^W)] \\ &= Prob[V_{1C} + \varepsilon_{1C} + max_j(V_{2j}^C + \varepsilon_{2j}^C) > V_{1W} + \varepsilon_{1W} + max_j(V_{2j}^W + \varepsilon_{2j}^W)] \\ &= \int I[V_{1C} + \varepsilon_{1C} + max_j(V_{2j}^C + \varepsilon_{2j}^C) > V_{1W} + \varepsilon_{1W} + max_j(V_{2j}^W + \varepsilon_{2j}^W)] f(\varepsilon) d\varepsilon \end{aligned}$$

where $I[.]$ is an indicator of whether the statement in brackets is true.

The integral can be approximated through simulation. For an accept-reject simulator: (1) Take a draw from $f(\varepsilon)$, with its components labeled $\varepsilon^r_{1C}, \varepsilon^{Cr}_{2j}$, etc. (2) Calculate $U^C_{2j} = V^C_{2j} + \varepsilon^{Cr}_{2j}$ for all j, determine the highest one and labeled it U^{Cr}_2. Similarly, calculate U^{Wr}_2. (3) Calculate the total utilities as $TU^r_C = V^r_{1C} + \lambda U^{Cr}_2$ and similarly for TU^r_W. (4) Determine whether $TU^r_C > TU^r_W$. If so, set $I^r = 1$. Otherwise, let $I^r = 0$. (5). Repeat steps 1-4 R times. The simulated probability of choosing college is $\tilde{P}_C = \sum_r I^r/R$.

Convenient error partitioning (as explained in section 1.2) can be utilized to obtain a smooth and more accurate simulator than accept-reject, provided that the integral over the first period errors has a closed form conditional on the second period errors. Suppose for example that ε_{1C} and ε_{1W} are iid extreme value. Label the second period errors collectively as ε_2 with any density $g(\varepsilon_2)$. Conditional on the second period errors, the probability of the student going to college is a standard logit model with an extra explanatory variable that captures the future effect of the current choice. That is,

$$P_C(\varepsilon_2) = \frac{e^{V_{1C}+U^C_2(\varepsilon_2)}}{e^{V_{1C}+U^C_2(\varepsilon_2)} + e^{V_{1W}+U^C_2(\varepsilon_2)}}$$

where $U^C_2(\varepsilon_2)$ is calculated from the second period errors as $U^C_2(\varepsilon_2) = max_j(V^C_{2j} + \varepsilon^C_{2j})$, and similarly for $U^W_2(\varepsilon_2)$. The unconditional probability is then the integral of this logit formula over all possible values of the second-period errors:

$$P_C = \int P_C(\varepsilon_2) g(\varepsilon_2) d\varepsilon_2.$$

The probability is simulated as follows: (1) Take a draw from density $g(\cdot)$ and label it ε^r_2 . (2) Using this draw of the second period errors, calculate the utility that would be obtained from each possible job if the person went to college. That is, calculate $U^{Cr}_{2j} = V^C_{2j} + \varepsilon^{Cr}_{2j}$ for all j. (3) Determine the maximum of these utilities and label it U^{Cr}_2. This is the utility that the person would obtain in the second period if he went to college in the first period, based on this draw of the second-period errors. (4)-(5) Similarly, calculate U^{Wr}_{2j} $\forall j$ and then determine the maximum U^{Wr}_2. (6) Calculate the conditional choice probability for this draw as

$$P^r_C = \frac{e^{V_{1C}+U^{Cr}_2}}{e^{V_{1C}+U^{Cr}_2} + e^{V_{1W}+U^{Wr}_2}}.$$

(7) Repeat steps 1-6 many times, labeled $r = 1, \ldots, R$. (8) The simulated probability is $\tilde{P}_C = \sum_r P_C^r / R$.

If the second-period errors are also iid extreme value, then the probability of taking a particular job in the second period is standard logit. The probability of going to college and taking job j is

$$P_{Cj} = \left(\int \left[\frac{e^{V_{1C} + U_2^C(\varepsilon_2)}}{e^{V_{1C} + U_2^C(\varepsilon_2)} + e^{V_{1W} + U_2^C(\varepsilon_2)}} \right] g(\varepsilon_2) d\varepsilon_2 \right) \left(\frac{e^{V_{2j}^C}}{\sum_k e^{V_{2k}^C}} \right)$$

The choice probabilities for the first period are simulated by taking draws of the second period errors, as described above with $g(\cdot)$ being the extreme value distribution. However, the probabilities for the second period are calculated exactly. The draws of the second-period errors are used only in calculating the first period probabilities, where they do not integrate out in closed form. The second-period errors integrate out of the second-period probabilities in a closed form, which is used to calculate the second-period probabilities exactly. Application to other distributions that allow correlation over alternatives, such as GEV or normal, is straightforward. Allowing the errors to be correlated over time can be accomplished with a joint normal distribution and simulation of both periods' probabilities.

7.7.2 Multiple periods

We first expand to three periods and then generalize to any number of periods. The model of college choice can be extended by considering retirement options. When a person reaches retirement age, there are usually several options available. He can continue working full time, or work part time and spend part of his retirement funds, or retire fully and collect social security and perhaps a pension. The person's income under these alternatives depends largely on the job that the person has held and the retirement plan that the job provided. Three periods are sufficient to capture the decision process. The person goes to college or not in the first period, chooses a job in the second period, and chooses among the available retirement-age options in the third period. The high school student knows, when deciding whether to go to college, that this decision will affect his job opportunities, which in turn will affect his retirement options. (This foreknowledge is starting to seem like a mighty big burden for a high school student, but let's proceed without getting too depressed.)

The set of retirement-age alternatives is labeled S with elements indexed by s. In the third period, the utility that the person obtains from alternative s if he went to college in the first period and had job j in the second period is U_{3s}^{Cj}. Conditional on these previous choices, the person chooses option s if $U_{3s}^{Cj} > U_{3t}^{Cj}$ for all $s \neq t$ and $s, t \in S$. Similar notation and behavior applies conditional on other choices in the first and second periods.

In the second period, the person recognizes that his job choice will affect his retirement-age options. He knows he will maximize among the available options when retirement age arrives. Suppose he chose college in the first period. In the second period, he knows that the utility he will obtain in the third period if he chooses job j is $max_s U_{3s}^{Cj}$. The total utility of choosing job j in the second period, given that he chose college in the first period, is therefore $TU_j^C = U_{2j}^C + \theta max^s U_{3s}^{Cj}$, where θ weights period three relative to period two. He chooses job j if $TU_j^C > TU_k^C$ for all $k \neq j$ and $j, k \in J$. Similar notation and behavior occurs if he chose to work in the first period.

Consider now the first period. He knows that, if he chooses college, he will choose the job that maximizes his utility from jobs conditional on going to college, and then will choose the retirement-age option that maximizes his utility conditional on that chosen job. The total utility from college is

$$\begin{aligned} TU_C &= U_{1c} + \lambda max_j TU_j^C \\ &= U_{1c} + \lambda max_j \left(U_{2j}^C + \theta max_s U_{3s}^{Cj} \right). \end{aligned}$$

This expression is similar to that in the two-period model except that it includes an additional layer of maximization: the maximization for the third period is contained in each maximization for the second period. A similar term gives the total utility of working in the first period, TU_W. The person chooses college if $TU_C > TU_W$.

This completes the description of the person's behavior. The researcher observes a portion of each utility function, U_{1C}, U_{1W}, U_{2j}^C and U_{2j}^W $\forall j \in J$ and U_{3s}^{Cj} and U_{3s}^{Wj} $\forall s \in S,\ j \in J$. The unobserved portions are collected labeled by the vector ε with density $f(\varepsilon)$. The probability that the person chooses college is

$$P_C = \int I(\varepsilon) f(\varepsilon) d\varepsilon$$

where

$$\begin{aligned} I(\varepsilon) &= 1 \\ &\text{if} \\ &\quad V_{1C} + \varepsilon_{1C} + \lambda max_j \left(V_{2j}^C + \varepsilon_{2j}^C + \theta max_s(V_{3s}^{Cj} + \varepsilon_{32}^{Cj}) \right) \\ &> V_{1W} + \varepsilon_{1W} + \lambda max_j \left(V_{2j}^W + \varepsilon_{2j}^W + \theta max_s(V_{3s}^{Wj} + \varepsilon_{32}^{Wj}) \right) \end{aligned}$$

This expression is the same as in the two-period model except that now the term inside the indicator function has an extra level of maximization. An accept-reject simulator is obtained by: (1) draw from $f(\varepsilon)$, (2) calculate the third period utility U_{3s}^{Cj} for each s, (3) identify the maximum over s, (4) calculate TU_{2j}^C with this maximum, (5) repeat steps 2-5 for each j and identify the maximum of TU_{2j}^C over j, (6) calculate TU_C using this maximum, (7) repeat steps 2-6 for TU_W, (8) determine whether $TU_C > TU_W$ and set $I = 1$ if it is, (9) repeat steps 1-8 many times and average the results. Convenient error partitioning can also be used. For example if all errors are iid extreme value, then the first-period choice probabilities, conditional on draws of the second- and third-period errors, are logit; the second-period probabilities, conditional on the third-period errors, are logit; and the third period probabilities are logit.

We can now generalize these concepts and introduce some widely-used terminology. Note that the analysis of the person's behavior and the simulation of the choice probabilities by the researcher start with the last period and work backwards in time to the first period. This process is called backwards recursion. Suppose there are J alternatives in each of T equal-length time periods. Let a sequence of choices up to period t be denoted $\{i_1, i_2, \cdots, i_t\}$. The utility that the person obtains in period t from alternative j is $U_{tj}(i_1, i_2, \cdots, i_{t-1})$, which depends on all previous choices. If the person chooses alternative j in period t, he will obtain this utility plus the future utility of choices conditioned on this choice. The total utility (current and future) that the person obtains from choosing alternative j in period t is $TU_{tj}(i_1, i_2, \cdots, i_{t-1})$. He chooses the alternative in the current period that provides the greatest total utility. Therefore the total utility he receives from his optimal choice in period t is $TU_t(i_1, i_2, \cdots, i_{t-1}) = max_j TU_{tj}(i_1, i_2, \cdots, i_{t-1})$. This total utility from the optimal choice at time t, TU_t, is called the valuation function at time t.

The person chooses optimally in the current period with knowledge that he will choose optimally in the future. This fact establishes a convenient relation between the valuation function in successive periods. In particular,

$$TU_t(i_1, \cdots, i_{t-1}) = max_j \left[U_{jt}(i_1, \cdots, i_{t-1}) + \delta TU_{t+1}(i_1, \cdots, i_t = j)\right]$$

where δ is a parameter that discounts the future. TU_{t+1} on the right hand side is the total utility that the person will obtain in period $t+1$ onward if he chooses alternative j in period t (i.e., if $i_t = j$). The equation states that the total utility that the person obtains from optimizing behavior from period t onwards, given previous choices, is the maximum over j of: the utility from j in period t plus the discounted total utility from optimizing behavior from period $t+1$ onwards conditional on choosing j in period t. This relation is Bellman's equation (1957) applied to discrete choice with perfect information.

$TU_{tj}(i_1, \cdots, i_{t-1})$ is sometimes called the conditional valuation function, conditional on choosing alternative j in period t. A Bellman equation also operates for this term:

$$TU_{tj}(i_1, \cdots, i_{t-1}) = U_{jt}(i_1, \cdots, i_{t-1} + \delta max_k \left[TU_{t+1,k}(i_1, \cdots, i_t = j)\right].$$

Since by definition $TU_t(i_1, \cdots, i_{t-1}) = max_j[TU_{tj}(i_1, \cdots, i_{t-1})]$, the Bellman equation in terms of conditional valuation function is equivalent to the Bellman equation in terms of the unconditional valuation function.

If T is finite, the Bellman equation can be applied with backwards recursion to calculate TU_{tj} for each time period. At $t = T$, there is no future time period, and so $TU_{Tj}(i_1, \cdots, i_{T-1}) = U_{Tj}(i_1, \cdots, i_{T-1})$. Then $TU_{T-1,j}(i_1, \cdots, i_{T-2})$ is calculated from $TU_{Tj}(i_1, \cdots, i_{T-1})$ using Bellman's equation. And so on forward to $t = 1$. Note that $U_{tj}(i_1, \cdots, i_{t-1})$ must be calculated for each t, j, and, importantly, for each possible sequence of past choices, $i_1, \cdots, i_{t-1}$. With J alternatives in T time periods, the recursion requires calculation of $(J^T)T$ utilities (that is, J^T possible sequences of choices, with each sequence containing T one-period utilities.) To simulate the probabilities, the researcher must calculate these utilities for each draw of unobserved factors. And these probabilities must be simulated for each value of the parameters in the numerical search for the estimates. This huge computational burden is called the curse of dimensionality and is the main stumbling block to application of the procedures with more than a few time periods and/or alternatives. We discuss in the next subsection procedures that have been suggested to avoid or mitigate this curse, after showing that the curse is even greater when uncertainty is considered.

7.7.3 Uncertainty about future impacts

In the analysis so far we have assumed that the decision-maker knows the utility for each alternative in each future time period and how this utility is affected by prior choices. Usually, the decision-maker does not possess such foreknowledge. A degree of uncertainty shrouds the future impacts of current choices.

The behavioral model can be adapted to incorporate uncertainty. For simplicity, return to the two-period model for our high school student. In the first period, the student does not know for sure the second-period utilities, U_{2j}^{C} and U_{2j}^{W} $\forall j$. For example, the student does not know, before going to college, how strong the economy, and hence his job possibilities, will be when he graduates. These utilities can be expressed as functions of unknown factors: $U_{2j}^{C}(e)$ where e refers collectively to all factors in period two that are unknown in period one. These unknown factors will become known (that is, will be revealed) when the student reaches the second period, but are unknown to the person in the first period. The student has a subjective distribution on e that reflects the likelihood that he ascribes to the unknown factors taking a particular realization in the second period. This density is labeled $g(e)$. He knows that, whatever realization of e actually occurs, he will, in the second period, choose the job that gives him the maximum utility. That is, he will receive utility $max_j U_{2j}^{C}(e)$ in the second period if he chooses college in the first period and the unknown factors end up being e. In the first period, when evaluating whether to go to college, he takes the expectation of this future utility over all possible realizations of the unknown factors, using his subjective distribution over these realizations. The expected utility that he will obtain in the second period if he chooses college in the first period is therefore $\int \left[max_j(U_{2j}^{C}(e))\right] g(e)d(e)$. The total expected utility from choosing college in the first period is then:

$$TEU_C = U_{1C} + \lambda \int \left[max_j(U_{2j}^{C}(e))\right] g(e)d(e)$$

TEU_W is defined similarly. The person chooses college if $TEU_C > TEU_W$. In the second period, the unknown factors become known, and the person chooses job j if he had chosen college if $U^C_{2j}(e*) > U^C_{2k}(e*)$ for all $k \neq j$, where $e*$ is the realization that actually occurred.

Turning to the researcher, we have an extra complication introduced by $g(e)$, the decision-maker's subjective distribution for unknown factors. In addition to not knowing utilities in their entirety, the researcher has only partial knowledge of the decision-maker's subjective probability $g(e)$. This lack of information is usually represented through parameterization. The researcher specifies a density, labeled $h(e \mid \theta)$, that depends on unknown parameters θ. The researcher then assumes that the person's subjective density is the specified density evaluated at the true parameters θ^*. That is, the researcher assumes $h(e \mid \theta^*) = g(e)$. Stated more persuasively and accurately: the true parameters are, by definition, the parameters for which the researcher's specified density $h(e \mid \theta)$ becomes the density $g(e)$ that the person actually used. With a sufficiently flexible h, any g can be represented as h evaluated at some parameters, which are called the true parameters. These parameters are estimated along with the parameters that enter utility. (Other ways of representing the researcher's lack of knowledge about $g(e)$ can be specified; however, they are generally more complex.)

Utilities are decomposed into their observed and unobserved portions, with the unobserved portions collectively called ε with density $f(\varepsilon)$. The probability that the person goes to college is

$$\begin{aligned} P_C &= Prob(TEU_C > TEU_W) \\ &= \int \left[I(V_{1c} + \varepsilon_{1C} + \lambda \int \{max_j(V^C_{2j}(e) + \varepsilon^C_{2j}(e))\} h(e \mid \theta) d(e)) \right] f(\varepsilon) d\varepsilon. \end{aligned}$$

The probablity can be approximated by simulating the inside integral within the simulation of the outside integral. (1) Take a draw of ε. (2a) Take a draw of e from $h(e \mid \theta)$. (2b) Using this draw, calculate the term in squiggly brackets. (2c) Repeat steps 2a-b many times and average the results. (3) Using the value from 2c, calculate the term in square brackets. (4) Repeat steps 1-3 many times and average the results. As the reader can see, the curse of dimensionality grows worse.

Several authors have suggested ways to reduce the computational burden. Keane and Wolpin (1994) calculate the valuation function at selected realizations of the unknown factors and past choices; they then approximate the valuation function at other realizations and past choices through interpolating from the calculated valuations. Rust

(1997) suggests simulating future paths and using the average over these simulated paths as an approximation in the valuation function. Hotz and Miller (1993) and Hotz *et al.* (1993) show that there is a correspondence between the valuation function in each time period and the choice probabilities in future periods. This correspondence allows the valuation functions to be calculated with these probabilities instead of backwards recursion.

Each of these procedures has limitations and is applicable in only certain situations, which the authors themselves describe. As Rust (1994) has observed, it is unlikely that a general-purpose breakthrough will arise that makes estimation simple for all forms of dynamic optimization models. Inevitably the researcher will need to make tradeoffs in specifying the model to assure feasibility, and the most appropriate specification and estimation method will depend on the particulars of the choice process and the goals of the research. In this regard, I have found that two simplifications are very powerful in that they often provide a large gain in computational feasibility for a relatively small loss (and sometimes a gain) in content.

The first suggestion is for the researcher to consider ways to capture the nature of the choice situation with as few time periods as possible. Sometimes, in fact usually, time periods will need to be defined not by the standard markers, such as year or month, but rather in a way that is more structural to the decision process. For example, for the high school student deciding whether to go to college, it might seem natural to say that the student makes a choice each year among the jobs and schooling options that are available in that year, given his past choices. Indeed, this statement is true: the student does indeed make annual (or even monthly, weekly, daily) choices. However, such a model would clearly face the curse of dimensionality. In contrast, the specification that we discussed above involves only two time periods, or three if retirement is considered. Estimation is quite feasible for this specification. In fact, the two-period model might be more accurate than an annual model: students deciding on college probably think in terms of the college years and their post-college options, rather than trying to anticipate their future choices in each future year. McFadden and Train (1996) provide an example of how an dynamic optimization model with only a few well-considered periods can accurately capture the nature of the choice situation.

A second powerful simplification was first noted by Rust (1987).

Suppose that the factors that the decision-maker does not observe beforehand are also the factors that the researcher does not observe (either before or after), and that these factors are thought by the decision-maker to be iid extreme value. Under this admittedly restrictive assumption, the choice probabilities take a closed form that is easy to calculate. The result can be readily derived for our model of college choice. Assume that the student, when in the first period, decomposes second-period utility into an known and unknown part, e.g., $U^C_{2j}(e) = V^C_{2j} + e^C_{2j}$, and assumes that e^C_{2j} follows an extreme value distribution independent of all else. This unknown factor becomes known to the student in the second period, such that second-period choice entails maximization over known $U^C_{2j}\ \forall j$. However, in the first period it is unknown. Recall from section 3.5 that the expected maximum of utilities that are iid extreme value takes the familiar log-sum formula. In our context, this result means that

$$E\left(max_j(V^C_{2j} + \varepsilon^C_{2j})\right) \quad = \quad \alpha ln\left(\sum_j e^{V^C_{2j}}\right)$$

which we can label LS^C_2. LS^W_2 is derived similarly. The person chooses college if then

$$\begin{aligned} TEU_C &> TEU_W \\ U_{1C} + \lambda LS^C_2 &> U_{1W} + \lambda LS^W_2 \end{aligned}$$

Note that this decision-rule is in closed form: the integral over unknown future factors becomes the log-sum formula. Consider now the researcher. Each first-period utility is decomposed into an observed and unobserved part, $U_{1C} = V_{1C} + \varepsilon_{1C}$, $U_{1W} = V_{1W} + \varepsilon_{1W}$, and we assume that the unobserved portions are iid extreme value. For the second-period utilities, we make a fairly restrictive assumption. We assume that the part of utility that the researcher does not observe is the same as the part of utility that the student does not know beforehand. That is, we assume $U^C_{2j} = V^C_{2j} + \varepsilon^C_{2j}\ \forall j$, where the researcher's ε^C_{2j} is the same as the student's e^C_{2j}. Under this assumption, the researcher can calculate the log-sum term for future utility, LC^C_2 and LS^W_2, exactly, since these terms depend only on the observed portion of utility in the second period, $V^C_{2j}\ \forall j$, which is observed by the researcher and known beforehand by the decision-maker. The probability of the student choosing college is then

$$\begin{aligned}
P_C &= Prob(TEU_C > TEU_W) \\
&= Prob(U_{1C} + \lambda LS_2^C > U_{1W} + \lambda LS_2^W \\
&= Prob(V_{1C} + \varepsilon_{1C} + \lambda LS_2^C > V_{1W} + \varepsilon_{1W} + \lambda LS_2^W \\
&= \frac{e^{V_{1C}+\lambda LS_2^C}}{e^{V_{1C}+LS_2^C} + e^{V_{1W}+\lambda LS_2^W}}.
\end{aligned}$$

The model takes the same form as the upper part of a nested logit model: the first-period choice probability is the logit formula with a log-sum term included as an extra explanatory variable. Multiple periods are handled the same way as multi-level nested logits.

It is doubtful that the researcher, in reality, observes everything that the decision-maker knows beforehand. However, the simplification that arises from this assumption is so great, and the curse of dimensionality that would arise otherwise is so severe, that proceeding as if it were true is perhaps worthwhile in many situations.

Part II

Estimation

Chapter 8

Numerical Maximization

8.1 Motivation

Most estimation involves maximization of some function, such as the likelihood function, the simulated likelihood function, or squared moment conditions. This chapter describes numerical procedures that are used to maximize a likelihood function. Analogous procedures apply when maximizing other functions.

Knowing and being able to apply these procedures is critical in our new age of discrete choice modeling. In the past, researchers adapted their specifications to the few convenient models that were available. These models were included in commercially available estimation packages, such that the researcher could estimate the models without knowing the details of how the estimation was actually performed from a numerical perspective. The thrust of the wave of discrete choice methods is to free the researcher to specify models that are tailor-made to her situation and issues. Exercising this freedom means that the researcher will often find herself specifying a model that is not exactly the same as any in commercial software. The researcher will need to write special code for her special model.

The purpose of this chapter is to assist in this exercise. Though not usually taught in econometrics courses, the procedures for maximization are fairly straightforward and easy to implement. Once learned, the freedom they allow is invaluable.

8.2 Notation

The log-likelihood function takes the form $LL(\beta) = \sum_{n=1}^{N} lnP_n(\beta)/N$ where $P_n(\beta)$ is the probability of the observed outcome for decision-maker n, N is the sample size, and β is a $K \times 1$ vector of parameters. In this chapter, we divide the log-likelihood function by N, such that LL is the average log-likelihood in the sample. Doing so does not affect the location of the maximum (since N is fixed for a given sample) and yet facilitates interpretation of some of the procedures. All the procedures operate the same whether or not the log-likelihood is divided by N. The reader can verify this fact as we go along by observing that N cancels out of the relevant formulas.

The goal is to find the value of β that maximizes $LL(\beta)$. In terms of Figure 8.1, the goal is to locate $\hat{\beta}$. Note in this figure that LL is always negative, since the likelihood is a probability between 0 and 1 and the log of any number between 0 and 1 is negative. Numerically, the maximum can be found by "walking up" the likelihood function until no further increase can be found. The researcher specifies starting values β_0. Each iteration, or step, moves to a new value of the parameters at which $LL(\beta)$ is higher than at the previous value. Denote the current value of β as β_t, which is attained after t steps from the starting values. The question is: what is the best step we can take next, that is, what is the best value for β_{t+1}?

The gradient at β_t is the vector of first derivatives of $LL(\beta)$ evaluated at β_t:

$$g_t = \left(\frac{\partial LL(\beta)}{\partial \beta} \right)_{\beta_t}.$$

This vector tells us which way to step in order to go up the likelihood function. The Hessian is the matrix of second derivatives:

$$H_t = \left(\frac{\partial g_t}{\partial \beta'} \right)_{\beta_t} = \left(\frac{\partial^2 LL(\beta)}{\partial \beta \partial \beta'} \right)_{\beta_t}.$$

The gradient has dimension $K \times 1$ and the Hessian is $K \times K$. As we will see, the Hessian can help us to know *how far* to step, given that the gradient tells us *in which direction* to step.

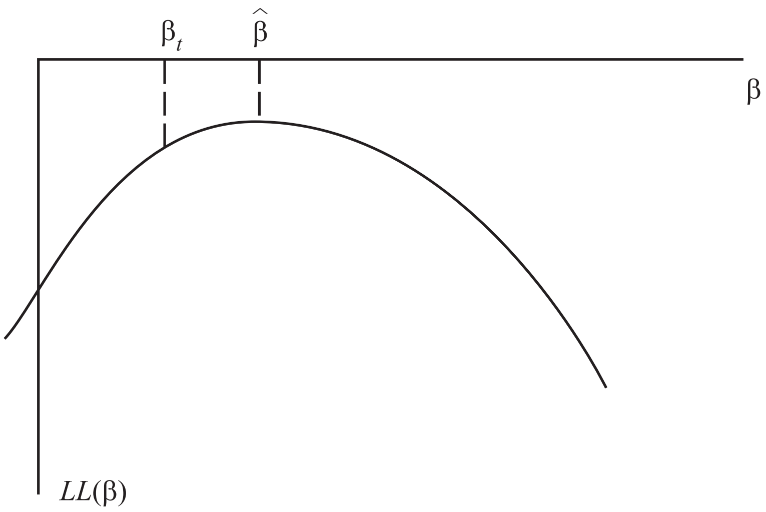

Figure 8.1: Maximum likelihood estimate.

8.3 Algorithms

Of the numerous maximization algorithms that have been developed over the years, I describe below only the most prominent, with an emphasis on the pedagogical value of the procedures as well as their practical use. Readers who are induced to explore further will find the treatments by Judge et al. (1985, Appendix B) and Ruud (2000) rewarding.

8.3.1 Newton-Raphson

To determine the best value of β_{t+1}, take a second order Taylor's approximation of $LL(\beta_{t+1})$ around $LL(\beta_t)$:

$$LL(\beta_{t+1}) = LL(\beta_t) + (\beta_{t+1} - \beta_t)'g_t + \frac{1}{2}(\beta_{t+1} - \beta_t)'H_t(\beta_{t+1} - \beta_t). \quad (8.1)$$

Now find the value of β_{t+1} that maximizes this approximation to $LL(\beta_{t+1})$:

$$\frac{\partial LL(\beta_{t+1})}{\partial \beta_{t+1}} = g_t + H_t(\beta_{t+1} - \beta_t) = 0.$$

$$H_t(\beta_{t+1} - \beta_t) \;\; = \;\; -g_t$$

$$\begin{aligned}(\beta_{t+1} - \beta_t) &= -H_t^{-1} g_t \\ \beta_{t+1} &= \beta_t + (-H_t^{-1}) g_t.\end{aligned}$$

The Newton-Raphson procedure uses this formula. The step from the current value of β to the new value is $(-H_t^{-1})g_t$, that is, the gradient vector premultiplied by the negative of the inverse of the Hessian.

This formula is intuitively meaningful. Consider $K = 1$, as illustrated in Figure 8.2. The slope of the log-likelihood function is g_t. The second derivative is the Hessian H_t, which is negative for this graph since the curve is drawn to be concave. The negative of this negative Hessian is positive and represents the degree of curvature. That is, $(-H_t)$ is the positive curvature. Each step of β is the slope of the log-likelihood function divided by its curvature. If the slope is positive, β is raised as in the first panel, and β is lowered if the slope if negative as in the second panel. The curvature determines how large a step is made. If the curvature is great, meaning that the slope changes quickly as in the first panel of Figure 8.3, then the maximum is likely to be close and so a small step is taken (dividing the gradient by a large number gives a small number). Conversely, if the curvature is small, meaning that the slope is not changing much, then the maximum seems to be further away and so a larger step is taken.

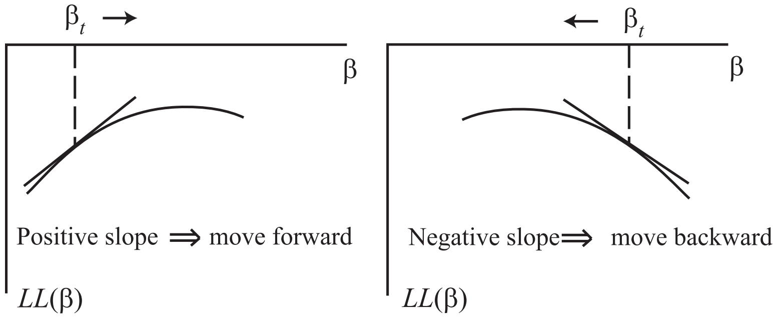

Figure 8.2: Direction of step follows the slope.

Three issues are relevant to the Newton-Raphson (N-R) procedure.

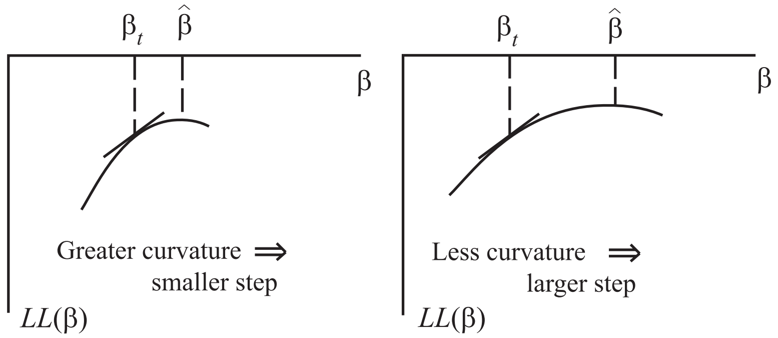

Figure 8.3: Step size is inversely related to curvature.

Quadratics

If $LL(\beta)$ were exactly quadratic in β, then the N-R procedure would reach the maximum in one step from any starting value. This fact can easily be verified with $K = 1$. If $LL(\beta)$ is quadratic, then it can be written as:

$$LL(\beta) = a + b\beta + c\beta^2.$$

The maximum is:

$$\begin{aligned} \frac{\partial LL(\beta)}{\partial \beta} &= b + 2c\beta = 0 \\ \hat{\beta} &= -\frac{b}{2c}. \end{aligned}$$

The gradient and Hessian are $g_t = b + 2c\beta_t$ and $H_t = 2c$, and so N-R gives us:

$$\begin{aligned} \beta_{t+1} &= \beta_t - H_t^{-1} g_t \\ &= \beta_t - \frac{1}{2c}(b + 2c\beta_t) \\ &= \beta_t - \frac{b}{2c} - \beta_t \\ &= -\frac{b}{2c} = \hat{\beta}. \end{aligned}$$

Most log-likelihood functions are not quadratic, and so the N-R procedure takes more than one step to reach the maximum. However, knowing how N-R behaves in the quadratic case helps in understanding its behavior with non-quadratic LL, as we will see below.