Obtaining Analytic Derivatives for a Class of Discrete-Choice Dynamic Programming Models [*]

Curtis Eberwein[†] John C. Ham [‡]

June 5, 2007

Abstract

This paper shows how to recursively calculate analytic first and second derivatives of the likelihood function generated by a popular version of a discrete-choice, dynamic programming model, allowing for a dramatic decrease in computing time used by derivative-based estimation algorithms. The derivatives also are very useful for finding the exact maximum of the likelihood function, for de-bugging complicated program code, and for estimating standard errors.

JEL classification: C4, C5, C6

[*] John Ham would like to thank the NSF for financial support. The authors would like to thank Donghoon Lee, Holger Sieg and Kenneth Wolpin for helpful comments. All mistakes are ours and the opinions expressed here in no way reflect those of the NSF.

[†] Center for Human Resource Research, Ohio State University, 921 Chatham Lane, Suite 100, Columbus, OH 43221, E-mail: ceberw@postoffice.chrr.ohio-state.edu

[‡] Professor, Department of Economics, University of Southern California, E-mail: johnham@usc.edu

1 Introduction

This paper shows how to calculate the analytic derivatives of the likelihood function with respect to model parameters for a class of discrete-choice dynamic programming models. The model is one where the stochastic component of utility depends on an iid extreme value (temporal) term (Rust (1987)) and an individual-specific (permanent) component (Heckman and Singer (1984)). This structure has been used by Van Der Klaauw (1996), Arcidiacono, Sieg and Sloan (2007), and Liu, Mroz and Van der Klaauw (2004).

We can think of several reasons having these derivatives is important. First, these dramatically reduce the number of function evaluations and computation time necessary to estimate the model. For example, suppose our model contains 30 parameters. For each candidate for the parameters, most maximization methods require the value of the function and the first derivatives. If we use (the more accurate) two-sided derivatives, this involves 61 function evaluations in total, while one-sided derivatives require 31 function evaluations. Using analytic first derivatives requires the equivalent of only two function evaluations (i.e. one for the function, and one for the derivatives), drastically cutting the computer time used at each iteration. This saving of computer time is especially important for the estimation of structural models, which is one of the few remaining areas in empirical work where computational demands restrict the type of models we can estimate.

Analytical first and second derivatives also aid in calculating the standard errors of parameter estimates. Standard practice in structural estimation is to use minus the outer product of the gradient using numeric derivatives to obtain an estimate of the second derivative matrix. Having analytic first and second derivatives improves on this in two ways. First, having these allows one to obtain a sandwich estimator that is robust to non i.i.d. sampling schemes. Second, while numerical derivatives are very close to the true derivatives for most parameter vectors, they can be quite different close to the optimum.[1] Thus, the outer product of the gradient based on numeric first derivatives may provide relatively noisy estimates of the outer product of the analytic first derivatives.

Third, analytic first and second derivatives can be useful in debugging complicated programs to estimate structural models. For example, if numeric and analytic derivatives (calculated away from the optimum) are quite

[1] We have found this to be true in previous applications. The reason seems to be because the true derivatives are zero at the optimum, so the error in the numeric derivatives becomes large relative to the magnitude of the true derivatives near an optimum.

close, one can be quite confident that both the first derivatives subroutine and the function subroutine are programmed correctly. Moreover, if the derivatives do not agree for certain parameters, one often will find the programming error by focusing on these parameters. Alternatively, if the analytic second derivatives and the numeric derivatives of the analytic first derivatives agree, one can be reasonably certain that the first derivative subroutine and the second derivative subroutine are correct. Fourth, having analytic second derivatives can help one obtain an optimum of a relatively flat likelihood function since it enables one to use a second derivative maximization routine such as GRADX (Goldfeld, Quand, Trotter). Fifth, analytical first derivatives can help one get closer to the actual optimum, which is important for standard errors. Here the idea is that as one approaches the optimum, numeric first derivatives become increasingly noisy estimates of the analytic first derivatives, and thus convey less useful information for maximization than analytic derivatives.

The paper proceeds as follows. Section 2 outlines the widely-used model we consider. Section 3 generates the likelihood function for our model. We note that the value of the likelihood can be obtained (recursively) in closed form. In Section 4 we consider the analytic first derivatives of the log likelihood, and show that they can be obtained recursively with a similar order of complexity to that for obtaining the value of the likelihood function. In Section 5 we show that analytic second derivatives can be obtained in a similar fashion. Section 6 concludes the paper.

2 The Model

We assume there are I mutually exclusive, collectively exhaustive choices that an individual chooses among over T periods of time.[2]

The temporal utility function at time t for alternative i is given by:

$$u_i(s(t), \theta_{ik}, \epsilon_{it}) = g_i(s(t), \theta_{ik}) + \epsilon_{it}. \tag{1}$$

Here, $g_i()$ is a continuously differentiable function, $s(t)$ is a state vector (observed by the econometrician and the individual making the choices), θ_{ik} is a permanent heterogeneity term with K points of support (Heckman and Singer (1984)), and ϵ_{it} is an extreme-value error term.

We assume the realization of θ_{ik} is observed by the individual, but not the econometrician. Associated with each point of support is an I-tuple of

[2] We assume T is finite. Of course, one can allow T to tend to infinity to approximate arbitrarily closely an infinite horizon dynamic program.

values.[3] That is, let $\bar{\theta}_1 = (\theta_{11}, ..., \theta_{1I})$, ... $\bar{\theta}_K = (\theta_{K1}, ..., \theta_{KI})$ and $Pr(\bar{\theta} = \bar{\theta}_k) = P_k$, $(k < K)$ with $Pr(\bar{\theta} = \bar{\theta}_K) = 1 - \sum_{J=1}^{K-1} P_j$. We assume the econometrician seeks to estimate the $\bar{\theta}_k$ and their probabilities of occurring, as well as the number of points of support, K.

The error term ϵ_{it} is a temporal shock to the utility of choosing alternative i in period t. It is assumed to be independent across alternatives and time. The individual observes the current vector of these shocks, but not future values. The econometrician observes neither. The probability distribution function for each of these shocks is given by:

$$F(\epsilon_{it}) = \exp[-e^{-\tau(\epsilon_{it}+c/\tau)}]. \tag{2}$$

That is, the ϵ_{it} are extreme-value errors.[4] The number c is chosen so that the errors are mean zero (i.e. c is Euler's Constant).

Note that while ϵ_{it} is assumed to be additive in the temporal utility, θ_{ik} is only assumed to enter the temporal utility in a manner that will allow for differentiability.

Given these assumptions, the value function in the final period, T, is:

$$V[s(T), \bar{\theta}_k, \epsilon_T] = \max_{i \in I}\{g_i(s(T), \theta_{ik}) + \epsilon_{iT}\}, \tag{3}$$

where ϵ_t is the vector of realized temporal shocks to utility in any period t.

Since the temporal shocks are extreme value, the expectation of this (prior to observing ϵ_T) is given by (Rust (1987)):

$$EV[s(T), \bar{\theta}_k, \epsilon_T] = \frac{1}{\tau} \ln[\sum_{i \in I} e^{\tau g_i(s(T), \theta_{ik})}]. \tag{4}$$

For any $t < T$ we can recursively define the value function as:

$$\begin{aligned} V(s(t), \bar{\theta}_k, \epsilon_t) = &\max_{i \in I}\{g_i(s(t), \theta_{ik}) + \epsilon_{it} \\ &+ \beta E[V(s(t+1), \bar{\theta}_k, \epsilon_{t+1}) | d_i(t) = 1]\}. \end{aligned} \tag{5}$$

Here, $d_i(t) = 1$ if and only if alternative i is chosen in period t ($d_i(t) = 0$ otherwise) and $\beta \in (0, 1)$ is the discount factor. Note the above allows $s(t + 1)$ to depend on choices made by the individual up to period t.

[3] Other methods of estimating the unobserved heterogeneity can easily be incorporated, such as the one-factor loading structure, e.g. Eberwein, Ham, and LaLonde (1997).

[4] In the above, $\tau > 0$ is a scale parameter. Generally, this will not be empirically identified in a discrete-choice model and could be set to equal, say, one. We do not normalize this since it may be necessary to adjust its value to avoid underflow or overflow problems.

Define the alternative specific value as:

$$V_i[s(t), \bar{\theta}_k, \epsilon_t] = g_i(s(t), \theta_{ik}) + \epsilon_{it} + \beta E[V(s(t+1), \bar{\theta}_k, \epsilon_{t+1})|d_i(t) = 1]. \quad (6)$$

And:

$$\tilde{V}_i[s(t), \theta_k] = V_i[s(t), \bar{\theta}_k, \epsilon_t] - \epsilon_{it}. \quad (7)$$

Then:

$$EV(s(t), \bar{\theta}_k, \epsilon_t) = \frac{1}{\tau} \ln[\sum_{i \in I} e^{\tau \tilde{V}_i(s(t), \bar{\theta}_k)}]. \quad (8)$$

Thus, the value function can be calculated in closed form and is given recursively by:

$$V(s(t), \bar{\theta}_k, \epsilon_t) = \max_{i \in I}\{\tilde{V}_i(s(t), \bar{\theta}_k) + \epsilon_{it}\}, \quad (9)$$

where:

$$\tilde{V}_i(s(t), \bar{\theta}_k) = g_i(s(t), \theta_{ik}) + \beta\{\frac{1}{\tau} \ln[\sum_{j \in I} e^{\tau \tilde{V}_j(s(t+1), \bar{\theta}_k)}]|d_i(t) = 1\}. \quad (10)$$

Noting the term to the right of β is zero for $t = T$ this recursively defines the value function for all states and all periods in closed form.

3 The Likelihood

Each observation will consist of vectors $\bar{s}$ and $\bar{d}$ which give, respectively, $s(t)$ and i such that $d_i(t) = 1$ for $t \in \{1, 2, ..., N\}$ where $N \leq T$ is the number of periods observed. Since the temporal shocks are extreme value, for any point of support, k, of the heterogeneity distribution, the likelihood of the observation is given by:

$$L(\bar{s}, \bar{d}|\bar{\theta}_k) = \prod_{t=1}^{N} \frac{\sum_{i \in I} d_i(t) e^{\tau \tilde{V}_i(s(t), \bar{\theta}_k)}}{\sum_{j \in I} e^{\tau \tilde{V}_j(s(t), \bar{\theta}_k)}}. \quad (11)$$

The overall likelihood for an individual is then given by:

$$L(\bar{s}, \bar{d}) = \sum_{k \in K} P_k L(\bar{s}, \bar{d}|\bar{\theta}_k). \quad (12)$$

In practice one would parameterize $P_k = e^{\gamma_k}/\sum e^{\gamma_j}$ with $\gamma_K = 0$ and estimate the γ's instead of the P's.

The above gives (recursively) the likelihood (and thus the log-likelihood) in closed form.

4 Analytic First Derivatives of the Log Likelihood

This section shows how to derive the derivatives of the likelihood with respect to the parameters being estimated. We first focus on a generic parameter λ_1 which influences one or more of the functions $g_i(s(t), \theta_{ik})$ and assume the derivatives of these functions with respect to λ_1 are known (λ_1 can be one of the elements of some $\bar{\theta}_k$). This will be true for virtually any empirical specification.

From (11) the log-likelihood for any point of support, k, of the unobserved heterogeneity is:

$$\ln[L(\bar{s}, \bar{d}|\bar{\theta}_k)] = \sum_{t=1}^{N}[\tau \sum_{i\in I} d_i(t)\tilde{V}_i(s(t), \bar{\theta}_k) - \ln(\sum_{j\in I} e^{\tau \tilde{V}_j(s(t),\bar{\theta}_k)})]. \tag{13}$$

Using this, we have:

$$\frac{\partial \ln L(\bar{s}, \bar{d}|\bar{\theta}_k)}{\partial \lambda_1} = \tau \sum_{t=1}^{N}\{\sum_{i\in I}[d_i(t) - z_i(s(t), \bar{\theta}_k)]\frac{\partial \tilde{V}_i(s(t), \bar{\theta}_k)}{\partial \lambda_1}\}, \tag{14}$$

where:

$$z_i(s(t), \bar{\theta}_k) = \frac{e^{\tau \tilde{V}_i(s(t),\bar{\theta}_k)}}{\sum_{j\in I} e^{\tau \tilde{V}_j(s(t),\bar{\theta}_k)}}. \tag{15}$$

Thus, to get the derivatives of the likelihood function, we need the derivatives of the $\tilde{V}_i$. Note that:

$$\tilde{V}_i(s(T), \bar{\theta}_k) = g_i(s(T), \theta_{ik}), \tag{16}$$

so we have:

$$\frac{\partial \tilde{V}_i(s(T), \bar{\theta}_k)}{\partial \lambda_1} = \frac{\partial g_i(s(T), \theta_{ik})}{\partial \lambda_1}. \tag{17}$$

For $t < T$:

$$\tilde{V}_i(s(t), \bar{\theta}_k) = g_i(s(t), \theta_{ik}) + \beta[\frac{1}{\tau} \ln(\sum_{j \in I} e^{\tau \tilde{V}_j(s(t+1), \bar{\theta}_k)}) | d_i(t) = 1]. \tag{18}$$

Then:

$$\begin{aligned} \frac{\partial \tilde{V}_i(s(t), \bar{\theta}_k)}{\partial \lambda_1} =& \frac{\partial g_i(s(t), \theta_{ik})}{\partial \lambda_1} \\ &+ \beta[\sum_{j \in I} z_j(s(t+1), \bar{\theta}_k) \frac{\partial \tilde{V}_j(s(t+1), \bar{\theta}_k)}{\partial \lambda_1} | d_i(t) = 1]. \end{aligned} \tag{19}$$

Thus, one can build the derivatives of the $\tilde{V}_i$ recursively working backward from the end of the planning horizon in much the same way as value functions are calculated.

The strategy to calculate the derivatives is as follows. Use (17) and (19) to calculate the derivatives of the $\tilde{V}_i$ at each state point that could be reached. Having calculated these, next use them to calculate (14) along the observed path of the state and choices for the individual. The derivative of the likelihood for the individual is then:

$$\frac{\partial L(\bar{s}, \bar{d})}{\partial \lambda_1} = \sum_{k \in K} P_k L(\bar{s}, \bar{d} | \bar{\theta}_k) \frac{\partial \ln L(\bar{s}, \bar{d} | \bar{\theta}_k)}{\partial \lambda_1}. \tag{20}$$

The derivative of the log likelihood is thus:

$$\frac{\partial \ln L(\bar{s}, \bar{d})}{\partial \lambda_1} = \frac{1}{L(\bar{s}, \bar{d})} \frac{\partial L(\bar{s}, \bar{d})}{\partial \lambda_1}. \tag{21}$$

The derivatives, written out in closed form, would be hopelessly complicated. But, as the above shows, calculating these recursively is on a similar order of complexity as calculating value functions recursively.

If we estimate the parameters γ_k defined above, it is easy to show that:

$$\frac{\partial P_k}{\partial \gamma_q} = [1(q = k) - P_k] P_q, \tag{22}$$

where 1() is the indicator function and equals 1 if its argument is true, zero otherwise. Then:

$$\frac{\partial L(\bar{s}, \bar{d})}{\partial \gamma_q} = \sum_{k \in K} \frac{\partial P_k}{\partial \gamma_q} L(\bar{s}, \bar{d} | \bar{\theta}_k), \tag{23}$$

and the derivatives of the log likelihood are obtained by dividing by the likelihood.

5 Analytic Second Derivatives

In this section we derive the analytic second derivatives of the log likelihood. Let λ_1 and λ_2 be parameters of the model. Differentiating (21) with respect to λ_2 yields:

$$\frac{\partial^2 \ln L(\bar{s},\bar{d})}{\partial\lambda_2\partial\lambda_1} = \frac{1}{L(\bar{s},\bar{d})}\frac{\partial^2 L(\bar{s},\bar{d})}{\partial\lambda_2\partial\lambda_1} - \frac{1}{L(\bar{s},\bar{d})^2}\frac{\partial L(\bar{s},\bar{d})}{\partial\lambda_1}\frac{\partial L(\bar{s},\bar{d})}{\partial\lambda_2}. \tag{24}$$

We have already shown how to derive all the terms in (24) except the mixed partial, so we need only derive these to complete this section.

If $\lambda_1 = \gamma_q$ and $\lambda_2 = \gamma_s$, then differentiating (23), using (22) yields:

$$\frac{\partial^2 L(\bar{s},\bar{d})}{\partial\gamma_s\partial\gamma_q} = \sum_{k\in K}[1(q=k)\frac{\partial P_q}{\partial\gamma_s} - P_q\frac{\partial P_k}{\partial\gamma_s} - P_k\frac{\partial P_q}{\partial\gamma_s}]L(\bar{s},\bar{d}|\bar{\theta}_k). \tag{25}$$

If $\lambda_1 = \gamma_q$ and $\lambda_2 \notin \{\gamma_1, ..., \gamma_{K-1}\}$, differentiate (23) to get:

$$\frac{\partial^2 L(\bar{s},\bar{d})}{\partial\lambda_2\partial\gamma_q} = \sum_{k\in K}\frac{\partial P_k}{\partial\gamma_q}\frac{\partial L(\bar{s},\bar{d}|\bar{\theta}_k)}{\partial\lambda_2}. \tag{26}$$

The only remaining case is $\lambda_1, \lambda_2 \notin \{\gamma_1, ..., \gamma_{K-1}\}$. Using (20):

$$\begin{aligned}\frac{\partial^2 L(\bar{s},\bar{d})}{\partial\lambda_2\partial\lambda_1} = &\sum_{k\in K} P_k L(\bar{s},\bar{d}|\bar{\theta}_k)\{\frac{\partial \ln L(\bar{s},\bar{d}|\bar{\theta}_k)}{\partial\lambda_2}\frac{\partial \ln L(\bar{s},\bar{d}|\bar{\theta}_k)}{\partial\lambda_1} \\ &+ \frac{\partial^2 \ln L(\bar{s},\bar{d}|\bar{\theta}_k)}{\partial\lambda_2\partial\lambda_1}\}.\end{aligned} \tag{27}$$

Again, we have shown how to calculate all terms except the mixed partial. Differentiating (14) we get:

$$\begin{aligned}\frac{\partial^2 \ln L(\bar{s},\bar{d}|\bar{\theta}_k)}{\partial\lambda_2\partial\lambda_1} = \tau\sum_{t=1}^{N}\{\sum_{i\in I}[(d_i(t) - z_i(s(t),\bar{\theta}_k))\frac{\partial^2\tilde{V}_i(s(t),\bar{\theta}_k)}{\partial\lambda_2\partial\lambda_1} \\ - \frac{\partial z_i(s(t),\bar{\theta}_k)}{\partial\lambda_2}\frac{\partial\tilde{V}_i(s(t),\bar{\theta}_k)}{\partial\lambda_1}]\}.\end{aligned} \tag{28}$$

From the definition of $z_i(s(t), \bar{\theta}_k)$:

$$\frac{\partial z_i(s(t), \bar{\theta}_k)}{\partial \lambda_2} = \tau z_i(s(t), \bar{\theta}_k) \sum_{j \in I} [1(j = i) - z_j(s(t), \bar{\theta}_k)] \frac{\partial \tilde{V}_j(s(t), \bar{\theta}_k)}{\partial \lambda_2}. \tag{29}$$

To complete the derivation, we need the mixed partial on the right-hand side of (28). Differentiating (19) we have:

$$\begin{aligned}\frac{\partial^2 \tilde{V}_i(s(t), \bar{\theta}_k)}{\partial \lambda_2 \partial \lambda_1} =& \frac{\partial^2 g_i(s(t), \theta_{ik})}{\partial \lambda_2 \partial \lambda_1} + \\ & \beta\{\sum_{j \in I} [\frac{\partial z_j(s(t+1), \bar{\theta}_k)}{\partial \lambda_2} \frac{\partial \tilde{V}_j(s(t+1), \bar{\theta}_k)}{\partial \lambda_1} + \\ & z_j(s(t+1), \bar{\theta}_k) \frac{\partial^2 \tilde{V}_j(s(t+1), \bar{\theta}_k)}{\partial \lambda_2 \partial \lambda_1}]|d_i(t) = 1\}.\end{aligned} \tag{30}$$

Note that the term to the right of β is zero when $t = T$, so we can calculate this directly at T. But then we can calculate this for $T - 1$ and, by backward induction, for all t. This completes the derivation of the analytic second derivatives.

6 Conclusion

In this paper we show how to recursively calculate analytic first and second derivatives for a popular specification of a structural discrete choice model. Obtaining these derivatives is no more difficult than recursively calculating the value of the likelihood function. Our approach will drastically reduce the computing and debugging time necessary for estimation routines for this model that use derivatives. Our approach also makes it easier to get closer to the exact optimum of the function. Finally, our approach will also aid in obtaining asymptotic standard errors for parameter estimates of the model, independently of whether one uses a derivative based algorithm to estimate the model.

References

Arcidiacono, P., H. Sieg and F. Sloan (2007), “Living Rationally Under the Volcano? An Empirical Analysis of Heavy Drinking and Smoking.” *International Economic Review*, **48**, 37–65.

Eberwein, C., J. Ham and R. LaLonde (1997), "The Impact of Being Offered and Receiving Classroom Training on the Employment Histories Of Disadvantaged Women: Evidence from Experimental Data," *The Review of Economic Studies*, **64**, 655–682.

Heckman, J. and B. Singer (1984), "Econometric Duration Analysis", *Journal of Econometrics*, **24**, 63-132.

Liu, H., T. Mroz and W. Van der Klaauw (2004), "Maternal Employment, Migration, and Child Development." Manuscript, East Carolina University.

Rust, J. (1987), "Optimal Replacement of GMC Bus Engines: An Empirical Analysis of Harold Zurcher", *Econometrica*, **55**, 999–1033.

Van Der Klaauw, W. (1996), "Female Labour Supply and Marital Status Decisions: A Life-Cycle Model", *Review of Economic Studies*, **63**, 199–235.