Web Appendix

Estimation of Dynamic Discrete Choice Models by Maximum Likelihood and the Simulated Method of Moments*

Philipp Eisenhauer
The University of Chicago

James J. Heckman
The University of Chicago
Center for the Economics of Human Development
American Bar Foundation

Stefano Mosso
The University of Chicago

October 1, 2014

*For further information or questions and suggestions, please contact us at info@policy-lab.org. We are grateful to George Yates for his help in preparing this appendix.

Contents

A Identification

We establish that our model is semi-parametrically identified. Our estimated model of schooling restricts agents to binomial choices at each decision node and there is no role for time. However, we provide identification results for a broader class of models. We allow for multinomial choices and introduce time $t \in \mathcal{T} = \{1, \ldots, T\}$. The model in the paper is a special case of our more general analysis. In this more flexible model, earnings functions are specified by:

$$Y(t,s) = \mu_{t,s}(X(t,s)) + \theta' \alpha_{t,s} + \epsilon(t,s),$$

let $p(t,s) = \theta' \alpha_{t,s} + \epsilon(t,s)$. The costs functions are specified by:

$$C(t,s',s) = K_{t,s',s}(Q(t,s',s)) + \theta' \varphi_{t,s',s} + \eta(t,s',s),$$

let $w(t,s',s) = \theta' \varphi_{t,s',s} + \eta(t,s',s)$. Finally, the measurement functions are specified by:

$$M(j) = \mu_j(X(j)) + \theta' \gamma_j + \nu(j),$$

let $e(j) = \theta' \gamma_j + \nu(j)$.

The observed components are determined by covariates $X(t,s) \in \mathcal{X}(t,s)$ for earnings, $Q(t,s',s) \in \mathcal{Q}(t,s',s)$ for costs, and $X(j) \in \mathcal{X}(j)$ for measurements. We show that all functions $\mu_{t,s}(X(t,s))$, $K_{t,s',s}(Q(t,s',s))$, $\mu_j(X(j))$ and all distributions $F_{P(t,s)}(p(t,s))$ of unobservables for outcome equations, all distributions $F_{W(t,s',s)}(w(t,s',s))$ of the unobservables in all costly exits from each state, and all distributions $F_{E(j)}(e(j))$ of the unobservables in all measurement equations are identified for any t, s', s, and j. We extend the results from Heckman and Navarro (2007) to a context of recurring states and multinomial transitions. To simplify notation we remove individual subscripts and consider vectors of individual observations indexed over t and s. Variables without arguments refer to any t, s, j, and i.

Define $U(t',\omega \,|\, \mathcal{I}(t,s)) = -K_{t',\omega,s}(Q(t',\omega,s)) + \mathbb{E}[V(t',\omega) \,|\, \mathcal{I}(t,s)]$ and consider the difference:

$$\Delta[t',\omega \big|\, \mathcal{I}(t,s)] = (U(t',\omega \,|\, \mathcal{I}(t,s)) - w(t',\omega,s)) - \max_{\substack{\sigma \in \Omega(t,s) \\ \sigma \neq \omega}} (U(t',\sigma \,|\, \mathcal{I}(t,s)) - w(t',\sigma,s)),$$

such that state ω is picked whenever $\Delta[t',\omega \big|\, \mathcal{I}(t,s)] > 0$. This condition defines a partition in the space of the unobservables such that state ω is selected.

Theorem 1. *Assume that:*

(i) P, W, *and* E *are continuous random variables with mean zero, finite variance, and support* $Supp(P) \times Supp(W) \times Supp(E)$. *Assume that the cumulative distribution function of* W *is strictly increasing over its full support for any* t *and* s.

(ii) $X, Q \perp\!\!\!\perp (P, W, E)$ *for all* t *and* s.

(iii) $Supp(\mu(X), \mu_j(X), U(Q)) = Supp(\mu(X)) \times Supp(\mu_j(X)) \times Supp(U(Q))$.

(iv) $Supp(-W) \subseteq Supp(U(Q))$ *for any* t *and* s.

Then $\mu_{t,s}(X(t,s))$ *is identified for any* t *and* s, $\mu_j(X(j))$ *is identified for all* j, *and the joint distribution* $F_{P(t,s),E(j)}(p(t,s),e(j))$ *is identified for any* t, s, j.

Proof. Conditions (iii) and (iv) guarantee that there exist sets $\overline{\mathcal{Q}}(t,s',s)$ such that

$$\lim_{Q(t,s',s) \to \overline{\mathcal{Q}}(t,s',s)} \mathrm{P}(\Delta[t',s'] > 0) = 1.$$

In the limit sets, we can form:

$$Pr[p(t,s) < Y(t,s) - \mu_{t,s}(X(t,s)), e(j) < M(j) - \mu_j(X(j)) \,|\, X(j) = x(j),\, X(t,s) = x(t,s)] = \\ = F_{P(t,s),E(j)}(Y(t,s) - \mu_{t,s}(x(t,s)), M(j) - \mu_j(x(j))),$$

and then we can trace out the whole distribution $F_{P(t,s),E(j)}(p(t,s),e(j))$ by independently varying the points of evaluation. □

Whenever the limit set condition is not satisfied in the analyzed sample, then identification relies either on the assumption that in large samples such limit sets exist, or it is conditional on a subset and only bounds for model parameters can be recovered. Notice that the plausibility of these conditions depends on the postulated model. In particular, the richer the specification for the set of feasible future states $\mathcal{S}^f(t,s)$ and the finer the time partition for the model, the harder it is to have this condition satisfied in the data. Fewer observations will populate each state in any given finite sample. Given the above theorem, which mimics Theorem 4 in Heckman and Navarro (2007), we can identify the joint distribution of outcomes across different states s and times t using factor analysis as described in the aforementioned paper. Factor analysis also allows to identify the factor loadings $(\alpha_{t,s}, \gamma_j)$ and to separately identify the marginal distributions of the factors θ and the marginal distribution of the idiosyncratic shocks $\epsilon(t,s)$ and $\nu(j)$ for any t, s, and j. Note that the measurement system is not needed for identification of the factor distributions if the state space is sufficiently large (the number of states plus the number of transitions is greater than $2N+1$ when N is the number of factors). However, it increases efficiency and aids in the interpretation of the factors, e.g., as cognitive and non-cognitive abilities.

Theorem 2. *Assume that:*

(i) Conditions (i) to (iv) of Theorem 1 are satisfied.

(ii) $K_{t,s}(Q(t,s',s))$ *is a continuous function for any t and any s.*

(iii) $Q(t,s',s) \in \mathcal{Q}$*, a common set over t and s.*

(iv) For each transition remaining in the current state is always a costless option. For an agent in state s in t: $K_{t',s',s}(Q(t',s',s)) + w(t',s',s) = 0$ *if* $s' = s$.

(v) For all alternatives $\omega \in \Omega(t,s)$ *there exist a coordinate of* $Q(t',\omega,s)$ *that possesses an everywhere positive Lebesgue density conditional on the other coordinates and it is such that* $K_{t',\omega,s}(Q(t',\omega,s))$ *is strictly increasing in this coordinate.*

(vi) $U(t',\omega \mid \mathcal{I}(t,s))$ *belongs to the class of Matzkin (1993) functions according to her Lemmas 3 and 4.*

Then we identify the function $K_{t,\omega,s}(Q(t,\omega,s))$*, the marginal distribution of the unobservable portion of the cost*

functions $F_{W(t,\omega,s)}(w(t,\omega,s))$, and exploiting the factor structure representations, the factor loadings $\varphi_{t,\omega,s}$ and marginal distribution of the idiosyncratic shocks in the costs functions $F_{H(t,\omega,s)}(\eta(t,\omega,s))$ for all transitions.

Proof. Consider all final transitions. We define transitions to be final when they lead to final states. A state s is defined as final if $\Omega(t,s) = \{s\}$ for all t. No choice is left to the agent but to remain in the current state. Recall that remaining in the current state involves no costs. For any final state $\omega \in \Omega(t,s)$ we have:

$$\begin{aligned} U(t',\omega \,|\, \mathcal{I}(t,s)) &= -K_{t',\omega,s}(Q(t',\omega,s)) + \mathbb{E}[V(t',\omega) \,|\, \mathcal{I}(t,s)] \\ &= -K_{t',\omega,s}(Q(t',\omega,s)) + \mathbb{E}[(\mu_{t',\omega}(X(t',\omega)) + p(t',\omega)) \,|\, \mathcal{I}(t,s)] \\ &= -K_{t',\omega,s}(Q(t',\omega,s)) + \mu_{t',\omega}(X(t',\omega)) + \mathbb{E}[p(t,\omega) \,|\, \Delta[t',\omega \,|\, \mathcal{I}(t,s)] > 0,\ \mathcal{I}(t,s)] \\ &= -K_{t',\omega,s}(Q(t',\omega,s)) + \mu_{t',\omega}(X(t',\omega)) + \theta'\alpha_{t,\omega}. \end{aligned}$$

Notice that $\mu_{t',\omega}(X(t',\omega)) + \theta'\alpha_{t,\omega}$ is known by Theorem 1 and due to the factor structure assumption. Thus we can identify the cost equation $K_{t',\omega,s}(Q(t',\omega,s))$. Imposing restrictions on the generality of the cost function $K_{t',\omega,s}(Q(t',\omega,s))$ is necessary such that $U(t',\omega \,|\, \mathcal{I}(t,s))$ satisfies (ii), (v), and (iv). Standard arguments from Matzkin (1993) guarantee identification of the function $K_{t,s',s}(Q(t,s',s))$. We do not have to worry about the fact that only differences in utilities are identified in her setup as by (iii), we always have an alternative which implies zero costs. We can also identify the distribution $F_{W(t',\omega,s)}(w(t',\omega,s))$ for any final states. Exploiting the factor structure we can then identify the joint distribution $F_{W(t',\omega,s),P(t',\omega,s),E(j)}(w(t',\omega,s), p(t',\omega,s), e(j))$ for all final transitions and by isolating the dependency between unobservables, we identify the marginal distribution $F_{H(t,\omega,s)}(\eta(t,\omega,s))$ for each final transition. Once these are obtained, by backward induction all expected value functions are identified and therefore all $K_{t,s',s}(Q(t,s',s))$ and $F_{W(t',\omega,s),P(t',\omega,s),E(j)}(w(t',\omega,s), p(t',\omega,s), e(j))$ for any transition and all marginal distributions $F_{H(t,\omega,s)}(\eta(t,\omega,s))$ for any transition are identified. Note that linearity does not fulfill the necessary conditions and only allows for identification up to scale. We therefore need to consider the case separately where the scale of the cost function is not identified. □

Theorem 3. *Assume that:*

(i) Conditions of Theorem 1 and 2 are satisfied, but for the fact that the scale of $K_{t,s',s}(Q(t,s',s))$ is not

identified as when it is linear.

(ii) (a) In any final state, $\mathcal{X}(t,s)\setminus\mathcal{Q}(t,s's)$ is not empty and $\mu_{t,s}(X(t,s))$ has an additive component which depends only on variables in $\mathcal{X}(t,s)\setminus\mathcal{Q}(t,s's)$. Alternatively, (b) there is a coordinate of the vector $Q(t,s',s)$ such that $K_{t,s',s}(Q(t,s's))$ is additively separable in that coordinate and it has a known coefficient.

Then the scale of $K_{t,s',s}(Q(t,s',s))$ is determined.

Proof. Assumption (ii.a) guarantees that there is a component which can be identified in the outcome equations by the limit sets argument and that can be independently varied from other elements in $U(t,s)$. Applying (ii.b) implies that the scale is known. Notice that the expected value function has an equivalent role as one of the variables in the set defined by (ii.a) for any non final transition, provided that the discount rate is known. Otherwise, if the discount rate is not known and therefore appears as a coefficient in front of $U(t',\omega)$ for future accessible states, we require exclusion restrictions of the type in (ii) in at least one non final transition to identify it. □

Following the analysis of Heckman and Navarro (2007), we can identify the discount rate under the same conditions given there.

B Data Description

Our baseline data is the NLSY79 (Bureau of Labor Statistics, 2001). We restrict our sample to white males only. We construct longitudinal schooling histories by compiling all information on school attendance, including self-reports and the high school survey. We then check the compatibility of all the information for each individual within and across time. In the presence of contradictions, we review all information for the questionable observation and try to identify the source of the error and correct it. If impossible, we drop the observation. Finally, we impose the structure of our decision tree on the agents' educational histories. We ignore any form of adult education.

We use the following set of observables: annual earnings, current geographic location, small child in household, number of siblings, mother's and father's education, dummy variables for marriage status, intact families in 1979, south at age 14, and urban area at age 14.

We impute missing values. When dealing with time constant covariates, imputation is straightforward. If information on time varying covariates is missing for only a few years, we use a three year moving average for continuous covariates and the last value for discrete variables. Otherwise the agent is dropped from our sample. If annual earnings are missing for a limited time only, we impute them using a three year moving average.

We use tuition data for two- and four-year colleges from the Integrated Postsecondary Education Data System (IPEDS). We carefully construct state averages. We ensure comparability of the tuition data over time and address the change in the definitions in 1986.[1] We only use tuition from public universities. We construct local economic conditions such as hourly wages and unemployment using the Current Population Survey (CPS) data by state, level of education, ethnicity, and gender. We merge all datasets using the NLSY Geocode Data.

[1]We thank Amanda Agan for spotting the inconsistency and suggesting the solution to it in accordance with suggestions received from the statisticians at the National Center for Education Statistics.

C Rates of Return, Option Values, and Regret

Table 1 presents internal rates of return for selected comparisons of schooling levels. For definition of this traditional concept, see Heckman et al. (2006). We compare the recorded earnings streams until age 45. We therefore consider earnings in all states up to the one in the first column. Missing earnings are set to zero, unless during high school enrollment. There we impute a three year moving average.

Table 1: Internal Rates of Return

All			
High School Graduation	vs.	High School Dropout	215%
Early College Graduation	vs.	Early College Dropout	24%
Early College Graduation	vs.	High School Graduation (cont'd)	19%
Late College Dropout	vs.	High School Graduation (cont'd)	10%
Late College Graduation	vs.	High School Graduation (cont'd)	17%
Late College Dropout	vs.	High School Graduation (cont'd)	16%

Notes: The calculation is based on 1,407 individuals in the observed data.

The Mincer rate of return is 11.6%.

Table 2 reports the median *ex ante* net returns to education by treatment status. We condition on agents that actually visit the relevant decision state. The treated choose the transition to the state in the first column.

Table 2: Net Returns

State	All	Treated	Untreated
High School Finishing	64%	75%	-27%
Early College Enrollment	-3%	24%	-28%
Early College Graduation	50%	82%	-44%
Late College Enrollment	-21%	22%	-38%
Late College Graduation	10%	62%	-51%

Notes: We simulate a sample of 50,000 agents based on the estimates of the model.

Table 3 reports the average *ex ante* gross returns to education by treatment status. We condition on agents that actually visit the relevant decision state. The treated choose the transition to the state in the first column.

Table 3: Gross Returns

State	All	Treated	Untreated
High School Finishing	27%	29%	16%
Early College Enrollment	14%	20%	8%
Early College Graduation	75%	84%	49%
Late College Enrollment	29%	28%	29%
Late College Graduation	24%	36%	9%

Notes: We simulate a sample of 50,000 agents based on the estimates of the model.

Table 4 shows the percentage of agents experiencing regret, i.e., those agents for which the *ex post* and *ex ante* returns do not agree in sign. We condition on agents that actually visit the relevant decision state. The treated choose the transition to the state in the first column.

Table 4: Regret

State	All	Treated	Untreated
High School Finishing	7%	4%	24%
Early College Enrollment	15%	28%	2%
Early College Graduation	29%	33%	19%
Late College Enrollment	21%	27%	19%
Late College Graduation	27%	34%	18%

Notes: We simulate a sample of 50,000 agents based on the estimates of the model.

Table 5 reports the option value contribution, i.e., the relative share of the option value in the overall value of each state. We condition on agents that actually visit the relevant decision state. The treated choose the transition to the state in the first column.

Table 5: Option Value Contribution

State	All	Treated	Untreated
High School Finishing	7%	8%	2%
Early College Enrollment	30%	37%	23%
Late College Enrollment	17%	24%	15%

Notes: We simulate a sample of 50,000 agents based on the estimates of the model.

D Approximation Error in Adaptive Gauss-Hermite Quadrature

We present an analysis of the accuracy of the Gauss-Hermite quadrature method in the context of our model.[2] We consider a simple model with three states and a single transition. This section is self-contained and the notation is independent from the rest of the paper.

Denote the start state as S with two possible exits, states Z and Q. The transition from S to Q has an associated cost equation while the transitions from S to Z is costless. Suppose further that state S has no contemporaneous value, while Q and Z are characterized by an associated outcome equation. Furthermore, suppose that the model specifies one unobserved normally distributed factor and one measurement equation. All unobserved components are assumed to be normally distributed with mean zero. The components of the likelihood therefore are the following.

- The factor density:
$$f_1(\theta) = \frac{1}{\sigma_1}\phi\left(\frac{\theta}{\sigma_1}\right). \tag{1}$$

- The distribution of the random disturbance for the outcome equations in state Q and Z. Let i in $\{2,3\}$ index states $\{Q,Z\}$. States 2 and 3 are characterized by a single linear outcome equation whose disturbance is specified by the density:
$$f_i(\epsilon_i \mid x_i,\theta) = \frac{1}{\tau_i}\phi\left(\frac{y_i - x_i'\beta_i - \alpha_i\theta}{\tau_i}\right) \tag{2}$$
which can be rewritten emphasizing its interpretation as a density for the factor viewed as a random effect on each outcome:
$$f_i(\epsilon_i \mid x_i,\theta) = \frac{1}{|\alpha_i|\frac{\tau_i}{|\alpha_i|}}\phi\left(-\frac{\theta - \frac{y_i - x_i'\beta_i}{\alpha_i}}{\frac{\tau_i}{\alpha_i}}\right) = \frac{1}{|\alpha_i|}\frac{1}{\sigma_i}\phi\left(\frac{\theta-\mu_i}{\sigma_i}\right)$$
where $\mu_i \equiv \frac{y_i - x_i'\beta_i}{\alpha_i}$ and $\sigma_i \equiv \frac{\tau_i}{|\alpha_i|}$.

[2]We thank George Yates for his help in preparing this appendix.

- The density of the random disturbance, ϵ_4, in the measurement equation:

$$f_4(\epsilon_4 \mid x_4, \theta) = \frac{1}{\tau_4}\phi\left(\frac{m_4 - x_4'\beta_4 - \alpha_4\theta}{\tau_4}\right) \tag{3}$$

  which can also be rewritten emphasizing its interpretation as a density for the factor viewed as a random effect on the measurement:

$$f_4(\epsilon_4 \mid x_4, \theta) = \frac{1}{|\alpha_4|}\frac{1}{\sigma_4}\phi\left(\frac{\theta - \mu_4}{\sigma_4}\right)$$

  where $\mu_4 \equiv \frac{y_4 - x_i'\beta_4}{\alpha_4}$ and $\sigma_4 \equiv \frac{\tau_4}{|\alpha_4|}$.

- The density of the cost disturbance in the equations determining the costs of going from S to Q. Given that Q and Z are both final states, denoting by C the variables determining the cost equation, the difference in their values is given by:

$$\Delta V_{QZ} = V_Q - V_Z = (x_Q'\beta_Q + \alpha_Q\theta) - (x_C'\beta_C + \alpha_C\theta + \epsilon_5) - (x_Z'\beta_Z + \alpha_Z\theta) \tag{4}$$

$$= (x_Q'\beta_Q - x_C'\beta_C - x_Z'\beta_Z) + (\alpha_Q - \alpha_C - \alpha_Z)\theta \tag{5}$$

  Define the vectors:

$$x_5 \equiv [x_Q \quad x_C \quad x_Z] \qquad \beta_5 \equiv [\beta_Q \quad \beta_C \quad \beta_Z]$$

  and the scalar:

$$\alpha_5 \equiv \alpha_Q - \alpha_C - \alpha_Z.$$

  We then have that the probability of an agent of selecting one of the final states j in $\{Q, Z\}$ is given by:

$$P(j \mid x_5, \theta) = g(\theta \mid x_5) = \begin{cases} \Phi\left(\frac{x_5'\beta_5 + \alpha_5\theta}{\tau_5}\right) & \text{if } j = Q \text{ is picked,} \\ 1 - \Phi\left(\frac{x_5'\beta_5 + \alpha_5\theta}{\tau_5}\right) & \text{if } j = Z \text{ is picked.} \end{cases} \tag{6}$$

Notice that the likelihood is now a multivariate normal distribution times the CDF component where

everything is a function of the unknown factor and of the covariates:

$$\begin{aligned} L(\boldsymbol{\Psi} \mid \boldsymbol{X}) &= \int_{\mathbb{R}} \left[\prod_{k=1}^{4} f_k(\theta) \right] g(\theta) d\theta \\ &= \frac{1}{|\alpha_2 \alpha_3 \alpha_4|} \int_{\mathbb{R}} \frac{1}{(2\pi)^{k/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} (\theta - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\theta - \boldsymbol{\mu}) \right) g(\theta) d\theta \end{aligned} \tag{7}$$

where:

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & \sigma_3 & 0 \\ 0 & 0 & 0 & \sigma_4 \end{bmatrix}.$$

The above multivariate normal distribution can be factored into a single distribution. This implies that the above likelihood can be written as:

$$\begin{aligned} L(\boldsymbol{\Psi} \mid \boldsymbol{X}) &= \frac{1}{|\alpha_2 \alpha_3 \alpha_4|} \int_{\mathbb{R}} \left[\prod_{k=1}^{4} \frac{1}{\sigma_k} \phi\left(\frac{\theta - \mu_k}{\sigma_k} \right) \right] g(\theta) d\theta \\ &= \kappa \int_{\mathbb{R}} \left[\frac{1}{\sigma_*} \phi\left(\frac{\theta - \mu_*}{\sigma_*} \right) \right] g(\theta) d\theta \end{aligned} \tag{8}$$

where κ, μ_*, and σ_* can be found by simple algebra.

By use of this example we can characterize the approximation error in adopting the adaptive Gauss-Hermite quadrature method to perform the integral specified in the above likelihood. Suppose the agent chooses cost-exit Q when exiting state S. Using the formula for the Gauss-Hermite quadrature (Judd, 1998) and the standard mathematical notation $(\Phi \circ \varphi)(\theta)$ for the function composition $\Phi(\varphi(\theta))$, the probability of the agent's observed decision is the smooth function:

$$G(\theta) = g(\theta) = \Phi\left(\frac{x_5' \beta_5 + \alpha_5 \theta}{\tau_5} \right) = (\Phi \circ \varphi)(\theta), \tag{9}$$

where

$$\varphi(\theta) \equiv r_5 \theta + s_5 \text{ with } r_5 \equiv \frac{\alpha_5}{\tau_5} \text{ and } s_5 \equiv \frac{x_5' \beta_5}{\tau_5}. \tag{10}$$

Agent specific parameters μ_*, σ_* are specified by the above multivariate normal distribution, whence

the integral in (8) is

$$\int_{\mathbb{R}} \frac{1}{\sigma_*}\phi\left(\frac{\theta-\mu_*}{\sigma_*}\right) G_1(\theta)d\theta = \int_{-\infty}^{\infty} \frac{e^{-y^2}}{\sqrt{\pi}} \left(\Phi \circ \varphi \circ A_H\right)(y|\mu_*,\sigma_*)\, dy$$
$$\approx \sum_{n=1}^{N} \vartheta_n \Phi\left(\sqrt{2}t_5 P_n + u\right), \tag{11}$$

where $A_H(y|\mu_*,\sigma_*) \equiv \left(\sqrt{2}\sigma_* y + \mu_*\right)$, $(\varphi \circ A_H)(y|\mu_*,\sigma_*) = r_5\left(\sqrt{2}\sigma_* y + \mu_*\right) + s_5 = \sqrt{2}t_5 y + u$ by standard formulas for the Gauss-Hermite quadrature (Judd, 1998) and (10) with

$$t_5 \equiv r_5\sigma_* = \frac{\alpha_5\sigma_*}{\tau_5} \text{ and } u \equiv r_5\mu_* + s_5 = \frac{x_5'\beta_5 + \alpha_5\mu_*}{\tau_5}. \tag{12}$$

Lemma 1. *Approximation* (11) *is exact when* $\alpha_5 = 0$.

Proof. By (10), $\alpha_5 = 0 \Rightarrow r_5 = 0 \Rightarrow G(\theta) = \Phi(s_5)$. So, mathematically, when $\alpha_5 = 0$,

$$\int_{\mathbb{R}} \frac{1}{\sigma_*}\phi\left(\frac{\theta-\mu_*}{\sigma_*}\right) G(\theta)d\theta = \Phi(s_5)\int_{\mathbb{R}} \frac{1}{\sigma_*}\phi\left(\frac{\theta-\mu_*}{\sigma_*}\right) d\theta = \Phi(s_5) = \Phi\left(\frac{x_5'\beta_5}{\tau_5}\right). \tag{13}$$

Also, by (12), (10), $\alpha_5 = 0 \Rightarrow (t_5 = 0 \text{ and } u = s_5)$. So, algorithmically, using quadrature formulas,

$$\int_{\mathbb{R}} \frac{1}{\sigma_*}\phi\left(\frac{\theta-\mu_*}{\sigma_*}\right) G(\theta)d\theta \approx \sum_{n=1}^{N} \vartheta_n \Phi\left(\sqrt{2}t_5 P_n + u\right) = \Phi(u)\sum_{n=1}^{N} \vartheta_n = \Phi(s_5). \tag{14}$$

By Lemma 1, in analysis of approximation error, we may assume $\alpha_5 \neq 0$. In the following discussion, we develop intuition about the accuracy of approximation (11) when $\alpha_5 \neq 0$. Define

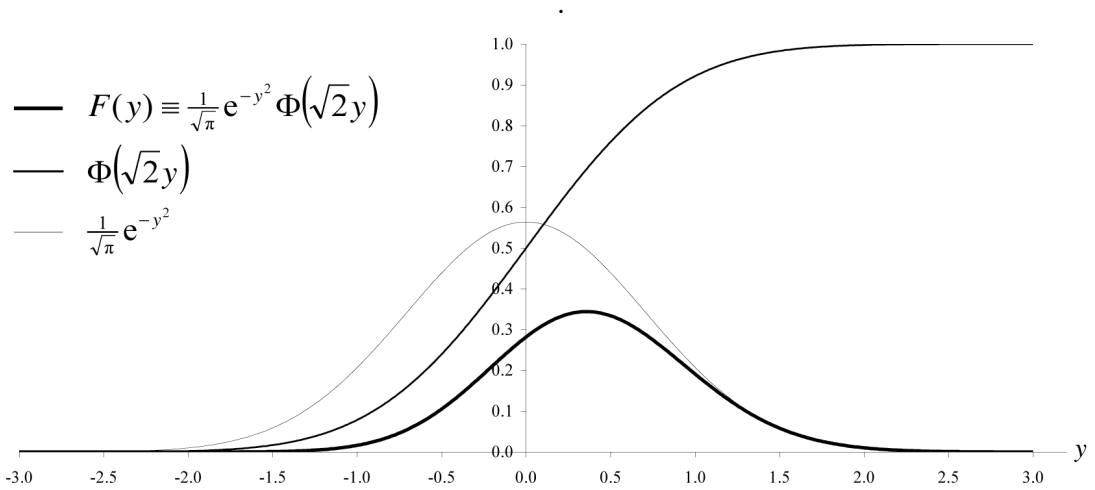
$$F(y) \equiv \frac{1}{\sqrt{\pi}} e^{-y^2} (\Phi \circ \varphi \circ A_H)(y|\mu_*,\sigma_*) \equiv \frac{1}{\sqrt{\pi}} e^{-y^2} \Phi(\sqrt{2}t_5 y + u) \tag{15}$$

where the equalities follow from (10) and (12).

$F(y)$ is the integrand in the middle expression of (11), the likelihood integrand (for the simple model) transformed for adaptive Gauss-Hermite quadrature. Therefore by (8), κ times the integral of $F(y)$ over $\mathbb{R}$ is the likelihood of an observed agent experience, conditional on the factor. Figure 1 is

a plot of $F(y)$ and its multiplier components as defined by (15), with $t_5 = 1$ and $u = 0$. The figure and the others below are presented to support initial intuition about the geometry of the likelihood form in the simple model with one agent decision.

Figure 1: Plot of the transformed likelihood integrand components of (15) with parameters $t_5 = 1$ and $u = 0$



Note: The complete integral of $F(y) = 0.5$ for all t_5 when $u = 0$. See Lemma 2 below.

Notice that:

$$\frac{1}{\sqrt{\pi}} e^{-y^2} = \frac{1}{\sqrt{2\pi}\frac{1}{\sqrt{2}}} e^{-\frac{1}{2}\left(y/\frac{1}{\sqrt{2}}\right)^2} = \sqrt{2}\phi(\sqrt{2}y) = \text{the pdf of the } N\left(0, \frac{1}{\sqrt{2}}\right) \text{ distribution.} \tag{16}$$

Thus,

$$\int_{\mathbb{R}} F(y)dy = \int_{\infty}^{\infty} \frac{e^{-y^2}}{\sqrt{\pi}} \Phi(\sqrt{2}t_5 y + u)dy = \sqrt{2} \int_{-\infty}^{\infty} \phi(\sqrt{2}y)\Phi(\sqrt{2}t_5 y + u)dy. \tag{17}$$

Then,

$$z \equiv \sqrt{2}y \Rightarrow \int_{\mathbb{R}} F(y)dy = \sqrt{2} \int_{\infty}^{\infty} \phi(z)\Phi(t_5 z + u)\frac{dz}{\sqrt{2}} = \int_{\infty}^{\infty} \phi(z)\Phi(t_5 z + u)dz. \tag{18}$$

Lemma 2. *If $u = 0$ then, for all $t_5 \in \mathbb{R} \setminus \{0\}$, (a) $\int_{\mathbb{R}} F(y)dy = \frac{1}{2}$ and (b) approximation* (11) *is exact for all $N > 0$.*

Proof. Pick any $t_5 \neq 0$ and note that, by (18) and (11),

$$u = 0 \Rightarrow \int_{\mathbb{R}} F(y)dy = \int_{-\infty}^{\infty} \phi(z)\Phi(t_5 z)dz \approx \sum_{n=1}^{N} \vartheta_n \Phi(\sqrt{2}t_5 P_n) \tag{19}$$

Also recall two symmetry properties of the standard normal distribution. For all $x \in \mathbb{R}$,

$$\begin{aligned} \Phi(-x) &= 1 - \Phi(x) \\ \phi(-x) &= \phi(x). \end{aligned} \tag{20}$$

Then (a) in Lemma 2 is proved by a simple calculation using (19) and (20). By (19):

$$\int_{\mathbb{R}} F(y)dy = \int_{-\infty}^{0} \phi(z)\Phi(t_5 z)dz + \int_{0}^{\infty} \phi(z)\Phi(t_5 z)dz. \tag{21}$$

To the first integral on the right in (21) apply change of variables $s \equiv -z \geq 0$. Using (20):

$$\int_{-\infty}^{0} \phi(z)\Phi(t_5 z)dz = -\int_{\infty}^{0} \phi(-s)\Phi(-t_5 s)ds = \int_{0}^{\infty} \phi(s)[1 - \Phi(t_5 s)]ds = \frac{1}{2} - \int_{0}^{\infty} \phi(s)\Phi(t_5 s)ds. \tag{22}$$

Substituting (22) into (21), we see immediately: $\int_{\mathbb{R}} F(y)dy = \frac{1}{2}$. (a) is proved.

Finally, (b) is proved using (19), (20), and properties of Gauss-Hermite points and weights W_N. When $N = 1$ is immediate:

$$W_1 = \{(0, \sqrt{\pi})\} \Rightarrow \sum_{n=1}^{N} \vartheta_n \Phi(\sqrt{2}t_5 P_n) = \Phi(0) = \frac{1}{2}. \tag{23}$$

When $N > 1$, we use properties of points and weights in W_N for all N. Specifically, (b.1) the points in W_N, zeros of the Hermite polynomial of degree N, occur in pairs $\{\pm P_m : m = 1, \ldots, \lfloor \frac{N}{2} \rfloor\}$ equal but opposite in sign. If N is odd, the last point $P_0 = 0$. Moreover, (b.2) for each m indexing a pair $\pm P_m$, the

weights $\{W_m^{\pm} := 1, \ldots, \lfloor \frac{N}{2} \rfloor\}$ in W_N are equal and positive; that is, $W_m^+ = W_m^- \equiv W_m > 0$ for each n. If N is odd, $W_0 = \sqrt{\pi} - 2\sum_{m=1}^{\frac{N-1}{2}} W_m$. Thus, when N is even,

$$\sum_{n=1}^{N} \vartheta_n \Phi(\sqrt{2}t_5 P_n) = \sum_{m=1}^{\frac{N}{2}} \vartheta_m \left[\Phi(\sqrt{2}t_5 P_m) + \Phi(-\sqrt{2}t_5 P_m)\right] = \sum_{m=1}^{\frac{N}{2}} \vartheta_m = \frac{1}{2};$$

and when N is odd:

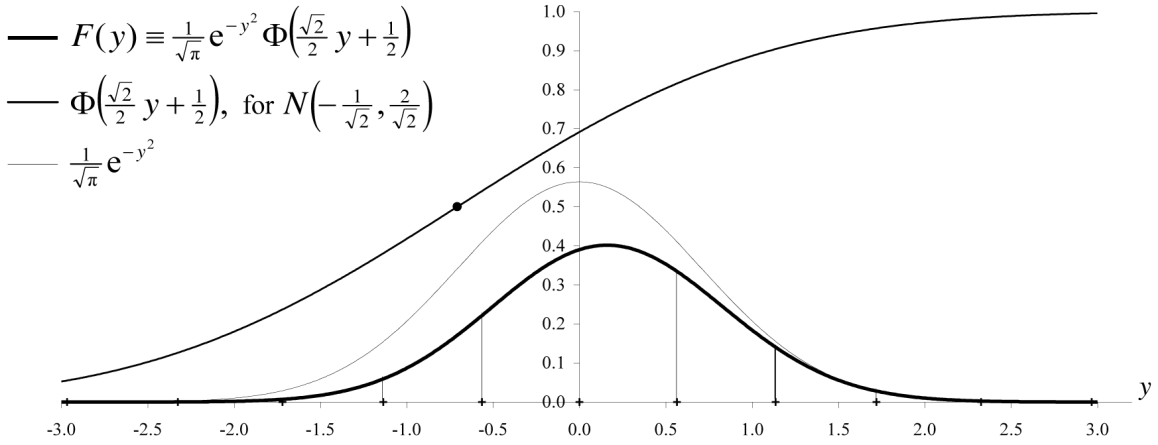
$$\sum_{n=1}^{N} \vartheta_n \Phi(\sqrt{2}t_5 P_n) = \sum_{m=1}^{\frac{N-1}{2}} \vartheta_m = \frac{1}{2}\left[1 - 2\sum_{m=1}^{\frac{N-1}{2}} \vartheta_m\right] + \sum_{m=1}^{\frac{N-1}{2}} \vartheta_m = \frac{1}{2}.$$

By Lemma 2, in analysis of approximation error, *we may assume* $u \neq 0$. We next develop intuition about the accuracy of approximation (11) in the general case $\alpha_5 \neq 0 \neq u$. Analogous to the pdf in (16), note that:

$$\Phi(\sqrt{2}t_5 y + u) = \Phi\left(\frac{t_5}{|t_5|}\frac{y - \left(-\frac{u}{\sqrt{2}t_5}\right)}{\frac{1}{\sqrt{2}|t_5|}}\right). \tag{24}$$

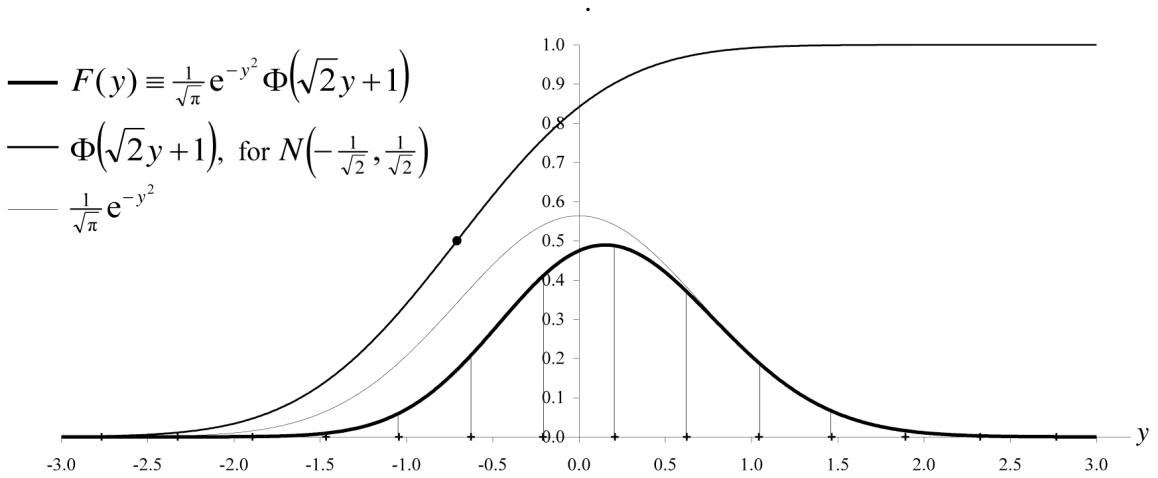
The normal CDF component of $F(y)$ has mean $-\frac{u}{\sqrt{2}t_5}$ and standard deviation $\frac{1}{\sqrt{2}|t_5|}$. Characteristics of $F(y)$ components (16) and (24) are primary features of adaption to each agent. The original integrand density for $N(\mu_*, \sigma_*)$ is transformed to a fixed density in (16) centered at the origin. The original CDF is translated and rescaled to compensate as in (24). By (12), a small standard deviation σ_* may imply a small t_5 which implies the CDF (24) in $F(y)$ has a large standard deviation. Untransformed density with narrow peak is replaced by a pdf of fixed peak-width $> 2\sqrt{2}$. Quadrature points populate expanded support of the transformed integrand. Figures 2 - 7 below illustrate the integrand $F(y)$ in (15) for selected values of t_5 and u. When u and t_5 are positive, the CDF component mean is left of the origin by (24).

Figure 2: Plot of transformed likelihood integrand components in (15) with parameters $t_5 = 0.5 = u$



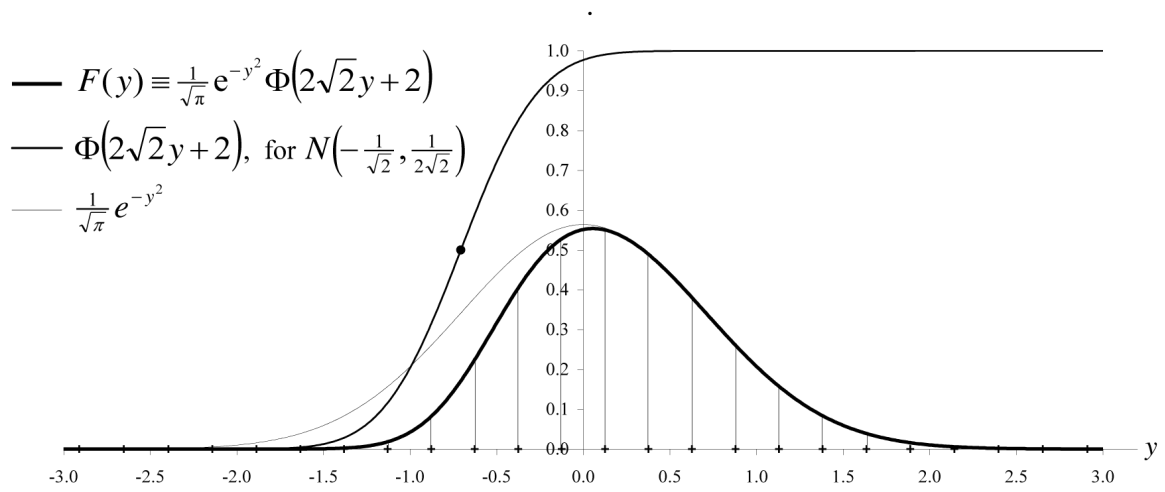
Note: The integral of $F(y) \approx 0.672639576990712$ with $N = 15$. See Table 6 below.

Figure 3: Plot of transformed likelihood integrand components in (15) with parameters $t_5 = 1.0 = u$



Note: The integral of $F(y) \approx 0.760249938906524$ with $N = 28$. See Table 7 below.

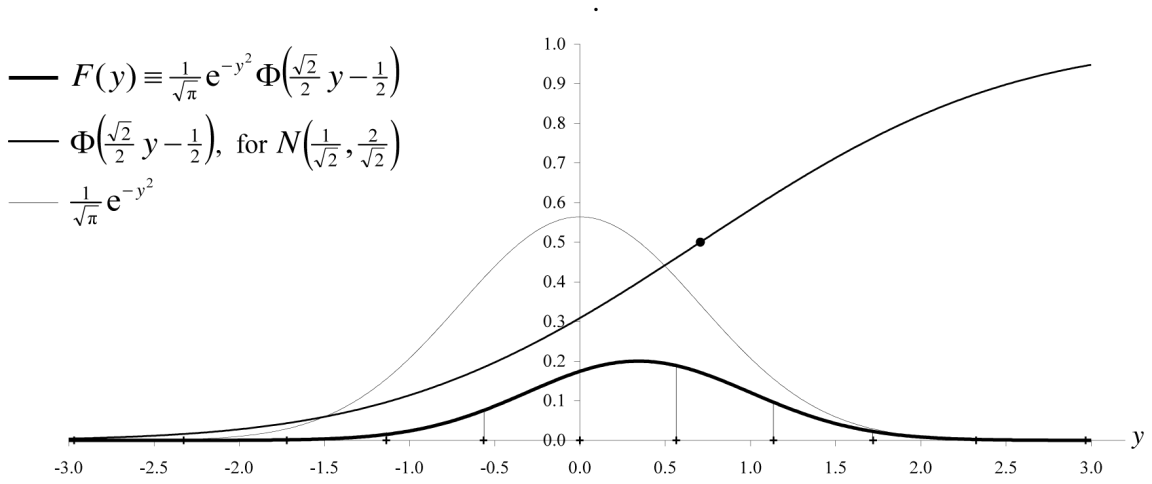
Figure 4: Plot of transformed likelihood integrand components in (15) with parameters $t_5 = 2.0 = u$



Note: The integral of $F(y) \approx 0.814453315238651$ with $N = 78$. See Table 8 below.

When $u < 0$ and $t_5 > 0$, the CDF component mean is right of the origin by (24):

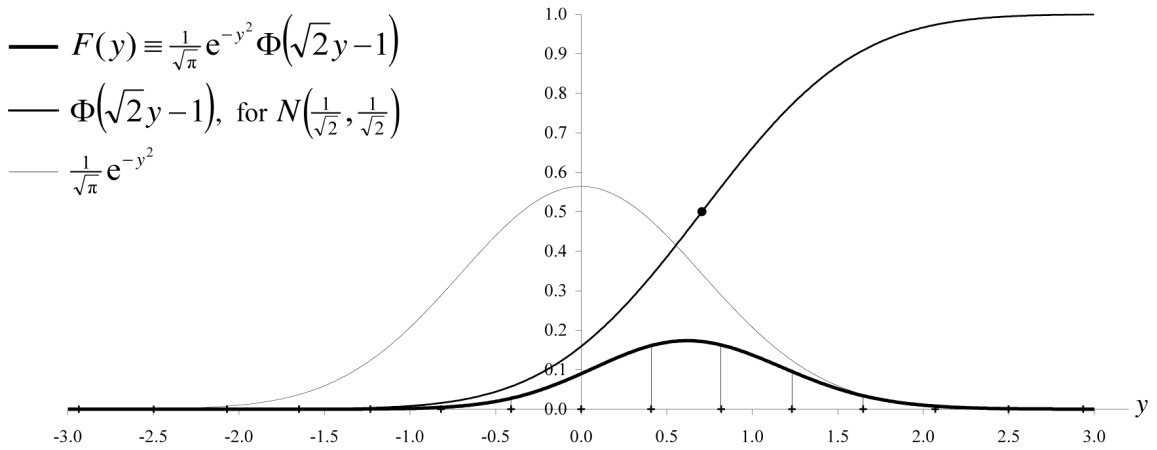
Figure 5: Plot of transformed likelihood integrand components in (15) with parameters $t_5 = 0.5 = -u$



Note: The integral of $F(y) \approx 0.327360423009288$ with $N = 15$. See Table 9 below.

Figure 6: Plot of transformed likelihood integrand components in (15) with parameters $t_5 = 1.0 = -u$

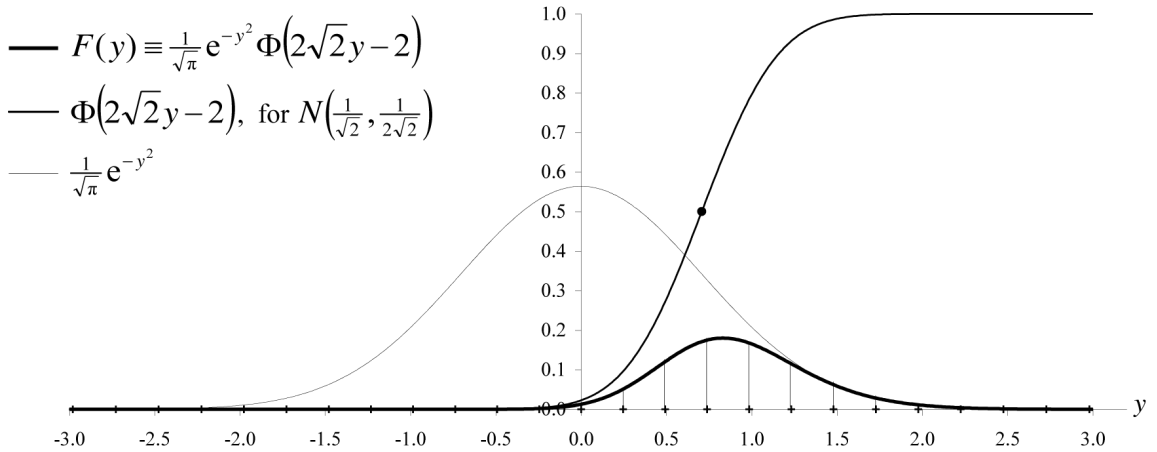
.



Note: The integral of $F(y) \approx 0.239750061093477$ with $N = 29$. See Table 10 below.

Figure 7: Plot of transformed likelihood integrand components in (15) with parameters $t_5 = 2.0 = -u$

.



Note: The integral of $F(y) \approx 0.185546684761349$ with $N = 81$. See Table 11 below.

The integral of the transformed likelihood integrand $F(y)$, plotted in Figures 2 - 7, may be approximated by (11). For each selected pair of parameters (u, t_5) in Figures 2 - 7, Tables 6 - 11 below describe accuracy of the quadrature sum in (11) as a function of the number of quadrature points N. Calculations for each table below were used to position the vertical lines under the curve of $F(y)$ in each figure above. For the N of the last row in each table, quadrature points in W_N that fall in the interval $[-3, +3]$ are at the position of those vertical lines and small y-axis-ticks in the figure corresponding to

the table. In the tables, the quadrature sum in (11) that approximates $\int F(y)dy$ is denoted:

$$\begin{aligned} W\sum_N(\Phi|u,t_5) = \sum_{n=1}^{N}\vartheta_n(\Phi\circ A_H)(\Theta_n|u,t_5) &= \sum_{n=1}^{N}\vartheta_n\Phi(\sqrt{2}t_5\Theta_n+u) \\ &= \sum_{n=1}^{N}\vartheta_n(\Phi\circ\varphi\circ A_H)(\Theta_n|\mu_*^A,\sigma_*) \\ &= W\sum_N(G_1^A|\mu_*^A,\sigma_*). \end{aligned} \tag{25}$$

Let $H\sum_N(\Phi|u,t_5) = \sum_{n=1}^{N}\vartheta_n\Phi\left(\sqrt{2}t_5P_n+u\right)$, then the error of approximation (11) is simply the difference:

$$\begin{aligned} E_N(u,t_5) &= \int_{\mathbb{R}} F(y)dy - H\sum_N(\Phi|u,t_5) \\ &= \int_{\mathbb{R}}\frac{1}{\sqrt{\pi}}e^{-y^2}\Phi\left(\sqrt{2}t_5y+u\right)dy - \sum_{n=1}^{N}\vartheta_n\Phi\left(\sqrt{2}t_5P_n+u\right). \end{aligned} \tag{26}$$

For each table row, $\int F(y)dy$ was calculated by Romberg integration on the interval $[-9,+9]$. Romberg calculation was conducted in 19-digit hardware precision. Iterative refinement of the interval partition terminated with the *relative* error of polynomial interpolation $< 10^{-16}$. Since approximation error $E_N(u,t_5)$ does not directly reveal the number of significant digits in the approximation, define *the relative error at N points* of approximation (11):

$$\mathrm{R}E_N(u,T_5) \equiv \left|\frac{\int_{\mathbb{R}}\frac{1}{\sqrt{\pi}}e^{-y^2}\Phi\left(\sqrt{2}t_5y+u\right)dy - \sum_{n=1}^{N}\vartheta_n\Phi\left(\sqrt{2}t_5\Phi_n+u\right)}{\int_{\boldsymbol{R}}\frac{1}{\sqrt{\pi}}e^{-y^2}\Phi\left(\sqrt{2}t_5y+u\right)dy}\right|. \tag{27}$$

Then the *decimal significance at N points* of approximation (11) is:

$$S_n(u,t_5) = -\log_{10}\mathrm{R}E_N(u,t_5) \tag{28}$$

$$= \log_{10}\left|\frac{\int_{\mathbb{R}}\frac{1}{\sqrt{\pi}}e^{-y^2}\Phi\left(\sqrt{2}t_5y+u\right)dy}{\int_{\mathbb{R}}\frac{1}{\sqrt{\pi}}e^{-y^2}\Phi\left(\sqrt{2}t_5y+u\right)dy - \sum_{n=1}^{N}\vartheta_n\Phi\left(\sqrt{2}t_5P_n+u\right)}\right|. \tag{29}$$

In the tables, $S_N(u,t_5)$ rounded *down* to the nearest integer is the number of significant decimal digits

in $\mathrm{H}\sum_N(\Phi|u,t_5)$ relative to the Romberg value for $\int F(y)dy$. For given (u,t_5), each table lists values of $\mathrm{H}\sum_N(\Phi|u,t_5)$ and the signed integer (26) for selected increasing N until $\mathrm{R}E_N(u,t_5) < 10^{-15}$. Thus, table lengths vary with the efficiency of Gauss-Hermite quadrature at each (u,t_5). The last two columns show $\mathrm{R}E_N(u,t_5)$ with $S_N(u,t_5)$ rounded down the the nearest tenth. When $\mathrm{H}\sum_N(\Phi|u,t_5)$ equals the Romberg value, both rounded to 15 digits, the value of the quadrature sum is shown in bold-face.

When u and t_5 are positive:

Table 6: Accuracy of N-point adaptive Gauss-Hermite quadrature with parameters $t_5 = 0.5 = u$.

N	$\mathrm{W}\sum_N(\Phi\|0.5,0.5)$	$E_N(0.5,0.5)$	$\mathrm{R}E_N(0.5,0.5)$	$S_N(0.5,0.5)$
4	0.672618686869557	$2.08901211543 \times 10^{-5}$	$3.10569313327 \times 10^{-5}$	4.5
8	0.672639574765081	2.2256304×10^{-9}	$3.10569313327 \times 10^{-9}$	8.4
12	0.672639576990494	2.172×10^{-13}	3.229×10^{-13}	12.4
15	**0.672639576990712**	-2×10^{-15}	3×10^{-15}	15.5

In W_{15}, the extreme points are ± 4.49999070730939 with weight $8.58964989963327 \times 10^{-10}$.

Table 7: Accuracy of N-point adaptive Gauss-Hermite quadrature with parameters $t_5 = 1.0 = u$.

N	$\mathrm{W}\sum_N(\Phi\|1.0,1.0)$	$E_N(1.0,1.0)$	$\mathrm{R}E_N(1.0,1.0)$	$S_N(1.0,1.0)$
4	0.758944432021307	$1.3055068852163 \times 10^{-3}$	$1.7172074845465 \times 10^{-3}$	2.7
8	0.760251305281224	$-1.3663747004 \times 10^{-6}$	$1.7972703849 \times 10^{-6}$	5.7
12	0.760250049464872	$-1.105583487 \times 10^{-7}$	$1.454236865 \times 10^{-7}$	6.8
16	0.760249940494693	$-1.5881696 \times 10^{-9}$	2.0890098×10^{-9}	8.6
20	0.760249938922564	-1.60404×10^{-11}	2.10989×10^{-11}	10.6
24	0.760249938906643	-1.196×10^{-13}	1.573×10^{-13}	12.8
28	0.760249938906524	-4×10^{-16}	6×10^{-16}	15.2

In W_{28}, the extreme points are ± 6.59160544236774 with weight $6.43254743880186 \times 10^{-20}$.
Note: Romberg value = **0.760249938906523** at relative tolerance 10^{-16} rounded to 15 digits.

Table 8: Accuracy of N-point adaptive Gauss-Hermite quadrature with parameters $t_5 = 2.0 = u$.

N	$\mathrm{W}\sum_N(\Phi\|2.0, 2.0)$	$E_N(2.0, 2.0)$	$\mathrm{R}E_N(2.0, 2.0)$	$S_N(2.0, 2.0)$
4	0.816631474537427	$-2.1781592987761 \times 10^{-3}$	$2.6743820155460 \times 10^{-3}$	2.5
8	0.818015114411898	$-3.5617991732470 \times 10^{-3}$	$4.3732392104063 \times 10^{-3}$	2.3
12	0.814804231765483	$-3.509165268317 \times 10^{-4}$	$4.308614382998 \times 10^{-4}$	3.3
16	0.814430926294541	$2.23889441103 \times 10^{-5}$	$2.74895364675 \times 10^{-5}$	4.5
20	0.814438558202192	$1.47570364595 \times 10^{-5}$	$1.81189470083 \times 10^{-5}$	4.7
24	0.814450180176825	$3.1350618260 \times 10^{-6}$	$3.8492836451 \times 10^{-6}$	5.4
28	0.814452900521139	$4.147175118 \times 10^{-7}$	$5.091974015 \times 10^{-7}$	6.2
32	0.814453293777743	$2.14609082 \times 10^{-8}$	$2.63500778 \times 10^{-8}$	7.5
36	0.814453322229591	$-6.9909400 \times 10^{-9}$	8.5835981×10^{-9}	8
40	0.814453318135879	$-2.8972278 \times 10^{-9}$	3.5572668×10^{-9}	8.4
44	0.814453315930169	$-6.915176 \times 10^{-10}$	8.490576×10^{-10}	9
48	0.814453315365336	$-1.266851 \times 10^{-10}$	1.555462×10^{-10}	9.8
52	0.814453315256783	-1.81315×10^{-11}	2.22622×10^{-11}	10.6
56	0.814453315240331	-1.6799×10^{-12}	2.0627×10^{-12}	11.6
60	0.814453315238586	6.56×10^{-14}	8.05×10^{-14}	13
64	0.814453315238569	8.20×10^{-14}	1.006×10^{-13}	12.9
68	0.814453315238625	2.61×10^{-14}	3.20×10^{-14}	13.4
72	0.814453315238645	6.2×10^{-15}	7.6×10^{-15}	14.1
76	0.814453315238650	1.2×10^{-15}	1.5×10^{-15}	14.8
78	**0.814453315238651**	5×10^{-16}	7×10^{-16}	15.1

In W_{78}, the extreme points are ± 11.7257979195159 with weight $7.6920667316977 \times 10^{-61}$.

When u is negative and t_5 is positive:

Table 9: Accuracy of N-point adaptive Gauss-Hermite quadrature with parameters $t_5 = 0.5 = -u$.

N	$\mathrm{W}\sum_N(\Phi\| -0.5, 0.5)$	$E_N(-0.5, 0.5)$	$\mathrm{R}E_N(-0.5, 0.5)$	$S_N(0.5, 0.5)$
4	0.327381313130443	$-2.08901211543 \times 10^{-5}$	$6.38138262476 \times 10^{-5}$	4.1
8	0.327360425234919	$-2.2256304 \times 10^{-9}$	6.7987154×10^{-9}	8.1
12	0.327360423009506	-2.172×10^{-13}	6.635×10^{-13}	12.1
15	0.327360423009288	2×10^{-16}	6×10^{-16}	15.2

In W_{15}, the extreme points are ± 4.49999070730939 with weight $8.58964989963327 \times 10^{-10}$. Note: Romberg value = **0.327360423009289** at relative tolerance 10^{-16} rounded to 15 digits.

Table 10: Accuracy of N-point adaptive Gauss-Hermite quadrature with parameters $t_5 = 1.0 = -u$.

N	$\mathrm{W}\sum_N(\Phi\vert -1.0, 1.0)$	$E_N(-1.0, 1.0)$	$\mathrm{R}E_N(-1.0, 1.0)$	$S_N(1.0, 1.0)$
4	0.241055567978693	$-1.3055068852163 \times 10^{-3}$	$5.4452828051931 \times 10^{-3}$	2.2
8	0.239748694718776	$1.3663747004 \times 10^{-6}$	$5.6991630956 \times 10^{-6}$	5.2
12	0.239749950535128	$1.105583487 \times 10^{-7}$	$4.611400232 \times 10^{-7}$	6.3
16	0.239750059505307	1.5881696×10^{-9}	6.6242717×10^{-9}	8.1
20	0.239750061077436	1.60404×10^{-11}	6.69048×10^{-11}	10.1
24	0.239750061093357	1.196×10^{-13}	4.988×10^{-13}	12.3
28	0.239750061093476	4×10^{-16}	1.8×10^{-15}	14.7
29	**0.239750061093477**	0	2×10^{-16}	15.6

In W_{29}, the extreme points are ± 6.72869519860885 with weight $1.02934180872194 \times 10^{-20}$.

Table 11: Accuracy of N-point adaptive Gauss-Hermite quadrature with parameters $t_5 = 2.0 = -u$.

N	$\mathrm{H}\sum_N(\Phi\vert -2.0, 2.0)$	$E_N(-2.0, 2.0)$	$\mathrm{R}E_N(-2.0, 2.0)$	$S_N(2.0, 2.0)$
4	0.183368525462573	$2.1781592987761 \times 10^{-3}$	$1.17391442567546 \times 10^{-2}$	1.9
8	0.181984885588102	$3.5617991732470 \times 10^{-3}$	$1.91962425943005 \times 10^{-2}$	1.7
12	0.185195768234517	$3.509165268317 \times 10^{-4}$	$1.8912573257942 \times 10^{-3}$	2.7
16	0.185569073705459	$-2.23889441103 \times 10^{-5}$	$1.206647488156 \times 10^{-4}$	3.9
20	0.185561441797808	$-1.47570364595 \times 10^{-5}$	$7.95327411995 \times 10^{-5}$	4
24	0.185549819823175	$-3.1350618260 \times 10^{-6}$	$1.68963505334 \times 10^{-5}$	4.7
28	0.185547099478861	$-4.147175118 \times 10^{-7}$	$2.2351114078 \times 10^{-6}$	5.6
32	0.185546706222257	$-2.14609082 \times 10^{-8}$	$1.156631186 \times 10^{-7}$	6.9
36	0.185546677770409	6.9909400×10^{-9}	$3.76775258 \times 10^{-8}$	7.4
40	0.185546681864121	2.8972278×10^{-9}	$1.56145489 \times 10^{-8}$	7.8
44	0.185546684069831	6.915176×10^{-10}	3.726921×10^{-10}	8.4
48	0.185546684634664	1.266851×10^{-10}	6.827669×10^{-10}	9.1
52	0.185546684743217	1.81315×10^{-11}	9.77195×10^{-11}	10
56	0.185546684759669	1.6799×10^{-12}	9.0540×10^{-12}	11
60	0.185546684761414	-6.56×10^{-14}	3.536×10^{-13}	12.4
64	0.185546684761431	-8.2×10^{-14}	4.417×10^{-13}	12.3
68	0.185546684761375	-2.61×10^{-14}	1.405×10^{-13}	12.8
72	0.185546684761355	-6.2×10^{-15}	3.33×10^{-14}	13.4
76	0.185546684761350	-1.2×10^{-15}	6.6×10^{-15}	14.1
80	0.185546684761349	-2×10^{-16}	1.1×10^{-15}	14.9
81	**0.185546684761349**	1×10^{-16}	7×10^{-16}	15.1

In W_{81}, the extreme points are ± 11.9681194448687 with weight $2.45280551389805 \times 10^{-63}$.

E Details on the ML Program

We present details on the implementation of our program for the ML estimation (DDC-ML). The program is written in C. In the DDC-ML program, all value functions are exactly calculated according to the appropriate information set by use of backward induction. The program can simulate any model it can estimate and vice-versa. After estimation, simulation can be used for goodness-of-fit testing and counterfactual policy design.

The main task of the DDC-ML program is the maximization of the likelihood function defined in equation (10). The DDC-ML program uses the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Press et al., 1992). We focus on four aspects of the numerical implementation that make the DDC-ML program the benchmark for scientific computation: (1) accuracy of integration, (2) accuracy of normal CDF evaluation, (3) overall precision, and (4) program flexibility.

E.1 Accuracy of Integration

The most important innovation of the DDC-ML program with respect to the current literature is the possibility to specify a set of factors for the unobserved heterogeneity in earnings and cost equations. The additional specification of a measurement system facilitates their meaningful interpretation.

The introduction of factors in the model adds the complication of integrating the entire likelihood with respect to their distribution. Provided that a reliable estimator of the normal CDF is available, complications arise as the support of the likelihood integral is potentially unbounded on the real line. While the econometrician has control over the scaling of the covariates, there is no such control over the standard deviation of the factor distribution which is updated at each step of the maximization routine.

In the DDC-ML program, the default integration method is adaptive Gauss-Hermite quadrature. This strategy is robust against the change of the integrand shape as the location of the quadrature points is adapted at each iteration of the optimizer (Davis and Rabinowitz, 1984; Judd, 1998). In Web Appendix D, we test the reliability of the Gauss-Hermite quadrature against the more accurate, but more time consuming Romberg integration (Romberg, 1955). Overall, our investigation reveals 15

digits of precision of a 40-point Gauss-Hermite quadrature in our model setup.

We use an additional defensive strategy. At each optimization step, we test the accuracy of integration for the particular values of structural parameters by the calculation of a test integral with known value. We integrate a normal pdf (with all candidate standard deviations) over its full support and expect a result very close to one. We test two alternative quadrature methods in the following order until the required precision is achieved: (1) Gauss-Hermite, (2) Gauss-Legendre. If both fail, we use Romberg integration and iterate until polynomial interpolation across successive interval partitions achieves the requested precision.

E.2 Accuracy of Normal CDF Evaluation

The product of transition probabilities is a key element in the likelihood calculation. They characterize the model's dynamic structure. Their reliable calculation is necessary to ensure the overall accuracy of our program. During optimization, normal CDF evaluations are potentially required anywhere on the real line. Small values for standard deviations $\hat{\sigma}$ can have severe consequences. Let z denote the index of observed characteristics in the transition probabilities. At each iteration the argument of the choice probability $\frac{z}{\hat{\sigma}}$ can take an almost arbitrarily large magnitude. Positive ratios are not a problem as $\Phi(\frac{z}{\hat{\sigma}})$ can be set to one without causing particular imprecision. However, negative values are a problem because setting $\Phi(\frac{z}{\hat{\sigma}}) = 0$ is a non-trivial work-around as $\ln(\Phi(\frac{z}{\hat{\sigma}}))$ is needed in the final likelihood calculation. In the DDC-ML program, the normal probability calculation is done via the algorithm proposed by Marsaglia (2004) extended to C long double precision.[3] This extended Marsaglia algorithm allows calculation of the standard normal CDF on the domain $[-150, +150]$ with about 16 digits of precision, falling to about 15 digits near -150.

[3]This extended hardware precision is defined in two ways (compilers differ in their interpretation of the format): with 80 bits, precision of approximately 19 decimal digits, magnitude range around $[10^{-4198}, 10^{+4197}]$. With 128 bits (quad precision), precision of approximately 34 decimal digits the magnitude range is unchanged.

E.3 Overall Precision

The calculation of the objective function must be accurate enough so that changes in function values between two points in the parameter space are sufficiently large. As the optimum is approached, the difference in successive function values becomes smaller and smaller and the function values agree in an increasing number of leading significant digits. Subtracting one from the other results in increasingly lower-order digits. Low-order digits are the least accurate, but the sign of the difference must be correct so that the optimizer's updating step is correct. This problem worsens if the surface of the likelihood function is nearly flat around the solution. Experience with our DDC-ML program shows that our 15 digits precision in the likelihood function value is required for 3-digit precision in MLE point estimates.

E.4 Program Flexibility

Estimation of complex structural models that are often characterized by hundreds of parameters is a labor-intensive process. The ability of the econometrician to control every step of the estimation is a key ingredient to a well designed program.

The DDC-ML program supports such a step-by-step sequence by reporting: (a) ultimately, a local optimum for a defensible model with indicators of quality for the convergence such as condition number of the Hessian and test of accuracy of the integral calculations; (b) along the way of estimation, failure to make progress to convergence with indicators of the problem(s), enhanced by a flexible specification syntax that enables the econometrician to easily adjust along the path model structure or data or parameters governing the calculations (such as the maximum numbers of iterations to be taken, the required norm of the gradient to claim convergence, the maximal step that the BFGS algorithm is allowed to take, etc.) in response to diagnosis of indicated problems.

Estimation of flexibly defined dynamic discrete-choice models is beset with potential problems. A well designed program must give to the econometrician the possibility to diagnose such problems in order to properly address them. Many common problems are sample-size dependence, rare cells, bad starting values, outlier agents, low data variability, poor data-scaling, unidentified structure, and

many others. There is not a single well-defined dimension along which models can be ranked in terms of the difficulty of their estimation. For example, a model specifying a small number of agent states is typically untroubled by rare cells in the data, but if the model also specifies many latent factor effects, then problems arise in sample-size dependence and low data-variability.

In the DDC-ML program, each problem manifests at a point in the sequence where details are displayed to support diagnosis, and easy-to-use syntax enables rapid trials of hypotheses for correcting the problem. In particular, the DDC-ML program allows for (a) model specification and data syntax checks for coherence in the specified model with the given data, (b) checks of potential causes (such as bad starting values or empirical underidentification) which might cause singularity in the Hessian matrix, (c) accuracy, logic, and data-preparation checks performed via a comparison between the calculation of analytic and finite difference likelihood derivatives whose divergence is often due to problems (such as incorrectly scaled covariates), (d) early-stage estimation checks by initial low-ambition trials of estimator progress. In each of the above sequenced steps, the program reports success or reports problem indicators. To respond to problems, calculation results and accuracy indicators are printed in detail, and every detail of configuration syntax and data file formats and structure (current parameter vector and the BFGS approximation of the matrix of second derivatives) is viewable and editable by the user (for example, to allow subsequent estimation of new parameters previously fixed at a given value).

Once estimation is underway, the sequence can be fine-grained under control from the econometrician. The econometrician can then judge the quality of the progress of the estimation step by step by using all indicators of potential problems offered by the program. Details of judgment and proper use of the program from the econometrician arising in this fine-grained realm are numerous; the totality of them defines a craft, an art, a style of dynamic discrete choice model estimation. That sequence stands in stark contrast to a push-button one-step concept that is pass/fail for convergence and too often, in practice, positions the user with little ability to judge quality (accuracy and precision) of estimation results (point and interval estimates) and little diagnostic support for estimation failure. This last point is crucial, as it is a fundamental characteristic of the estimation of complex structural models. As much as the economist's judgment is required in evaluating each modeling assumption, the economist's and programmer's judgment are also of crucial importance in the process of estimation.

The evaluation of a converging estimation run might only partially rely on objective measures such as condition numbers. The best evaluation comes from the econometrician's judgment who can evaluate if the estimator path was solid enough to trust the obtained results. For these reasons, it is often extremely hard to perform multiple serial runs of estimation (as would be required for a bootstrap exercise for example) not only because each of these runs often takes a long time, but because each of them cannot be automatized and requires the econometrician's judgment to evaluate its reliability.

References

BUREAU OF LABOR STATISTICS, *NLS Hanbook 2001: The National Longitudinal Surveys* (Washington, DC: U.S. Department of Labor, 2001).

DAVIS, P. J. AND P. RABINOWITZ, *Methods of Numerical Integration* (New York, NY: Academic Press, 1984).

HECKMAN, J. J., L. J. LOCHNER AND P. E. TODD, "Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond," in E. A. Hanushek and F. Welch, eds., *Handbook of the Economics of Education*volume 1 (Amsterdam, Netherlands: Elsevier, 2006), 307–458.

HECKMAN, J. J. AND S. NAVARRO, "Dynamic Discrete Choice and Dynamic Treatment Effects," *Journal of Econometrics* 136 (February 2007), 341–396.

JUDD, K. L., *Numerical Methods in Economics* (Cambridge, MA: MIT Press, 1998).

MARSAGLIA, G., "Evaluating the Normal Distribution," *Journal of Statistical Software* 11 (July 2004), 1–11.

MATZKIN, R. L., "Nonparametric Identification and Estimation of Polychotomous Choice Models," *Journal of Econometrics* 58 (July 1993), 137–168.

PRESS, W. H., B. P. FLANNERY, S. A. TEUKOLSKY AND W. T. VETTERLING, *Numerical Recpies in Fortran 77: The Art of Scientific Computing* (New York, NY: Cambridge University Press, 1992).

ROMBERG, W., "Vereinfachte Numerische Integration," *Det Kongelige Norske Videnskabers Selskab Forhandlinger* 28 (1955), 30–36.