A generalized endogenous grid method for discrete-continuous choice

Fedor Iskhakov[*] John Rust[†] Bertel Schjerning[‡]

July 12, 2012

PRELIMINARY DRAFT, PLEASE DO NOT QUOTE WITHOUT PERMISSION

Abstract

This paper extends the endogenous grid method (2006 “The method of endogenous gridpoints for solving dynamic stochastic optimization problems”, Economic Letters) for models with sequential discrete and continuous choice. Unlike existing generalizations, we propose solution algorithm that inherits both advantages of the original method, namely it avoids all root finding operations, and it also efficiently deals with restrictions on the continuous decision variable. To further speed up the solution, we perform the inevitable optimization across discrete decisions as a more efficient computation of upper envelope of a set of piece-wise linear functions. We formulate the algorithm relying as little as possible on a particular model specification, and precisely define the class of dynamic stochastic optimal control problems it can be applied to. We illustrate our algorithm using a finite horizon model of retirement behavior with consumption-savings decisions and borrowing constraints, and show that in comparison to traditional approaches the proposed method runs at least an order of magnitude faster to deliver the same precision of the solution. To implement the method we develop a generic software package that includes pseudo-language for easy model specification and computational modules which support both shared memory and cluster parallelization. The package is wrapped in a Matlab class and incurs low start-up cost to the user. The software package is accessible in the public domain.

Keywords: Discrete and continuous choice, dynamic structural model, consumption and savings, discrete labour supply.

JEL codes: C63

[*] Corresponding author. Center of Study of Choice (CenSoC), Faculty of Business, University of Technology, Sydney Level 4, 645 Harris St. Ultimo, NSW 2007, email: `fedor.iskhakov@uts.edu.au`

[†] Department of Economics, Georgetown University, Washington, DC, phone: (301) 801-0081, email: `jrust@technoluddites.com`

[‡] Department of Economics, University of Copenhagen and Centre for Applied Microeconometrics (CAM), Øster Farimagsgade 5, building 26, DK-1353 Copenhagen K, Denmark

1 Introduction

Solving dynamic stochastic optimization problems is rarely feasible analytically, and numerical solutions are usually computationally intensive. Traditionally numerical solutions are obtained through backward induction (or value function iterations) which require solving for the optimum value of control in each time period and in each point of the state space of the original problem. The state space is usually discretized beforehand with an exogenously fixed grid. The endogenous grid method (EGM) proposed by Carroll [2006] avoids internal optimizations by letting the grid over the state space vary from one time period to another. The essence of the EGM method is to guess an optimal value of control and back out the values of the state variables for which the guess would indeed be optimal. Repeating this procedure until the whole state space is sufficiently explored in each time period gives the same mapping of points of the state space to the optimal decisions as in the traditional solution framework. However, by replacing iterative optimization routine with a single shot algorithm, the EGM approach leads to substantial decrease in runtime, especially in large scale problems.

Carroll [2006] presents EGM as a solution method for a standard problem of maximizing discounted utility of consumption subject to intertemporal budget constraint, for which he provides both microeconomic and macroeconomic interpretations. Under the micro interpretation, future consumption is uncertain due to shocks to labor income, and the budget constraint reflects the dynamics of liquid assets with positive returns on savings and in presence of borrowing constraint. Under the representative agent macro interpretation, future consumption is uncertain due to shocks to aggregate productivity, and the budget constraint reflects the dynamics of depreciating capital. In both interpretations the model boils down to an optimal control problem with one continuous state and one continuous decision variable.

The purpose of this paper is to generalize the EGM method to make it suitable for a larger class of dynamic stochastic optimization problems, namely those which include additional discrete and continuous state variables and more importantly additional discrete choices. The intended application for our modification of EGM is a more involved microeconometric model with discrete labor supply choices accompanying the consumption-savings process covered by Carroll's micro interpretation. We use a simple model from this class for numerical illustrations and comparisons to the traditional solution algorithm.

To solve the consumption-savings model, Carroll's EGM method proceeds in the following way. In the terminal period consumption is identical to total resources in that period, which forms the base for backward induction. Further, for given value of end-of-period assets, next period cash-in-hand is computed using intertemporal budget constraint, then optimal consumption is found from already known consumption function for next time period, and finally the Euler equation is used to recover optimal consumption in the current period corresponding to the chosen value of end-of-period assets. These two values combine to the amount of total assets for which the posited level of consumption is indeed optimal. Assuming the decision maker has a strictly concave utility function, the Euler equation can be solved analytically using an inverted marginal utility function, which allows the method to avoid traditional iterative root finding algorithms altogether.

The main idea of our generalization of EGM method is to apply Carroll's original algorithm (which we refer to as *EGM step*) for each discrete decision available in current time period. This produces a consumption function defined over the endogenous grid for each current discrete decision. Along with the consumption function we also compute discrete-decision-specific value functions which are also defined over endogenous grids. In order to recover the discrete policy function which takes the form of one or several threshold values of total resources where the decision maker shifts from one discrete decision to the other, the computed value functions must be compared over all values of total resources. We employ a fast algorithm for computing the upper envelope of a collection of piece-wise linear functions to speed up this step by avoiding the unification of the endogenous grids and excessive interpolations. After the thresholds are computed, the resulting consumption function for the period is constructed from the corresponding decision-specific consumption functions calculated earlier with the EGM step. As a result, the consumption function appears to have discontinuities associated with points where there are kinks in the value function. These kinks arise on the upper envelope value function, where discrete-decision-specific value functions cross each other.

When the EGM step is performed again in preceding time periods, the kinks in value function translate into the intervals of total resources where the solution to the Euler equation is not unique. In such non-convex regions solving the first order condition is not sufficient for finding the optimal value of consumption, and slower global optimization methods were used in previous literature to deal with this problem. We develop a new approach to handling non-convex regions which is also based on fast upper envelope algorithm. We describe our algorithm in detail in section 3.

EGM has already been generalized for the case of additional state variables and additional controls by Barillas and Fernandez-Villaverde [2007] who propose an iterative scheme in which EGM step is alternated with traditional value function iteration step. During EGM step the objective function is optimized only with respect to one continuous decision variable while keeping all other policy functions (decision rules) fixed. The latter are updated at each value function iteration step. Barillas and Fernandez-Villaverde use standard representative agent neoclassical growth model with endogenous continuous labour supply to illustrate their method, and show the speed-up of the solution to be of two orders of magnitude compared to the traditional approach.

Our generalization of EGM, although similar in spirit, differs from Barillas and Fernandez-Villaverde [2007] in several important aspects. First, Barillas and Fernandez-Villaverde tailor their algorithm for infinite horizon problems, which allows them to use EGM directly in place of value function iterations in the search for the fixed point that solves the Bellman equation. Restricting attention to infinite horizon problems allows for an accurate initial guess of a value function, using the analytic steady state solution for the deterministic version of the model. As Barillas and Fernandez-Villaverde note[1], for some models (including the one they use to illustrate their method) running EGM with one control fixed to its deterministic steady state value, and performing just a single value function iteration afterwards may give nearly as accurate a solution as the traditional method. In contrast, our method is developed with finite horizon problems in mind, thus there is no gain in the efficiency of the method due to a possibly better initial guess for the value function. Calculations start at terminal period from the same point (terminal utilities) as they would in the traditional backward induction approach. Yet, our method can be easily extended to infinite horizon.[2]

Second, Barillas and Fernandez-Villaverde [2007] cannot avoid root-finding operations completely. Besides optimizations with respect to some decision variables which are inevitable in value function iterations, in their method a non-linear equation has to be numerically solved during each transition between the two types of iterations. In our algorithm we make use of the fact that value function iterations when all decision variables are discrete are not only faster because internal optimization tasks reduce to calculating maximums across finite sets of points, but also they can be "vectorized" when the upper envelope of a set of decision-specific value functions is calculated instead. The gain in computational speed is due to the fact that not all decision-specific value functions have to be compared in each state point in order to calculate the upper envelope. In addition, we develop an efficient algorithm for combining the different

[1] Section 3.2, page 2701.

[2] Although the question of choice of the initial guess for policy functions will then arise.

grids these value functions are defined over (due to the use of EGM) which avoids multiple re-interpolations. In effect, our algorithm avoids all root-finding operations, except the complicated cases of utility maximization in terminal period[3].

Fella [2011] provides a generalization of the endogenous grid point method for non-concave problems which includes the specification we consider here. Fella proposed a different method to solve an example model of durable goods purchases with switching costs. The key idea of Fella's method is to identify the regions of the problem where first order conditions are not sufficient and run auxiliary numerical optimizations to verify whether the solution found by EGM step is indeed a global maximum. Contrary to his approach, we show that identifying exact boundaries of such non-convex regions is not necessary and that global maxima can be found with a much less time consuming computation of the upper envelope over the auxiliary value functions produced by each solution of the first order condition.

The rest of the paper is organized as follows. We start with setting up a simple illustrative model of optimal retirement and show how it can be best solved using a modification of the traditional dynamic programming approach, that iteratively solves a "kinked" Euler equation. The "kinked" Euler equation arises when there is a combination of a discrete retirement decision and a continuous consumption-saving choice. Then, unlike much of the previous literature including both Carroll [2006], Fella [2011] and Barillas and Fernandez-Villaverde [2007] who center their algorithms around particular model specifications, in section 3 we set up an abstract stochastic control problem and present our solution method separately from the details of any particular economic model. In section 4 we present numerical results and compare generalized EGM to traditional solution methods for our illustrative problem, and then we conclude.

2 Illustrative model

2.1 Model of retirement

Consider a model with a single state variable M_t which denotes total liquid resources (wealth) available in the beginning of period t, and two choices: how much to consume c_t and whether or not to work d_t. Let $d_t = \mathbb{W}$ denote the choice to work and $d_t = \mathbb{R}$ denote the choice to retire.

We assume that if someone chooses to work they receive a fixed (non-random) wage y together with an age-dependent additive disutility (or cost) of working w_t. Let the time constant discount be $\beta \in (0,1)$[4]. We assume that there is a potentially random return $R \geq 0$ to savings, but in this model we do not consider portfolio allocation choice focusing on one discrete decision to retire[5].

We consider a version of the problem where retirement is assumed to be an absorbing state, i.e. once retired we rule out the possibility of subsequent labor market entry. The value function for a person who has not yet retired $V_t(M, \mathbb{W})$ is given by

$$V_t(M, \mathbb{W}) = \max \left[V_t(M, \mathbb{R}), \max_{0 \leq c_t \leq M} \left[u(c_t) - w_t + \beta E V_{t+1} \left(R \left(M + y - c_t \right), \mathbb{W} \right) \right] \right], \tag{1}$$

[3] When terminal utility is increasing monotonically, this maximization is trivial because in presence of borrowing constraint consumption is bounded from above.

[4] The closed-form solutions provided below can be extended to the case where β is age-dependent which could reflect age-variation due to mortality.

[5] The solutions below can be extended to allow returns of the form $R(\mu)$ which depend on a parameter $\mu \in (0,1)$ that can be viewed as a portfolio allocation decision between a riskless security (Treasury bills) and risky securities (stock portfolio).

where $V_t(M, \mathbb{R})$ is the value function for a retiree given by

$$V_t(M, \mathbb{R}) = \max_{0 \le c_t \le M} \left[u(c_t) + \beta E V_{t+1} \left(R \left(M - c_t \right), \mathbb{R} \right) \right]. \tag{2}$$

For the class of constant relative risk averse utility functions, $u(c) = (c^\rho - 1)/\rho$ for $\rho \in [0,1)$ (with $u(c) = \log(c)$ when $\rho = 0$), Hakansson [1970], extending Phelps [1962], derived closed-form solutions for optimal consumption $c_t(M)$ and for $V_t(M, \mathbb{R})$, $t \in \{1, \ldots, T\}$, where T is the upper bound on lifespan. We have for $\rho \in (0,1)$

$$V_t(M, \mathbb{R}) = \left[\frac{M^\rho}{\rho} \right] \left(\sum_{i=0}^{T-t} K^i \right)^{(1-\rho)} - \frac{1}{\rho} \left(\sum_{i=0}^{t} \beta^i \right) \tag{3}$$

$$c_t(M, \mathbb{R}) = M \left(\sum_{i=0}^{T-t} K^i \right)^{-1}, \tag{4}$$

where $K = (\beta R^\rho)^{1/(1-\rho)}$, so $K = \beta$ when $\rho = 0$. Further, for $\rho = 0$ we have[6]

$$V_t(M, \mathbb{R}) = \log(M) \left(\sum_{i=0}^{T-t} \beta^i \right) + A_{T-t} \tag{5}$$

where

$$A_t = \left(\sum_{i=0}^{t} i \beta^i \right) [\log(R) + \log(\beta)] - \log \left(\sum_{i=0}^{t} \beta^i \right) \left(\sum_{i=0}^{t} \beta^i \right).$$

The optimal retirement threshold $\overline{M}_t$ is the value of M that makes the person indifferent between retiring and not retiring at age t

$$V_t(\overline{M}_t, \mathbb{R}) = V_t(\overline{M}_t, \mathbb{W}). \tag{6}$$

If $w_t > 0$ (i.e. there is a positive disutility from working), it will be optimal for a person of age t to retire if $M \ge \overline{M}_t$ and work otherwise. We will have a non-convex kink in the value function for working $V_t(M, \mathbb{W})$ at the point $\overline{M}_t$ since we have from (1)

$$V_t(M, \mathbb{W}) = \max \left[V_t \left(M, \mathbb{R} \right), V_t \left(M, \mathbb{W} \right) \right] \tag{7}$$

and we can show that in this problem the two functions will only intersect once at $\overline{M}_t$ with $V_t(M, \mathbb{W}) > V_t(M, \mathbb{R})$ for $M < \overline{M}_t$ and $V_t(M, \mathbb{W}) < V_t(M, \mathbb{R})$ for $M > \overline{M}_t$.

Let $c_t(M, \mathbb{R})$ be the optimal consumption of a retiree of age t. This function is given by

$$c_t(M, \mathbb{R}) = argmax_{0 \le c \le M} \left[u(c) + \beta E V_{t+1} \left(R \left(M - c_t \right), \mathbb{R} \right) \right]. \tag{8}$$

Let the optimal consumption of a individual who is still working (not yet retired) be denoted $c_t(M, \mathbb{W})$. This function is composed of two components corresponding to the two discrete choices available to a worker, namely

$$c_t(M, \mathbb{W}) = \begin{cases} c_t(M, \mathbb{W}, \mathbb{W}) & \text{if } M \le \overline{M}_t, \\ c_t(M, \mathbb{W}, \mathbb{R}) & \text{if } M > \overline{M}_t, \end{cases} \tag{9}$$

where $c_t(M, \mathbb{W}, \mathbb{W})$ is a consumption function of the worker who decides to continue working, given by

$$c_t(M, \mathbb{W}, \mathbb{W}) = argmax_{0 \le c \le M} \left[u(c_t) - w_t + \beta E V_{t+1} \left(R \left(M + y - c_t \right), \mathbb{W} \right) \right], \tag{10}$$

[6] Equation (5) can be derived by L'Hôpital's rule from (3) in the limit as $\rho \to 0$

and $c_t(M, \mathbb{W}, \mathbb{R})$ is a consumption function of the worker who decides to retire, given by

$$c_t(M, \mathbb{W}, \mathbb{R}) = argmax_{0 \le c \le M} \left[u(c_t) - w_t + \beta E V_{t+1} \left(R \left(M - c_t \right), \mathbb{R} \right) \right], \tag{11}$$

which is identical to optimal consumption of a retiree because disutility of work w_t is independent of consumption, i.e. $c_t(M, \mathbb{W}, \mathbb{R}) = c_t(M, \mathbb{R})$.

The optimal consumption function of a worker $c_t(M, \mathbb{W})$ has a discontinuity at $\overline{M}_t$, and for some small $\xi > 0$ $c_t(\overline{M}_t - \xi, \mathbb{W}) > c_t(\overline{M}_t + \xi, \mathbb{W})$. This follows from the fact that the kink in the value function $V_t(M, \mathbb{W})$ at $\overline{M}_t$ results from the maximization in (7) of two concave functions, and thus can only be downward pointing. The derivative of the value function $V_t'(M, \mathbb{W})$ makes a discontinuous jump at this point, i.e. $V_t'^{-}(\overline{M}_t, \mathbb{W}) < V_t'^{+}(\overline{M}_t, \mathbb{W})$, where V'^{-} and V'^{+} denote left and right hand derivatives correspondingly.

It will be an important test to check if the solution method can accurately determine the optimal retirement thresholds $\overline{M}_t$ and capture the discontinuity in the optimal consumption function $c_t(M, \mathbb{W})$ at these points.

2.2 Solving the model using the "kinked" Euler equation

Numerical dynamic programming is traditionally used to solve models similar to the retirement model laid out in previous section. The method proceeds backwards from the terminal period T by iteratively solving the Bellman equations (1) and (2). Optimization in (1) and (2) can be performed directly using various standard numerical maximization/minimization methods, but to make a closer comparison to EGM, we adopt a numerical approach based on first order conditions. Therefore, we start with deriving the "kinked" Euler equation for the retirement problem.

For $M \le \overline{M}_t$ it is optimal for the worker to continue working, and we have the following first order condition for (1) holding for each M_t in this region

$$u'\left(c_t\left(M, \mathbb{W}\right)\right) = \beta \int_0^\infty R V_{t+1}'\left(R\left(M + y - c_t\left(M, \mathbb{W}\right)\right), \mathbb{W}\right) F(dR), \tag{12}$$

where $F(R)$ is cumulative distribution function for the returns R.

However we have

$$V_{t+1}'(M, \mathbb{W}) = \begin{cases} V_{t+1}'(M, \mathbb{W}) & \text{if } M \le \overline{M}_{t+1}, \\ V_{t+1}'(M, \mathbb{R}) & \text{if } M > \overline{M}_{t+1}, \end{cases} \tag{13}$$

Via the Envelope Theorem, we have

$$V_t'(M, \mathbb{W}) = \begin{cases} \beta \int_0^\infty R V_{t+1}'\left(R\left(M + y - c_t\left(M, \mathbb{W}\right)\right), \mathbb{W}\right) F(dR), & \text{if } M \le \overline{M}_t, \\ \beta \int_0^\infty R V_{t+1}'\left(R\left(M - c_t\left(M, \mathbb{R}\right)\right), \mathbb{R}\right) F(dR), & \text{if } M > \overline{M}_t. \end{cases} \tag{14}$$

However, for the first order condition for post-retirement consumption, we have

$$u'\left(c_t\left(M, \mathbb{R}\right)\right) = \beta \int_0^\infty R V_{t+1}'\left(R\left(M - c_t\left(M, \mathbb{R}\right)\right), \mathbb{R}\right) F(dR), \tag{15}$$

Substituting equations (12) and (15) into the equation (14) above, we get

$$V_t'(M, \mathbb{W}) = \begin{cases} u'(c_{t+1}(M, \mathbb{W})) & \text{if } M \le \overline{M}_t, \\ u'(c_{t+1}(M, \mathbb{R})) & \text{if } M > \overline{M}_t. \end{cases} \tag{16}$$

Now substituting formula (16) into the Euler first order condition for optimal pre-retirement consumption (12) we obtain

$$\begin{aligned}u'\left(c_t\left(M,\mathbb{W}\right)\right) &= \beta\int_0^\infty Ru'(c_{t+1}(R\left(M+y-c_t\left(M,\mathbb{W}\right)\right),\mathbb{W}))\cdot\mathbb{I}\{R\left(M+y-c_t\left(M,\mathbb{W}\right)\right)\leq\overline{M}_{t+1}\}F(dR)\\ &+ \beta\int_0^\infty Ru'(c_{t+1}(R\left(M-c_t\left(M,\mathbb{W}\right)\right),\mathbb{R}))\cdot\mathbb{I}\{R\left(M-c_t\left(M,\mathbb{R}\right)\right)>\overline{M}_{t+1}\}F(dR). \qquad (17)\end{aligned}$$

This is the appropriate "kinked Euler equation" to be solved to determine optimal consumption in the pre-retirement phase of the dynamic programming problem. Note that with a change of variables $q=F(R)$, we can write the kinked Euler equation (17) as

$$\begin{aligned}u'\left(c_t\left(M,\mathbb{W}\right)\right) &= \beta\int_0^{\overline{q}_t} F^{-1}(q)u'\left(c_{t+1}\left(F^{-1}(q)\left(M+y-c_t\left(M,\mathbb{W}\right)\right),\mathbb{W}\right)\right)dq\\ &+ \beta\int_{\overline{q}_t}^1 F^{-1}(q)u'\left(c_{t+1}\left(F^{-1}(q)\left(M-c_t\left(M,\mathbb{W}\right)\right),\mathbb{R}\right)\right)dq \qquad (18)\end{aligned}$$

where the threshold $\overline{q}_t$ is given by

$$\overline{q}_t = F\left(\frac{\overline{M}_{t+1}}{M+y-c_t(M,\mathbb{W})}\right). \qquad (19)$$

As long as distribution of returns is not degenerate, the resulting "kinked" Euler equation (18) is continuous and smooth in $c_t(M,\mathbb{W})$ in spite of the discontinuity in the consumption function $c_{t+1}(M,\mathbb{W})$ at $M=\overline{M}_{t+1}$. In fact, it is straightforward to define the *Euler residual function* as the difference between left hand side and right hand side of the equation (18) and write down the derivative of this function. Then solving the Euler equation amounts to finding zeros of the Euler residual function, and with analytical derivatives available, this task can be efficiently performed by Newton's Method.

The difficulty that remains and has to be addressed separately in the non-convex problems with discontinuous drops in optimal consumption similar to the one we consider here, is the fact that Euler equation is only a necessary condition and solutions may not be unique. Even in the case of random returns when Euler residual function is smooth, it is not necessarily monotone, and thus there may be several solutions to the Euler equation. In fact, in the simple retirement model in the period $T-3$ there are two solutions for the Euler equation for worker for some levels of wealth M_{T-3}. In such cases, globally optimal level of consumption can be recovered by comparing value functions for each of the solutions of the Euler equation. This considerably complicates the numerical solution procedure not only because of the need to compare multiple solutions of the Euler equation, but also because of the necessity to find *all* possible solutions before determining which one results in global optimum at each level of of wealth.

3 Generalized EGM algorithm for discrete-continuous problems

In this section we provide formal description of the class of the models that can be solved with generalized EGM method, make a compact presentation of the solution algorithm itself and discuss its interior machinery in more detail.

3.1 Abstract model to be solved by generalized EGM method

The generalized EGM algorithm is designed to solve a particular dynamic stochastic optimal control problem in discrete time with one continuous and one discrete decision variables. The model has the form

$$max_{\delta \in \mathfrak{F}} \left\{ E \left[\sum_{t=T_0}^{T} \left(\prod_{\tau=T_0}^{t} \beta_\tau \right) \cdot u\left(c_t, d_t, st_t\right) \right] \right\} \tag{20}$$

$$\begin{aligned} \text{s.t. } M_{t+1} &= \mathcal{R}_{t+1}\left(A_t, st_t, d_t, st_{t+1}, \xi_{t+1}\right) && (21) \\ M_t &= A_t + c_t && (22) \\ A_t &\geqslant A_0, && (23) \end{aligned}$$

where notations are as follows:

$u\left(c_t, d_t, st_t\right)$ is an instantaneous utility at period t, which is dependent on consumption $c_t = M_t - A_t$, discrete decision d_t and a vector of additional state variables st_t;

$M_t \in \mathbb{R}^1$ is a uni-dimensional continuous state variable which is given special role in the solution algorithm, with standard microeconomic interpretation of total resources (money-at-hand) available for consumption in period t;

$A_t \in \mathbb{R}^1$ is a scalar continuous decision variable which is interpreted as *end-of-period* resources remaining after within-period consumption in accordance with (22), which is subject to credit constraint (23);

$d_t \in \left\{d^{(1)}, .., d^{(D)}\right\}$ is the discrete decision variable which has D possible values;

$\delta = \left\{\delta_{T_0}, .., \delta_T\right\}$ is a set of decision rules (policy functions) of the form $\delta_t : (M_t, st_t) \rightarrow (A_t, d_t)$ which map points of the state space into choices at each time period, and are jointly chosen from the class of feasible decision rules $\mathfrak{F}$;

β_τ is exogenous discount factor which may be time-dependent, for example reflecting mortality probabilities;

$\mathcal{R}_{t+1}\left(A_t, st_t, d_t, st_{t+1}, \xi_{t+1}\right)$ is the intertemporal budget constraint which describes how the next period total resources M_{t+1} depend on current period savings A_t given the transition from st_t to st_{t+1} and the discrete decision d_t;

ξ_{t+1} is an idiosyncratic shock in period $t + 1$ that affects the budget constraint.

The intertemporal budget constraint for the retirement model is given by $\mathcal{R}_{t+1}\left(A_t, st_t, d_t, st_{t+1}, \xi_{t+1}\right) = \xi_{t+1}A_t + y \cdot \mathbb{I}\left\{d_t = \mathbb{W}\right\}$, where $\xi_{t+1} = R$ is the time-independent stochastic return on savings. It is also straightforward to verify that with different definitions for $\mathcal{R}_{t+1}\left(A_t, st_t, d_t, st_{t+1}, \xi_{t+1}\right)$ the problem (20) nests both micro and macro specifications in Carroll [2006].

The solution method relies on the following properties *assumed* about this problem:

A1. The instantaneous utility $u\left(c, d_t, st_t\right)$ is twice continuously differentiable, concave, and has a monotonic derivative with respect to c (i.e. $\frac{\partial^2 u}{\partial c \partial c} < 0$);

A2. The decision space contains one scalar continuous choice variable A_t and one scalar discrete choice variable d_t [7];

A3. The structure of the constraints (21-23) holds, thus singling out one continuous state variable M_t for which the stochastic motion rule is given by (21) and which enters the utility function only through $c_t = M_t - A_t$;

[7] Because vector values of any multinomial discrete decision can always be "re-coded" into the single set of values, there is no loss of generality in assuming d_t to be scalar.

A4. The decision A_t is restricted to $[A_0, M_t]$ which ensures through (22) that $c_t \geq 0$;

A5. The transition probabilities (densities) for the state process $\{st_t\}$ used in calculating the expectation in (20) are independent of M_t, namely $P\left(st_{t+1}|st_t, M_t, A_t, d_t\right) = P\left(st_{t+1}|st_t, A_t, d_t\right)$.

3.2 General layout of the solution algorithm

According to the Principle of Optimality, the problem (20) can be written in recursive form as

$$V_t\left(M_t, st_t\right) = \begin{cases} max_{\delta_t}\left\{u\left(M_t - A_t, d_t, st_t\right) + \beta_{t+1}E\left[V_{t+1}\left(\mathcal{R}_{t+1}\left(A_t, .., \xi_{t+1}\right), st_{t+1}\right)\right]\right\}, & 1 \leqslant t < T, \\ max_{\delta_t}\left\{u\left(M_t - A_t, d_t, st_t\right)\right\}, & t = T, \end{cases}$$
$$\text{s.t. } A_t \geqslant A_0,\ M_1 \text{ fixed},\ st_1 \text{ fixed}. \tag{24}$$

$V_t\left(M_t, st_t\right)$ is the value function of the problem. In the absence of discrete component in the decision rule δ_t first order conditions for (24) in combination with envelope theorem would lead to the single Euler equation that constitutes the foundation of EGM in Carroll's original 2006 paper. But because the maximand in (24) is not differentiable with respect to d_t, it is not possible to apply Carroll's argument in full and derive the set of Euler equations to characterize the solution to the problem (20). Instead, we rely on the result from Clausen and Strub which generalizes the argument we used in section 2.2 to derive the kinked Euler equation in the optimal retirement problem.

Theorem 3 in Clausen and Strub establishes that when the utility function is differentiable with respect to consumption (assumption A.1) the value function is also differentiable at interior optimal choices. Moreover, the Euler equation with respect to the continuous control variable holds as long as no constraints bind. The intuition behind their result is the following. Because value function $V_t\left(M_t, st_t\right)$ is in fact an upper envelope of the discrete choice specific value functions

$$V_t^{d_t}\left(M_t, st_t\right) = max_{A_t^{d_t}\left(M_t, st_t\right)}\left\{u\left(M_t - A_t, d_t, st_t\right) + \beta_{t+1}E\left[V_{t+1}\left(\mathcal{R}_{t+1}\left(A_t, .., \xi_{t+1}\right), st_{t+1}\right)\right]\right\} \tag{25}$$

(for $1 \leqslant t < T$), and as long as there are no kinks in $V_t^{d_t}\left(M_t, st_t\right)$ caused by binding constraints, it may only have downward pointing kinks which cannot be even local maxima. In other words, the decision maker never chooses a savings level where he is indifferent between the discrete choices.

Therefore we base our generalization on the fact that the Euler equation with respect to the continuous control remains the necessary condition for optimality, and thus the EGM step can be carried through conditional on the discrete choices. The Euler equation with respect to the continuous control can be derived similarly to the "kinked" Euler equation in the optimal retirement problem. For problem (20) it takes the form

$$\begin{cases} \frac{\partial u}{\partial c}\left(c_t, d_t, st_t\right) = \beta_{t+1}E\left[\frac{\partial \mathcal{R}_{t+1}}{\partial A_t} \cdot \frac{\partial u}{\partial c}\left(c_{t+1}, d_{t+1}, st_{t+1}\right)|A_t, st_t\right], & 1 \leqslant t < T, \\ \frac{\partial u}{\partial c}\left(c_t, d_t, st_t\right) = \frac{\partial u}{\partial c}\left(c_T, d_T, st_T\right) = 0, & t = T, \end{cases} \tag{26}$$

where the arguments of the intertemporal budget constraint function $\mathcal{R}_{t+1}\left(A_t, st_t, d_t, st_{t+1}, \xi_{t+1}\right)$ are dropped to simplify exposition.

The general layout of the algorithm is presented in Table 1; each column represents one of a series of nested loops, and each row represents a distinct operation within a corresponding loop. When $t = T$ (upper panel in the table) there is no subsequent period, and the problem can be reformulated as a static maximization of utility of consumption under borrowing constraint. For

Table 1: Schematic layout of the generalized EGM algorithm

<table>
<tr><td rowspan="3">In the last period when $t = T$</td><td rowspan="2">For each st_T</td><td colspan="3">Choose some initial grid over M_T</td></tr>
<tr><td colspan="3">Compute optimal consumption c_T using second line in Euler equation (26) and restriction $0 \leqslant c_T \leqslant M_T - A_T$</td></tr>
<tr><td colspan="4">Output optimal consumption and savings functions $c_T\left(M_T, st_T\right)$ and $A_T\left(M_T, st_T\right) = M_T - c_T\left(M_T, st_T\right)$, as well as value functions $V_t\left(M_T, st_T\right) = u\left(c_T, d_T, st_T\right)$ for each st_T, and proceed with next iteration of t</td></tr>
<tr><td rowspan="9">For t from $T-1$ to 1</td><td rowspan="8">For each st_t</td><td rowspan="6">For each d_t</td><td rowspan="4">For a sequence of guesses of A_t generated such that the endogenous grids $M_t^{grid}(d_t, st_t)$ overlap for all discrete decisions</td><td>Compute left hand side of the Euler equation $E\left[\frac{\partial \mathcal{R}_{t+1}}{\partial A_t} \cdot \frac{\partial u}{\partial c}\left(c_{t+1}, d_{t+1}, st_{t+1}\right) \vert A_t, st_t\right]$ and the expected value function $E\left[V_{t+1}\left(\mathcal{R}_{t+1}, st_{t+1}\right) \vert A_t, st_t\right]$ using
• transition probabilities for the state process $P\left(st_{t+1} \vert st_t, A_t, d_t\right)$
• probability distribution of shock ξ_{t+1}
• optimal consumption $c_{t+1}(M_{t+1}, st_{t+1})$ and optimal discrete choice $d_{t+1}(M_{t+1}, st_{t+1})$ computed on the previous iteration of t
• value function $V_{t+1}\left(M_{t+1}, st_{t+1}\right)$ computed on the previous iteration of t</td></tr>
<tr><td>Compute the optimal consumption c_t at t for given A_t using the Euler equation (26) and the inverse of the marginal utility function</td></tr>
<tr><td>Add another point $M_t = A_t + c_t$ to the (decision specific) endogenous grid $M_t^{grid}(d_t, st_t)$</td></tr>
<tr><td>Compute the d_t-specific value function at the new grid point $V^{d_t}\left(M_t, st_t\right) = u\left(c_t, d_t, st_t\right) + \beta_{t+1} E\left[V\left(M_{t+1}, st_{t+1}\right) \vert A_t, st_t\right]$</td></tr>
<tr><td colspan="2">Compute the “secondary” upper envelopes in the non-concave regions (where more than one solution to the Euler equation is found for the same values of total resources M_t) by dropping the endogenous points corresponding to local maxima of the value function.</td></tr>
<tr><td colspan="2">Save to memory the calculated d_t-specific optimal consumption function $c_t^{d_t}\left(M_t^{grid}(d_t, st_t), st_t\right)$ and d_t-specific value function $V_t^{d_t}\left(M_t^{grid}(d_t, st_t), st_t\right)$ defined over decision specific endogenous grid $M_t^{grid}(d_t, st_t)$</td></tr>
<tr><td colspan="3">Compute the upper envelope of decision specific value functions $V_t^{d_t}\left(M_t^{grid}(d_t, st_t), st_t\right)$ while simultaneously constructing the unified endogenous grid $M_t^{grid}(st_t)$ and optimal discrete decision rule $d_t(M_t, st_t)$</td></tr>
<tr><td colspan="3">Using the optimal discrete decision rule and the corresponding d_t-specific optimal consumption functions, find the unified optimal consumption function $c_t\left(M_t^{grid}(st_t), st_t\right)$ and value function $V_t\left(M_t^{grid}(st_t), st_t\right)$.</td></tr>
<tr><td colspan="4">Output optimal consumption and savings functions $c_t\left(M_t, st_t\right)$ and $A_t\left(M_t, st_t\right) = M_t - c_t\left(M_t, st_t\right)$, the optimal discrete decision rule $d_t(M_t, st_t)$ and value functions $V_t\left(M_t, st_t\right)$ for each st_t, and proceed with next iteration of t</td></tr>
</table>

all utility functions which are monotonically increasing in consumption ($\frac{\partial u}{\partial c}(c_t, d_t, st_t) > 0$) the optimal consumption and savings at the terminal period is just $c_T = M_T - A_0$ and $A_T = A_0$. In such case the only equation to be solved numerically is eliminated and the method avoids root finding operations completely.

The lower panel of Table 1 contains a set of nested loops that have to be run in order to find optimal behavior in all the rest of time periods. The outer-most and the next loop are standard in backwards induction solution approaches for Markovian decision problems, but the content of the latter is specific for our generalization of EGM.

The main principles are the following. The core EGM step is performed conditional on each current period discrete decision d_t to produce the d_t-specific endogenous grid $M_t^{grid}(d_t, st_t) \in \mathbb{R}^{n(d_t, st_t)}$ and optimal consumption rule $c_t\left(M_t^{grid}(d_t, st_t), st_t\right)$ defined over this grid. The standard EGM step is augmented with the calculation of the d_t-specific value functions $V_t^{d_t}\left(M_t^{grid}(d_t, st_t), st_t\right)$, performed in parallel at very small marginal cost. It is worth noting that instead of the notion of a fixed grid over end-of-period total resources A_t we adopt the notion of a sequence of guesses of A_t, which are taken one by one and fed into the EGM step. This is important because the proper sequence of guesses may differ for different d_t, and therefore we develop an adaptive algorithm to generate it in order to ensure that d_t-specific endogenous grids overlap over the interval of interest of M_t.

Once EGM step is performed for each value of current discrete decision d_t, d_t-specific value functions $V_t^{d_t}\left(M_t^{grid}(d_t, st_t), st_t\right)$ must be compared to reveal the ranges of M_t where each discrete option is optimal. This comparison may be hard and time consuming due to the fact that the functions to be compared are defined over generally unknown d_t-specific grids $M_t^{grid}(d_t, st_t)$. One extreme case is when for some d'_t and d''_t the grids do not overlap, i.e. $max\left\{M_t^{grid}(d'_t, st_t)\right\} < min\left\{M_t^{grid}(d''_t, st_t)\right\}$. The adoptive algorithm for generating the sequences of A_t mentioned above is designed to rule out this case and ensure that all endogenous grids overlap in the range of the initial grid over M_t chosen at $t = T$.

The computational burden of comparison of the d_t-specific value functions $V_t^{d_t}\left(M_t^{grid}(d_t, st_t), st_t\right)$ is also due to the fact that functions of interest are defined over different grids. The brute force approach implied in previous papers [Barillas and Fernandez-Villaverde, 2007, Fella, 2011] requires, first, that period t unified endogenous grid $M_t^{grid}(st_t)$ is fixed, second, that all value functions at the nodes of this new grid are interpolated, and third, that maximum is found at each point. Additional steps are required to compute the exact boundaries of the regions of optimality for each of the discrete decisions d_t. We propose an algorithm for computation of the the upper envelope of piece-wise linear functions which works across the whole range of M_t thus "vectorizing" the comparison. The exact values of switching between different discrete decisions are also found naturally in the computation of upper envelope. Theorem 2 in Pach and Sharir [1989] provides the upper bound on the number of linear segments in the upper envelope, namely $O\left\{D \cdot (n-1) \cdot \alpha\left(D \cdot (n-1)\right)\right\}$, where $\alpha(\bullet)$ is extremely slowly growing inverse Ackermann function[8] and n is the largest number of points in the endogenous grids.[9]

Our algorithm for the computation of the upper envelope is based on the idea of re-utilizing the existing grid points to avoid unnecessary interpolations and the insight that many comparisons can be skipped for inferior functions. The algorithm achieves linear running time in the number of endogenous grid points.

[8] For all practical purposes it can be assumed that $\alpha\left(D(n-1)\right) < 5$.

[9] Algorithm 1 and Theorem 2.1 in [Edelsbrunner, 1989] demonstrates how this boundary is achieved for piece-wise linear planes with respect to both time and storage.

Finally, the optimal consumption rule $c_t\left(M_t^{grid}(st_t), st_t\right)$ and value function $V_t\left(M_t^{grid}(st_t), st_t\right)$ (in next to last row in Table 1) are also constructed as part of the upper envelope calculation.

3.3 Borrowing constraints

Borrowing constraints present both theoretical and numerical problems, but similarly to the original EGM by Carroll [2006] our generalized method deals very effectively with both of them.

The theoretical difficulty is due to the fact that Euler equation (26) is only necessary as long as the constraint $A_t \geqslant A_0$ is not binding. Carroll [2006] deals with it by running the EGM iteration for the guess $A_t = A_0$ and finding the threshold M_0 where the decision maker is at the verge of becoming credit constrained. All points $M_t > M_0$ where the constraint is relaxed are reconstructed in the EGM step using Euler equation, and for all points $M_t < M_0$ the credit constraint binds, implying that $A_t = A_0$ and $c_t = M_t - A_0$ from (22). So, if optimal consumption is graphed against M_t, Carroll [2006] simply connects the left-most point recovered in EGM step to the point $(A_0, 0)$.

Our generalization of EGM applies exactly the same approach when it comes to the calculation of optimal consumption rule $c_t^{d_t}\left(M_t^{grid}(d_t, st_t), st_t\right)$ during the EGM step. The first point in the construction of d_t-specific endogenous grid $M_t^{grid}(d_t, st_t)$ is manually set to be A_0 with corresponding consumption $c_t^{d_t}(A_0, st_t) = 0$. The EGM step starts with current savings equal to A_0 and recovers the second point in the d_t-specific endogenous grid which is denoted by

$$M_0^{d_t}(st_t) = A_0 + c_t^{d_t}\left(M_0^{d_t}(st_t), st_t\right).$$

For all values of M_t between A_0 and $M_0^{d_t}(st_t)$ the decision maker is credit constrained and saves exactly A_0. This is due to monotonicity of optimal savings function.

Numerical difficulties arise when the d_t-specific value functions $V_t^{d_t}(M_t, st_t)$ are calculated on $M_t^{grid}(d_t, st_t)$ in the proximity of A_0. For any utility function that assigns negative infinity to zero consumption the value functions become increasingly steep at the left side of the grid, and their piece-wise linear approximations become increasingly rough. As a result, the upper envelope of these approximations becomes excessively complex on the left end of the grid, and the algorithm finds lots of switching between different discrete decisions on very small intervals close to A_0. However, we are able to suppress this numerical noise using the following property of the d_t-specific value functions $V_t^{d_t}(M_t, st_t)$.

Denote $M_0^{d_t}(st_t)$ as the level of total resources that is returned by the EGM step when it is called with A_0, to the left of which the decision maker is credit constrained when making choice d_t. Denote $EV_0^{d_t}(st_t)$ as the expected choice-specific value function corresponding to no savings ($A_t = A_0$), i.e.

$$EV_0^{d_t}(st_t) = E\left[V_{t+1}\left(\mathcal{R}_{t+1}\left(A_0, st_t, d_t, st_{t+1}, \xi_{t+1}\right), st_{t+1}\right)\right].$$

Then for $M_t < M_0^{d_t}(st_t)$ the d_t-specific value function $V_t^{d_t}(M_t, st_t)$ is given by

$$V_t^{d_t}(M_t, st_t) = u_1(M_t - A_0, st_t) + \left[u_2(d_t, st_t) + \beta_{t+1} EV_0^{d_t}(st_t)\right], \tag{27}$$

where we assume that the utility function is additively separable in consumption and d_t, i.e. $u(M_t - A_0, d_t, st_t) = u_1(M_t - A_0, st_t) + u_2(d_t, st_t)$. Note that the first term in the right hand side of (27) is independent of d_t and the second term is independent of M_t. The latter implies that for any $M_t < M_0^{d_t}(st_t)$, the d_t-specific value function $V_t^{d_t}(M_t, st_t)$ can be calculated exactly by adding a constant $u_2(d_t, st_t) + \beta_{t+1} EV_0^{d_t}(st_t)$ to the utility function. But moreover, the former implies that the collection of d_t-specific value functions $V_t^{d_t}\left(M_t, st_t\right)$ in the proximity of A_0 appears to be a collection of functions $u_1(c_t, st_t)$ with different d_t-specific vertical shifts. In other words, the d_t-specific value functions $V_t^{d_t}\left(M_t, st_t\right)$ do not intersect for any $M_t < M_0(st_t) = min_{d_t}\left\{M_0^{d_t}(st_t)\right\}$.

Our algorithm for upper envelope computation exploits these properties, disregarding the interval $(A_0, M_0\left(st_t\right))$ completely, and is capable of computing analytical d_t-specific value functions on the intervals $\left(M_0\left(st_t\right), M_0^{d_t}\left(st_t\right)\right)$ to avoid any numerical noise. The resulting upper envelope $V_t\left(M_t, st_t\right)$ (in the third row and next to last row in Table 1) is given with a piecewise linear approximation $V_t\left(M_t^{grid}(st_t), st_t\right)$ to the right of the threshold $M_0(st_t)$, and can be calculated exactly to the left of it using

$$V_t\left(M_t, st_t\right) = u\left(M_t - A_0, d_t, st_t\right) + \beta_{t+1} max_{d_t}\left\{EV_0^{d_t}(st_t)\right\}. \tag{28}$$

The use of an analytical functional form for the value function for $M_t < M_0(d_{t-1}, st_t)$ on the next t iteration also substantially the accuracy of the numerical solution.

3.4 Parallelization

Our generalization of EGM allows for efficient parallelization on the most computationally intensive level of the algorithm.

The general scheme for parallelization of backwards induction computations is to parallelize the within time period operations, and then gather and redistribute to all nodes the results of each time period iteration. In other words, the parallelization is across the state space, with synchronization occurring at the end of each time period of the backward induction iterations. This is necessary because entities to be computed in time period t in general may depend on the value in all points of the state space in period $t+1$.

Our algorithm is compatible with this general scheme, and is parallelized on the level of second column in Table 1 with synchronization at the end of each time period iteration.

The prospects for deeper parallelization is less obvious. Even though the next nested loop over d_t can be parallelized, the consequent upper envelope algorithm is inherently sequential. Yet, because the dimensionality of the state space st_t is the main factor underlying computational complexity of EGM, we expect the method to give good scalability in practical applications even without any deeper parallelization.

4 Comparisons of the solution methods

4.1 Computation time and accuracy of solution

The table below previews the numerical comparison of the two solution algorithms for the retirement model solved for 50 time periods using 5000 grid points. The model can be solved very accurately with much fewer points, but we have chosen 5000 grid points to scale up the problem and thus better highlight the differences in computational time.

Table 2: Computation time and accuracy of solution

	Traditional Euler	EGM
Running time	37 sec.	0.11 sec.
Max abs error, $c_t^\star$	5e-9	4e-14
Mean abs error, $c_t^\star$	1.4e-12	1.5e-14
Max abs error, $V_t(M, \mathbb{R})$	39.466	15.163
Mean abs error, $V_t(M, \mathbb{R})$	2.5e-02	3.2e-02

5 Conclusions

<Speed-up with no loss of accuracy>

The clear limitation of our extension of EGM is the restriction that there is a single continuous decision variable c_t. While other continuous decisions could in principle be discretized, so that our proposed discrete-continuous adoption of EGM will apply to any desired degree of approximation, it does not seem to be an attractive strategy because of the loss of accuracy in discrete representation of the policy function.

<Discuss typical models for which this method is optimal: large number of discrete choices, but not discretized continuous choice (curse of dimensionality kills when accuracy of discretization increases)>

References

Francisco Barillas and Jesus Fernandez-Villaverde. A generalization of the endogenous grid method. *Journal of Economic Dynamics and Control*, 31(8):2698–2712, August 2007. URL http://ideas.repec.org/a/eee/dyncon/v31y2007i8p2698-2712.html.

Christopher D. Carroll. The method of endogenous gridpoints for solving dynamic stochastic optimization problems. *Economics Letters*, 91(3):312–320, June 2006. URL http://ideas.repec.org/a/eee/ecolet/v91y2006i3p312-320.html.

Andrew Clausen and Carlo Strub. Envelope theorems for non-smooth and non-concave optimization. Preliminary and incomplete, version of February 14, 2011.

Herbert Edelsbrunner. The upper envelope of piecewise linear functions: Tight bounds on the number of faces. *Discrete &amp; Computational Geometry*, 4:337–343, 1989. ISSN 0179-5376. URL `http://dx.doi.org/10.1007/BF02187734`. 10.1007/BF02187734.

Giulio Fella. A generalized endogenous grid method for non-concave problems. *School of Economics and Finance, Queen Mary University of London Working Paper*, N. 677, 2011.

Nils H. Hakansson. Optimal investment and consumption strategies under risk for a class of utility functions. *Econometrica*, 38-5:587–607, 1970.

János Pach and Micha Sharir. The upper envelope of piecewise linear functions and the boundary of a region enclosed by convex plates: Combinatorial analysis. *Discrete &amp; Computational Geometry*, 4:291–309, 1989. ISSN 0179-5376. URL `http://dx.doi.org/10.1007/BF02187732`. 10.1007/BF02187732.

Edmund Phelps. The accumulation of risky capital: A sequential utility analysis. *Econometrica*, 30-4:729–743, 1962.