EGM^n: The Sequential Endogenous Grid Method

Alan Lujan[a,b,*]

[a] *Johns Hopkins University, Krieger School of Arts and Sciences Washington, DC,*
[b] *Econ-ARK*

ARTICLE INFO

Keywords:
endogenous grid method
dynamic programming
interpolation
computational economics

ABSTRACT

Heterogeneous agent models with multiple decisions face a computational dilemma. Joint optimization over all choices is slow; strong separability restrictions speed computation but rule out economically interesting interactions. This paper resolves this tension through sequential decomposition. Economically simultaneous decisions need not be computationally joint. Decomposing problems into sequential stages allows each stage to exploit the Endogenous Grid Method's efficiency, chaining speed gains across decisions. The resulting non-rectilinear grids require interpolation methods that respect their structure. We develop ENGINE (ENdogenous Grid INterpolation and Extrapolation), a sequential interpolation algorithm requiring no preprocessing and no triangulation, which exploits the index structure inherited from exogenous grids. The combined method avoids optimization at stages where separable utility or invertible transitions permit EGM inversion. Benchmarks show 2-3x speedups over existing curvilinear interpolation, with simpler implementation and natural parallelization.

1. Introduction

Solving heterogeneous agent models with multiple decisions has traditionally required nested optimization, which is slow, or restrictive assumptions about preferences that limit economic relevance. Models featuring idiosyncratic risk, borrowing constraints, and rich household heterogeneity in the tradition of Aiyagari (1994) and Huggett (1993) have become central to macroeconomic analysis (Krueger, Mitman and Perri, 2016). The Endogenous Grid Method transformed computation of consumption-savings problems by inverting first-order conditions, but extending this insight to problems with multiple continuous choices and high-dimensional state spaces has proved difficult, with earlier attempts requiring either full triangularity of the FOC system (Iskhakov, 2015) or hybrid methods that embed optimization within EGM loops (Druedahl and Jørgensen, 2017).[1]

A household choosing consumption, labor, and portfolio allocation makes these decisions based on the same information, but the computational problem can be decomposed into sequential stages. This paper shows how to structure such decompositions so that most stages admit an EGM inversion, with remaining stages solved by reduced-dimension optimization, accumulating efficiency gains across decisions.

1.1. Background and literature

Dynamic heterogeneous agent models require solving high-dimensional optimization problems at each point in a potentially large state space. Standard grid search approaches impose real computational costs: a

I would like to thank Chris Carroll, Matthew White, and Fedor Iskhakov for their helpful comments and suggestions. The remaining errors are my own. This paper was awarded the Outstanding Graduate Student Paper Award at the 29th International Conference on Computing in Economics and Finance (CEF 2023, Nice, France). Presentations at the Johns Hopkins University Macro Brownbag (2024), the International Association for Applied Econometrics annual conference (2024, Thessaloniki, Greece), and the University of Texas at Austin MA in Economics Program 10-Year Reunion Celebration (2024) improved the exposition and applications. The author received support through the Econ-ARK project funded by a grant from the Alfred P. Sloan Foundation during the development of this work. All figures and other numerical results were produced using the `Econ-ARK/HARK` toolkit (Carroll, Kaufman, Kazil, Palmer and White, 2018).

*Corresponding author
`alujan@jhu.edu` (A. Lujan)
`https://advanced.jhu.edu/directory/alan-lujan/` (A. Lujan)
ORCID(s): `0000-0002-5289-7054` (A. Lujan)

[1] In companion work, Lujan (2026) extends EGM to Epstein-Zin preferences via a power transformation, complementing the multi-decision approach developed here.

three-choice problem with N grid points per dimension requires $O(N^3)$ evaluations per state, and this cost multiplies across all state space points.[2] Carroll (2006) introduced the Endogenous Grid Method (EGM) to accelerate solution of dynamic stochastic consumption-savings problems. Starting from a grid of post-decision states, EGM inverts the first-order condition to recover the optimal consumption policy and the corresponding pre-decision state, converting problems that required hours of computation into tractable exercises. Originally restricted to models with a single control and state variable, EGM was extended by Barillas and Fernández-Villaverde (2007) to multiple controls (labor-leisure choice) in a neoclassical growth model, and subsequent work expanded its applicability further.[3]

The decade after Carroll's original method saw EGM's scope broaden along two fronts: handling structural complications (non-convexities, constraints, discrete choices) and scaling to multidimensional problems.

On the structural side, handling non-convexities and constraints required new techniques. The Generalized Endogenous Grid Method (GEGM) of Fella (2014) relaxes continuity requirements by evaluating candidate solutions from first-order conditions in overlapping regions. Occasionally binding constraints among endogenous variables, as in durable goods models with collateral requirements, were addressed by Hintermaier and Koeniger (2010). Building on both advances, Druedahl and Jørgensen (2017) introduce G2EGM, which extends the generalized approach to multiple controls and states.

Discrete choices and non-standard preferences posed different challenges. How should EGM handle jumps between discrete alternatives? Iskhakov et al. (2017) incorporate discrete choices through extreme value errors, though their application remains restricted to single control and state variables; Clausen and Strub (2020) formalize the conditions under which this works and analyze nested solution approaches. On the preference side, Hallengreen, Jørgensen and Olesen (2024) develop iEGM, extending EGM to utility functions lacking closed-form inverse marginal utility by using numerical inversion. Monte Carlo evidence from Jørgensen (2013) confirms that EGM dominates both standard value function iteration and MPEC in speed and robustness for structural estimation.

The frontier that matters most for this paper is the multivariate one. White (2015) formalized conditions under which EGM applies to multidimensional problems and developed interpolation methods for the curvilinear grids that result. Iskhakov (2015) characterize a triangular structure in systems of first-order conditions that permits sequential solution by substitution without root-finding. Triangular EGM requires triangularity of the entire FOC system; EGM^n instead requires sequential invertibility of Euler equations at each stage, permitting mixed structures where some stages use separable utility, others use invertible transitions, and others use standard optimization. Druedahl (2021) provides a guide to solving non-convex consumption-saving models, combining an upper envelope algorithm with the nested EGM approach of Druedahl and Jørgensen (2017). Nested methods reduce computational complexity relative to joint optimization but still embed optimization or root-finding within outer EGM loops, retaining some grid search overhead. Ludwig and Schön (2018) propose Delaunay triangulation for interpolating on the curvilinear endogenous grids that arise in such problems, though triangulation costs can offset the gains from avoiding grid search.

A parallel line of research exploits first-order conditions without explicit EGM structure. Maliar and Maliar (2013) develop the Envelope Condition Method, which shares EGM's strategy of avoiding numerical optimization, though the method's forward-solution structure restricts it to infinite-horizon problems. Arellano, Maliar, Maliar and Tsyrennikov (2016) apply envelope condition methods to sovereign default models, achieving EGM-like efficiency gains. Mendoza and Villalvazo (2020) propose a fixed-point iteration algorithm (FiPIt) that avoids both root-finding and irregular interpolation in models with occasionally binding constraints, offering an alternative to EGM when constraint structures differ.

Machine learning offers a different path to high-dimensional problems, particularly models with aggregate uncertainty where the wealth distribution becomes a state variable (Krusell and Smith, 1998). Scheidegger and Bilionis (2019) apply Gaussian Process Regression to compute global solutions for high-dimensional stochastic problems. Maliar, Maliar and Winant (2021) use neural networks to approximate systems of equations characterizing dynamic economic models. Azinovic, Gaegauf and Scheidegger (2022) develop deep

[2] See Maliar and Maliar (2014) for a thorough survey of numerical methods for such problems.

[3] See Barillas and Fernández-Villaverde (2007); Maliar and Maliar (2013); Fella (2014); Iskhakov, Jørgensen, Rust and Schjerning (2017).

equilibrium nets that train neural networks to satisfy all equilibrium conditions along simulated paths, demonstrating applicability to life-cycle models with heterogeneity. These machine learning approaches trade the interpretability of explicit policy functions for scalability to very high dimensions, whereas EGM^n maintains interpretable intermediate value functions at each decision stage. In a complementary direction, Bayer, Luetticke, Weiss and Winkelmann (2026) develop an endogenous gridpoint method for distributional dynamics (DEGM), applying endogenous grid ideas to track the wealth distribution in heterogeneous agent models with aggregate risk, achieving an order of magnitude speedup over histogram methods.

1.2. Methodology and contributions

EGM^n (Sequential EGM) decomposes multidecision problems into stages that each admit an EGM inversion, passing efficiency gains forward from one stage to the next. Each stage applies EGM when the subproblem structure permits inversion of first-order conditions; when a subproblem lacks the required separability or invertibility, standard optimization methods apply. The method requires continuous choices satisfying smooth regularity conditions (Section 5). Compared to nested approaches that embed optimization within EGM loops, EGM^n requires stronger structural assumptions (separable utility or independent transitions), but avoids optimization at applicable stages and exposes interpretable intermediate value functions at each stage. ENGINE addresses the resulting curvilinear interpolation challenge, achieving 2-3x speedups over existing methods (Section 4).

We develop the method through two models of increasing complexity: Section 2 uses a three-choice labor-consumption-portfolio problem to illustrate sequential decomposition, and Section 3 extends to a health investment model where two persistent state variables generate genuinely two-dimensional curvilinear grids. The interpolation challenge these grids pose motivates ENGINE (Section 4), while Section 5 formalizes the separability and invertibility conditions that determine when each stage admits EGM inversion. Section 6 discusses limitations and extensions.

2. The Sequential Endogenous Grid Method

Models where households simultaneously choose consumption, labor supply, and portfolio allocation present a computational challenge. Joint optimization over all three choices requires evaluating a three-dimensional optimization at every state space point. Imposing strong separability restrictions speeds computation but may rule out economically interesting preference specifications. Sequential decomposition exploits partial separability: when decisions are separable in stages but not necessarily globally, each stage can be solved efficiently through EGM inversion.

2.1. Problem setup

The baseline problem for demonstrating the Sequential Endogenous Grid Method (EGM^n) is a discrete time version of Bodie, Merton and Samuelson (1992) where a consumer has the ability to adjust their labor as well as their consumption in response to financial risk. This model serves as an ideal pedagogical example for three reasons: it features three simultaneous decisions (labor, consumption, portfolio), it admits a natural sequential ordering where each stage sheds a state variable, and two of three stages permit EGM inversion while the portfolio stage requires optimization. The objective consists of maximizing the present discounted lifetime utility of consumption and leisure over a finite horizon of $T+1$ periods, where $t=0$ denotes the beginning of the planning horizon (the first period of economic life) and $t=T$ denotes the terminal period.

$$\mathrm{V}_0(B_0,\theta_0) = \max \mathbb{E}_0 \left[\sum_{n=0}^{T} \beta^n \mathrm{u}(C_n, Z_n) \right]. \tag{1}$$

where V_0 is the lifetime value function at the start of period zero, B_0 is beginning-of-period bank balances, θ_0 is the transitory wage shock, and $\mathbb{E}_0$ denotes the expectation conditional on information available at the start of the planning horizon.

In particular, this example makes use of a utility function adapted from Bodie et al. (1992), with additively separable utility of consumption and leisure that is homogeneous of degree $1-\rho$ in permanent income (ensuring a balanced growth path):

$$\mathrm{u}(C, Z) = u(C) + h(Z) = \frac{C^{1-\rho}}{1-\rho} + (\nu \mathbf{P})^{1-\rho} \cdot \frac{Z^{1-\zeta}}{1-\zeta}, \tag{2}$$

where $\zeta > 0$ determines the curvature of leisure utility (and thus the Frisch elasticity of labor supply), the term $(\nu\mathbf{P})^{1-\rho}$ with $\nu > 0$ scales leisure utility to have the same curvature and growth properties as consumption utility, and $\mathbf{P}$ is permanent income. This scaling follows Mertens and Ravn (2011) and ensures that both utility components are homogeneous of degree $1-\rho$ in $\mathbf{P}$, permitting normalization.[4] The use of additively separable utility is deliberate: it enables multiple EGM steps in the solution process. Dividing through by $\mathbf{P}^{1-\rho}$ and defining normalized consumption $c = C/\mathbf{P}$, the normalized period utility becomes

$$\mathrm{u}(c, z) = \frac{c^{1-\rho}}{1-\rho} + \nu^{1-\rho} \cdot \frac{z^{1-\zeta}}{1-\zeta}. \tag{3}$$

For the remainder of the analysis, we work with these normalized variables.

This model represents a consumer who begins the period with a level of bank balances b_t and a given wage offer θ_t. Simultaneously, they are able to choose consumption, labor intensity, and a risky portfolio share with the objective of maximizing their utility of consumption and leisure, as well as their future wealth.

Expressing the problem in normalized recursive form[5] makes the stationarity of the decision rules apparent. The household solves

$$\begin{aligned}
\mathrm{v}_t(b_t, \theta_t) = & \max_{\{c_t, z_t, \varsigma_t\}} \mathrm{u}(c_t, z_t) + \beta \mathbb{E}_t \left[\Gamma_{t+1}^{1-\rho} \mathrm{v}_{t+1}(b_{t+1}, \theta_{t+1}) \right] \\
& \text{s.t.} \\
\ell_t &= 1 - z_t \\
m_t &= b_t + \theta_t \ell_t \\
a_t &= m_t - c_t \\
\mathbb{R}_{t+1} &= \mathsf{R} + (\mathbf{R}_{t+1} - \mathsf{R})\varsigma_t \\
b_{t+1} &= a_t \mathbb{R}_{t+1} / \Gamma_{t+1},
\end{aligned} \tag{4}$$

where β is the discount factor, Γ_{t+1} is the permanent income growth factor, R is the risk-free gross return, $\mathbf{R}_{t+1}$ is the stochastic risky asset return, $\mathbb{R}_{t+1}$ is the portfolio return, and non-negativity constraints $c_t \geq 0$, $z_t \in [0, 1]$, and $\varsigma_t \in [0, 1]$ restrict feasible choices. The terminal condition is $\mathrm{v}_{T+1} \equiv 0$, so the agent in the final period T consumes all remaining resources and chooses full leisure. Throughout, we assume standard constraint qualifications hold such that interior solutions satisfy first-order conditions.[6] The constraints define a sequence of state transitions: labor supply ℓ_t determines market resources m_t (bank balances plus labor income), consumption determines liquid savings a_t, and the portfolio choice ς_t induces a stochastic return $\mathbb{R}_{t+1}$ that yields next period's normalized bank balances b_{t+1}.

Although the household makes all three decisions simultaneously from an economic perspective, the dependence structure permits sequential solution. The labor-leisure choice determines market resources;

[4] An alternative formulation expresses preferences in terms of disutility of labor as $h(\ell) = -\zeta \frac{\ell^{1+\nu}}{1+\nu}$, which gives $h'(\ell) = -\zeta \ell^{\nu}$ and $h'^{-1}(x) = (-x/\zeta)^{1/\nu}$ for $x < 0$. This formulation does not support a balanced growth path because labor disutility is not homogeneous of degree $1-\rho$ in permanent income.

[5] Following Carroll (2009). Since both utility components are homogeneous of degree $1-\rho$ in $\mathbf{P}$, the level utility can be written as $\mathbf{P}_t^{1-\rho} \left[\frac{c_t^{1-\rho}}{1-\rho} + \nu^{1-\rho} \frac{z_t^{1-\zeta}}{1-\zeta} \right]$, and dividing the Bellman equation by $\mathbf{P}_t^{1-\rho}$ yields the stationary normalized recursive formulation.

[6] Specifically: (i) utility and value functions are twice continuously differentiable in the interior of the constraint set; (ii) the Inada conditions $\lim_{c \to 0} u'(c) = \infty$ and $\lim_{c \to \infty} u'(c) = 0$ hold, ensuring interior solutions away from zero consumption; and (iii) constraint sets are convex with non-empty interior. These conditions ensure first-order conditions are necessary for optimality at interior solutions. Strict concavity of the utility and value functions further ensures sufficiency: any point satisfying the FOC is a global maximum.

given those resources, the consumption-saving choice determines liquid assets; given liquid assets, the portfolio choice follows. This natural ordering reflects the problem's information flow rather than introducing artificial timing. Each stage uses information from subsequent stages (through continuation values) while shedding state variables that later stages do not require.

The sequential decomposition begins at the start of the period with the labor-leisure problem. To distinguish value functions at different stages, we introduce stage superscripts: the original problem has $\mathrm{v}_t \equiv v_t^0$, representing the value at the first decision stage (labor-leisure), and each subsequent stage v_t^i represents the value function after making decisions at stages $0, 1, \ldots, i-1$. At this stage, the household observes bank balances b_t and the wage offer θ_t, choosing leisure to maximize the sum of current leisure utility and the continuation value from market resources m_t:

$$\begin{aligned} v_t^0(b_t, \theta_t) &= \max_{z_t} h(z_t) + v_t^1(m_t) \\ &\text{s.t.} \\ z_t &\in [0,1] \\ \ell_t &= 1 - z_t \\ m_t &= b_t + \theta_t \ell_t. \end{aligned} \tag{5}$$

Once market resources are realized, the pure consumption-saving problem determines how to allocate m_t between current consumption and liquid assets. The state space has been reduced to a single dimension since the wage offer no longer matters:

$$\begin{aligned} v_t^1(m_t) &= \max_{c_t} u(c_t) + v_t^2(a_t) \\ &\text{s.t.} \\ c_t &\in [0, m_t] \\ a_t &= m_t - c_t. \end{aligned} \tag{6}$$

The final stage allocates liquid savings a_t between risk-free and risky assets. This portfolio problem involves no within-period utility, only the expected continuation value from next period's bank balances:

$$\begin{aligned} v_t^2(a_t) &= \max_{\varsigma_t} \beta \mathbb{E}_t \left[\Gamma_{t+1}^{1-\rho} v_{t+1}^0(b_{t+1}, \theta_{t+1}) \right] \\ &\text{s.t.} \\ \varsigma_t &\in [0,1] \\ \mathbb{R}_{t+1} &= \mathsf{R} + (\mathbf{R}_{t+1} - \mathsf{R})\varsigma_t \\ b_{t+1} &= a_t \mathbb{R}_{t+1} / \Gamma_{t+1}. \end{aligned} \tag{7}$$

The sequential formulation follows the nested approaches of Clausen and Strub (2020) and Druedahl (2021) but composes EGM inversions without embedding optimization. Each choice is self-contained in a subproblem, with the structure chosen to minimize state variables at each stage. The sequential formulation is mathematically equivalent to the original joint problem by Bellman's principle of optimality Bellman (1957): no uncertainty resolves between subproblems within a single period. From the agent's information set at time t, all three decisions are made before any time-$t+1$ shocks realize. The expectation operator appears only in the final subproblem, ensuring identical information across decisions. If shocks were to resolve between stages, agents would have access to information unavailable in the original formulation, changing the problem's solution.

Limiting cases confirm that the decomposition nests simpler models. When $\nu \to 0$, leisure utility vanishes and the labor-leisure stage degenerates: the agent works full time regardless of wages, reducing the problem to the standard two-stage consumption-portfolio EGM of Carroll (2006). When $\beta \to 0$, the fully impatient agent consumes all resources immediately and the portfolio stage becomes irrelevant, recovering the static consumption problem. These extremes verify that the sequential structure reduces to well-understood special cases, each lacking one of the three dimensions of choice.

2.2. Sequential solution

Not every subproblem admits an EGM solution. The portfolio stage illustrates this limitation. By assigning leisure utility to the labor-leisure stage and consumption utility to the consumption-savings stage, we exhaust the separable utility components. No separable utility term remains for the risky share; it affects utility only through future wealth, not through contemporaneous utility. This subproblem requires standard convex optimization rather than EGM inversion.

Restating the problem in compact form gives

$$v_t^2(a_t) = \max_{\varsigma_t} \beta \mathbb{E}_t \left[\Gamma_{t+1}^{1-\rho} v_{t+1}^0 \left(a_t(\mathsf{R} + (\mathbf{R}_{t+1} - \mathsf{R})\varsigma_t)/\Gamma_{t+1}, \theta_{t+1} \right) \right]. \tag{8}$$

The first-order condition with respect to the risky portfolio share, after dividing through by $\beta a_t > 0$ (which is positive at interior solutions, so that the first-order conditions hold with equality), is

$$\mathbb{E}_t \left[\Gamma_{t+1}^{-\rho} \frac{\partial v_{t+1}^0}{\partial b} \left(b_{t+1}, \theta_{t+1} \right) \left(\mathbf{R}_{t+1} - \mathsf{R} \right) \right] = 0. \tag{9}$$

Finding the optimal risky share requires numerical root-finding of this condition. The envelope condition is

$$(v_t^2)'(a_t) = \beta \mathbb{E}_t \left[\Gamma_{t+1}^{-\rho} \frac{\partial v_{t+1}^0}{\partial b} \left(b_{t+1}, \theta_{t+1} \right) \mathbb{R}_{t+1} \right]. \tag{10}$$

This completes the portfolio stage solution. The post-decision value v_t^2 captures the expectation over future shocks; Section 3 denotes this quantity w_t to emphasize its post-decision interpretation following Powell (2011).[7]

The consumption-saving EGM follows Carroll (2006) but we cover it for exposition. The consumption-savings subproblem in compact form, substituting the market resources constraint and ignoring the no-borrowing constraint for now, is:

$$v_t^1(m_t) = \max_{c_t} u(c_t) + v_t^2(m_t - c_t). \tag{11}$$

The first-order condition with respect to c_t yields the familiar Euler equation:

$$u'(c_t) = (v_t^2)'(m_t - c_t) = (v_t^2)'(a_t). \tag{12}$$

The marginal utility of consuming one more unit today must equal the discounted expected marginal value of saving that unit for tomorrow, here expressed through the post-decision value function $(v_t^2)'$.

Inverting this equation is the first of three EGM inversions. Inversion requires strict monotonicity of u', which holds under CRRA preferences since $u''(c) = -\rho c^{-\rho-1} < 0$ for $\rho > 0$, ensuring a one-to-one mapping between marginal utility and consumption levels. We adopt the notational convention that bracketed variables (e.g., [a], [m]) denote exogenous grids, while gothic (fraktur) letters (e.g., $\mathfrak{c}$, $\mathfrak{m}$) denote endogenous quantities constructed by inverting first-order conditions. For single-variable value functions, primes denote ordinary derivatives; Section 3 switches to superscript notation for partial derivatives of multivariate functions. The inverted policy function is

$$\mathfrak{c}_t(a_t) = u'^{-1}\left((v_t^2)'(a_t) \right). \tag{13}$$

[7] An alternative formulation defines $v_t^2(a_t) = \max_{\varsigma_t} \tilde{v}_t^1(a_t, \varsigma_t)$ where $\tilde{v}_t^1$ embeds the expectation. Both marginal value functions $\partial \tilde{v}_t^1/\partial a$ and $\partial \tilde{v}_t^1/\partial \varsigma$ can then be computed in one expectation step, avoiding redundant integration.

Given the utility function above, the marginal utility of consumption and its inverse are

$$u'(c) = c^{-\rho} \qquad u'^{-1}(x) = x^{-1/\rho}. \tag{14}$$

Carroll (2006) demonstrates that an exogenous grid of [a] points yields the unique $\mathfrak{c}_t([\mathrm{a}])$ that optimizes the consumption-saving problem. Using the market resources constraint, the exact amount of market resources consistent with this consumption-saving decision is

$$\mathfrak{m}_t([\mathrm{a}]) = \mathfrak{c}_t([\mathrm{a}]) + [\mathrm{a}]. \tag{15}$$

This $\mathfrak{m}_t([\mathrm{a}])$ is the "endogenous" grid that is consistent with the exogenous decision grid [a]. Given a $(\mathfrak{m}_t([\mathrm{a}]), \mathfrak{c}_t([\mathrm{a}]))$ pair for each $a \in [\mathrm{a}]$, an interpolating consumption function can be constructed for market resources values not on the endogenous grid.

The envelope condition pins down the marginal value of market resources at the optimum:

$$(v_t^1)'(m_t) = (v_t^2)'(a_t) = u'(c_t). \tag{16}$$

This chain of equalities links stages: the labor-leisure stage needs $(v_t^1)'$, which it obtains directly from the consumption-saving solution without additional computation.

The labor-leisure subproblem can be restated more compactly by substituting $\ell_t = 1 - z_t$ and $m_t = b_t + \theta_t \ell_t$ directly into the objective:

$$v_t^0(b_t, \theta_t) = \max_{z_t} h(z_t) + v_t^1(b_t + \theta_t(1 - z_t)). \tag{17}$$

Here market resources $m_t = b_t + \theta_t(1 - z_t)$ appear inside the value function, so the trade-off between leisure and labor income is mediated entirely through v_t^1.

The first-order condition with respect to leisure is

$$h'(z_t) = (v_t^1)'(m_t)\theta_t. \tag{18}$$

This equates the marginal utility of leisure to the opportunity cost of forgone labor income θ_t, scaled by the marginal utility of consumption $(v_t^1)'(m_t) = u'(c_t)$.

The marginal utility of leisure and its inverse are

$$h'(z) = \nu^{1-\rho} z^{-\zeta} \qquad h'^{-1}(x) = (x/\nu^{1-\rho})^{-1/\zeta} \quad (x > 0). \tag{19}$$

Using an exogenous grid of market resources [m] and wage shocks $[\theta]$, leisure follows as

$$\mathfrak{z}_t([\mathrm{m}], [\theta]) = h'^{-1}\left((v_t^1)'([\mathrm{m}])[\theta]\right). \tag{20}$$

However, agents with high market resources m_t and low wage offers θ_t may find the unconstrained optimum violates the feasibility constraint $z_t \in [0, 1]$. When this occurs, the solution is projected onto the constraint boundary, defining the constrained optimal function $\hat{\mathfrak{z}}_t([\mathrm{m}], [\theta])$ as

$$\hat{\mathfrak{z}}_t([\mathrm{m}], [\theta]) = \max\left\{\min\left\{\mathfrak{z}_t([\mathrm{m}], [\theta]), 1\right\}, 0\right\}. \tag{21}$$

This projection onto the feasible set ensures that the leisure choice respects the time endowment constraint.[8] In practice, the projection primarily addresses the upper bound $z_t = 1$ (full leisure, zero

[8] With the functional form $h'(z) = \nu^{1-\rho} z^{-\zeta}$, the Inada condition $\lim_{z \to 0} h'(z) = \infty$ ensures the lower bound $z_t = 0$ never binds for finite wages and marginal values. At the upper bound $z_t = 1$, the agent chooses full leisure when $h'(1) = \nu^{1-\rho} \geq (v_t^1)'(m_t)\theta_t$, which occurs for sufficiently low wages or high wealth.

labor supply), which binds for wealthy agents facing low wages. The lower bound $z_t = 0$ rarely binds due to the Inada condition on leisure utility. In regions where constraints bind, the Kuhn-Tucker conditions replace the unconstrained first-order condition. Interpolation must handle potential non-differentiabilities at constraint boundaries by sorting the endogenous grid points and applying piecewise interpolation with separate segments on either side of each binding constraint, though these kink regions typically affect only small portions of the state space.

Labor follows as $\mathfrak{l}_t(m_t, \theta_t) = 1 - \hat{\mathfrak{z}}_t(m_t, \theta_t)$. For each θ_t and m_t on the exogenous grid, the endogenous grid of bank balances is $\mathfrak{b}_t(m_t, \theta_t) = m_t - \theta_t \mathfrak{l}_t(m_t, \theta_t)$.

The envelope condition then provides the marginal value of bank balances. At interior solutions where $z_t \in (0, 1)$, the first-order condition $h'(z_t) = (v_t^1)'(m_t)\theta_t$ holds, giving

$$\frac{\partial v_t^0}{\partial b}(b_t, \theta_t) = (v_t^1)'(m_t) = h'(z_t)/\theta_t. \tag{22}$$

At corner solutions ($z_t = 0$ or $z_t = 1$), the envelope theorem still yields $\partial v_t^0/\partial b = (v_t^1)'(m_t)$, but the second equality does not hold because the first-order condition is replaced by the Kuhn-Tucker inequality.

The resulting endogenous grid for the labor-leisure problem is curvilinear rather than rectilinear, requiring specialized interpolation methods discussed in Section 4.[9]

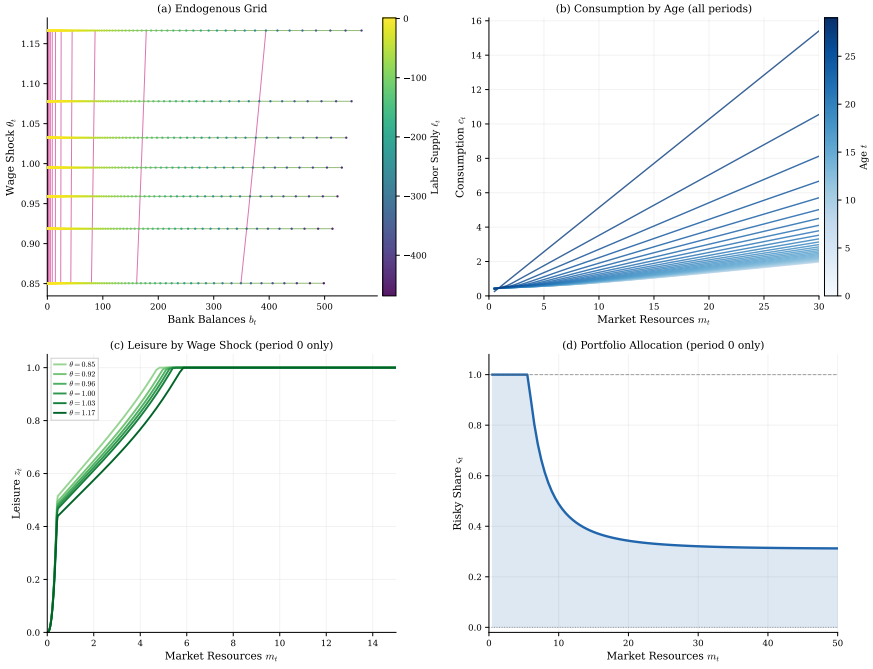

Figure 1: Sequential decomposition produces well-behaved policy functions for the labor-portfolio model ($T = 30$, parameters in footnote).[11] Panel (a) shows the curvilinear endogenous grid from the labor-leisure stage at $t = 0$, colored by optimal labor supply; the grid's shape is the direct output of EGM inversion, not a post-processing step. Panels (b), (c), and (d) confirm that the three-stage EGM recovers smooth, monotone policy functions. Panel (b) shows that the marginal propensity to consume is higher at low wealth and declines as precautionary savings motives weaken, consistent with buffer-stock theory. Panel (c) shows that lower wages induce more leisure at $t = 0$, since the opportunity cost of not working falls. Panel (d) shows the risky portfolio share at $t = 0$: wealthier households bear more equity risk, a pattern that would be difficult to obtain from a joint three-dimensional optimization without the sequential decomposition.

[9]The labor-leisure problem could be solved using simpler interpolation methods since the grid warping occurs along only one dimension. Curvilinear Grid Interpolation is used here to demonstrate the sequential decomposition that is the essence of EGM^n and to illustrate CGI in a transparent setting before encountering the two-dimensional curvilinear grids of Section 3.

Figure 1 shows the full solution: panel (a) displays the curvilinear endogenous grid colored by labor supply, panels (b-d) show the consumption, leisure, and portfolio policy functions. The labor-portfolio example demonstrates EGM^n's core principle: decompose economically simultaneous decisions into computationally sequential stages, composing EGM inversions without optimization. Sequential decomposition exploited separability in utility (leisure) and transitions (portfolio returns). The labor-leisure stage used EGM inversion; the portfolio stage required convex optimization. Each stage required at most one post-decision state variable, keeping dimensionality manageable. The resulting curvilinear grid from the labor-leisure inversion requires specialized interpolation, addressed in Section 4. Retirement planning with multiple accounts requires tracking several state variables simultaneously across stages, presenting a more demanding test of the method's applicability.

3. The EGM^n in Higher Dimensions

The labor-portfolio problem in Section 2 features at most one post-decision state variable per stage, keeping dimensionality manageable. Problems where multiple state variables persist across stages present a more demanding test: two state variables must be tracked simultaneously through the EGM inversion. Health investment, durable goods choices, and human capital accumulation all exhibit this structure. The health investment model requires the risk-aversion coefficient $\rho < 1$ for technical reasons detailed below; it serves to illustrate interpolation properties of curvilinear EGM grids rather than realistic calibration. The labor-portfolio model of Section 2 imposes no such restriction. The health investment problem demonstrates that EGM^n extends to such settings, though the interpolation challenge intensifies: endogenous grids become curvilinear, requiring specialized interpolation methods.

3.1. A health investment model

For a demonstration of multivariate EGM^n, we turn to a consumption-saving model with health investment adapted from White (2015), which itself builds on Ludwig and Schön (2018). This model tests EGM^n more severely than the labor-portfolio example: two continuous state variables (wealth and health) persist across EGM stages, generating genuinely two-dimensional curvilinear grids. White (2015) designed this model specifically to demonstrate curvilinear interpolation for EGM, so comparing ENGINE against it on the same problem provides a direct test of interpolation methods. In this model, the agent makes decisions about consumption and health investment, subject to wage rate uncertainty, health depreciation risk, and mortality risk that depends on health status.

Each period t, the agent observes their market resources m_t and health capital h_t. They choose consumption c_t and health investment n_t to maximize lifetime utility, where $t = 0$ denotes the beginning of the planning horizon and $t = T$ denotes the terminal period. Consumption yields CRRA utility, while health investment produces additional health capital through a concave production function. Health capital serves two purposes: it reduces mortality risk and increases labor income through higher productivity.

The recursive problem is:

$$
\begin{aligned}
\mathrm{v}_t(m_t, h_t) &= \max_{c_t, n_t} u(c_t) + \beta \mathcal{S}_t \mathbb{E}_t \left[\mathrm{v}_{t+1}(m_{t+1}, h_{t+1})\right] \\
\text{s.t.} \quad & c_t \geq 0, \quad n_t \geq 0 \\
H_t &= h_t + \mathrm{g}(n_t) \\
a_t &= m_t - c_t - n_t \\
\mathcal{S}_t &= 1 - \mathsf{D}_t/(1 + H_t) \\
y_{t+1} &= \omega_{t+1} H_t \\
h_{t+1} &= (1 - \delta_{t+1}) H_t \\
m_{t+1} &= \mathsf{R} a_t + y_{t+1},
\end{aligned}
\tag{23}
$$

where the utility function and health production function are:

$$u(c) = \frac{c^{1-\rho}}{1-\rho}, \qquad \mathrm{g}(n) = \frac{\gamma}{\alpha} n^{\alpha}. \tag{24}$$

The notation is as follows: H_t denotes post-investment health capital, a_t is end-of-period assets, $\mathcal{S}_t$ is the survival probability (with the maximum death probability D_t applying when health is zero, i.e., $H_t = 0$), δ_{t+1} is the stochastic health depreciation rate, ω_{t+1} is the stochastic wage rate, and y_{t+1} is labor income. The health production scale factor $\gamma > 0$ and the health production elasticity $0 < \alpha < 1$ determine diminishing returns to health investment. The terminal condition is $\mathrm{v}_{T+1} \equiv 0$, so the agent in the final period consumes all remaining resources. The agent trades off current consumption utility against the survival probability $\mathcal{S}_t$, which depends on health through D_t: higher health investment raises $\mathcal{S}_t$ and thereby increases the weight placed on future value.[12]

This model requires $\rho < 1$ so that utility is non-negative for all $c > 0$, and that there is positive probability of zero wage ($\omega = 0$). These restrictions ensure that the first-order conditions are necessary and sufficient to characterize the solution. The $\rho < 1$ restriction implies low risk aversion (below log utility), which limits the model to less risk-averse agents than standard calibrations; the health model serves to illustrate the interpolation properties of curvilinear EGM grids rather than to represent empirically plausible household behavior. The labor-portfolio model of Section 2 imposes no such restriction and uses $\rho = 5$. As a limiting case, when the health production elasticity $\alpha \to 1$ (linear production), the marginal product g' becomes constant, so the EGM inversion at the health investment stage yields a unique solution independent of the state variables, and the problem collapses to a standard consumption-savings problem with an exogenous health process.[13]

3.2. Sequential EGM solution

EGM^n decomposes a problem with multiple controls into a sequence of subproblems, each handling a single control variable. This problem splits into three stages: a health investment stage, a consumption stage, and a post-decision stage that handles expectations. Using the stage superscript convention from Section 2, $\mathrm{v}_t \equiv v_t^0$ is the value at the first decision stage (health investment). For the post-decision stage, we use w_t rather than v_t^2 (the notation of Section 2) to emphasize that it represents a post-decision value function (capturing expectations over future shocks) rather than a decision stage. This notation follows Powell (2011) where w denotes the post-decision state value.[14] Since all value functions in this section have multiple arguments, we write partial derivatives as superscripts ($\mathrm{w}_t^a \equiv \partial \mathrm{w}_t / \partial a$) rather than the prime notation of Section 2; the Appendix uses subscript notation (v_{lH}^1) for second-order terms.

The post-decision stage represents the value of end-of-period states (a_t, H_t) before the realization of next period's shocks:

$$\mathrm{w}_t(a_t, H_t) = \beta \left(1 - \frac{\mathsf{D}_t}{1 + H_t}\right) \mathbb{E}_t \left[\mathrm{v}_{t+1}(m_{t+1}, h_{t+1})\right], \tag{25}$$

where $y_{t+1} = \omega_{t+1} H_t$, $h_{t+1} = (1 - \delta_{t+1}) H_t$, and $m_{t+1} = \mathsf{R} a_t + y_{t+1}$. Conceptualizing this subproblem as a separate stage allows the function w_t to be constructed once and used in prior optimization problems without repeated expectation calculations.[15]

[12] The concave production function $\mathrm{g}(n) = \frac{\gamma}{\alpha} n^{\alpha}$ with $0 < \alpha < 1$ exhibits diminishing returns to health investment. The marginal product $\mathrm{g}'(n) = \gamma n^{\alpha - 1}$ is strictly decreasing and satisfies the Inada conditions $\lim_{n \to 0} \mathrm{g}'(n) = \infty$ and $\lim_{n \to \infty} \mathrm{g}'(n) = 0$, ensuring interior solutions for health investment when resources permit.

[13] With $\rho > 1$, utility is negative and the product $\mathcal{S}_t \cdot \mathbb{E}[\mathrm{v}_{t+1}]$ would reward mortality. The restriction mirrors White (2015) and is specific to this health formulation, not to EGM^n generally. Models with mortality risk and $\rho > 1$ can use a bequest motive or utility differences; see De Nardi (2004).

[14] The labor-portfolio model in Section 2 places the discount factor β inside v_t^2, which similarly absorbs the expectation. The w_t notation makes this structure explicit.

[15] Computing the post-decision value function separately improves computational efficiency by avoiding redundant expectation calculations. The marginal value functions $\mathrm{w}_t^a(a_t, H_t) \equiv \partial \mathrm{w}_t / \partial a$ and $\mathrm{w}_t^H(a_t, H_t) \equiv \partial \mathrm{w}_t / \partial H$ can be computed once on an exogenous grid and interpolated as needed by earlier stages. Note that w_t^H includes both the survival probability benefit of health (through $\partial \mathcal{S}_t / \partial H_t = \mathsf{D}_t / (1 + H_t)^2$) and the continuation value benefit (through next period's health).

After making their health investment decision, the agent has liquid wealth l_t and post-investment health H_t. The consumption problem is:

$$\begin{aligned} v_t^1(l_t, H_t) &= \max_{c_t} u(c_t) + \mathrm{w}_t(a_t, H_t) \\ &\text{s.t.} \quad c_t \geq 0 \\ a_t &= l_t - c_t. \end{aligned} \tag{26}$$

Post-investment health H_t passes through this stage unaffected because it enters the continuation value but is not changed by the consumption decision. For each fixed value of H_t, this is a standard one-dimensional consumption-saving problem that can be solved via EGM.[16]

At the beginning of the period, the agent has market resources m_t and health capital h_t. The health investment problem is:

$$\begin{aligned} v_t^0(m_t, h_t) &= \max_{n_t} v_t^1(l_t, H_t) \\ &\text{s.t.} \quad n_t \geq 0 \\ l_t &= m_t - n_t \\ H_t &= h_t + \mathrm{g}(n_t). \end{aligned} \tag{27}$$

This stage has no direct utility payoff, but the agent's choice of health investment affects both their liquid wealth (through the budget) and their post-investment health (through the production function). The differentiable and invertible health production function g enables an EGM step. The constraint $n_t \geq 0$ is handled by the Inada condition $\lim_{n \to 0} \mathrm{g}'(n) = \infty$: the marginal product of health investment grows without bound near zero, so interior solutions obtain whenever the marginal value of health is positive and the agent has resources to invest.[17]

3.2.1. The consumption EGM

The consumption stage applies the same EGM inversion as Section 2, now with health H_t passing through as a parameter. On an exogenous grid of post-decision states $([\mathrm{a}], [\mathrm{H}])$, the first-order condition $u'(c_t) = \mathrm{w}_t^a(a_t, H_t)$ inverts to

$$\mathfrak{c}_t([\mathrm{a}], [\mathrm{H}]) = \left(\mathrm{w}_t^a([\mathrm{a}], [\mathrm{H}])\right)^{-1/\rho}. \tag{28}$$

The endogenous liquid wealth grid is $\mathfrak{l}_t([\mathrm{a}], [\mathrm{H}]) = [\mathrm{a}] + \mathfrak{c}_t([\mathrm{a}], [\mathrm{H}])$, and the envelope conditions provide the marginal values needed by the health investment stage:

$$\begin{aligned} (v_t^1)^l(l_t, H_t) &= \mathrm{w}_t^a(a_t, H_t) = u'(c_t), \\ (v_t^1)^H(l_t, H_t) &= \mathrm{w}_t^H(a_t, H_t). \end{aligned} \tag{29}$$

The health investment stage uses EGM with a differentiable transition. The problem in compact form is:

$$v_t^0(m_t, h_t) = \max_{n_t} v_t^1(m_t - n_t, h_t + \mathrm{g}(n_t)). \tag{30}$$

[16] Health H_t enters the continuation value but not the consumption first-order condition, so for each fixed H_t this parallels the consumption-savings stage in Section 2. Strict concavity of u and w_t ensures joint strict concavity of $v_t^1(l_t, H_t)$, guaranteeing uniqueness and valid (non-folding, monotonic) EGM grids.

[17] Unlike the consumption stage, which exploits separable utility, this stage exploits a differentiable and invertible transition function. The health production function g plays a role analogous to the matching function in pension deposit problems: both provide the mathematical structure needed for EGM inversion without requiring separable utility.

The first-order condition with respect to n_t is:

$$-(v_t^1)^l(l_t, H_t) + (v_t^1)^H(l_t, H_t)\mathrm{g}'(n_t) = 0. \tag{31}$$

Rearranging the first-order condition yields:

$$\mathrm{g}'(n_t) = \frac{(v_t^1)^l(l_t, H_t)}{(v_t^1)^H(l_t, H_t)}. \tag{32}$$

Since $\mathrm{g}'(n) = \gamma n^{\alpha-1}$ is invertible, optimal health investment follows:[18]

$$\mathfrak{n}_t([\mathrm{l}], [\mathrm{H}]) = \left(\frac{(v_t^1)^l([\mathrm{l}], [\mathrm{H}])}{\gamma(v_t^1)^H([\mathrm{l}], [\mathrm{H}])}\right)^{1/(\alpha-1)}. \tag{33}$$

The transition equations then recover the endogenous pre-decision states:

$$\begin{aligned} \mathfrak{m}_t([\mathrm{l}], [\mathrm{H}]) &= [\mathrm{l}] + \mathfrak{n}_t([\mathrm{l}], [\mathrm{H}]) \\ \mathfrak{h}_t([\mathrm{l}], [\mathrm{H}]) &= [\mathrm{H}] - \mathrm{g}(\mathfrak{n}_t([\mathrm{l}], [\mathrm{H}])). \end{aligned} \tag{34}$$

The envelope conditions provide:[19]

$$\begin{aligned} (v_t^0)^m(m_t, h_t) &= (v_t^1)^l(l_t, H_t) \\ (v_t^0)^h(m_t, h_t) &= (v_t^1)^H(l_t, H_t). \end{aligned} \tag{35}$$

Starting from a regular rectilinear grid of ([a], [H]), the health investment EGM step produces a curvilinear endogenous grid of $(\mathfrak{m}_t, \mathfrak{h}_t)$. Figure 2 illustrates this transformation.

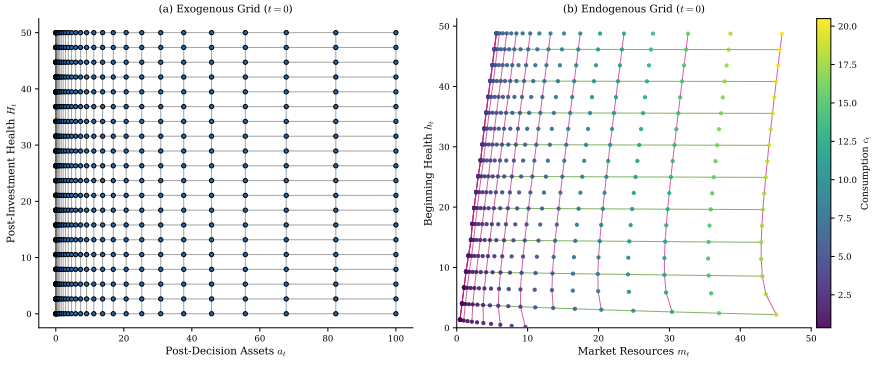

Figure 2: EGM grid transformation for the health investment model at $t = 0$ (the first period of economic life). Panel (a) shows the regular exogenous grid of post-decision assets ([a]) and post-investment health ([H]). Panel (b) shows the curvilinear endogenous grid of market resources ($\mathfrak{m}_t$) and beginning-of-period health ($\mathfrak{h}_t$), with points colored by optimal consumption. The endogenous grid preserves the topological ordering of the exogenous grid despite substantial warping near the constraint boundary, confirming that ENGINE's index-based interpolation remains valid.[21]

Despite this warping, the endogenous grid maintains the topological structure of the exogenous grid: neighboring points in the (l, H) grid map to neighboring points in the (m, h) grid. This property enables efficient interpolation, as discussed in Section 4.

[18] Strict monotonicity of g' ensures the inverse is well-defined. For the power function, $\mathrm{g}''(n) = \gamma(\alpha-1)n^{\alpha-2} < 0$ since $\alpha < 1$, confirming strict monotonicity. The inverse is $(\mathrm{g}')^{-1}(y) = (y/\gamma)^{1/(\alpha-1)}$.

[19] The envelope conditions follow from the fact that both m_t and h_t enter the value function only through the post-decision states l_t and H_t. The first-order condition ensures that the marginal effects of adjusting n_t cancel out, leaving only the direct effects of the state variables on continuation value.

4. Multivariate Interpolation on Curvilinear Grids

EGM's efficiency comes from working on exogenous grids where calculations are straightforward, then using the inverted Euler equation to recover endogenous state variables. But this efficiency creates a problem. The resulting endogenous grid is curvilinear rather than rectilinear, and standard multilinear interpolation, which assumes each dimension varies independently, cannot handle it. What interpolation methods can respect the actual geometry of these grids while exploiting the topological structure that persists despite geometric distortion?

This section presents ENGINE (ENdogenous Grid INterpolation and Extrapolation), which relies on the index structure that endogenous grids inherit from exogenous grids despite geometric warping. ENGINE reduces multidimensional interpolation to a sequence of one-dimensional interpolations along these inherited indices, using the fact that row k remains row k after warping and neighbors remain neighbors, even when geometric regularity (uniform spacing, rectilinear alignment) is lost.

Consider a curvilinear grid indexed by $j \in \{1, \ldots, J\}$ and $k \in \{1, \ldots, K\}$, where each "row" (fixed k) contains J points. Standard curvilinear interpolation locates the containing quadrilateral cell via sector-walking, then solves for normalized coordinates within that cell (a quadratic equation in 2D, Newton iteration in higher dimensions). ENGINE sidesteps both operations by reframing the problem: instead of asking "which cell contains this point?", ENGINE asks "where does a vertical line at x^* intersect each row?" For each of the K rows, this is a one-dimensional interpolation problem: find where x^* falls among the J x-coordinates in that row, then interpolate the y-coordinate to produce an intermediate point. A final 1D interpolation across these K intermediate points yields the answer. The approach works because EGM preserves index structure: point (j, k) in the exogenous grid maps to a geometrically warped but topologically consistent location in the endogenous grid, so that row k remains row k and neighbors remain neighbors.

4.1. Regularity conditions

The interpolation challenge arises because first-order conditions induce nonlinear mappings from exogenous to endogenous grids. Highly nonlinear or non-monotonic Euler equations can distort grid structure beyond the regular spacing assumed by standard methods. Binding constraints compound the problem by creating kinks in policy functions, introducing additional irregularity.

The pure consumption-savings problem illustrates the basic phenomenon. An exogenous grid of post-decision liquid assets [a] maps through the inverted Euler equation to an endogenous grid of market resources [m] with different spacing; the nonlinearity of marginal utility ensures this mapping is non-uniform. In one dimension, that is no obstacle: non-uniform linear interpolation handles the distortion.

Higher-dimensional problems inherit this warping in each dimension simultaneously, potentially destroying the regular structure assumed by standard multilinear interpolation. The degree of structural preservation determines the appropriate interpolation method. Figure 2 in the previous section illustrates this transformation for the health investment model.

Previous approaches to this problem include Delaunay triangulation (Ludwig and Schön, 2018) and curvilinear interpolation (White, 2015). Delaunay triangulation constructs a mesh of simplices over the scattered points and performs barycentric interpolation within each simplex. While general and robust, triangulation is computationally expensive, often requiring $O(N \log N)$ for construction alone, and must be recomputed when grids change during iterative solution procedures. The curvilinear method takes advantage of the grid structure to map irregular quadrilateral sectors to the unit square for bilinear interpolation; sector location uses a visibility walk algorithm, while coordinate calculation in 2D employs a closed-form quadratic solution (3D and higher require Newton iteration).

ENGINE requires structural conditions on the endogenous grid, but these conditions are not additional assumptions: they follow from the same economic structure that makes EGM applicable. Theorem 4.2 below proves that any grid produced by an EGM inversion automatically satisfies ENGINE's requirements, provided the underlying value function is strictly concave and smooth. This applies to standard EGM (Carroll, 2006), nested EGM (Druedahl, 2021), and the EGM^n of this paper alike. However, discrete choices pose a problem: when optimal policies jump discontinuously between discrete alternatives, the resulting grids violate monotonicity (IPM) and ENGINE does not apply. Such problems require upper envelope techniques as in G2EGM or DCEGM. The Endogenous Grid Method works when Euler equations are invertible, value

functions are strictly concave, and both are twice continuously differentiable. These properties generate well-behaved endogenous grids automatically. A grid satisfying these conditions is *homeomorphic* to the rectangular index grid: the EGM mapping is a continuous bijection with continuous inverse from index space to physical space.[22]

The conditions separate into two categories: geometric regularity (ensuring a well-defined grid) and algorithmic efficiency (ensuring ENGINE operates efficiently). We state these conditions formally, then prove that grids satisfying them support ENGINE interpolation, and finally show that EGM-generated grids satisfy these conditions automatically.

Nonlinear EGM mappings can distort cells severely enough to invert them. The Definition 4.1 condition, the classical validity condition from structured mesh generation (Knupp, 1999), ensures each cell maintains proper orientation.

Definition 4.1 (Fold-Free Cell)**.** *Each grid cell with corners* (x_{jk}, y_{jk})*,* $(x_{j+1,k}, y_{j+1,k})$*,* $(x_{j,k+1}, y_{j,k+1})$*,* $(x_{j+1,k+1}, y_{j+1,k+1})$ *is fold-free (also called valid or orientation-preserving) if the bilinear Jacobian determinant is positive at all four corners. Writing the bilinear map in parametric coordinates* $(s,t) \in [0,1]^2$*, the Jacobian* $\mathbf{J}(s,t) = \frac{\partial x}{\partial s}\frac{\partial y}{\partial t} - \frac{\partial x}{\partial t}\frac{\partial y}{\partial s}$ *must satisfy* $\mathbf{J}(0,0) > 0$*,* $\mathbf{J}(1,0) > 0$*,* $\mathbf{J}(0,1) > 0$*, and* $\mathbf{J}(1,1) > 0$*. At the lower-left corner, this reduces to the discrete Jacobian:*

$$\mathbf{J}_{jk} \equiv (x_{j+1,k} - x_{jk})(y_{j,k+1} - y_{jk}) - (x_{j,k+1} - x_{jk})(y_{j+1,k} - y_{jk}) > 0 \tag{36}$$

The remaining three corner conditions use the cell's edge vectors evaluated at the corresponding vertices. Since the bilinear Jacobian is itself bilinear in (s,t)*, it attains its minimum at a corner; four-corner positivity therefore ensures* $\mathbf{J}(s,t) > 0$ *throughout the cell, guaranteeing that the cell does not self-intersect or fold over itself.*[23]

Fold-free ensures local cell validity, but ENGINE also requires that bracketing along rows and columns be well-defined. Definition 4.2 ensures coordinates increase monotonically along each index direction.

Definition 4.2 (Index-Preserved Monotonicity (IPM))**.** *A grid* $\{(x_{jk}, y_{jk}) : j = 1, \ldots, J,\ k = 1, \ldots, K\}$ *satisfies Index-Preserved Monotonicity if:*

1. ***Row monotonicity:*** *For each fixed* k*,* $x_{j+1,k} > x_{jk}$ *for all* $j \in \{1, \ldots, J-1\}$
2. ***Column monotonicity:*** *For each fixed* j*,* $y_{j,k+1} > y_{jk}$ *for all* $k \in \{1, \ldots, K-1\}$

Strict inequalities ensure non-degenerate grid spacing. Weak IPM (allowing equalities) accommodates constraint regions where optimal policies are constant, but requires care to avoid degenerate cells.[24]

Together, Definition 4.1 and Definition 4.2 ensure the grid mapping is a homeomorphism, guaranteeing interpolation correctness. The final condition, Definition 4.3, ensures efficiency: it bounds how fast bracketing indices change across rows, enabling ENGINE to use binary search with bounded search ranges.

Definition 4.3 (Index-Preserved Order (IPO))**.** *For a grid satisfying IPM, let* $j^*(k; x)$ *denote the bracketing index in row* k *for query* x*, defined by* $x_{j^*,k} \le x < x_{j^*+1,k}$*. Index-Preserved Order with constant* $\alpha \ge 1$ *then ensures that this bracketing index changes at a bounded rate across rows:*

[22] IPM and IPO are new terms introduced here. The standard characterization of a valid curvilinear coordinate system is that the grid mapping is a homeomorphism.

[23] The discrete Jacobian $\mathbf{J}_{jk}$ equals twice the signed area of the triangle formed by vertices (x_{jk}, y_{jk}), $(x_{j+1,k}, y_{j+1,k})$, and $(x_{j,k+1}, y_{j,k+1})$. Equivalently, it is the cross product of edge vectors $\mathbf{e}_1 = (x_{j+1,k} - x_{jk}, y_{j+1,k} - y_{jk})$ and $\mathbf{e}_2 = (x_{j,k+1} - x_{jk}, y_{j,k+1} - y_{jk})$. Positive $\mathbf{J}_{jk}$ means these vectors form a right-handed (counterclockwise) pair; negative indicates the cell has folded (inverted orientation); zero indicates degeneracy. The four-corner condition is standard in the structured mesh generation literature (Knupp, 1999) and is stronger than requiring positivity only at the lower-left corner. For EGM-generated grids, Theorem 4.2 establishes that the continuous Jacobian $\det(\mathbf{J}) \ge 1$ throughout each cell, automatically satisfying the four-corner condition.

[24] When constraint regions produce constant policies along a grid dimension (e.g., zero consumption at the borrowing constraint), adjacent grid points may have identical coordinates. The implementation handles this by (1) skipping zero-width intervals during bracket search, (2) using one-sided limits at degeneracy boundaries, and (3) treating constant-policy regions as a single effective grid point. These cases are rare for well-designed grids that concentrate resolution in economically active regions.

$$|j^*(k_1; x) - j^*(k_2; x)| \leq \alpha \cdot |k_1 - k_2| \tag{37}$$

for all query points x and row indices k_1, k_2. Smaller α indicates a smoother grid where brackets change slowly. The symmetric condition holds for column indices.

When $\alpha = 1$, adjacent rows' brackets differ by at most one; when $\alpha = J$, IPO imposes no constraint and ENGINE degenerates to linear search along each row, yielding $O(K \cdot J)$ per query. ENGINE is built around IPO: binary search on rows bounds the search range for j^* at each probe by α.

The IPO constant α relates to the Lipschitz constant L of the EGM mapping. To see this, consider how the bracketing index $j^*(k; x)$ changes across rows. The EGM mapping transforms exogenous grid points to endogenous coordinates. If this mapping has Lipschitz constant L, then $|x_{j,k_1} - x_{j,k_2}| \leq L \cdot |k_1 - k_2| \cdot \Delta_k$ for row spacing Δ_k. For a fixed query x, the bracket j^* can therefore shift by at most $L \cdot |k_1 - k_2| \cdot \Delta_k / \Delta_j$ indices when moving from row k_1 to k_2, where Δ_j is the column spacing. Thus $\alpha = L \cdot \Delta_k / \Delta_j$. Strict concavity of the value function bounds L (the second derivative controls how fast optimal policies change), so for smooth EGM problems, α is typically small.[25]

In practice, α can be estimated empirically after solving by computing $\max_{k_1,k_2} |j^*(k_1; x) - j^*(k_2; x)| / |k_1 - k_2|$ over a dense set of test points. For the models in this paper, $\alpha \leq 2$ on all grids tested, and empirical query times scale as $O(\log J + \log K)$ rather than the worst-case $O(\log J \cdot \log K)$.[26]

With the regularity conditions defined, we now establish their implications. Fold-free and IPM together guarantee that any query point has exactly one location in the grid, so interpolation is well-defined (Theorem 4.1).

Theorem 4.1 (Grid Homeomorphism)**.** *A curvilinear grid satisfying IPM and fold-free is homeomorphic to the rectangular index grid. That is, the mapping $\Phi : \{1, \ldots, J\} \times \{1, \ldots, K\} \to \mathbb{R}^2$ defined by $\Phi(j, k) = (x_{jk}, y_{jk})$ is injective, and extending Φ via bilinear interpolation within each cell yields a homeomorphism from $[1, J] \times [1, K]$ to the grid's image. (See Section A.1 for proof.)*

The homeomorphism ensures correctness, but we also need efficiency. The geometric conditions connect to concrete algorithmic guarantees (Corollary 4.1.1): IPO bounds how fast brackets change, enabling ENGINE to locate cells without exhaustive search.

Corollary 4.1.1 (Applicability)**.** ***ENGINE***
Under IPM and fold-free, the curvilinear grid is homeomorphic to the rectangular index grid. Adding IPO with constant α ensures ENGINE's sequential linear interpolation is both well-defined and efficient:

1. ***Well-defined:*** *IPM ensures coordinate-wise bracketing succeeds along each index slice, with row monotonicity guaranteeing that each slice intersects any vertical line at most once. Combined with fold-free (non-inverted cells), Pass 1 interpolation yields unique intermediate values.*
2. ***Efficient:*** *The IPO constant α controls how quickly brackets change across rows. ENGINE achieves $O(\log J \cdot \log K)$ complexity per query; for small α (smooth problems), empirical performance approaches $O(\log J + \log K)$.*[27]

The preceding results establish what conditions enable ENGINE. The practical question is whether these conditions impose additional restrictions on practitioners. Standard EGM assumptions (strict concavity, smoothness, monotone marginal utility) imply ENGINE's conditions automatically (Theorem 4.2): no additional restrictions arise.

[25] Economically, α becomes large when policy functions change rapidly relative to grid spacing, which occurs near occasionally binding constraints or when the value function has high curvature (strong precautionary motives, tight borrowing limits). In such regions, the EGM mapping stretches or compresses the grid non-uniformly. Standard practice in computational economics uses log-spaced or double-log-spaced grids that concentrate resolution where policies vary most, which keeps α moderate. Highly anisotropic grids (very different spacing in different dimensions) can also inflate α; balanced grid design mitigates this.

[26] The implementation uses a configurable curvature bound (default $\alpha = 2$) that controls the search margin: when probing row k after previously probing row k', the j-bracket search range is bounded by $\pm(\alpha \cdot |k - k'| + 1)$ indices around the previous bracket.

[27] When conditions are partially violated, ENGINE degrades gracefully. If IPM fails but fold-free holds, linear search replaces binary search, yielding $O(K \log J)$. Only when fold-free fails (grid folds over itself) must one resort to unstructured methods.

Theorem 4.2 (Conditions Follow from). ***ENGINEEGM***
Consider a one-dimensional EGM problem where the value function v *is strictly concave and twice continuously differentiable, and the marginal utility* u' *is strictly decreasing. Then the endogenous grid produced by EGM is fold-free and satisfies IPM and IPO with* α *bounded by the curvature of* v*. For multidimensional EGM problems, the same conclusion holds under the additional assumption that the continuation value function exhibits non-negative cross-partial derivatives (complementarity between state variables, e.g.,* $v^1_{lH} \geq 0$ *in the health model). ENGINE's regularity and efficiency conditions are consequences of EGM's applicability conditions, not independent restrictions. (See Section A.2 for proof.)*

The health investment model of Section 3 illustrates these results: consumption and health investment policies are monotonic (IPM), policy functions vary smoothly (IPO), and the EGM mapping is locally invertible (fold-free). Violations of these conditions typically signal either numerical issues or economic features requiring special treatment.

Occasionally binding constraints, such as borrowing limits, merit explicit discussion. These constraints introduce kinks in policy functions but preserve monotonicity, satisfying IPM, IPO, and fold-free. EGM handles such constraints naturally: the exogenous grid defines the unconstrained (interior) region, and the constrained region is constructed by prepending a point at the constraint boundary (e.g., $c = 0$ at $m = 0$ for a liquidity constraint). The kink occurs where the policy transitions from corner to interior solution, but on either side of this boundary the policy remains monotonic in wealth. Grid cells straddling the kink may exhibit elevated α locally, but ENGINE handles this gracefully by widening the search range. Discrete choices, by contrast, create discontinuities that violate these conditions.

4.2. The ENGINE algorithm

ENGINE exploits the index correspondence between endogenous and exogenous grids: each point $(m_t^{(j,k)}, h_t^{(j,k)})$ in the endogenous grid corresponds to indices (j, k) from the exogenous grid $(l_t^{(j)}, H_t^{(k)})$. This structure enables efficient cell location and interpolation without explicit geometric constructions.

There exists a homeomorphism between the curvilinear grid in physical space and the rectilinear grid in index space. Figure 3 illustrates this correspondence: the warped grid on the left maps bijectively and continuously to the regular index grid on the right, with the inverse also continuous, preserving neighborhood relationships. ENGINE works in index space where bracketing is efficient, then inverts the homeomorphism to recover physical coordinates.

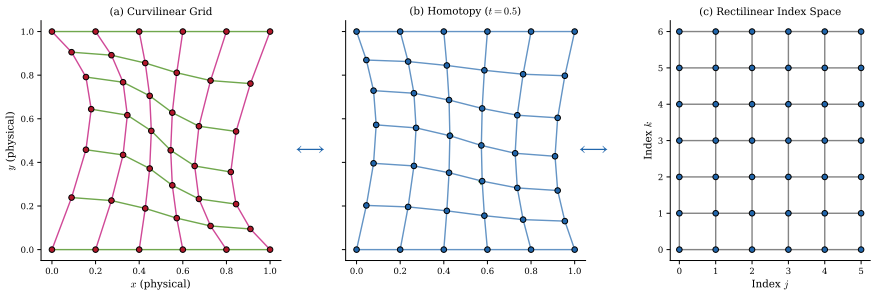

Figure 3: The homeomorphism from physical to index space preserves cell topology despite substantial geometric warping (left: curvilinear physical grid; right: rectilinear index grid). This structure is what makes ENGINE possible: because neighbors remain neighbors after warping, binary search in index space locates the correct cell without geometric operations, reducing cell location from $O(J + K)$ sector-walking to $O(\log J \cdot \log K)$ binary search.

The full algorithm is provided in the Section A.4. Figure 4 illustrates this process: the vertical line at $x = x^*$ intersects each k-slice of the curvilinear grid, with Pass 1 finding the intermediate y-coordinates at these intersections. Pass 2 then interpolates across these intermediate values to produce the final result.

The choice of dimension order (rows versus columns) can affect accuracy when grid warping is highly anisotropic. For the health investment problem, interpolating first along the liquid wealth dimension (which varies more smoothly) and then along the health dimension typically yields better results.

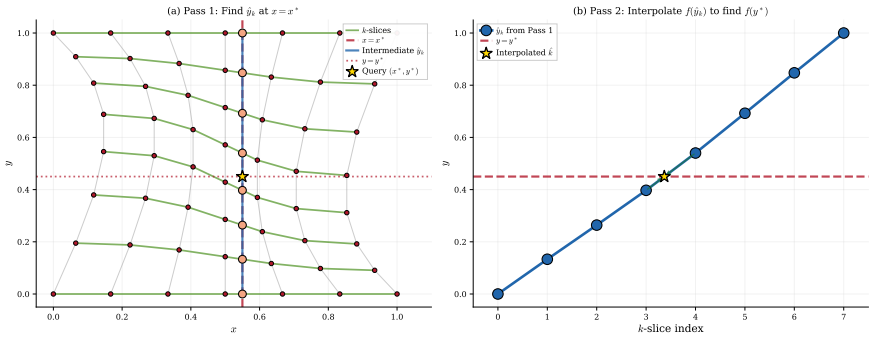

Figure 4: ENGINE reduces 2D interpolation to sequential 1D operations. Pass 1 (panel a): the vertical line at $x = x^*$ intersects each k-slice at intermediate coordinates $\hat{y}_k$ (circles), each requiring only a 1D bracket search. Pass 2 (panel b): a single 1D interpolation across the intermediate points $\{\hat{y}_k\}$ yields the final value at the query point (star). No cell search, coordinate inversion, or preprocessing is required.

ENGINE sidesteps the cell search and coordinate inversion required by curvilinear methods, requiring only sequential 1D interpolations. The tradeoff is that ENGINE requires IPM and IPO in addition to fold-free, but for EGM-generated grids these conditions hold automatically.

For sorted query points, an additional optimization applies: given sorted queries $\{q_i\}$ and sorted grid coordinates $\{x_j\}$, a "walking pointer" bracket index j only advances (never backtracks) as queries are processed in order. Finding the enclosing interval for query q_i requires checking whether $q_i < x_{j+1}$; if so, interpolation occurs within $[x_j, x_{j+1}]$; otherwise, j advances until the condition holds. This reduces per-query complexity from $O(\log J)$ to $O(1)$ when queries are monotonically ordered, as occurs with meshgrid evaluation.

For a query point (x^*, y^*), the interpolated value is:

$$\hat{f}(x^*, y^*) = \texttt{ENGINE}_k\Big(y^*;\ \{\hat{y}_k\}_{k=1}^{K},\ \{\hat{f}_k\}_{k=1}^{K}\Big), \tag{38}$$

where $\hat{y}_k$ and $\hat{f}_k$ are the intermediate coordinates and values from Pass 1.

The preceding theorems established that ENGINE is well-defined and efficient on EGM-generated grids. Decomposing 2D interpolation into sequential 1D operations does not sacrifice accuracy: the method achieves the same second-order convergence as standard bilinear interpolation (Proposition 4.1).

Proposition 4.1 (Sequential Interpolation Accuracy)**.** *For a $J \times K$ grid satisfying IPM, IPO, and fold-free with maximum cell diameter h and a twice-differentiable function f, the two-pass sequential linear interpolation achieves:*

$$|f(x^*, y^*) - \hat{f}| = O(h^2)\|D^2 f\|_\infty \tag{39}$$

where $\|D^2 f\|_\infty = \sup_{x,y} \max\{|f_{xx}|, |f_{xy}|, |f_{yy}|\}$ denotes the supremum norm of the Hessian.

On rectangular grids, this method is equivalent to standard bilinear interpolation. On curvilinear grids satisfying the regularity conditions, the sequential passes provide a smooth approximation that maintains second-order accuracy for functions with bounded second derivatives.[28]

With accuracy established, Proposition 4.2 formalizes the complexity analysis. IPO bounds how fast brackets change across rows, enabling ENGINE to reuse bracket information from previous probes rather than searching from scratch.

[28] Here h denotes the maximum cell diameter. The $O(h^2)$ rate for piecewise linear interpolation is standard (de Boor, 2001). On curvilinear grids the bound holds when the grid mapping Jacobian is bounded away from zero and infinity, which fold-free and IPM ensure. Composing two $O(h^2)$ interpolations preserves this order because intermediate values have bounded variation when the original function is smooth.

Proposition 4.2 (Complexity)**.** ***ENGINE***
For a $J \times K$ grid satisfying fold-free, IPM, and IPO with constant α, ENGINE achieves:

- ***Single query:*** $O(\log J \cdot \log K)$
- ***Regridding to*** $P \times Q$ ***grid:*** $O(PQ \log J \cdot \log K)$

(See Section A.3 for proof.)

Remark 4.1 (Empirical Complexity)**.** *For smooth EGM problems where the IPO constant α is small, empirical performance is closer to $O(\log J + \log K)$ per query because bracket updates require only $O(1)$ work at each probe in practice. The gap between the proven $O(\log J \cdot \log K)$ and the observed $O(\log J + \log K)$ is striking; we suspect a tighter bound holds but have not found the right potential function argument to establish it formally.*

The method offers several computational advantages: natural parallelization across query points (each query is independent), efficient memory access patterns from sequential array traversal, and no preprocessing or Newton iteration in any dimension.[29]

4.2.1. Extension to higher dimensions

ENGINE extends naturally to N dimensions via N sequential passes. For a problem with indices $(j_1, j_2, \ldots, j_N)$:

1. Pass 1 interpolates along the j_1 direction for each combination of $(j_2, \ldots, j_N)$.
2. Pass 2 interpolates along the j_2 direction using the results from Pass 1.
3. Continue until Pass N produces the final interpolated value.

The total complexity for a single query scales as $O(\sum_{n=1}^{N-1}(\prod_{m=n+1}^{N} J_m) \log J_n + \log J_N)$ when using the naive approach that evaluates all intermediate slices. Binary search optimizations analogous to the 2D case can reduce this, but the gains depend on higher-dimensional analogues of IPO that bound bracket displacement across multiple index dimensions simultaneously.[30]

4.3. Comparison with alternative methods

The three approaches to multi-dimensional EGM interpolation differ in their algorithmic strategy for handling curvilinear endogenous grids:

Method	Setup	Query Cost	Interpolation	Grid Requirements
ENGINE (this paper)	$O(1)$	$O(\log J \cdot \log K)$	Sequential linear	Fold-free + IPM + IPO(α)
Curvilinear (White, 2015)	$O(1)$	$O(J + K)$	Bilinear	Fold-free only
Delaunay (Ludwig and Schön, 2018)	$O(N \log N)$	$O(\log N)$	Piecewise linear	None

ENGINE achieves $O(\log J \cdot \log K)$ per query, which is asymptotically faster than curvilinear's $O(J + K)$ sector-walking. For smooth EGM problems with small IPO constant α, empirical performance approaches $O(\log J + \log K)$. Benchmarks show 2-3x speedups depending on grid size, with larger gains at larger grids.

[29] The implementation uses NumPy (Harris, Millman, van der Walt, Gommers, Virtanen, Cournapeau, Wieser, Taylor, Berg, Smith, Kern, Picus, Hoyer, van Kerkwijk, Brett, Haldane, Del Río, Wiebe, Peterson, Gérard-Marchant, Sheppard, Reddy, Weckesser, Abbasi, Gohlke and Oliphant, 2020) with Numba (Lam, Pitrou and Seibert, 2015) just-in-time compilation within the Econ-ARK HARK toolkit (Carroll et al., 2018). The 2-3x speedups reflect both algorithmic gains (reduced complexity from $O(J + K)$ to $O(\log J \cdot \log K)$ per query) and cache-friendly sequential access patterns. When multiple functions share the same endogenous grid, the cell-location work from Pass 1 is reused across all functions.

[30] ENGINE's sequential approach avoids the exponential complexity of triangulation-based methods in higher dimensions. For a three-dimensional problem with J^3 grid points, Delaunay triangulation requires $O(J^3 \log J)$ construction time. The naive ENGINE approach requires $O(J^2 \log J)$ per query; with appropriate regularity conditions, binary search can reduce this toward $O(\log^3 J)$, though the conditions ensuring this are more complex than the 2D case.

Figure 5 illustrates how these three methods partition the state space. Bilinear interpolation identifies the containing quadrilateral cell and computes normalized coordinates within it. Delaunay triangulation constructs a mesh of triangles and uses barycentric coordinates. ENGINE processes the grid slice-by-slice, interpolating along indexed rows before combining intermediate results.

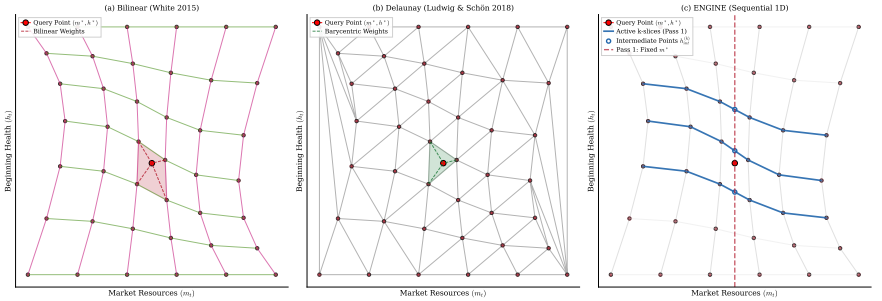

Figure 5: All three methods correctly partition the state space, but ENGINE's slice-based structure is cheaper to traverse. Bilinear interpolation searches for the containing quadrilateral (panel a), Delaunay triangulation constructs simplices from scratch (panel b), and ENGINE interpolates sequentially along indexed k-slices (highlighted curves) before combining intermediate points (circles) to reach the query point (star) in panel (c). The index structure inherited from the exogenous grid makes panel (c) reducible to binary search, while panels (a) and (b) require geometric operations without that structure.

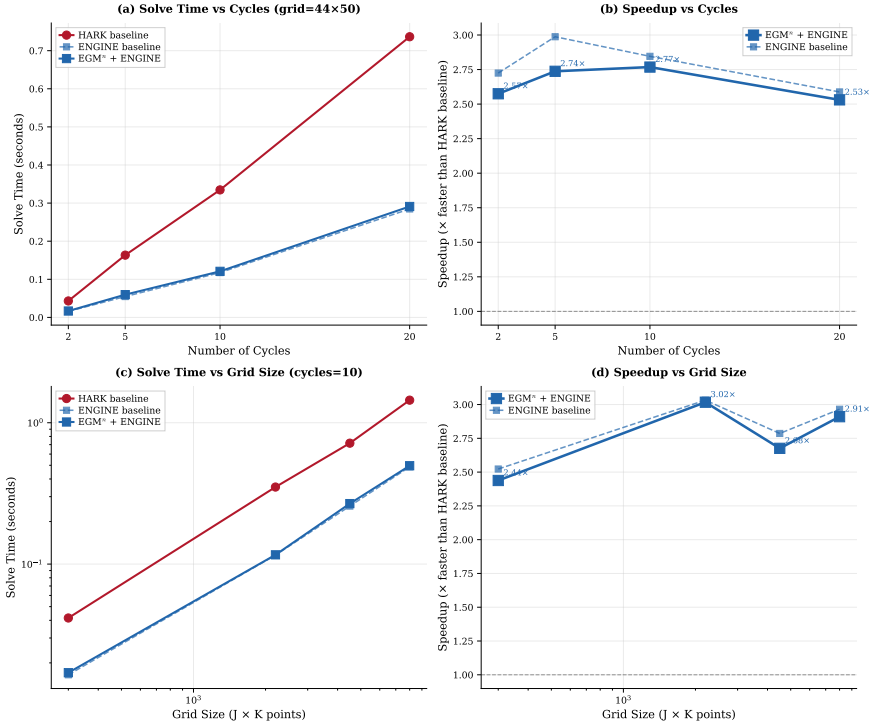

Figure 6: Performance comparison on the health investment model. Panels (a) and (b) show solve time and speedup versus the number of backward induction cycles at fixed grid size (44×50). Panels (c) and (d) show solve time and speedup versus grid size at fixed cycle count (10). EGM^n + ENGINE achieves 2-3x speedups over the curvilinear interpolation baseline of White (2015) depending on grid size.

Figure 6 compares ENGINE against the curvilinear interpolation of White (2015) across both time horizon (cycles) and grid resolution.[31] The benchmarks use the health investment model of Section 3 with $\rho = 0.5$, providing a direct comparison on the same model class for which curvilinear interpolation was developed. As noted in Section 3, the $\rho < 1$ restriction is specific to this health formulation; the benchmarks illustrate interpolation performance on curvilinear grids rather than an economically calibrated model. Speedups scale with grid size as the complexity advantage materializes; accuracy matches standard bilinear interpolation per Proposition 4.1.

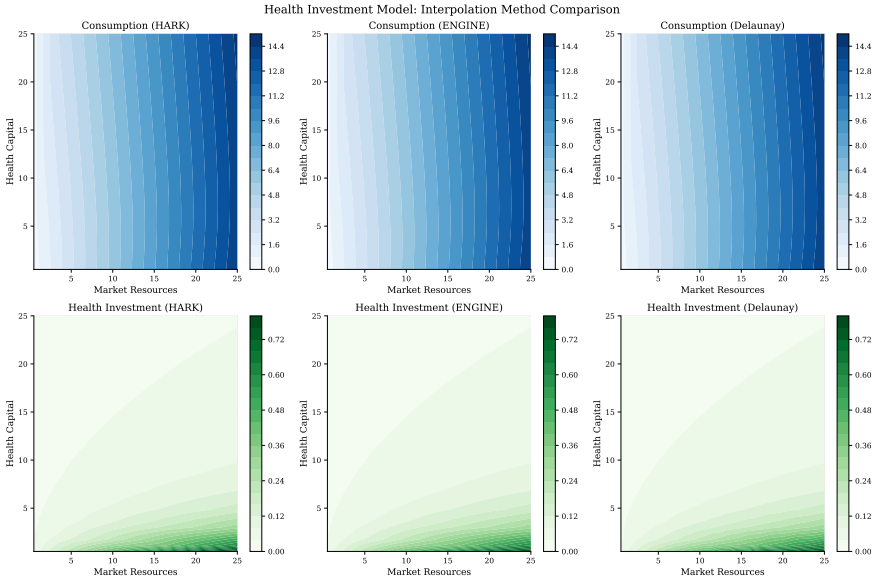

Figure 7: Policy functions for the health investment model computed using different interpolation methods. The top row shows the consumption function $\mathrm{c}(m, h)$ and the bottom row shows the health investment function $\mathrm{n}(m, h)$. All three interpolation methods (curvilinear interpolation of White (2015), ENGINE, and Delaunay triangulation) produce visually indistinguishable policy surfaces: maximum pointwise differences are below 10^{-4} at all grid points tested.

For the health investment model, the dominant cost components in the backward induction loop are: expectation calculations (computing marginal value functions on the exogenous grid via quadrature), interpolation (evaluating value and marginal value functions on the curvilinear endogenous grid), and EGM inversion (recovering endogenous states from first-order conditions). At typical grid sizes (44×50), interpolation accounts for roughly 40-50% of solve time, expectations for 30-40%, and EGM inversion for the remainder. ENGINE's speedup therefore translates most directly into the interpolation component, explaining why total solve-time speedups (2-3x) are smaller than the asymptotic complexity ratio might suggest.

ENGINE provides efficient interpolation when grids satisfy fold-free, IPM, and IPO. By Theorem 4.2, any EGM inversion under standard conditions (strict concavity, smoothness) produces such grids automatically. The next section addresses a separate question: when can a multidecision problem be decomposed into EGM^n stages? The separability and invertibility conditions there determine when EGM^n applies, completing the practical toolkit for applying these methods to new problems.

5. Conditions for Sequential Decomposition

When does EGM^n apply? The labor-portfolio problem (Section 2) and health investment problem (Section 3) illustrated two key structures: separable utility functions and invertible transitions. This section formalizes these requirements and provides practical guidance for decomposing new problems. The

[31] Both ENGINE and the baseline use the HARK toolkit's (Carroll et al., 2018) Numba-compiled interpolation routines, so the comparison is between equivalently compiled codebases.

examples demonstrated specific instances; we now characterize the general conditions that make sequential decomposition with EGM inversion possible.

5.1. Splitting the problem into subproblems

Decomposition begins by counting independent control variables. A problem with n control variables typically decomposes into n subproblems. One must take care not to double-count: the consumption-savings choice ($c+a=m$) represents one decision, not two. Similarly, labor-leisure is a single choice despite involving two variables.

Each control variable requires an enabling structure that permits EGM inversion. Two mathematical structures suffice: separable, differentiable, and invertible utility functions (as in the leisure utility of Section 2); or differentiable and invertible transition functions (as in the health production function of Section 3). Match each control variable to its enabling structure. The labor-portfolio example features additive utility separability: leisure utility enables the labor-leisure EGM step, consumption utility enables the consumption-savings EGM step. When no structure applies (as in the portfolio choice stage), standard optimization is required.

The ordering of subproblems should shed state variables early. Poor sequencing propagates unnecessary state variables through later stages. In the consumption-leisure-portfolio problem, placing labor-leisure first resolves the wage rate before the consumption stage, keeping that subproblem one-dimensional. Choosing consumption first would force the labor decision to track both bank balances and wages, doubling its dimensionality.[32]

5.2. Formal conditions

We now abstract from specific models to general conditions. The notation shifts to uppercase calligraphic functions (V, W, U, T, G) for generic value, continuation, utility, transition, and shock functions, with generic state variables x, y, s in place of the model-specific variables: consumption c, assets a, and cash-on-hand m of preceding sections.

Consider a utility function of the form

$$\mathrm{U}(\mathbf{a}) = \mathrm{u}_{-i}(\mathbf{a}^{-i}) + \mathrm{u}_i(a^i), \tag{40}$$

where $\mathbf{a} = (a^1, \ldots, a^n)$ is the vector of control variables, a^i is the i-th control variable, and $\mathbf{a}^{-i}$ is the vector of all control variables except the i-th one. This utility function is separable in the control variable corresponding to index i.

$$\begin{aligned} \mathrm{V}(x,s) = \max_{\mathbf{a}\in\Gamma(x,s)} \; & \mathrm{U}(\mathbf{a}) + \beta\mathbb{E}\left[\mathrm{V}_{+1}(x',s')|y,s\right] \\ \text{s.t.} \\ y &= \mathrm{T}(x,\mathbf{a}) \\ x' &= \mathrm{G}(y,s) \end{aligned} \tag{41}$$

We collect the discounted continuation value into a single function:

$$\mathrm{W}(y,s) = \beta\mathbb{E}\left[\mathrm{V}_{+1}(\mathrm{G}(y,s),s')|y,s\right]. \tag{42}$$

Substituting, the problem reduces to:

[32] To see this concretely, the consumption subproblem would become two-dimensional: $v^0(b,\theta) = \max_c \mathrm{u}(c) + v^1(b',\theta)$ subject to $b' = b - c \geq -\theta$, requiring interpolation on a (b,θ) grid instead of just b. The labor-leisure subproblem would then have the additional constraint: $v^1(b',\theta) = \max_z h(z) + v^2(a)$ subject to $0 \leq z \leq 1$ and $a = b' + \theta(1-z) \geq 0$. The poor ordering forces us to carry the wage state through both stages, doubling the dimensionality of the first stage.

$$\begin{aligned} \mathrm{V}(x,s) = & \max_{\mathbf{a}\in\Gamma(x,s)} \mathrm{U}(\mathbf{a}) + \mathrm{W}(y,s) \\ & \text{s.t.} \\ & y = \mathrm{T}(x,\mathbf{a}) \end{aligned} \tag{43}$$

When post-decision states are multidimensional, with components y^j for $j = 1,\ldots,n$ and transition functions T^j mapping controls to each component (here superscripts on y and T index components, not derivatives), the first-order condition (FOC) with respect to control a^i (assuming an interior solution) becomes

$$\frac{\partial \mathrm{U}(\mathbf{a})}{\partial a^i} + \sum_{j=1}^{n} \frac{\partial \mathrm{W}(y,s)}{\partial y^j} \frac{\partial \mathrm{T}^j(x,\mathbf{a})}{\partial a^i} = 0. \tag{44}$$

The requirement $\frac{\partial \mathrm{T}^j(x,\mathbf{a})}{\partial a^i} = 0$ for $j \neq i$ (separability in the transition) is sufficient to solve for a^i independently. This condition ensures that control a^i affects only the i-th post-decision state, decoupling the FOC from other transition equations. When transition separability fails, the first-order conditions couple across stages, requiring simultaneous solution of all controls — exactly the computational burden that sequential EGM is designed to avoid.[33]

Under transition separability, the first-order condition for stage i simplifies to:

$$\frac{\partial \mathrm{U}(\mathbf{a})}{\partial a^i} + \frac{\partial \mathrm{W}(y,s)}{\partial y^i} \frac{\partial \mathrm{T}^i(x,\mathbf{a})}{\partial a^i} = 0. \tag{45}$$

Each stage's optimality condition depends only on its own control and the continuation value, not on controls chosen at other stages.

5.2.1. Separable utility

Once the problem is split into subproblems, the Endogenous Grid Method applies to each applicable subproblem while iterating backwards from the terminal period. The EGM step applies when there is a separable, differentiable and invertible utility function in the subproblem or when there is a differentiable and invertible transition in the subproblem.

To see when the method fails, consider non-separable CES preferences $\mathrm{U}(c,z) = \frac{(c^\theta z^{1-\theta})^{1-\rho}}{1-\rho}$. The first-order condition (FOC) for consumption involves leisure: $\frac{\partial \mathrm{U}}{\partial c} = \theta c^{\theta(1-\rho)-1} z^{(1-\theta)(1-\rho)}$. Inverting for c requires knowing z, and the symmetric condition for z requires knowing c. The FOCs are coupled, so neither can be solved in isolation by EGM inversion. Such problems require joint root-finding (NEGM) or upper envelope techniques (G2EGM).

Consider a generic subproblem with a differentiable and invertible utility function:

$$\begin{aligned} \mathrm{V}(x) = & \max_{a\in\Gamma(x)} \mathrm{U}(x,a) + \mathrm{W}(y) \\ & \text{s.t.} \\ & y = \mathrm{T}(x,a) \end{aligned} \tag{46}$$

where $\mathrm{W}(y) = \beta\mathbb{E}[\mathrm{V}_{+1}(y)]$ is the continuation value. For an interior solution, the first-order condition is

$$\frac{\partial \mathrm{U}(x,a)}{\partial a} + \mathrm{W}'(y) \frac{\partial \mathrm{T}(x,a)}{\partial a} = 0. \tag{47}$$

[33] Separability is sufficient but not strictly necessary. Alternative structures may also permit independent solution: for example, a triangular system where T^j depends on $a^1,\ldots,a^j$ but not on $a^{j+1},\ldots,a^n$ allows sequential substitution, solving for a^1 first, then a^2 given a^1, and so on. However, separability is the most common and practically verifiable condition.

When corner solutions occur (e.g., a at constraint boundaries), the unconstrained optimum from inverting the first-order condition must be projected onto the feasible set, as demonstrated in Section 2 for the leisure choice.

The question is when this first-order condition can be inverted to yield the optimal control directly. Proposition 5.1 gives the standard case: separable utility with a transition linear in the control, which covers most consumption-savings applications.

Proposition 5.1 (Separable Utility)**.** *Suppose* $\mathrm{U}(x, a) = \mathrm{u}_{-}(x) + \mathrm{u}(a)$ *is additively separable in the state* x *and control* a*, so the marginal utility* $\phi(a) \equiv \mathrm{u}'(a)$ *depends only on the control. For interior solutions where* ϕ *is strictly monotone so that* ϕ^{-1} *exists and is single-valued,*[34] *the first-order condition can be inverted. When the transition is linear in the control,* $\frac{\partial \mathrm{T}}{\partial a} = k$ *for some* $k \neq 0$ *that does not depend on the control variable (though it may depend on state variables known when solving the subproblem), the optimal control on an exogenous grid of post-decision states* $[\mathrm{y}]$ *is*

$$\mathfrak{a}([\mathrm{y}]) = \phi^{-1}\left(-k \cdot \mathrm{W}'([\mathrm{y}])\right). \tag{48}$$

Separability is essential: since x *is the endogenous variable being recovered,* ϕ *must not depend on* x*. The pre-decision state then follows from inverting the transition:* $\mathfrak{x}([\mathrm{y}]) = \mathrm{T}^{-1}([\mathrm{y}], \mathfrak{a}([\mathrm{y}]))$*. Strict concavity of* u *ensures uniqueness.*

The linearity assumption $\frac{\partial \mathrm{T}}{\partial a} = k$ is satisfied by standard budget constraints appearing in consumption-savings models. For example, $a = m - c$ implies $\frac{\partial a}{\partial c} = -1$, so $k = -1$. Similarly, the labor income constraint $m = b + \theta \cdot \ell$ has $\frac{\partial m}{\partial \ell} = \theta$, which varies with the wage shock but is constant with respect to the control ℓ. Linearity in the control allows the first-order condition to be inverted analytically, which is the computational step that replaces root-finding. State-dependent coefficients (like θ) are permissible provided they are known when solving the subproblem.

Using an exogenous grid of the post-decision state y, the optimal decision rule a follows at each point on the grid. This inversion replaces a numerical root-find with a single function evaluation, the core speed advantage of EGM. The monotonicity requirement ensures that the mapping from the post-decision state to the control is well-defined, while concavity guarantees uniqueness of the optimal decision at each grid point.

Proposition 5.1 covers the standard case where utility is separable. But some subproblems have no within-period utility at all, only transitions affecting multiple post-decision states. Health investment in Section 3 is such a case: there is no immediate utility from health spending, only effects on future health and wealth. Proposition 5.2 shows that EGM still applies when transitions are invertible, even without separable utility. Consider a problem with two endogenous state variables and two post-decision states:

$$\begin{aligned} \mathrm{V}(x_1, x_2, s) = & \max_{a \in \Gamma(x_1, x_2, s)} \mathrm{W}(y_1, y_2, s) \\ & \text{s.t.} \\ & y_1 = \mathrm{T}_1(x_1, a) \\ & y_2 = \mathrm{T}_2(x_2, a) \end{aligned} \tag{49}$$

where the continuation value is

$$\mathrm{W}(y_1, y_2, s) = \beta \mathbb{E}\left[\mathrm{V}_{+1}(\mathrm{G}_1(y_1, s), \mathrm{G}_2(y_2, s), s') | y_1, y_2, s\right]. \tag{50}$$

The first-order condition becomes

$$\frac{\partial \mathrm{W}(y_1, y_2, s)}{\partial y_1} \cdot \frac{\partial \mathrm{T}_1(x_1, a)}{\partial a} + \frac{\partial \mathrm{W}(y_1, y_2, s)}{\partial y_2} \cdot \frac{\partial \mathrm{T}_2(x_2, a)}{\partial a} = 0. \tag{51}$$

[34] Strict monotonicity of the marginal utility ensures that the inverse function is well-defined and single-valued. This condition is satisfied by standard utility functions like CRRA utility where $\mathrm{u}'(c) = c^{-\rho}$ is strictly decreasing in consumption.

Proposition 5.2 (Invertible Transitions)**.** *Suppose both transitions are additively separable in the control:*

$$\mathrm{T}_1(x_1, a) = f_1(x_1) + k \cdot a, \quad \mathrm{T}_2(x_2, a) = f_2(x_2) + \mathrm{g}(a) \tag{52}$$

where $k \neq 0$ is constant, f_1 and f_2 are invertible, g' is strictly monotone, and $\partial \mathrm{W}/\partial y_2 \neq 0$ (the marginal value of the second post-decision state is non-zero, since otherwise the first-order condition degenerates and does not pin down the control). Then the first-order condition yields

$$a = \mathrm{g}'^{-1}\left(-k \cdot \frac{\partial \mathrm{W}(y_1, y_2, s)/\partial y_1}{\partial \mathrm{W}(y_1, y_2, s)/\partial y_2}\right), \tag{53}$$

where strict monotonicity of g' ensures existence and uniqueness of the inverse.[35]

Additive separability in both transitions allows the derivative with respect to a to not depend on the state variables x_1 or x_2, which have not yet been recovered when solving the first-order condition on the exogenous grid of post-decision states. Once a is obtained from the inversion, the pre-decision states follow via $x_1 = f_1^{-1}(y_1 - k \cdot a)$ and $x_2 = f_2^{-1}(y_2 - \mathrm{g}(a))$. The formulation where one state variable enters linearly (e.g., $\mathrm{T}_1 = x_1 - a$ with $f_1(x_1) = x_1$ and $k = -1$) is a common special case.

This additive separability defines what Iskhakov (2015) calls "triangular" structure in transitions. The triangular property enables the multidimensional EGM to proceed without root-finding: when the FOC system has triangular form, each control can be recovered sequentially by substitution rather than by solving a coupled nonlinear system. Their approach solves problems where the entire multidimensional structure is triangular, enabling simultaneous solution of all choices with applications to wealth accumulation, health capital, and portfolio problems. Together, the two propositions characterize the practical scope of EGM^n: sequential EGM applies whenever each subproblem satisfies either separable utility (Proposition 1) or invertible transitions (Proposition 2), and the decomposition ordering can be chosen to satisfy these conditions stage by stage.

6. Conclusion

EGM^n applies when the problem satisfies a few structural conditions. Each condition has a natural economic interpretation: separability reflects that the marginal utility of one choice does not depend on the level of another; triangularity reflects that each decision's consequences flow forward in time without looping back.

First, the utility function must be additively separable across controls, or the transition functions linking controls to post-decision states must be differentiable and invertible. Second, choices must be continuous with smooth resulting policy functions; discrete choices require G2EGM or the Discrete-Continuous EGM (DCEGM). Third, the first-order conditions for different controls must decouple; coupled FOCs require NEGM or G2EGM.

When EGM^n applies, ENGINE handles interpolation on the resulting curvilinear grids with no preprocessing and natural parallelization. The ordering of subproblems should minimize the information set passed forward: the labor-portfolio example sequences (leisure-labor, consumption-savings, portfolio) so each stage sheds state variables before the next. When subproblems are independent, as with consumption and health investment affecting different state dimensions, ordering is immaterial.

The two examples confirm that economically simultaneous decisions need not be computationally joint: the labor-portfolio problem eliminates optimization at two of three stages, and the health model avoids root-finding entirely. When separable structure permits decomposition, most stages admit EGM inversion, passing efficiency gains forward through the stage sequence. ENGINE (ENdogenous Grid INterpolation and Extrapolation) addresses the interpolation challenge by exploiting the index correspondence inherited from exogenous grids, achieving 2-3x speedups over existing curvilinear interpolation.

A multistage problem can mix stages satisfying separable utility (Proposition 5.1) with stages satisfying triangular transitions (Proposition 5.2), and even include stages solved by standard optimization, expanding

[35] Strict monotonicity of g' ensures the inverse is well-defined. In the health investment example, the production function $\mathrm{g}(n) = \frac{\gamma}{\alpha} n^{\alpha}$ satisfies this property with $\mathrm{g}'(n) = \gamma n^{\alpha - 1}$ strictly decreasing for $\alpha < 1$.

applicability beyond problems that require joint inversion of coupled first-order conditions across all decisions. The same smoothness conditions that enable EGM inversion also determine ENGINE's efficiency: strict concavity bounds the Lipschitz constant of the EGM mapping, which in turn bounds the IPO constant α that controls interpolation complexity. Problems with smooth value functions typically have small α, enabling ENGINE's $O(\log J \cdot \log K)$ complexity (with empirical performance approaching $O(\log J + \log K)$ for smooth problems).

Relative to NEGM (Druedahl, 2021), which nests root-finding or optimization within EGM loops, EGM^n avoids all numerical solution steps when its structural assumptions hold; and relative to triangular EGM (Iskhakov, 2015), which requires triangularity of the entire FOC system, EGM^n requires only sequential invertibility at each stage, permitting mixed structures. G2EGM (Druedahl and Jørgensen, 2017) handles discrete choices and non-convexities through upper envelope techniques, a domain where EGM^n does not apply. ENGINE complements this by providing efficient interpolation with no preprocessing, simple implementation, and efficient memory access patterns.

6.0.1. Limitations

Problems with non-separable utility (e.g., consumption-leisure complementarity where FOCs are coupled) or correlated transitions require alternative methods such as G2EGM or NEGM. ENGINE's regularity conditions (fold-free, IPM, IPO) follow automatically from EGM's standard regularity assumptions (strict concavity, smoothness), so they impose no additional burden on applicable problems. However, problems with discrete choices or severe non-convexities violate EGM's assumptions and hence ENGINE's conditions, requiring alternative methods; occasionally binding constraints typically preserve monotonicity and remain tractable. ENGINE's multilinear interpolation faces a curse of dimensionality: each interpolation requires 2^d cell vertices, making problems with $d \geq 5$ state variables computationally challenging regardless of the interpolation method.

Several extensions merit future investigation. Combining EGM^n stages with G2EGM's upper envelope techniques could handle problems where some decisions are separable while others involve discrete choices. Adaptive grid refinement could concentrate computational effort in regions where policy functions exhibit rapid variation. The sequential structure also facilitates parallelization: independent EGM inversions across grid points can be distributed across processors with minimal coordination cost.

Applications that become newly tractable include high-frequency portfolio rebalancing with transaction costs and multiple asset classes, multiperiod health insurance choices with endogenous health states and moral hazard, and durable goods decisions with both extensive and intensive margins. The efficiency gains from avoiding triangulation setup costs and exploiting efficient memory access patterns make structural estimation via simulated method of moments or maximum likelihood more practical, particularly when the model must be solved thousands of times during optimization. Some models are simply waiting for a faster solver. Models of consumption response to fiscal stimulus (Kaplan and Violante, 2014), wealth accumulation over the life cycle (Cagetti, 2003), and household balance sheet dynamics (Fagereng, Holm and Natvik, 2021) all feature multiple interacting decisions that can benefit from sequential decomposition when their structure permits.

The EGM^n and ENGINE algorithms are available as open-source software, allowing researchers to apply sequential decomposition and index-based interpolation to their own multidimensional lifecycle models.[36]

6.1. Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the author used Claude (Anthropic) in order to edit and proofread the manuscript and to assist with coding. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

CRediT authorship contribution statement

Alan Lujan: Conceptualization, Methodology, Software, Writing – original draft, Visualization.

[36]The algorithms are implemented in the `ConsSequentialEGMSolver` and `CurvilinearInterp` classes within the Econ-ARK HARK toolkit (Carroll et al., 2018).

References

Aiyagari, S.R., 1994. Uninsured Idiosyncratic Risk and Aggregate Saving. The Quarterly Journal of Economics 109, 659–684. doi:10.2307/2118417.

Arellano, C., Maliar, L., Maliar, S., Tsyrennikov, V., 2016. Envelope condition method with an application to default risk models. Journal of Economic Dynamics and Control 69, 436–459. doi:10.1016/j.jedc.2016.05.016.

Azinovic, M., Gaegauf, L., Scheidegger, S., 2022. Deep equilibrium nets. International Economic Review 63, 1471–1525. doi:10.1111/iere.12575.

Barillas, F., Fernández-Villaverde, J., 2007. A generalization of the endogenous grid method. Journal of Economic Dynamics and Control 31, 2698–2712. doi:10.1016/j.jedc.2006.08.005.

Bayer, C., Luetticke, R., Weiss, M., Winkelmann, Y., 2026. An Endogenous Gridpoint Method for Distributional Dynamics. Journal of Monetary Economics 158, 103895. doi:10.1016/j.jmoneco.2026.103895.

Bellman, R., 1957. Dynamic Programming. Princeton University Press, Princeton, NJ.

Bodie, Z., Merton, R.C., Samuelson, W.F., 1992. Labor supply flexibility and portfolio choice in a life cycle model. Journal of Economic Dynamics and Control 16, 427–449. doi:10.1016/0165-1889(92)90044-F.

de Boor, C., 2001. A Practical Guide to Splines. Revised ed., Springer, New York. doi:10.1007/978-1-4612-6333-3.

Cagetti, M., 2003. Wealth accumulation over the life cycle and precautionary savings. Journal of Business & Economic Statistics 21, 339–353. doi:10.1198/073500103288619007.

Carroll, C., Kaufman, A., Kazil, J., Palmer, N., White, M., 2018. The econ-ARK and HARK: Open source tools for computational economics, in: Akici, F., Lippa, D., Niederhut, D., Pacer, M. (Eds.), Proceedings of the Python in Science Conference, SciPy, Austin, Texas. pp. 25–30. doi:10.25080/majora-4af1f417-004.

Carroll, C.D., 2006. The method of endogenous gridpoints for solving dynamic stochastic optimization problems. Economics Letters 91, 312–320. doi:10.1016/j.econlet.2005.09.013.

Carroll, C.D., 2009. Precautionary saving and the marginal propensity to consume out of permanent income. Journal of Monetary Economics 56, 780–790. doi:10.1016/j.jmoneco.2009.06.016.

Clausen, A., Strub, C., 2020. Reverse calculus and nested optimization. Journal of Economic Theory 187, 105019. doi:10.1016/j.jet.2020.105019.

De Nardi, M., 2004. Wealth Inequality and Intergenerational Links. The Review of Economic Studies 71, 743–768. doi:10.1111/j.1467-937X.2004.00302.x.

Druedahl, J., 2021. A Guide on Solving Non-convex Consumption-Saving Models. Computational Economics 58, 747–775. doi:10.1007/s10614-020-10045-x.

Druedahl, J., Jørgensen, T.H., 2017. A general endogenous grid method for multi-dimensional models with non-convexities and constraints. Journal of Economic Dynamics and Control 74, 87–107. doi:10.1016/j.jedc.2016.11.005.

Fagereng, A., Holm, M.B., Natvik, G.J., 2021. Mpc Heterogeneity and Household Balance Sheets. American Economic Journal: Macroeconomics 13, 1–54. URL: https://www.aeaweb.org/articles?id=10.1257/mac.20190211, doi:10.1257/mac.20190211.

Fella, G., 2014. A generalized endogenous grid method for non-smooth and non-concave problems. Review of Economic Dynamics 17, 329–344. doi:10.1016/j.red.2013.07.001.

Hallengreen, A., Jørgensen, T.H., Olesen, A.M., 2024. The Endogenous Grid Method without Analytical Inverse Marginal Utility. Working Paper 24-11. University of Copenhagen, Center for Economic Behavior. URL: https://www.econ.ku.dk/cebi/publikationer/working-papers/CEBI_WP_11-24.pdf, doi:10.2139/ssrn.4830404.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., Del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585, 357–362. doi:10.1038/s41586-020-2649-2.

Hintermaier, T., Koeniger, W., 2010. The method of endogenous gridpoints with occasionally binding constraints among endogenous variables. Journal of Economic Dynamics and Control 34, 2074–2088. doi:10.1016/j.jedc.2010.05.002.

Huggett, M., 1993. The Risk-Free Rate in Heterogeneous-Agent Incomplete-Insurance Economies. Journal of Economic Dynamics and Control 17, 953–969. doi:10.1016/0165-1889(93)90024-M.

Iskhakov, F., 2015. Multidimensional endogenous gridpoint method: Solving triangular dynamic stochastic optimization problems without root-finding operations. Economics Letters 135, 72–76. doi:10.1016/j.econlet.2015.07.033. corrigendum: *Economics Letters* 150 (2017) 26, \doi10.1016/j.econlet.2016.11.002.

Iskhakov, F., Jørgensen, T.H., Rust, J., Schjerning, B., 2017. The endogenous grid method for discrete-continuous dynamic choice models with (or without) taste shocks. Quantitative Economics 8, 317–365. doi:10.3982/QE643.

Jørgensen, T.H., 2013. Structural estimation of continuous choice models: Evaluating the EGM and MPEC. Economics Letters 119, 287–290. doi:10.1016/j.econlet.2013.02.027.

Kaplan, G., Violante, G.L., 2014. A Model of the Consumption Response to Fiscal Stimulus Payments. Econometrica 82, 1199–1239. doi:10.3982/ECTA10528.

Knupp, P.M., 1999. Winslow Smoothing on Two-Dimensional Unstructured Meshes. Engineering with Computers 15, 263–268. doi:10.1007/s003660050021. see also Knupp and Steinberg (1993) *Fundamentals of Grid Generation* for comprehensive treatment of Jacobian-based grid validity conditions.

Krueger, D., Mitman, K., Perri, F., 2016. Macroeconomics and household heterogeneity. Elsevier. volume 2. pp. 843–921. doi:10.1016/bs.hesmac.2016.04.003.

Krusell, P., Smith, A.A., 1998. Income and Wealth Heterogeneity in the Macroeconomy. Journal of Political Economy 106, 867–896. doi:10.1086/250034.

Lam, S.K., Pitrou, A., Seibert, S., 2015. Numba: a LLVM-based Python JIT compiler, in: Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC, Association for Computing Machinery, Austin, Texas. pp. 1–6. doi:10.1145/2833157.2833162.

Ludwig, A., Schön, M., 2018. Endogenous Grids in Higher Dimensions: Delaunay Interpolation and Hybrid Methods. Computational Economics 51, 463–492. doi:10.1007/s10614-016-9611-2.

Lujan, A., 2026. The Endogenous Grid Method for Epstein-Zin Preferences. doi:10.48550/arXiv.2601.04438.

Maliar, L., Maliar, S., 2013. Envelope condition method versus endogenous grid method for solving dynamic programming problems. Economics Letters 120, 262–266. doi:10.1016/j.econlet.2013.04.031.

Maliar, L., Maliar, S., 2014. Chapter 7 - Numerical Methods for Large-Scale Dynamic Economic Models. Elsevier. volume 3 of *Handbook of Computational Economics*. pp. 325–477. doi:10.1016/B978-0-444-52980-0.00007-4.

Maliar, L., Maliar, S., Winant, P., 2021. Deep learning for solving dynamic economic models. Journal of monetary economics 122, 76–101. doi:10.1016/j.jmoneco.2021.07.004.

Mendoza, E.G., Villalvazo, S., 2020. Fipit: A simple, fast global method for solving models with two endogenous states & occasionally binding constraints. Review of Economic Dynamics 37, 81–102. doi:10.1016/j.red.2020.01.001.

Mertens, K., Ravn, M.O., 2011. Understanding the aggregate effects of anticipated and unanticipated tax policy shocks. Review of Economic Dynamics 14, 27–54. doi:10.1016/j.red.2010.07.004.

Powell, W.B., 2011. Approximate Dynamic Programming: Solving the Curses of Dimensionality. Wiley Series in Probability and Statistics. 2nd ed., Wiley, Hoboken, NJ. doi:10.1002/9781118029176.

Scheidegger, S., Bilionis, I., 2019. Machine learning for high-dimensional dynamic stochastic economies. Journal of Computational Science 33, 68–82. doi:10.1016/j.jocs.2019.03.004.

White, M.N., 2015. The method of endogenous gridpoints in theory and practice. Journal of Economic Dynamics and Control 60, 26–41. doi:10.1016/j.jedc.2015.08.001.

A. Appendix: Proofs

Notation. In the proofs below, subscripts on value functions denote partial derivatives (e.g., $v^1_{lH} \equiv \partial^2 v^1/\partial l \partial H$). Superscripts serve two roles: on stage-numbered functions (v^1, v^2), superscripts are stage labels and subscripts always denote partials; on w, which carries no stage number, superscripts denote partial derivatives directly ($\mathrm{w}^a \equiv \partial \mathrm{w}/\partial a$, $\mathrm{w}^{aH} \equiv \partial^2 \mathrm{w}/\partial a \partial H$). This follows the convention in the main text (Section 3).

A.1. Proof of Grid Homeomorphism Theorem

Proof of Grid Homeomorphism Theorem. The proof establishes that the bilinear extension $\tilde{\Phi} : [1, J] \times [1, K] \to \mathbb{R}^2$ is a homeomorphism by showing it is a continuous injection from a compact set to a Hausdorff space.

Step 1 (Cell-level injectivity). Within each cell $[j, j+1] \times [k, k+1]$, the bilinear map in parametric coordinates $(s, t) \in [0, 1]^2$ takes the form $x(s,t) = (1-s)(1-t)x_{jk} + s(1-t)x_{j+1,k} + (1-s)t\, x_{j,k+1} + st\, x_{j+1,k+1}$, and similarly for y. The partial derivative $\partial x/\partial s = (1-t)(x_{j+1,k} - x_{jk}) + t(x_{j+1,k+1} - x_{j,k+1})$ is a convex combination of the row increments $x_{j+1,k} - x_{jk}$ and $x_{j+1,k+1} - x_{j,k+1}$, both strictly positive by IPM row monotonicity. Hence $\partial x/\partial s > 0$ throughout the cell. Similarly, $\partial y/\partial t > 0$ by column monotonicity.

The Jacobian determinant $\mathbf{J}(s,t) = \frac{\partial x}{\partial s}\frac{\partial y}{\partial t} - \frac{\partial x}{\partial t}\frac{\partial y}{\partial s}$ is bilinear in (s, t), so it attains its minimum at a corner of $[0, 1]^2$. Since all four corner Jacobians are positive by the fold-free condition (Definition 4.1), $\mathbf{J}(s, t) > 0$ throughout the cell. We establish cell-level injectivity in two steps. First, the boundary of the cell maps injectively: each edge is the image of an affine map with strictly positive slope (since $\partial x/\partial s > 0$ along horizontal edges and $\partial y/\partial t > 0$ along vertical edges), so the boundary image is a simple closed quadrilateral by IPM. Second, an orientation-preserving bilinear map that is injective on the boundary with $\mathbf{J} > 0$ throughout is injective on the entire cell (Knupp, 1999).

Step 2 (Global injectivity). We show that distinct points in $[1, J] \times [1, K]$ map to distinct points in $\mathbb{R}^2$.

Points in the same cell: Injective by Step 1.

Points in non-adjacent cells: For any fixed x-coordinate x_0 in the domain, the interpolated y-value on row boundary k at $x = x_0$ is strictly increasing in k: since $\partial y/\partial t > 0$ in every cell, the y-value along any vertical line through the grid increases monotonically across row boundaries. Therefore, two points in distinct row bands $[k, k+1]$ and $[k', k'+1]$ with $|k - k'| \geq 2$ sharing the same image x-coordinate must have distinct image y-coordinates. Similarly, cells in different column bands with $|j - j'| \geq 2$ have strictly separated x-ranges by the symmetric argument using $\partial x/\partial s > 0$.

Points in adjacent cells: Adjacent cells sharing an edge have identical bilinear maps along that shared edge (both cells use the same vertex coordinates). The interiors of adjacent cells cannot overlap: for cells sharing a vertical edge at column $j+1$, the left cell satisfies $x(1,t) = (1-t)\,x_{j+1,k} + t\,x_{j+1,k+1}$ while the right cell satisfies $x(0,t) = (1-t)\,x_{j+1,k} + t\,x_{j+1,k+1}$, the boundary values inherited from the shared column vertices. Since $\partial x/\partial s > 0$ in both cells, the left cell's interior lies strictly to the left of the shared edge and the right cell's interior lies strictly to the right.

For cells sharing a horizontal edge at row $k+1$, a symmetric argument using $\partial y/\partial t > 0$ separates the interiors vertically. For diagonally adjacent cells (sharing only a vertex at $(j+1, k+1)$), the two cells differ in both row and column index. Their x-ranges may overlap, but only at the single shared vertex: the left cell's x-values at row $k+1$ end at $x_{j+1,k+1}$, while the right cell's x-values at row k start at $x_{j+1,k}$. Since $\partial y/\partial t > 0$ separates their y-ranges at any shared x-coordinate (the lower cell's maximum y at the boundary equals the upper cell's minimum y), the interiors are disjoint.

Together, $\tilde{\Phi}$ is a continuous injection from the compact set $[1, J] \times [1, K]$ to $\mathbb{R}^2$. Since continuous bijections from compact spaces to Hausdorff spaces are homeomorphisms, the grid mapping is a homeomorphism onto its image.

□

A.2. Proof of EGM Regularity Theorem

Proof of . **EGM Regularity Theorem**

Fold-free (1D case): For the consumption-savings EGM, the inversion $\mathfrak{m}(a) = a + \mathfrak{c}(a)$ where $\mathfrak{c}(a) = (u')^{-1}((v^2)'(a))$ is a composition of smooth functions when u and v^2 are twice differentiable. By the inverse function theorem, the mapping is locally diffeomorphic wherever its Jacobian is non-singular. Strict concavity of v^2 ensures $(v^2)'' < 0$, and strict concavity of u ensures $u'' < 0$, so $\partial\mathfrak{c}/\partial a = \frac{(v^2)''(a)}{u''(\mathfrak{c}(a))} > 0$ (both numerator and denominator are negative, yielding a positive ratio). Thus $\partial\mathfrak{m}/\partial a = 1 + \partial\mathfrak{c}/\partial a > 1 > 0$, establishing positive Jacobian.

Fold-free (2D case): For the health investment EGM, the mapping $\Phi : (l, H) \mapsto (m, h)$ is defined by $m = l + \mathfrak{n}(l, H)$ and $h = H - \mathrm{g}(\mathfrak{n}(l, H))$. The Jacobian matrix is:

$$\mathbf{J} = \begin{pmatrix} 1 + \partial\mathfrak{n}/\partial l & \partial\mathfrak{n}/\partial H \\ -\mathrm{g}' \cdot \partial\mathfrak{n}/\partial l & 1 - \mathrm{g}' \cdot \partial\mathfrak{n}/\partial H \end{pmatrix} \tag{54}$$

The determinant is $\det(\mathbf{J}) = (1 + \partial\mathfrak{n}/\partial l)(1 - \mathrm{g}' \cdot \partial\mathfrak{n}/\partial H) + \mathrm{g}' \cdot \partial\mathfrak{n}/\partial H \cdot \partial\mathfrak{n}/\partial l = 1 + \partial\mathfrak{n}/\partial l - \mathrm{g}' \cdot \partial\mathfrak{n}/\partial H$.

To establish $\det(\mathbf{J}) > 0$, we derive the partial derivatives of $\mathfrak{n}$ explicitly. The first-order condition $\mathrm{g}'(\mathfrak{n}) = v^1_l / v^1_H$ implicitly defines $\mathfrak{n}(l, H)$. Differentiating both sides with respect to l:

$$\mathrm{g}'' \cdot \frac{\partial\mathfrak{n}}{\partial l} = \frac{v^1_{ll} \cdot v^1_H - v^1_l \cdot v^1_{lH}}{(v^1_H)^2} \tag{55}$$

Solving for $\partial\mathfrak{n}/\partial l$:

$$\frac{\partial\mathfrak{n}}{\partial l} = \frac{v^1_{ll} \cdot v^1_H - v^1_l \cdot v^1_{lH}}{\mathrm{g}'' \cdot (v^1_H)^2} \tag{56}$$

Similarly, differentiating with respect to H:

$$\frac{\partial\mathfrak{n}}{\partial H} = \frac{v^1_{lH} \cdot v^1_H - v^1_l \cdot v^1_{HH}}{\mathrm{g}'' \cdot (v^1_H)^2} \tag{57}$$

Joint strict concavity of v^1 ensures the Hessian is negative definite: $v^1_{ll} < 0$, $v^1_{HH} < 0$, and $v^1_{ll} v^1_{HH} - (v^1_{lH})^2 > 0$. Since $\mathrm{g}'' < 0$ (concave production) and $v^1_H > 0$, we can sign the derivatives. The condition $v^1_H > 0$ holds because $v^1_H = \mathrm{w}^H$, which is positive: higher post-investment health H increases both the survival probability $\mathcal{S} = 1 - \mathsf{D}/(1+H)$ and labor income $y_{t+1} = \omega_{t+1} H$, both of which raise continuation value.

We assume $v^1_{lH} \geq 0$, meaning wealth and health are complements or independent in the continuation value. This is imposed as an assumption for the health model. Economic intuition suggests health capital

raises the marginal value of wealth (through higher future income and survival probability), but a formal derivation from the model's primitives would require differentiating through the expectation and survival probability, which we do not pursue here. If this complementarity condition fails ($v^1_{lH} < 0$, so that health and wealth are substitutes), $\partial\mathfrak{n}/\partial H$ can become positive, causing column monotonicity to fail and the endogenous grid to fold, which violates IPM.[37]

For $\partial\mathfrak{n}/\partial l$: the numerator $v^1_{ll} \cdot v^1_H - v^1_l \cdot v^1_{lH}$ has $v^1_{ll} < 0$ and $v^1_H > 0$, giving a negative first term; with $v^1_l > 0$ and $v^1_{lH} \geq 0$, the second term $-v^1_l \cdot v^1_{lH}$ is non-positive, so the numerator is negative. Dividing by $\mathrm{g}'' < 0$ yields $\partial\mathfrak{n}/\partial l > 0$, hence $1 + \partial\mathfrak{n}/\partial l > 1 > 0$.

For $\partial\mathfrak{n}/\partial H$: the numerator is $v^1_{lH} \cdot v^1_H - v^1_l \cdot v^1_{HH}$. With $v^1_{HH} < 0$ and $v^1_l > 0$, the second term $-v^1_l \cdot v^1_{HH} > 0$. If $v^1_{lH} \geq 0$, the first term is also non-negative, so the numerator is positive. Dividing by $\mathrm{g}'' < 0$ gives $\partial\mathfrak{n}/\partial H \leq 0$. Since $\mathrm{g}' > 0$, this yields $-\mathrm{g}' \cdot \partial\mathfrak{n}/\partial H \geq 0$, contributing non-negatively to $\det(\mathbf{J})$.

Combining: $\det(\mathbf{J}) = 1 + \partial\mathfrak{n}/\partial l - \mathrm{g}' \cdot \partial\mathfrak{n}/\partial H \geq 1 + 0 + 0 = 1 > 0$, where the first inequality uses $\partial\mathfrak{n}/\partial l > 0$ and $-\mathrm{g}' \cdot \partial\mathfrak{n}/\partial H \geq 0$. The argument extends to higher dimensions by induction on the number of EGM stages, with each stage's Jacobian inheriting positivity from the composition of positive-Jacobian mappings. The inductive step requires that each stage's transition satisfies the same regularity conditions (strict concavity, complementarity) established above for the 2D case; the precise conditions on cross-partials become more restrictive as dimensionality increases.

IPM: Since $\mathfrak{m}'(a) = 1 + \partial\mathfrak{c}/\partial a > 1$ (established in the fold-free step), $\mathfrak{m}$ is strictly increasing. For adjacent exogenous grid points $a_1 < a_2$, we have $\mathfrak{m}(a_1) < \mathfrak{m}(a_2)$, preserving coordinate ordering.

For the 2D case, IPM requires: (i) row monotonicity, where for fixed H the coordinate m increases with l; and (ii) column monotonicity, where for fixed l the coordinate h increases with H. From the fold-free analysis above, $\partial m/\partial l = 1 + \partial\mathfrak{n}/\partial l > 1 > 0$, establishing row monotonicity. For column monotonicity, $h = H - \mathrm{g}(\mathfrak{n})$, so $\partial h/\partial H = 1 - \mathrm{g}' \cdot \partial\mathfrak{n}/\partial H$. Under the complementarity assumption ($v^1_{lH} \geq 0$), we showed $\partial\mathfrak{n}/\partial H \leq 0$, so $-\mathrm{g}' \cdot \partial\mathfrak{n}/\partial H \geq 0$ (since $\mathrm{g}' > 0$), yielding $\partial h/\partial H \geq 1 > 0$.

IPO: Twice-differentiability of v implies $(v^2)'$ is Lipschitz continuous. From $\mathfrak{c} = (u')^{-1}((v^2)')$, the chain rule gives $L_{\mathfrak{c}} = \sup|d\mathfrak{c}/da| = \sup|(v^2)''|/|u''|$, which is bounded when both second derivatives are bounded away from zero. Thus the EGM mapping $\mathfrak{m}(a) = a + \mathfrak{c}(a)$ is Lipschitz with constant $L = 1 + L_{\mathfrak{c}}$. Since v^2 is C^2 on a compact domain (by assumption), $|(v^2)''|$ is bounded by the extreme value theorem. CRRA utility bounds $|u''|$ away from zero on compact domains, so L is finite. For grid spacing Δ, adjacent points map to points at most $L\Delta$ apart. If adjacent rows have spacing Δ_j and adjacent columns have spacing Δ_k, the IPO constant is $\alpha = L \cdot \Delta_k/\Delta_j$; for uniform grids, $\alpha = L$.

For the 2D health investment EGM, the relevant Lipschitz constant is the cross-directional $L_H = \sup|\partial m_{\text{end}}/\partial H|$. This quantity is bounded under the same smoothness conditions assumed for the value function, and we verify numerically that α remains small in all computed examples.

Strict concavity bounds the second derivatives of v, which bounds L and hence α. Small α (smooth value function) means brackets change slowly across rows, yielding efficient $O(\log J \cdot \log K)$ complexity with empirical performance approaching $O(\log J + \log K)$ for typical problems. Thus, the same smoothness conditions that make EGM applicable also make ENGINE efficient.

□

A.3. Proof of ENGINE Complexity

Proof of . **ENGINE Complexity**

ENGINE locates the containing cell for a query (x^*, y^*) via binary search on rows, with each probe requiring a bracket search along that row.

Binary search structure. ENGINE performs binary search over the K rows to find the row band $[k^*, k^* + 1]$ satisfying $\hat{y}_{k^*} \leq y^* < \hat{y}_{k^*+1}$. Each probe at row k computes the intermediate coordinate $\hat{y}_k$ by interpolating along that row at $x = x^*$. The comparison $\hat{y}_k \lessgtr y^*$ determines whether to search higher or lower rows. This requires $O(\log K)$ probes. Binary search is valid because the interpolated y-coordinate

[37] By the envelope theorem, $(v^1)^l = \mathrm{w}^a$ and $(v^1)^H = \mathrm{w}^H$, so $v^1_{lH} = \partial/\partial H[\mathrm{w}^a(a^*(l,H),H)] = \mathrm{w}^{aH} + \mathrm{w}^{aa}\partial a^*/\partial H$, where superscripts on w denote partial derivatives. The first term $\mathrm{w}^{aH} \geq 0$ reflects health raising the return to saving through higher future income. The second term involves $\mathrm{w}^{aa} < 0$ (concavity in assets) multiplied by $\partial a^*/\partial H$, the response of optimal savings to health. The sign of this cross-partial depends on the balance of these two terms; we impose $v^1_{lH} \geq 0$ as an assumption rather than deriving it from primitives.

at any fixed x is strictly increasing in k: by fold-free ($\partial y/\partial t > 0$ in every cell) and the continuity of the bilinear extension, the y-value along any vertical line through the grid increases monotonically across row boundaries.

Cost per probe. Each probe must find the j-bracket satisfying $x_{jk} \le x^* < x_{j+1,k}$. The first probe uses binary search over J column indices, costing $O(\log J)$. Subsequent probes exploit the IPO property: if the previous probe at row k' found bracket j', then the bracket j^* at row k satisfies $|j^* - j'| \le \alpha|k - k'|$. Binary search within the constrained window of size $O(\alpha \cdot |k - k'|)$ costs $O(\log(\alpha \cdot |k - k'|)) = O(\log J)$ in the worst case. With $O(\log K)$ probes, each costing $O(\log J)$, the total complexity is $O(\log J \cdot \log K)$.

Empirical performance for small α. When $\alpha = O(1)$, the bracket window at each probe is bounded by $O(\alpha \cdot |k - k'|)$. As binary search on rows converges, the bracket window at each probe narrows: at depth i, probed rows are $O(K/2^i)$ apart, so the bracket window has size $O(\alpha K/2^i)$. For typical smooth EGM problems where $\alpha \le 2$, the bracket windows shrink rapidly enough that empirical per-probe costs are $O(1)$ after the first probe, yielding observed $O(\log J + \log K)$ total complexity. A formal proof of this tighter bound would require a potential function argument showing that the total bracket search work across all probes telescopes; the $O(\log J \cdot \log K)$ bound is what the analysis rigorously establishes.

Regridding complexity. For $P \times Q$ query points, each query is independent, giving $O(PQ \log J \cdot \log K)$. When queries are sorted (as in meshgrid evaluation), the walking-pointer optimization processes each row's queries in order, reducing the bracket search to $O(1)$ amortized per query within each row.

□

A.4. ENGINE Algorithm

Algorithm 1 ENGINE Two-Pass Interpolation

Require: Grid $\{(x_{jk}, y_{jk}, f_{jk})\}$ for $j = 1, \ldots, J$ and $k = 1, \ldots, K$; query point (x^*, y^*)
Ensure: Interpolated value $f(x^*, y^*)$

Pass 1: Binary search on rows
$k_l \leftarrow 1$, $k_h \leftarrow K$
$j_l \leftarrow \text{BinarySearch}(x^*, \{x_{j,k_l}\}_j)$ ▷ Initial bracket at bottom row
$\hat{y}_l \leftarrow \text{Interp1D}(x^*, x_{j_l,k_l}, x_{j_l+1,k_l}, y_{j_l,k_l}, y_{j_l+1,k_l})$
while $k_h - k_l > 1$ **do**
  $k_m \leftarrow \lfloor (k_l + k_h)/2 \rfloor$
  $j_m \leftarrow \text{BracketSearch}(x^*, \{x_{j,k_m}\}_j, j_l)$ ▷ Search near j_l via IPO
  $\hat{y}_m \leftarrow \text{Interp1D}(x^*, x_{j_m,k_m}, x_{j_m+1,k_m}, y_{j_m,k_m}, y_{j_m+1,k_m})$
  if $\hat{y}_m \le y^*$ **then**
    $k_l \leftarrow k_m$, $j_l \leftarrow j_m$, $\hat{y}_l \leftarrow \hat{y}_m$
  else
    $k_h \leftarrow k_m$
  end if
end while

Pass 2: Final interpolation
$j_h \leftarrow \text{BracketSearch}(x^*, \{x_{j,k_h}\}_j, j_l)$
$\hat{y}_h \leftarrow \text{Interp1D}(x^*, x_{j_h,k_h}, x_{j_h+1,k_h}, y_{j_h,k_h}, y_{j_h+1,k_h})$
$\hat{f}_l \leftarrow \text{Interp1D}(x^*, x_{j_l,k_l}, x_{j_l+1,k_l}, f_{j_l,k_l}, f_{j_l+1,k_l})$
$\hat{f}_h \leftarrow \text{Interp1D}(x^*, x_{j_h,k_h}, x_{j_h+1,k_h}, f_{j_h,k_h}, f_{j_h+1,k_h})$
$t \leftarrow (y^* - \hat{y}_l)/(\hat{y}_h - \hat{y}_l)$
return $(1 - t)\hat{f}_l + t\hat{f}_h$

The algorithm uses subscripts l, m, h for low, mid, and high indices respectively. Pass 1 performs binary search over rows to find the bounding row band $[k_l, k_h]$ satisfying $\hat{y}_l \le y^* < \hat{y}_h$. Each probe at row k finds the column bracket j such that $x_{jk} \le x^* < x_{j+1,k}$, then interpolates $\hat{y}$ along that row. The IPO property

bounds the bracket search range at each probe: consecutive probes have brackets within α positions of each other, so binary search within the constrained window costs $O(\log J)$ per probe in the worst case. Pass 2 interpolates function values at the bounding rows and combines them linearly.

For query points outside the grid domain, ENGINE performs linear extrapolation by extending boundary slopes. Each 1D interpolation step handles out-of-bounds queries by using the slope between the two nearest boundary points, ensuring continuous extension of the approximation beyond the grid's convex hull.

EGM applications require extrapolation: simulation may occasionally produce state values slightly outside the solution grid due to shock realizations or numerical precision. The extrapolation extends boundary slopes: for $x^* < x_1$, the value is $f_1 + (x^* - x_1)(f_2 - f_1)/(x_2 - x_1)$, and symmetrically for $x^* > x_J$. Extrapolation reliability depends critically on grid placement. Near binding constraints, policy functions exhibit high curvature or kinks, making linear extrapolation unreliable. In wealth-rich regions, by contrast, policies typically become smooth and approximately linear as precautionary motives diminish and consumption approaches permanent income rules. Standard practice therefore extends the asset grid well beyond the constraint region, typically to several multiples of target wealth (steady-state assets), ensuring that (i) interpolation handles the economically relevant constrained region exactly, and (ii) any extrapolation occurs only in the smooth unconstrained region where linear approximation is accurate. Extrapolation to economically invalid regions (e.g., negative assets when borrowing is constrained) produces meaningless results regardless of smoothness.