State Reduction and Second-order Perturbations of Heterogeneous Agent Models *

Michael Reiter, Institute for Advanced Studies, Vienna and NYU Abu Dhabi

August 8, 2023

Abstract

This paper develops a method to compute second-order perturbations of discrete-time heterogeneous agent models. It addresses the three main tasks to make second-order approximations tractable: state reduction, generating sufficient smoothness, and fast computation of the quadratic terms in the perturbation solution.

The method is applied to a model with divisible labor, one with indivisible labor, and to an OLG model with stochastic aging. Compared to a linearized solution, second-order perturbations achieve substantially higher accuracy if models are subject to large or medium-sized aggregate shocks. They also capture precautionary behavior with respect to aggregate shocks. A general method of state reduction is developed, called "conditional-expectations approach". In the example models, it performs better in terms of accuracy and reliability than alternative approaches.

JEL classification: C63, C68, E21
Keywords: heterogeneous agents; model reduction; perturbation methods

Address of the author:

Michael Reiter
Institute for Advanced Studies
Josefstädter Strasse 39
A-1080 Vienna
Austria

e-mail: michael.reiter@ihs.ac.at

*I am grateful to Christopher Carroll, David Childers, Ben Moll, Elisabeth Proehl, Akshay Shanker, Anthony Smith, and Pablo Winant for helpful comments and discussions on earlier versions of this paper. Financial support from the Anniversary Fund of the Austrian National Bank under Grant Number 17815 is gratefully acknowledged.

1 Introduction

The solution of heterogeneous agent models with incomplete markets poses important technical challenges, mainly because the underlying state space is very high-dimensional. While the method of Krusell and Smith (1998) has been a workhorse in this field for more than 20 years, several alternatives have been explored in recent years. Frequently used are variations of the linearization approach, either in state space form (Reiter 2009a; Reiter 2010a) or in sequence form (Boppart, Krusell, and Mitman 2018; Auclert, Bardóczy, Rognlie, and Straub 2021). In this approach, the solution of the individual problem is fully nonlinear in the individual states, the model is linearized in aggregate variables around the stationary state in the absence of aggregate shocks. This approach allows to solve a wide variety of heterogeneous agent models, but shares the general limitation of linearization methods, in particular it fails to account for the effect of *aggregate* uncertainty on individual behavior.

Given the success as well as the limitations of linearization, it is natural to proceed to higher-order perturbation solutions, at least to a second-order perturbation, providing a quadratic approximation around the steady state. To do this in the framework of discrete-time models, three problems must be solved. First, a systematic way is needed to reduce the dimensionality of the state space, since the number of coefficients of the approximation grows quadratically in the state dimension.Second, numerical approximations must be chosen such that the solution is smooth enough in the state variables for a perturbation to make sense. This problem already exists for linear perturbations, but becomes more severe when going to second-order. If the optimization problem of economic agents is non-convex, individual decisions are discontinuous and needed to be smoothed out in the aggregate. Third, one needs an efficient way to compute the large set of coefficients that describe the quadratic solution.

The main contribution of this paper is to present a method that successfully addresses these issues. The usefulness of the second-order perturbation is then investigated thorough accuracy checks on some example models. As a preliminary step, I present in Appendix A an improved version of the linear model reduction algorithm of Reiter (2010a), on which the second-order approximation builds.

For state reduction, I investigate two methods outlined in Reiter (2010a), namely principal component analysis (PCA) and the "conditional expectations approach" (CEA). They

use information from the linearized solution of the model; the results below will show that they are also useful for quadratic approximations. I compare these approaches to more traditional state selection methods such as using moments of the cross-sectional distribution, or the wealth of different cohorts in an OLG model.

To guarantee smoothness of the solution, I assume that economic agents face an i.i.d. shock with twice differentiable density function and bounded support in every period. This has several advantages. First, it helps to smooth the value function, integrated over the distribution of the shock, in the sense of being at least twice differentiable.[1] Second, it helps to smooth transition probabilities in the case where the cross-sectional distribution of capital is approximated by finite point masses. Third, it helps to ensure the existence of general equilibrium in models with non-convex choice problems. Discontinuous individual demand functions in combination with approximations of cross-sectional distributions by finite point masses imply discontinuous aggregate demand functions, which often prevents the existence of an equilibrium. Smooth i.i.d. shocks solve this problem by making aggregate quantities continuous.[2] One should acknowledge that the introduction of an i.i.d. shock is not just a change in the numerical approximation method, but a substantive change to the model itself. With the right choice of shock, it arguably makes the model more realistic.

To speed up the computation of the solution, an efficient implementation of the second-order differentiation of all model equations is needed. I implement the method of this paper in a Julia toolkit which includes fast automatic differentiation routines.[3] After differentiation, the coefficients of the second-order perturbation are determined by a system of linear equations. In heterogeneous agent models, this equation system can be huge. Using a recursive algorithm, one can exploit the sparsity of the system to obtain the

[1] I solve all models with the value function approach, which is the most general one. For the models the household optimization problem is convex, one can alternatively operate on the household Euler equation. Both approaches have advantages and disadvantages.

[2] Using i.i.d. shocks to smooth dynamics is a frequently used approach in many contexts, going back at least to Dotsey, King, and Wolman (1999). In the framework of linear perturbations, Childers (2018) shows that finite-dimensional approximations of the model converge to the linearized solution in function space if there is enough smooth noise in the model. The proof requires smooth noise that directly affects the state variable. I conjecture that the smooth shock used in this paper has the same effect.

[3] The current (preliminary) version of the toolkit can be obtained from the author on request. It replicates the results of this paper.

solution easily.

I apply the method to three models with heterogeneous households: a very standard model of heterogeneous households with endogenous, continuous labor choice; a model of indivisible labor very close to Chang and Kim (2007); and finally a model of twelve overlapping generations with stochastic aging. The main conclusions are as follows. The accuracy of the second-order approximation is at least one order of magnitude higher than that of a linearized solution if aggregate shocks are large. The CEA approach performs best among the state-reduction methods considered. In the models with a TFP shock only, the effect of aggregate uncertainty on household behavior in a heterogeneous agent model is of similar magnitude as in a representative agent RBC model. The necessary computing time is of the same order of magnitude for the second order perturbation as for the steady state and linearization.

1.1 Related literature

The model reduction presented here is based on the state reduction in the working paper Reiter (2010a), which has been applied in McKay and Reis (2016) and Reiter, Sveen, and Weinke (2013), among others. Appendix A completes the model reduction by adding optimal value function reduction. Over the last years, there has been growing interest in linearization and perturbation approaches to heterogeneous agent models. Ahn, Kaplan, Moll, Winberry, and Wolf (2018) develop a similar linearization technique for continuous time models. Boppart, Krusell, and Mitman (2018) show that there is an alternative way to compute a linearized solution, using the fact that the simulation path of a linearized model is just a linear superposition of impulse responses to "MIT shocks". Auclert, Bardóczy, Rognlie, and Straub (2021) provide an efficient implementation of this method. Using methods of functional analysis, Childers (2018) gives a theoretical foundation of finite-dimensional approximations to models with infinite-dimensional state space. He shows that there must be smooth noise of sufficient dimension in the model to make sure that the discretely approximated model converges to the solution of the continuous model. Although the i.i.d. noise introduced below does not have the same form as in Childers (2018), it should have a similar effect. Mertens and Judd (2017) perform the perturbation around the deterministic steady state with neither aggregate nor idiosyncratic shocks. Bhandari, Evans, Golosov, and Sargent (2021) derive a new perturbation method that

can be applied to certain problems of optimal policy with heterogeneous agents. Probably closest to the present paper is Gornemann, Kuester, and Nakajima (2021), who also use a second-order perturbation approach for solving a heterogeneous agent model. They use state reduction based on a PCA approach, and compute the second-order perturbation by a variation of the code of Schmitt-Grohé and Uribe (2004). In contrast to Gornemann, Kuester, and Nakajima (2021), the focus on the present paper is on the methodology, investigating different model reduction schemes, checking accuracy, and developing a fast iterative solution method.

Bilal (2023) uses mean-field theory to derive perturbation solutions for continuous-time models. Numerical examples are provided for first-order perturbations, formulas are also provided for second-order perturbations. This approach does not build on state aggregation, it exploits the fact that, in continuous time, the Jacobian of the value function with respect to the cross-sectional distribution can be efficiently computed using established techniques for partial differential equations. The relative advantages of continuous versus discrete-time methods will probably depend on the specific application and have yet to be explored in depth.

Beyond perturbation, a number of alternative approaches have been proposed. Similar to Den Haan (1997), Winberry (2018) uses a low-dimensional smooth approximation of cross-sectional distribution, to allow for higher-order perturbation solutions. Grand and Ragot (2022) propose an alternative way to reduce the state dimension with the aim of substantial simplification rather than maximal accuracy. Kubler and Scheidegger (2021) reduce the state space to a very small dimension using the concept of self-justified equilibria, which has a bounded-rationality interpretation. In contrast, my approach aims to provide an approximation that is as close as possible to the rational expectations equilibrium under full information.

2 Some Example Models

To illustrate the properties of the proposed method, I look at three different heterogeneous agent models, starting from a model with infinitely lived households and endogenous labor supply. It differs from a plain-vanilla RBC model only by introducing idiosyncratic shocks to household productivity (cf. Section 2.2). It has been known since Krusell and Smith (1998) that heterogeneity has little impact on aggregate variables in this simple setup.

I use this model specifically to see how the precautionary effect arising from aggregate uncertainty is affected by household heterogeneity.

The second model described in Section 2.2 is taken from Chang and Kim (2007). The continuous labor choice of the previous model is replaced by indivisible labor, where a household can work either zero hours or a fixed number of hours. Because of indivisibility, households face a non-convex optimization problem in this model. This raises additional problems for perturbation solutions which I will address. I will also test whether the perturbation solution can replicate the results from a Krusell-Smith solution, as reported in Takahashi (2014).

Finally I consider an OLG model with stochastic aging (cf. Section 2.3). Households differ by age and by wealth. Intra-cohort heterogeneity in wealth arises from different histories in both labor productivity and the speed of economic aging. This model shows a richer variation in the cross-sectional distribution, and is a good laboratory to investigate the usefulness of different types of state reduction.

2.1 Productivity and technology (all models)

Output Y is produced using a Cobb-Douglas production function in capital K and effective labor L

$$Y_t = Z_t K_{t-1}^{\alpha} L_t^{1-\alpha} \tag{1}$$

where total factor productivity Z follows the AR(1) process

$$Z_t = 1 + \rho_z \cdot (Z_{t-1} - 1) + \epsilon_{z,t}, \qquad \epsilon_{z,t} \sim (0, \sigma_z) \tag{2}$$

This process is formulated as linear in Z, not in $\log(Z)$, so that an increase in σ_z leaves the mean of Z unaffected, which facilitates the interpretation of precautionary effects.

Gross investment I increases the capital stock, subject to some possibly time-varying depreciation:

$$K_t = I_t + (1 - \delta_t) K_{t-1} \tag{3}$$

Net real interest rate r and wage per efficiency unit w are determined by their marginal productivities: $r_t = \alpha \cdot Z_t \left(\frac{K_{t-1}}{L_t}\right)^{\alpha-1} - \delta_t$, $w_t = (1-\alpha) Z_t \left(\frac{K_{t-1}}{L_t}\right)^{\alpha}$. The aggregate resource constraint is given by

$$Y_t = C_t + I_t \tag{4}$$

Table 1: Parameter values, quarterly frequency

Parameter	Meaning	Models	Value
α	capital share	all	0.36
$\bar{\delta}$	average depreciation rate	all	0.025
β	time discount factor	div.lab.	0.990
		indiv.labor	0.983
		OLG	0.995
η	weight leisure in util.	div.lab., OLG	1.5
B	disutility labor.	indiv.lab.	166.3
ρ_z	autocorr. TFP	all	0.95
σ_z	stdev. shock TFP	div.,indiv.labor	0.007
σ_z	stdev. shock TFP	OLG	0.005
ρ_δ	autocorr. depr.	OLG	0.50
σ_δ	stdev. shock deprec.rate	OLG	0.005
ρ_A	autocorr. slope prodtty.	OLG	0.95
σ_A	stdev. shock slope prodtty.	OLG	0.010
ρ_H	autocorr. individual prod.	all	0.929
σ_H	std.dev. indiv. productivity	all	0.227
σ_ξ	std.dev. i.i.d. shock	all	0.05

2.2 The infinite horizon models

The household problem

Households maximize $\mathrm{E}\sum_{t=0}^{\infty}\beta^t U(c_t, h_t)$ over consumption c_t and hours worked h_t, subject to the constraints

$$\begin{aligned} a_t &= (1+r_t)a_{t-1} + h_t x_t \xi_t w_t - c_t \\ a_t &>= \underline{a} \\ h_t &\in \mathcal{H} \end{aligned} \tag{5}$$

Here a_t denotes end-of-period assets and x_t is an individual productivity process, modeled as a finite Markov chain such that $\log(x)$ approximates an AR(1) process with persistence ρ_H and standard deviation σ_H. The i.i.d. component of individual productivity, ξ_t, has expected value 1 and standard deviation of σ_ξ, with a finite support and a density function

that is twice differentiable everywhere (cf. Online-appendix B.2). Factor prices w and r are functions of the aggregate state $S = (Z, D)$, where Z follows (2), and the law of motion for the cross-sectional distribution $D' = T(Z, D)$ is determined in general equilibrium.

In the **divisible-labor** model, current utility is given by $U(c, h) = \log(c) + \eta \log(1 - h)$ and hours are chosen from the set $\mathcal{H} = [0, 1]$. Labor supply is determined by the first order condition $\frac{u_L(C_t, h_t)}{u_C(C_t, h_t)} \geq w_t$ with equality if $h > 0$. For asset-rich households, the constraint $h \geq 0$ may be binding. In the **indivisible-labor** model, current utility is given by $U(c, h) = \log(c) - Bh$ and the choice set for labor has only two points, $\mathcal{H} = \{0, 1/3\}$. Whether a household works depends on the asset level as well as the persistent and the i.i.d. component of productivity. In this model, the aggregate number of hours is determined exclusively by the extensive margin. Online-appendix B.4 describes how to compute perturbations of the extensive margin effect.

The household value function satisfies the Bellman equation

$$V(a, x; S) = \int \max_{a' \geq \underline{a}, h \in \mathcal{H}} \Big\{ U\left((1 + r(S))a + w(S)h \cdot x \cdot \xi - a', h\right) + \beta \sum_{x'} \pi_x(x, x') \times \mathrm{E}_S\, V(a', x'; S') \Big\}\, dF(\xi) \quad (6)$$

The expectation operator E_S integrates over the distribution of the aggregate shocks, conditional on the current aggregate state S.

Aggregate variables are obtained from individual choices by integrating over the cross-sectional distribution D_t and the distribution $F(\xi)$ of the i.i.d. shock:

$$\begin{aligned} L_t &= \int \int x \cdot \xi \cdot h(a, x, \xi)\, dF(\xi)\, dD_t(a, x) \\ C_t &= \int \int c(a, x, \xi)\, dF(\xi)\, dD_t(a, x) \\ K_t &= \int a\, dD_t(a, x) \end{aligned}$$

Depreciation is assumed to be constant, $\delta_t = \bar{\delta}$. Parameters and their values are standard and listed in Table 1.

2.3 The OLG model

The OLG model is intended to illustrate the effects of more dimensions of heterogeneity, as well as the effects of a larger number of aggregate shocks. The model period is one year.

There are 12 cohorts of workers, meant to span the five-year periods from 20 to 80 years. Ageing is therefore modeled as a stochastic process, where workers move to the next cohort with probability 0.2. Workers of the first 8 cohorts are working, those of the last 4 cohorts are retired.

Demographics

Workers are born in cohort $\tau = 1$. Each period, a worker of any cohort $\tau \in (1, \ldots, 10)$ moves to cohort $\tau + 1$ with probability 0.2, and stays in cohort τ with probability 0.8. A worker in cohort $\tau = 11$ moves to cohort 12 with probability 0.1, dies with probability 0.1, and otherwise stays at $\tau = 11$. A worker in cohort 12 dies with probability 0.2, and stays in cohort $\tau = 12$ with probability 0.8. To formalize this, we define the probability π_s of surviving into the next period, moving from age τ to $\tau + 1$, as

$$\pi_s(\tau, \tau') = \begin{cases} 0.8 & \text{if} \quad \tau' = \tau \\ 0.2 & \text{if } \tau \leq 10, \quad \tau' = \tau + 1 \\ 0.1 & \text{if } \tau = 11, \quad \tau' = 12 \\ 0.0 & \text{else} \end{cases} \tag{7}$$

The death probability is $1 - \pi_s(\tau, \tau) - \pi_s(\tau, \tau + 1)$. With this demography, a constant fraction of workers is retired, namely 30.4 percent. A worker who dies is superseded by a new worker ("child"), born into cohort 1, who inherits the wealth that is left by the parent worker as well as the individual labor productivity of the parent.

Labor productivity and the value function

Individual labor productivity is the product of three factors. The n-state Markov process x and the i.i.d. shock ξ are exactly as in the infinite-horizon models. In addition, there is an age specific factor $X(\tau, t)$ with a slope that is subject to an aggregate shock,

$$X(\tau, t) = \exp\left(-(0.01 + \tilde{A}_t)(\tau - 6) - 0.02(\tau - 6)^2\right) \tag{8}$$

where

$$\tilde{A}_t = \rho_A \tilde{A}_{t-1} + \epsilon_A \qquad \epsilon_{A,t} \sim (0, \sigma_A) \tag{9}$$

This labor productivity profile is hump-shaped with quickly declining productivity in old age. The shock $\tilde{A}$ reduces productivity, affecting older cohorts more than younger ones. Notice that there is no pension system, households need to save for retirement.

Households discount future utility by the pure time discount factor β, multiplied with the survival probability. This implies that they do not value the utility of their offspring, bequests are therefore accidental. The household value function satisfies

$$V(a,x,\tau;S) = \int \max_{a'\geq \underline{a}, h\in[0,1]} \Big\{ \sum_{x',\tau'} \pi_x(x,x')\pi_s(\tau,\tau') \times \\ [U\,(c,,h) + \beta\, \mathrm{E}_S\, V(a',x',\tau';S')] \Big\}\, dF(\xi) \quad (10)$$

where consumption follows the budget constraint

$$c = (1+r(S))a + w(S)\cdot h \cdot x \cdot \xi \cdot X(\tau(S)) - a' \quad (11)$$

Again, the expectation operator E_S integrates over the distribution of the aggregate shocks, conditional on the current aggregate state S.

While retired, the productivity process x follows the same dynamics as during working age. It has no effect on the household's utility, but productivity is passed on to the child household.

Production

Technology is the same as in Section 2.2, but subject to a stochastic depreciation rate of capital, fluctuating around the mean value $\bar{\delta}$:

$$\delta_t = \bar{\delta} + \rho_\delta(\delta_{t-1} - \bar{\delta}) + \epsilon_{\delta,t} \qquad \epsilon_{\delta,t} \sim (0, \sigma_\delta) \quad (12)$$

In sum, the model now features three aggregate shocks: a shock to TFP, cf. Equ. (2); a shock to the age premium of labor efficiency, cf. Equ. (9); a shock to the depreciation rate of capital, cf. Equ. (12).

3 Solution Method

3.1 A class of heterogeneous agent models

In the following we assume that the state of each agent in the model can be described by a vector[4] of endogenous continuous individual state variables (such as different types of

[4]In the numerical examples, there is always one individual continuous state, namely financial wealth. Including additional continuous state variables is conceptually straightforward, but increases computational complexity.

assets), by an exogenous individual stochastic process which is approximated by a finite-state Markov chain, as well as, potentially, by a discrete state, taking on values from a finite set (such as employed/unemployed). Information about the heterogeneous agents is summarized in the cross-sectional distribution of individual states, denoted by D_t.[5]

We can now group the variables of the model as follows.

- D_t: the cross-sectional distribution of economic agents' individual states at the end of period t.
- q_t: the vector of endogenous aggregate states not included in D (such as past investment in a model with investment adjustment costs), also measured at the end of period t.
- z_t: the vector of exogenous driving forces, such as TFP.
- ε_t: the vector of current aggregate shocks.
- V_t: the value function of the agents.
- a_t: all the remaining aggregate variables.

The vector of all states predetermined at the beginning of period t is denoted by $S_{t-1} \equiv (D_{t-1}, q_{t-1}, z_{t-1})$. All time-$t$ variables are therefore a function of S_{t-1} and ε_t. We assume, however, that the distribution D_{t-1} directly affects individual behavior only through a set of aggregate variables that are part of a_t. In most cases, these will be prices such as the real wage or the real interest rate. A critical assumption about the model is that the vector a_t is only medium-sized, in contrast to the distribution D, which is high-dimensional. This is necessary for model reduction, cf. Appendices A.2 and A.3.

With these assumptions, the model equations can be written as follows:

$$\text{Exogenous dynamics:} \qquad z_t = \mathcal{Z}\left(z_{t-1}, \varepsilon_t\right) \tag{13a}$$

$$\text{Endogenous dynamics:} \qquad q_t = \mathcal{Q}\left(S_{t-1}, \varepsilon_t, a_t\right) \tag{13b}$$

$$\text{Aggregate equilibrium conditions:} \qquad 0 = \mathcal{A}\left(S_{t-1}, \varepsilon_t, a_t, \mathrm{E}_t\, a_{t+1}, \mathrm{E}_t\, V_{t+1}\right) \tag{13c}$$

$$\text{Bellman equations:} \qquad V_t = \mathcal{V}\left(a_t, \mathrm{E}_t\, a_{t+1}, \mathrm{E}_t\, V_{t+1}\right) \tag{13d}$$

[5]For simplicity we always talk about one cross-sectional distribution. The general case of several types of ex-ante different agents, with separate distributions, is subsumed in the above formulas simply by stacking the different distributions into one vector. The same applies to the value vector.

and the transition law of the cross-sectional distribution

$$D_t = \Pi\left(a_t, \mathrm{E}_t\, a_{t+1}, \mathrm{E}_t\, V_{t+1}\right) D_{t-1} \tag{14}$$

The vector of aggregate shocks ε_t is assumed to have bounded support (Jin and Judd 2002), being i.i.d. over time, with current covariance matrix Σ. The expectation E_t in (13) is over the aggregate shocks only. Integration over idiosyncratic shocks is finitely approximated and is implicit in the functions $\mathcal{Q}$, $\mathcal{A}$, $\mathcal{V}$ and Π (cf. Online-appendix B).

Notice that the optimal policies at grid points are not treated as variables of the system, they are considered to be a function of current states, aggregate variables and the expected value function of the next period. Optimal policies are computed in the solution process and integrated over D_{t-1} and the idiosyncratic shock ξ_t to obtain the aggregate variables a_t (for example aggregate labor supply) and the transition matrix Π. This has two advantages. First, it helps to keep the number of model variables as small as possible. Second, it gives the flexibility to compute the optimal policy at an endogenous set of grid points, in particular at the values of the i.i.d. shock where the optimal policy switches between different regimes. Differentiation of the optimal policy is described in the online-appendix B.4.

3.2 Outline of the solution algorithm

The model solution involves the following steps:

1. Finite approximation of the continuous theoretical model, cf. Section 3.3. Given a finite representation of the value function and the cross-sectional distribution, all variables can be stacked into the vector Θ, and the nonlinear model (13) can be written in compact form as

$$\mathcal{M}(S_{t-1}, \varepsilon_t, \Theta_t, \mathrm{E}_t\, \Theta_{t+1}) = 0 \tag{15}$$

   To simplify notation, we have assumed in (15) that every expression that appears in expectations is defined as a separate variable. For example, if the right hand side of an Euler equation contains the term $\mathrm{E}_t\left[(1 + r_{t+1})u'(c_{t+1})\right]$, we have assumed that a variable $rhs_t = (1 + r_t)u'(c_t)$ is defined, so that $\mathrm{E}_t\, rhs_{t+1}$ appears in the equation. The continuation values V_{t+1} appear naturally in this form.

2. Computing the stationary state $\bar{\Theta}$ of the discretized model without aggregate shocks, satisfying

$$\mathcal{M}(\bar{S}, 0, \bar{\Theta}, \bar{\Theta}) = 0 \tag{16}$$

where it is understood that $\bar{S}$ is the state vector in $\bar{\Theta}$. This step is standard.

3. Linearizing the system of model equations (15) in the set of variables $(S_{t-1}, \varepsilon_t, \Theta_t, \Theta_{t+1})$ around the deterministic steady state (Reiter 2009a). This is done by automatic differentiation (Griewank and Walther 2008) and is exact up to machine precision. What needs some explanation is the differentiation of the critical points in the state space, for example the point where an occasionally binding constraint starts binding, or the point where an agent switches from working to non-working in the model of indivisible labor. This is detailed in Online-appendix B.4.

4. Solving the linearized model using exact model reduction. This follows the working paper Reiter (2010a).[6] An improved version of the method is presented in Appendix A.

5. Replacing the distribution D_{t-1} in S_{t-1} by a set of statistics m_{t-1}, to get a reduced state vector $s_t \equiv (m_{t-1}, q_{t-1}, z_{t-1})$. Replacing D_{t-1} in Θ_t by m_t we get θ_t. The reduced model can be written as[7]

$$\mathcal{M}(s_{t-1}, \varepsilon_t, \theta_t, \mathrm{E}_t\, \theta_{t+1}) = 0 \tag{17}$$

To be feasible, the state vector s must be much smaller than the state in the lossless linear reduction of Step 4. Nevertheless, information from the linearized model solution is useful in finding a suitable s, as is explained in Section 3.4.

6. Computing a linear solution in the reduced state vector, cf. Section 3.5.2.

7. Differentiating the reduced model (17) twice in the variables $(s_{t-1}, \varepsilon_t, \theta_t, \mathrm{E}_t\, \theta_{t+1})$ and computing a second-order perturbation solution in the reduced state vector. Section 3.5.3 presents fast iterative algorithms to compute the quadratic terms in s_{t-1}

[6] Auclert, Bardóczy, Rognlie, and Straub (2021) provide an alternative method that would essentially yield to the same linearized solution. The advantage of exact model reduction is that it prepares the state reduction necessary for the second-order perturbation.

[7] Using the letter $\mathcal{M}$ in both (15) and (17) is a slight abuse of notation.

and ε_t. The terms accounting for precautionary behavior with respect to aggregate shocks are derived in Section 3.5.4.

8. Simulating the reduced model. Section 3.6 discusses different ways to simulate the second-order perturbation solution.

Steps 5–7 are the core contribution of this paper.

3.3 Discrete approximation of the model

In the models of Section 2, the individual state of an agent is described by a continuous state variable (beginning-of-period assets) and an exogenous discrete state, capturing productivity and age.[8] The aggregate state space therefore includes the cross-sectional distribution of assets, which is an infinite-dimensional object and must be finitely approximated. In the literature, this is often done either by a finite number of point masses (Young 2010) or by a histogram (Reiter 2009b), where it is assumed that the cross-sectional density is constant within histogram bins.[9] Perturbation around the steady state is based on the derivatives at the steady state, and one has to make sure that the transition dynamics of the points masses or histograms are smooth enough for the required derivatives to exist, and to give a meaningful approximation to fluctuations of realistic size. Moreover, in models with discrete choice such as the indivisible labor model, there are extensive margin effects arising from a continuous change of the threshold point where the optimal policy is switching between the different discrete choices. To address these issues, the idiosyncratic i.i.d. shock ξ with smooth density function was added to the agent problems in Section 2. This generates smooth transition probabilities between the points of a finite grid of capital, cf. Online-appendix C.2. Extensive margin effects arise from the movement of the thresholds in ξ where the change in the discrete choice occurs, cf. Online-appendix B.4.

The discrete approximation of the decision problem of the agents can be summarized as follows.

- The individual continuous state is approximated by a finite grid $[\bar{\kappa}_1, \ldots, \bar{\kappa}_{n_\kappa}]$.

[8] One can easily allow for an endogenous discrete state, such as employment status. Including additional continuous state variables is conceptually straightforward, but increases computational complexity.

[9] An alternative approach is to parameterize the density of the cross-sectional distribution by a smooth functional form (Den Haan 1997; Winberry 2018). The transition law is then nonlinear in the parameters of the distribution, which makes it difficult to use high-dimensional approximations.

- The individual exogenous state takes on values on the grid $\left[\bar{\zeta}_1, \ldots, \bar{\zeta}_{n_\zeta}\right]$.
- The cross-sectional distribution is given by the fraction of agents $D_{i,j}$ at each state $(\bar{\kappa}_i, \bar{\zeta}_j)$.
- For each $(\bar{\kappa}_i, \bar{\zeta}_j)$, the optimal policy is computed on a grid of i.i.d. shocks $\left[\bar{\xi}_1, \ldots, \bar{\xi}_{n_\xi}\right]$. If the policy regime changes between $\bar{\xi}_l$ and $\bar{\xi}_{l+1}$, the threshold point $\hat{\xi}$ is identified, and added to the ξ-grid at $(\bar{\kappa}_i, \bar{\zeta}_j)$.
- The value function at state $(\bar{\kappa}_i, \bar{\zeta}_j)$ is obtained by integrating the optimal value over the distribution of the i.i.d. shock ξ. The details of the integration are given in Online-appendix B.3.
- For each exogenous grid point $\bar{\zeta}_j$, the value function is interpolated as a cubic spline in the endogenous state κ.
- The cross-sectional average of any function $g(\kappa, \zeta, \xi)$ is obtained as $\sum_i^{n_\kappa} \sum_j^{n_\zeta} D_{i,j} \int_\xi g(\bar{\kappa}_i, \bar{\zeta}_j, \xi) \phi(\xi)\, d\xi$, where $\phi(\xi)$ is the density function of ξ, cf. Section B.2.
- The distribution dynamics is approximated by transition probabilities between grid points $(\bar{\kappa}_i, \bar{\zeta}_j)$. Because of the smooth i.i.d. shocks, these probabilities are differentiable, cf. Online-appendix C.2.

In all calculations, the optimal policy at (κ, ζ, ξ) is given implicitly by the Bellman equation with continuation value $\mathrm{E}_t\, V_{t+1}$.

3.4 Model reduction for the second-order approximation

3.4.1 Linear functions of the cross-sectional distribution

Appendix A shows how to reduce the dimension of the state vector such that the linear approximation in the reduced state is as accurate as the linear approximation in the full state, up to machine precision. This reduced state vector still contains hundreds of variables. Denote the number of state variables by n_s. In a second-order perturbation, the solution for each model variable has $1 + n_s + n_s(n_s + 1)/2$ parameters. Since the model contains a large number of variables, mostly from the approximation of the value function

of the heterogeneous agents,[10] a further significant reduction of the number of states is necessary. Nevertheless, the state reduction of Appendix A is a useful starting point for further reductions, as will be shown below.

For the quadratic approximation, I will replace the distribution D_t in the state vector by a set of linear functions m_t:

$$m_t = HD_t \tag{18}$$

Any variable in the quadratic approximation is then constructed as a function of the states $(m_{t-1}, q_{t-1}, z_{t-1}, \varepsilon_t)$. The matrix H should be chosen so that xm_t includes at least the variables that directly affect prices, typically the aggregate capital stock. Section 4.1 lists the types of additional statistics that are included in the reduced state vector.

3.4.2 Proxy distributions

To approximate the distribution dynamics (14) by an equation in the statistics m_t, I use the idea of a "proxy distribution" taken from Reiter (2010b). Again expressed as linear deviations from the steady state, the proxy distribution D_t^{pd} can be written as $D_t^{pd} = D^* + \Phi^{pd}(m_t - m_t^*)$ for some known matrix Φ^{pd}. It is natural to choose Φ^{pd} such that D_t^{pd} is the expectation of D_t conditional on $HD_t = m_t$ in the linearized model solution, assuming normally distributed shocks. This is given by[11]

$$\Phi^{pd} = \Sigma_D H'(H\Sigma_D H')^{-1} \tag{19}$$

We then replace (14) by

$$m_t = H\Pi\left(a_t, \mathrm{E}_t\, a_{t+1}, \mathrm{E}_t\, V_{t+1}\right)\Phi^{pd} m_{t-1} \tag{20}$$

so that the full model (15) is transformed into the reduced model (17). Replacing D by m, the state vector $S_{t-1} = (D_{t-1}, q_{t-1}, z_{t-1})$ is replaced by $s_{t-1} = (m_{t-1}, q_{t-1}, z_{t-1})$.

[10] For the linearized solution, Appendix A.2 derives a lossless reduction of the dimension of the value vector. Since this reduction is not lossless for the quadratic approximation, I keep the full value vector, to avoid additional sources of approximation error.

[11] Σ_D is a very large matrix, but can be represented in condensed form as $\Sigma_D = U\Sigma U'$, where U spans the ergodic set in which the model solution lies. U is obtained from a singular value decomposition of all the impulse responses to the different shocks of the model, and usually turns out to be of much smaller dimension than the cross-sectional distribution.

3.5 Perturbation solution in the reduced state

3.5.1 Notation for derivatives

To make the notation as compact as possible, we adopt the following conventions. Superscripts of functions or variables refer to components, subscripts denote derivatives. We use the Einstein summation convention, where the product of two terms with a common index, one of them as a superscript and the other one as a subscript, denotes the summation over this index. For example, $G^i_\alpha \tilde{x}^\alpha_t$ is to be understood as $\sum_\alpha G^i_\alpha \tilde{x}^\alpha_t$. To avoid ambiguities, it is necessary to specify the range over which the indices run. I distinguish the following types of indices:

- Greek letters α and β run over the elements of the time-t state vector $\tilde{x}_t = (\tilde{s}_{t-1}, \varepsilon_t)$.
- Greek letters γ and δ run over the elements of the predetermined states $\tilde{s}_t$ at the end of period t.
- Greek letters λ and μ run over the future shocks ε_{t+1}.
- Roman letters i and j, run over all time-t variables in $\tilde{\Theta}_t = [\tilde{s}_t; \tilde{y}_t; \tilde{V}_t]$
- The uppercase Roman letters I and J run over the elements of future variables $\tilde{\Theta}_{t+1}$.

3.5.2 Linear Terms

Before computing the second-order approximation, it is necessary to find the first-order terms in the reduced state space. Using the above notation, the linear perturbation in the reduced state can be written as

$$\tilde{\Theta}^i_t = G^i_\alpha \tilde{x}^\alpha_t \tag{21}$$

where superscript tilde denotes deviations from the deterministic steady state. The coefficients G^i_α have to be determined in the solution process. To allow for an iterative solution, we distinguish between the time-t coefficients G^i_α and the time-$(t+1)$ coefficients $\hat{G}^I_\alpha$, the hat denoting the approximation in the future period. Since $\mathrm{E}_t\, \varepsilon_{t+1} = 0$, expected future variables can then be written as a function of current states as

$$\mathrm{E}_t\, \tilde{\Theta}^I_{t+1} = \hat{G}^I_\gamma \tilde{s}^\gamma_t = \hat{G}^I_\gamma G^\gamma_\alpha \tilde{x}^\alpha_{t-1} \tag{22}$$

Therefore $\frac{\partial \tilde{\Theta}_t^i}{\partial \tilde{x}_t^\alpha} = G_\alpha^i$ and $\frac{\partial \mathrm{E}_t \tilde{\Theta}_{t+1}^I}{\partial \tilde{x}_t^\alpha} = \hat{G}_\gamma^I G_\alpha^\gamma$. Total differentiation of equation k of the equation system (15) at the deterministic steady state then yields the equilibrium conditions

$$\mathcal{M}_\alpha^k + \mathcal{M}_i^k G_\alpha^i + \mathcal{M}_I^k \hat{G}_\gamma^I G_\alpha^\gamma = 0 \tag{23}$$

where

$$\mathcal{M}_\alpha^k \equiv \frac{\partial \mathcal{M}^k}{\partial x_t^\alpha}, \qquad \mathcal{M}_i^k \equiv \frac{\partial \mathcal{M}^k}{\partial \Theta_t^i}, \qquad \mathcal{M}_I^k \equiv \frac{\partial \mathcal{M}^k}{\partial \Theta_{t+1}^I} \tag{24}$$

In the infinite-horizon solution, $G = \hat{G}$. Then (23) defines a quadratic equation system in the elements of G. There are at least three potential ways to compute the stable solution:

- Standard tools for solving quadratic matrix equations such as the QZ-algorithm. This is not feasible for very large models: even if the dimension of the state vector is reduced, the value function is still high-dimensional, cf. Footnote 10.

- Time iteration:

  1. Initialize $\hat{G}$.
  2. Given $\hat{G}$, solve (23) for G.
  3. Update $\hat{G} = G$.
  4. Iterate 2. and 3. until convergence.

  This iteration converges to the stable solution if the model has a unique stable solution (Rendahl 2017; Higham 2002). Notice that (23) is linear in the elements of G, but the coefficients of the linear system are changing in each step of the iteration through the updating of $\hat{G}$, which slows down the computations. However, the changing part is the matrix $\mathcal{M}_I^k \hat{G}_\gamma^I$, which only affects the n_s state variables and has the relatively small rank n_s. The changing linear system can therefore be efficiently handled by the Sherman-Morrison-Woodbury formula (Press, Flannery, Teukolsky, and Vetterling 1986, Section 2.7).

- Time iteration with lagged updating of the state transition G_α^γ. This is the same iteration as above, but solving

$$\mathcal{M}_\alpha^k + \mathcal{M}_i^k G_\alpha^i + \mathcal{M}_I^k \hat{G}_\gamma^I \hat{G}_\alpha^\gamma = 0 \tag{25}$$

rather than (23). In each step, the linear system has the same coefficients $\mathcal{M}_i^k$ and is sparse. There is no convergence proof for this iteration, but it convergences in all applications below and is very fast.

The interpretation of lagged updating is the following. Next period's expected continuation values are a function of end-of-period predetermined variables, which are determined by the aggregate state transition G_α^γ. Rather than solving for the current state transition in equilibrium, agents apply the transition function $\hat{G}_\alpha^\gamma$ obtained in the previous iteration. After convergence, the two coincide.

3.5.3 Second-order terms in the state variables

For the second-order approximation, we scale the covariance matrix of the shocks Σ by the factor σ. As usual, we proceed by deriving first the quadratic coefficients with respect to the state variables, taken at the deterministic steady state $\sigma = 0$. Derivatives with respect to σ are treated in Section 3.5.4.

For $\sigma = 0$, the quadratic approximation can be written as

$$\tilde{\Theta}_t^i = G_\alpha^i \tilde{x}_t^\alpha + \frac{1}{2} H_{\alpha\beta}^i \tilde{x}_t^\alpha \tilde{x}_t^\beta \tag{26}$$

so that second derivatives with respect to state variables are given by $\frac{\partial^2 \tilde{\Theta}_t^i}{\partial \tilde{x}_t^\alpha \partial \tilde{x}_t^\beta} = H_{\alpha\beta}^i$. At $\sigma = 0$ we have both $\mathrm{E}_t\, \varepsilon_{t+1}^\lambda = 0$ and $\mathrm{E}_t(\varepsilon_{t+1}^\lambda \varepsilon_{t+1}^\mu) = 0$, therefore

$$\begin{aligned}
\mathrm{E}_t\, \tilde{\Theta}_{t+1}^I &= G_\gamma^I \tilde{s}_t^\gamma + \frac{1}{2} \hat{H}_{\gamma\delta}^I \tilde{s}_t^\gamma \tilde{s}_t^\delta \\
\text{where} \quad \tilde{s}_t^\gamma &\equiv G_\alpha^\gamma \tilde{x}_t^\alpha + \frac{1}{2} H_{\alpha\beta}^\gamma \tilde{x}_t^\alpha \tilde{x}_t^\beta, \qquad \tilde{s}_t^\delta \equiv G_\alpha^\delta \tilde{x}_t^\alpha + \frac{1}{2} H_{\alpha\beta}^\delta \tilde{x}_t^\alpha \tilde{x}_t^\beta \\
\left. \frac{\partial^2 \mathrm{E}_t\, \tilde{\Theta}_{t+1}^I}{\partial \tilde{x}_{t-1}^\alpha \partial \tilde{x}_{t-1}^\beta} \right|_{\tilde{x}_t = 0} &= G_\gamma^I H_{\alpha\beta}^\gamma + \frac{1}{2} \hat{H}_{\gamma\delta}^I (G_\beta^\gamma G_\alpha^\delta + G_\alpha^\gamma G_\beta^\delta)
\end{aligned} \tag{27}$$

Notice that the components of G are already known from the linear solution, therefore there is no need to distinguish G and $\hat{G}$. In the second order step, we solve iteratively over H and $\hat{H}$, with $H = \hat{H}$ after convergence. Total differentiation of the k-th equation in (15)

with respect to the predetermined variables at the deterministic steady state gives[12]

$$R^k_{\alpha\beta} + \mathcal{M}^k_i H^i_{\alpha\beta} + \mathcal{M}^k_I \left[G^I_\gamma H^\gamma_{\alpha\beta} + \frac{1}{2}\hat{H}^I_{\gamma\delta}(G^\gamma_\beta G^\delta_\alpha + G^\gamma_\alpha G^\delta_\beta) \right] = 0 \tag{28}$$

where $R^k_{\alpha\beta}$ is defined as

$$\begin{aligned} R^k_{\alpha\beta} \equiv \mathcal{M}^k_{\alpha\beta} + \mathcal{M}^k_{\alpha i} G^i_\beta + \mathcal{M}^k_{\alpha I} G^I_\gamma G^\gamma_\beta + \mathcal{M}^k_{i\beta} G^i_\alpha + \mathcal{M}^k_{ij} G^i_\alpha G^j_\beta + \mathcal{M}^k_{iJ} G^i_\alpha G^j_\gamma G^\gamma_\beta \\ + \mathcal{M}^k_{I\beta} G^I_\gamma G^\gamma_\alpha + \mathcal{M}^k_{Ij} G^I_\gamma G^\gamma_\alpha G^j_\beta + \mathcal{M}^k_{IJ} G^I_\gamma G^\gamma_\alpha G^J_\delta G^\delta_\beta = 0 \end{aligned}$$

Notice that $R^k_{\alpha\beta}$ does not depend on the coefficients of the quadratic approximation, H, only on first order coefficients G which are given from the linear approximation in Section 3.5.2. Again, there are at least three potential ways to solve the equation system (28):

- Simultaneous linear equation system

  This is the way that standard software packages such as Dynare solve for the second order coefficients. With $H = \hat{H}$, (28) is a linear system in the elements of H:

  $$R^k_{\alpha\beta} + \mathcal{M}^k_i H^i_{\alpha\beta} + \mathcal{M}^k_I \left[G^I_\gamma H^\gamma_{\alpha\beta} + \frac{1}{2} H^I_{\gamma\delta}(G^\gamma_\beta G^\delta_\alpha + G^\gamma_\alpha G^\delta_\beta) \right] = 0$$

  In general, this equation system is dense and has approximate dimension $n_v \cdot n_s^2/2$, where n_v is the number of economic variables in the model, and n_s is the number of reduced state variables plus shocks. Since the storage requirement is quadratic and the computational effort of solving a dense linear system is cubic in the dimension of the system, this is infeasible or at least inefficient for very large models.

- Time iteration

  1. Set $\hat{H}^i_{\alpha\beta} = 0$ for all i, α and β.
  2. Given $\hat{H}$, and separately for each pair (α, β), solve (28) for the $H^i_{\alpha\beta}$. This is as a linear system in $H^i_{\alpha\beta}$. For any variable $\tilde{s}^\gamma \in \tilde{s}$, the coefficients of the linear system are $\mathcal{M}^k_\gamma + \mathcal{M}^k_I G^I_\gamma$. For any other variable, the coefficients are $\mathcal{M}^k_i$.

[12] Similar to (24), we use the notation

$$\mathcal{M}^k_{\alpha\beta} = \frac{\partial^2 \mathcal{M}^k}{\partial x^\alpha_t \partial x^\beta_t}, \qquad \mathcal{M}^k_{\alpha I} = \frac{\partial^2 \mathcal{M}^k}{\partial x^\alpha_t \partial \Theta^I_{t+1}}, \qquad \mathcal{M}^k_{iI} = \frac{\partial^2 \mathcal{M}^k}{\partial \Theta^i_t \partial \Theta^I_{t+1}}, \qquad \text{etc.}$$

for the partial derivatives of model equations.

3. Set $\hat{H}^i_{\alpha\beta} = H^i_{\alpha\beta}$ for all i, α and β.

4. Iterate 2. and 3. until convergence.

Notice that the coefficients of the linear system are the same in each iteration step.

- Time iteration with lagged update of the quadratic state transition $H^\gamma_{\alpha\beta}$:

1. Set $\hat{H}^i_{\alpha\beta} = 0$ for all i, α and β.

2. Given $\hat{H}$, and separately for each pair (α, β), solve

$$R^k_{\alpha\beta} + \mathcal{M}^k_i H^i_{\alpha\beta} + \mathcal{M}^k_I \left[G^I_\gamma \hat{H}^\gamma_{\alpha\beta} + \hat{H}^I_{\alpha\beta} (G^\alpha_\beta G^\beta_\alpha + G^\alpha_\alpha G^\beta_\beta)/2 \right] = 0 \tag{29}$$

for $H^i_{\alpha\beta}$. The linear system (29) differs from (28) only in replacing $G^I_\gamma H^\gamma_{\alpha\beta}$ by $G^I_\gamma \hat{H}^\gamma_{\alpha\beta}$.

3. Set $\hat{H}^i_{\alpha\beta} = H^i_{\alpha\beta}$ for all i, α and β.

4. Iterate 2. and 3. until convergence.

With lagged updating, the linear system in each step is sparser and therefore somewhat faster to solve. Lagged updating was converging in all applications below.

3.5.4 The effect of uncertainty

It is well known (Judd 1998; Schmitt-Grohé and Uribe 2004) that the first derivative of all policy functions with respect to σ, taken at the deterministic steady state $\sigma = 0$, is equal to zero. The same holds for the cross-derivatives with respect to a state and σ. This means that the effect of uncertainty on equilibrium variables in a second-order perturbation is proportional to the variance, not the standard deviation of the shock. This effect is given by a constant term for each variable, independent of the state of the economy. Denoting the constant term by $H^i_{\sigma\sigma}$, we can write the complete quadratic approximation as

$$\tilde{\Theta}^i_t = G^i_\alpha \tilde{x}^\alpha_t + \frac{1}{2} \left(H^i_{\alpha\beta} \tilde{x}^\alpha_t \tilde{x}^\beta_t + H^i_{\sigma\sigma} \sigma^2 \right) \tag{30}$$

so that $\frac{\partial^2 \tilde{\Theta}^i_t}{\partial \sigma^2} = H^i_{\sigma\sigma}$. The expected value of future variables is given by

$$\begin{aligned} \mathrm{E}_t \, \tilde{\Theta}^I_{t+1} &= G^I_\gamma \tilde{s}^\gamma_t + \frac{1}{2} \mathrm{E}_t \left(H^I_{\gamma\delta} \tilde{s}^\gamma_t \tilde{s}^\delta_t + H^I_{\gamma\lambda} \tilde{s}^\gamma_t \varepsilon^\lambda_{t+1} + H^I_{\lambda\mu} \varepsilon^\lambda_{t+1} \varepsilon^\mu_{t+1} + H^I_{\sigma\sigma} \sigma^2 \right) \\ &= G^I_\gamma \tilde{s}^\gamma_t + \frac{1}{2} \left(H^I_{\gamma\delta} \tilde{s}^\gamma_t \tilde{s}^\delta_t + H^I_{\lambda\mu} \Sigma_{\lambda\mu} \sigma^2 + H^I_{\sigma\sigma} \sigma^2 \right) \end{aligned} \tag{31}$$

where we use the abbreviation

$$\tilde{s}_t^\gamma \equiv G_\alpha^\gamma \tilde{x}_{t-1}^\alpha + \frac{1}{2}\left(H_{\alpha\beta}^\gamma \tilde{x}_{t-1}^\alpha \tilde{x}_{t-1}^\beta + H_{\sigma\sigma}^\gamma \sigma^2\right) \tag{32}$$

This implies $\frac{\partial^2 \mathrm{E}_t \tilde{\Theta}_{t+1}^I}{\partial \sigma^2} = G_\gamma^I H_{\sigma\sigma}^\gamma + H_{\lambda\mu}^I \Sigma_{\lambda\mu} + H_{\sigma\sigma}^I$. Differentiating the equation system (15) twice with respect to the scaling factor σ then gives

$$\mathcal{M}_i^k H_{\sigma\sigma}^i + \mathcal{M}_I^k \left[G_\gamma^I H_{\sigma\sigma}^\gamma + H_{\lambda\mu}^I \Sigma_{\lambda\mu} + H_{\sigma\sigma}^I\right] = 0 \tag{33}$$

With the $H_{\lambda\mu}^I$ given from the calculation in Section 3.5.3, Equ. (33) defines a linear equation system in the $H_{\sigma\sigma}^i$ that can be solved directly.

3.6 Model simulation

In the second-order perturbation solution, each variable is a quadratic function of the reduced state vector. Next period's reduced state is a quadratic function of the current reduced state and of the realization of next period's shock. Given a series of pseudo-random shocks, simulation can proceed as it is routinely done for example in Dynare for DSGE models. I call this **aggregate simulation**, and it is sufficient if only aggregate variables are needed.

However, this procedure is not suitable to simulate the cross-sectional distribution if the model is subject to aggregate shocks of realistic size. The main reason is that a linear or a quadratic approximation of individual decision rules is very likely to violate some constraints of the model. For example, saving choices will often violate borrowing constraints. A quadratic approximation of the cross-sectional distribution will often violate non-negativity constraints on densities.[13] The threshold points of decision rules will move between bins within the grid of i.i.d. shocks, in which case it is not clear how to perform the integration. To simulate the full cross-sectional distribution, a more complex calculation is required. A single step in the simulation proceeds as follows.

1. Given is the state (S_{t-1}, ε_t), which includes the cross-sectional distribution.

2. From (S_{t-1}, ε_t) obtain the reduced state (s_{t-1}, ε_t) by applying (18).

[13] A quadratic approximation of the log of densities avoids negative densities, but then densities will not add up to unity.

3. In the model solution, all time-t variables are a quadratic function of (s_{t-1}, ε_t). Use these quadratic expressions to compute

   - aggregate variables a_t
   - end-of-period aggregate states (m_t, z_t, k_t);
   - expected values $\mathrm{E}_t V_{t+1}$ and $\mathrm{E}_t a_{t+1}$ from end-of-period aggregate states, applying the covariance matrix Σ of the aggregate shocks.[14]

4. Using $\mathrm{E}_t V_{t+1}$ and the aggregate states, solve the individual optimization problem at all grid points and obtain the transition matrix Π as described in Online-appendix C.2.

5. Compute D_t from Equ. (14).

6. Draw a random shock ε_{t+1}; this completes the computation of the full state (S_t, ε_{t+1}).

I call this procedure **distribution simulation**. Notice that Step 4. in the simulation does not impose the equilibrium conditions for aggregate variables such as labor market clearing. If the aim is to maximize the precision of the model simulation, one should iterate over Steps 3. and 4 until the equilibrium conditions are satisfied, for example by a quasi-Newton method. In the accuracy checks in Section 4.3, I rather use the violations of market clearing as an accuracy measure.

4 Numerical Results

The following numerical results serve to measure the accuracy of quadratic perturbation solutions when models are subject to large aggregate shocks. A main focus is the comparison between different approaches to state reduction, which are explained in Section 4.1.

Exact approximation errors of the models with aggregate uncertainty cannot be computed, since an exact solution is not available. Therefore, Section 4.2 compares the deterministic part of the quadratic solution (cf. Section 3.5.3) to nonlinear perfect foresight paths in response to one-time shocks of different magnitude, which can be computed with high precision. Section 4.3 presents consistency checks of the stochastic solutions, which also serve to estimate the accuracy of the computed precautionary effect of aggregate uncertainty. Section 4.5 compares computing times for linear and for quadratic solutions.

[14] Since all variables are quadratic functions of the states, expected values only depend on the covariance matrix, not the exact distribution of the shocks.

4.1 State variables

In addition to the "minimal states" (aggregate capital, the driving processes and current shocks), I consider the following types of statistics for the reduced state vector.

- "MOM": additional moments (beyond the mean) of the cross-sectional distribution of capital, centered around the steady state of aggregate capital. These are only applied in the infinite-horizon models.
- "COH": the aggregate capital owned by each cohort or by adjacent groups of cohorts, applied in the OLG model.
- "PCA": the principal components of the covariance matrix of the cross-sectional distribution, obtained from the solution of the linearized model. The idea is to identify the linear combinations of the cross-sectional distribution that fluctuate most over the cycle, i.e., vectors h with $\|h\|_2 = 1$ such that the variance of $(h'D)$ is high. The principal components are the eigenvectors belonging to the largest eigenvalues of the covariance matrix Σ_D of the distribution D.
- "CEA": the leading elements of the reduced state vector of the linearized model. These are the leading singular vectors of the matrix that expresses the conditional expectations of future aggregate variables as linear functions of the current distribution. I therefore call this the "conditional expectations approach" (CEA) as explained in Appendix A.3.

The numerical experiments compare the performance of these statistics as state variables.

4.2 Accuracy of the deterministic quadratic solution

We measure the accuracy of the deterministic part of the quadratic solution by comparing the impulse response function to a one-time shock with the nonlinear perfect foresight path after this shock. The impulse response function is computed by the aggregate simulation approach (cf. Section 3.6). The perfect foresight path can be solved for with high precision.[15] Considering shocks of different size, we get an idea of the range of the state space for which a quadratic perturbation is a good approximation. As a benchmark, we also

[15] Auclert, Bardóczy, Rognlie, and Straub (2021) show an efficient way to compute the Jacobian of a perfect-foresight path at the steady state solution, which can be used to solve for nonlinear paths by a

consider the approximation error of the linearized solution with loss-less model reduction (cf. Appendix A).

The most important results for the three models are collected in Table 4.2. The table reports the maximum absolute deviation of the perturbation solution from the exact impulse response. We consider three aggregate variables: aggregate labor input, the change in aggregate capital ("investment"), and the level of the capital stock. The error in the capital stock is an indication of how approximation errors accumulate over time in the simulation. All numbers are expressed relative to the respective steady state value (for the change in capital, relative to the steady state level of capital).

Results are given for different model reduction types (column "Reduc") and different number of states ("#St"). We vary the size of the shock between one standard deviation ($Z_0 = 1$) and ten standard deviations ($Z_0 = 10$). For each shock size, the column "Neg" refers to a negative (i.e., contractionary) shock of this size, the column "NegPos" refers to the sum of the impulse responses to a negative shock and a positive shock of the same size. By summing over the responses to the negative and the positive shock, we eliminate the linear part in the solution and the approximation error, focusing on the quadratic and higher-order parts. The numbers in parentheses in the first line of every model report the maximum absolute impulse response, to which the maximum absolute approximation errors can be compared. By construction, the linear approximation ("LIN") explains nothing of the sum of positive and negative responses, the error in the column "NegPos" is therefore always equal to the maximum impulse response.

The first part of the table refers to the impulse responses to a one-time shock of minus one standard deviation in the **divisible-labor model**. As one can expect for such a small shock, the linear approximation already does a good job, with a maximal error of only half a percent of the impulse response. This is considerably better than the quadratic approximation "- 0", which only uses the minimal states (capital, lagged TFP and current shock), in this respect similar to the approach in Krusell and Smith (1998). With this shock size, the aggregation error is more severe than the linearization error. This is true for all three variables under a contractionary shock (column "Neg"). To improve on the linear solution, additional state variables have to be included in the quadratic solution. It turns out that two additional variables are sufficient. As additional states we either

quasi-Newton method. For the results below I compute the path by fixed point iteration with Anderson acceleration, which converges in a few steps.

Table 2: Accuracy of perturbation solutions: maximum absolute errors of impulse responses

				Labor		Investment		Capital	
Model	Z_0	Reduc	#St	Neg	NegPos	Neg	NegPos	Neg	NegPos
DivL	1	(ImpResp)		(3.7e-3)	(1.7e-5)	(5.8e-4)	(5.9e-7)	(4.9e-3)	(6.9e-6)
		LIN		1.69e-5	1.67e-5	5.85e-7	5.88e-7	6.88e-6	6.86e-6
		-	0	1.28e-4	2.60e-6	4.93e-6	1.58e-7	2.66e-5	2.30e-6
		MOM	2	7.88e-6	2.96e-6	2.19e-7	1.46e-7	3.73e-6	2.76e-6
		PCA	2	6.47e-6	2.61e-6	4.66e-7	1.47e-7	6.13e-6	2.35e-6
		CEA	2	2.34e-6	1.74e-6	2.11e-7	1.50e-7	2.85e-6	2.58e-6
	10	(ImpResp)		(3.9e-2)	(1.6e-3)	(5.8e-3)	(4.7e-5)	(4.8e-2)	(6.2e-4)
		LIN		1.68e-3	1.62e-3	4.71e-5	4.67e-5	6.19e-4	6.15e-4
		-	0	1.46e-3	1.35e-4	5.04e-5	6.20e-6	2.81e-4	1.42e-4
		MOM	2	1.44e-4	1.43e-4	4.42e-6	4.75e-6	7.49e-5	7.00e-5
		PCA	2	9.73e-5	9.55e-5	7.58e-6	3.28e-6	9.72e-5	9.78e-5
		CEA	2	8.52e-5	1.49e-5	1.75e-6	5.55e-7	9.76e-6	1.41e-5
IndL	10	(ImpResp)		(3.6e-2)	(2.1e-3)	(7.0e-3)	(5.1e-5)	(5.4e-2)	(6.4e-4)
		LIN		2.20e-3	2.13e-3	5.05e-5	5.14e-5	6.30e-4	6.39e-4
		-	0	1.01e-3	2.66e-4	3.70e-5	7.06e-6	1.17e-4	6.21e-5
		MOM	2	3.75e-4	3.60e-4	1.92e-5	1.64e-5	1.37e-4	1.50e-4
		PCA	2	5.47e-4	3.70e-4	1.75e-5	1.16e-5	1.55e-4	1.12e-4
		CEA	2	2.39e-4	6.39e-5	9.84e-6	7.14e-6	4.56e-5	4.57e-5
OLG	3	(ImpResp)		(1.3e-2)	(5.6e-3)	(1.3e-2)	(2.9e-4)	(2.1e-2)	(1.5e-3)
		LIN		5.64e-3	5.57e-3	2.91e-4	2.89e-4	1.45e-3	1.45e-3
		-	0	8.03e-4	3.26e-4	9.31e-5	1.44e-5	8.61e-4	1.01e-4
		COH	5	4.80e-4	3.60e-4	2.23e-5	1.09e-5	1.14e-4	1.49e-4
		PCA	4	4.21e-4	3.80e-4	6.37e-5	5.39e-5	1.41e-3	1.21e-3
		CEA	6	2.07e-4	1.41e-4	1.54e-5	1.13e-5	2.00e-5	1.46e-5
	10			(1.4e-2)	(6.5e-2)	(4.6e-2)	(3.2e-3)	(7.7e-2)	(1.6e-2)
		LIN		6.55e-2	6.49e-2	3.25e-3	3.24e-3	1.61e-2	1.63e-2
		CEA	6	5.15e-3	4.51e-3	1.43e-4	2.24e-4	4.08e-4	3.16e-4

Notes: Z_0: initial shock size; "Reduc": type of state reduction, cf. Section 4.1. #St: number additional states;

choose two additional cross-sectional moments (variance and skewness, "MOM"), the two most important principal components ("PCA") or the first two terms of the conditional expectations approach("CEA"). The latter achieves the lowest error for all three variables. Including more than two additional states brings no further improvement and can even lead to a deterioration. Detailed results for approximations with up to 10 additional variables are provided in the online-appendix E, Table 7.

It is instructive to study how the approximation changes when the size of the shock is multiplied by 10, shown in the next part of Table 4.2. This shock causes a recession with a drop in output and labor of about 9 and 4 percent, respectively. The error of the linear approximation increases quadratically, as one would expect from a linear approximation. This is reassuring: the linear perturbation solution has the expected properties, in spite of the discrete approximations underlying the approximation, even for very large shocks. In contrast, the error of the quadratic solution with minimal states increases linearly, because it is dominated by the aggregation error, which increases linearly in the deviation from the steady state. Adding more states, the role of the aggregation error is much reduced and the nonlinearity becomes more important. As a consequence, the error increases super-linearly in the shock size. With this large shock, the best quadratic solution achieves an increase in accuracy between one and two orders of magnitude over the linear solution.

As one would expect, the sum of impulse responses ("NegPos") increases about quadratically in the shock size. Again, the CEA approach performs better than the other state reduction methods, and the approximation error is only about 1 percent of the response. All in all, these results show that the quadratic perturbation provides high accuracy even for very large shocks in this model.

The **indivisible-labor model**, featuring a discontinuity in the individual labor and consumption function, is more difficult to solve accurately.[16] Table 4.2 again shows results for a shock of ten standard deviations. Accuracy of quadratic solutions deteriorates compared to the divisible-labor model by a factor of about four. Nevertheless, the quadratic solution improves on the linear solution by about one order of magnitude. What is most important, it shows that modeling the aggregate labor supply response by differentiating

[16]Takahashi (2014) found severe approximation errors in the original numerical results; extensive results in the online-appendix D show that our solution does not suffer from these problems, being close to the results of Takahashi.

the threshold points between working and non-working is successful. [17] Again, CEA turns out to be a reliable choice, doing well in all cases.

The **OLG model** poses additional challenges. The model is not calibrated to data, but is designed so as to be a stronger test case for state aggregation, by having additional heterogeneity through the OLG structure, and featuring three shocks with different properties. For that purpose, I choose standard deviations of 0.005 for the TFP shock, 0.01 for the shock to the slope of productivity with respect to age, and 0.005 for the shock to the depreciation factor. In the linearized model, these values yield a standard deviation of undetrended output of 4.77 percent and of 1.64 percent after detrending. About one third of this variation comes from the shock to depreciation, and about 20 percent come from the shock to the productivity slope.

The results in Table 4.2 refer to the case where all shocks occur simultaneously with either three or 10 times their standard deviation. The minimal state vector now includes 7 variables, namely aggregate capital as well as the lagged values of the three exogenous states and the three shocks. For the additional states we again compare the CEA and PCA approach, but also consider results for solutions that use the aggregate capital holdings of adjacent cohort as additional state variables ("COH"). The line "COH" with 11 additional states has the capital holdings of each of the cohorts 1 to 11 next to the aggregate capital stock. The line "COH" with 5 additional states adds the capital holdings of cohorts 1–2, 3–4, 5–6, 7–8 and 9–10. The line "COH" with 3 additional states adds the capital holdings of cohorts 1–3, 4–6, 7–9.

With shocks of three standard deviations, high accuracy can be achieved, improving over the linearized solution by more than one order of magnitude. For the sum of negative and positive responses, the error is somewhat more than 1 percent of the responses. Again, it turns out the CEA performs better than the alternatives. The upper panels in Figure 1 illustrate results for a large variety of state vectors, ranging up to 20 additional states for both PCA and CEA.[18] The graphs show very clearly that conditioning on more state variables does not necessarily increase accuracy.

The last part of the table, showing results for a simultaneous shock to all three processes of 10 standard deviations, points to the limits of the perturbation approach. Notice that the approximation in the linear approximation is greater than the impulse response itself. The

[17] The technicalities of the differentiation are described in the online-appendix B.4.

[18] More extensive results for all three models are collected in the online-appendix E.

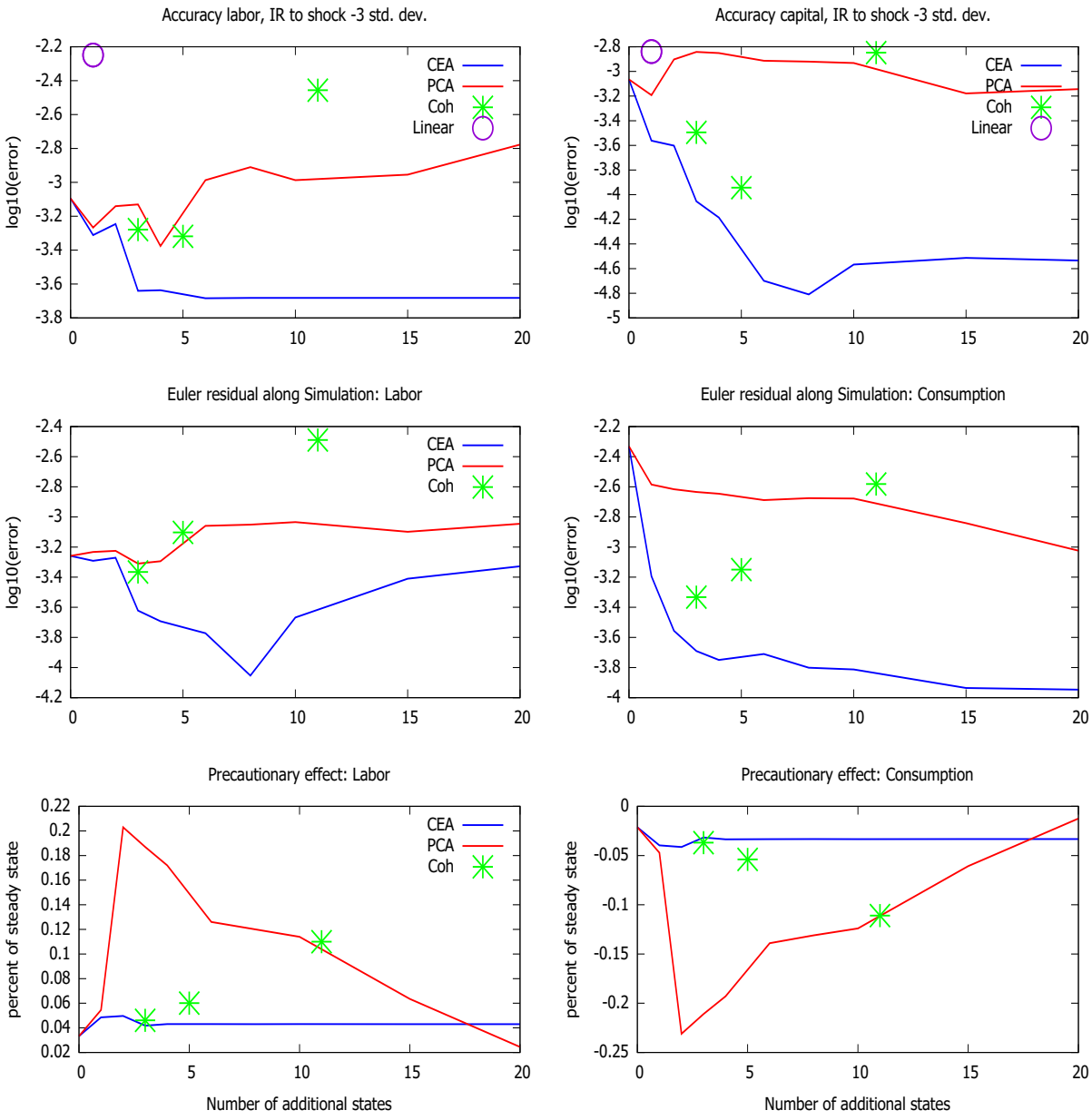

Figure 1: OLG Model, results for different state vectors

quadratic approximations improve greatly on the linear one, but the approximation error in labor is about 30 percent of the impulse response, so the approximation is not reliable. Increasing the number of states does not help, cf. the results in the online-appendix E. Shocks of this size probably require a global nonlinear solution strategy.

4.3 Accuracy of the precautionary term

In a second-order perturbation, the deterministic and the stochastic solution for any model variable only differ by a constant term that is proportional to the variance of the shocks. Table 3 presents these constants for aggregate labor and consumption for the three example models. As a benchmark, results are also given for a standard RBC model, using the same parameter values as the divisible-labor model.

Table 3: Precautionary effect

	RBC	Divisible Labor	Indiv. Labor	OLG
Labor	0.0117	0.0078	0.0068	0.0430
Consumption	-0.0109	-0.0086	-0.0080	-0.0333

Notes: in percent of steady state value.

We see that the differences between RBC and the infinite-horizon HA models are not large. Interestingly, the precautionary effect of *aggregate* shocks on aggregate consumption and labor is smaller in the divisible-labor than in the RBC model. This also depends on the wealth distribution: a more detailed analysis reveals that asset-rich households show very small precautionary behavior. In the OLG model, which is driven by different aggregate shocks, the precautionary effect is substantially larger.

The precautionary effect is related to the quadratic coefficients through formula (33), so that approximation errors in the quadratic terms will feed through to the precautionary term. I see no direct way to measure the approximation error in the precautionary effect unless a more precise solution is available. As a partial consistency check, one would require the precautionary term to be stable across quadratic solutions with different state vectors, at least for all those that achieve high accuracy. This turns out to be satisfied for the two infinite-horizon models. For the 19 different state vectors tried with the divisible-labor model, the precautionary effects on consumption and labor of all but 3 solutions lie

within plus or minus 10 percent of the median value. In the indivisible-labor model, it is in all but 4 solutions. This is not the case in the OLG model. The graphs at the bottom of Figure 1 show the precautionary effect in the OLG model for a large variety of state vectors. The estimated precautionary effect differs widely across the solutions based on the PCA approach. The positive result is that the estimated effect is almost constant for the CEA solutions with 6 or more additional states. These are also the solutions that achieve consistently high accuracy for the quadratic terms (cf. again the top panels in the figure). The reason why the precautionary term is more difficult to estimate in the OLG model is that it features three shocks. Two of them (TFP shock and shock on the productivity slope) generate a precautionary effect of the usual sign, negative on consumption and positive on labor. The variability of the third shock (to depreciation) has the opposite effect: positive on consumption and negative on labor.[19] This same problems can be expected to arise in DSGE models with several shocks.

As a further check, we also test accuracy of the stochastic solutions in a way similar to the common practice of computing Euler residuals at many points in the state space. More precisely, the residuals are computed as follows:

1. Simulate a long time series of all model variables by the distribution simulations, cf. Section 3.6. In each step, this includes the computation of the expected continuation value function.

2. From the continuation value function, compute optimal individual decisions, then integrate individual decisions to obtain aggregate consumption, labor, etc. Residuals are then defined as the difference between the aggregate values computed in this way and the values implied by the quadratic approximation.

The maximum of the residuals along the simulation paths are shown in the middle panels of Figure 1. Results are in line with the performance of the deterministic model. The CEA approach performs best, and optimal performance is approximately reached with 8 additional state variables. A further increase in the number of state variables has little effect.

[19]This may appear surprising, but it is also true in a simple RBC model with the same specification of the depreciation shock.

4.4 Summary of the accuracy results

Quadratic approximations are successful in providing solutions with high accuracy. For all the models considered here, they improve over the linear approximation by at least one order of magnitude for shocks of realistic size. To achieve this improvement, it is necessary to condition the solution on more than the "minimal states" (aggregate capital and the driving shocks), but adding 2–6 states turns out to be sufficient in our example models. For a given number of states, the most accurate results are usually achieved by the CEA approach, which includes those statistics of the cross-sectional distribution that are identified as being most useful for of prediction of future variables in the linearized solution. These variables turn out to be more relevant for the solution than the components of the distribution that account for most of the cyclical variation in the distribution, which is what PCA does.

It is important to note that an increase in the number of state variables may also lead to a reduction in accuracy. This happened in the infinite-horizon models when adding cross-sectional moments of capital beyond the third moment, or when adding the capital stock of each cohort in the OLG model. A likely explanation is the following.[20] Additional state variables are useful if they provide relevant information about the cross-sectional distribution. This information is conveyed through the "proxy distribution", which exploits the covariance of the state variables and the distribution in the linearized solution. If the correlation between some variables is substantially changed in the nonlinear solution, the information provided may become misleading.

4.5 Computational cost

The algorithm was coded in the programming language Julia. The divisible- and indivisible labor model can be easily done even on a small laptop, but the implementation of the OLG model needs somewhat more memory. For comparability across models, all computations were done on a Windows desktop with an AMD Ryzen 7-3700X 8-core CPU at 3.6 Ghz with 16GB memory. The computation times needed for the quadratic approximations are shown in Table 4.

[20]This problem is not due to collinearity issues, because the state variables are diagonalized to avoid collinearity.

Table 4: Computation times in seconds

	Divisible Labor		Indivisible Labor		OLG	
Solution	#states	seconds	#states	seconds	#states	seconds
SteadyState	1401	1.6	8501	17.7	16803	25.0
Linear	1401	3.0	8501	35.8	16803	57.2
Quadratic	5	2.7	5	29.5	7	85.6
Quadratic	10	4.7	10	47.2	10	113.3
Quadratic	15	6.9	15	70.0	15	141.5
Quadratic	20	10.3	20	102.3	20	188.5

Notes: Computation times exclude compilation time; timings are approximate and stochastic because of automatic garbage collection. The quadratic approximations use the CEA approach.

The table distinguishes three steps of the computation. First, solving for the steady state without aggregate shocks. Second, computing the linear solution, which involves the differentiation of the equation system, the computation of the loss-less model reduction of Appendix A and the model solution by QZ-decomposition. Third, computing the quadratic perturbation, which includes finding the proxy distribution, twice differentiating the reduced model, and performing the backward iterations detailed in Section 3.5. One should keep in mind that computing times depend on the exact implementation of the algorithms, and the time for any component may come down if further algorithmic improvements are found. Of course, lowering the grid sizes for capital and productivity would speed up computation, with a probably minor loss in accuracy.

The exact computing time depends on the dimension of the state vector, but for moderate dimension the costs are similar in magnitude to the cost of the linearized solution, including the steady state computation. All in all, computation time only increases about linearly in the dimension of the reduced state vector, if sufficient memory is available.

5 Conclusions

In this paper I have shown how to extend the linearization approach of Reiter (2009a) to a second-order perturbation. Applied to three different heterogeneous agent models, the

method achieves high accuracy, improving by an order of magnitude or more compared to linearization if the model is subject to large aggregate shocks. An iterative algorithm obtains the solution with computation times similar to the time required for steady state and first order approximation. The "conditional-expectations approach", building on Reiter (2010a), was proposed as a general approach to achieve a substantial reduction in the number of state variables with little loss in accuracy. In all the example models, this approach achieved higher accuracy than alternatives. The second-order approximation allows to compute the precautionary effect of aggregate uncertainty on behavior. In the example models featuring a TFP shock only, the precautionary effect is small, comparable to the effect in a standard RBC model. In the OLG model with several aggregate shocks, the precautionary effect is stronger.

A second-order perturbation solution in a reduced state space enables very fast simulation of the model and appears to be a promising method for simulation based estimation procedures. It could also be used as a starting point to obtain even more accurate approximations. One idea is to combine the first- and second-order terms with different state vectors, using linear approximation in a high dimensional state and quadratic approximation in a lower-dimensional state. This would keep both the aggregation error and the error from nonlinearity small. A second option is a hybrid approach, coupling quadratic perturbation with more general nonlinear transformations, as proposed by Judd (2002) and Fernandez-Villaverde and Rubio-Ramirez (2006). Exploring these options is left for future work.

A Loss-less Model Reduction for the Linear Solution

A.1 Model reduction: general outline

The vector Θ is of high dimension as it contains the cross-sectional distribution of wealth as well as the value function of households. The aim of this section is show how to reduce the dimension of the model with (almost) no loss in accuracy in the linearized version of the model. The engineering literature (Antoulas 2005) shows how to do model reduction if the model is already given in state space form. This theory cannot be applied directly, because the economic model first has to be solved before the state-space form is obtained.

We partition the linearized model, splitting the variables into three types: S_t denotes

the vector of state variables, and V_t denotes the vector of all variables that appear with time index $t+1$,[21] and y_t contains all other variables. All variables are expressed in deviations from deterministic steady state. While S_t and V_t can be very large vectors, it is essential that the dimension of y_t is small.

Partitioning the equation system of the general model in Section 3.1 in conformity to the variables, the linearization around the steady state can be written as

$$\begin{bmatrix} \Lambda_{ss} & 0 & 0 \\ \Lambda_{ys} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} S_{t-1} \\ y_{t-1} \\ V_{t-1} \end{bmatrix} + \begin{bmatrix} \Gamma_{ss} & \Gamma_{sy} & \Gamma_{sv} \\ \Gamma_{ys} & \Gamma_{yy} & \Gamma_{yv} \\ 0 & \Gamma_{vy} & \Gamma_{vv} \end{bmatrix} \begin{bmatrix} S_t \\ y_t \\ V_t \end{bmatrix} + \\ \mathrm{E}_t \begin{bmatrix} 0 & 0 & \Phi_{sv} \\ 0 & 0 & \Phi_{yv} \\ 0 & 0 & \Phi_{vv} \end{bmatrix} \begin{bmatrix} S_{t+1} \\ y_{t+1} \\ V_{t+1} \end{bmatrix} + \begin{bmatrix} \Psi_s \\ \Psi_y \\ 0 \end{bmatrix} \varepsilon_t = 0 \quad (34)$$

Only the variables V appear with time index $t+1$, only S appears with time index $t-1$, and these two groups do not overlap. Only y, not S enters the equations for V. This can be easily achieved by adding auxiliary variables to y. We assume that Γ_{ss}, Γ_{yy}, and Γ_{vv} are invertible. We further assume that Γ_{ss} and Γ_{vv} are sparse (often diagonal), so that $\Gamma_{ss}^{-1}\Lambda_{ss}$ etc. can be easily computed.

Using the regularity of Γ_{ss} and Γ_{vv}, we can rewrite (34) as

$$\begin{bmatrix} \Gamma_{ss}^{-1}\Lambda_{ss} \\ \Lambda_{ys} \\ 0 \end{bmatrix} S_{t-1} + \begin{bmatrix} I & \Gamma_{ss}^{-1}\Gamma_{sy} & \Gamma_{ss}^{-1}\Gamma_{sv} \\ \Gamma_{ys} & \Gamma_{yy} & \Gamma_{yv} \\ 0 & \Gamma_{vv}^{-1}\Gamma_{vy} & I \end{bmatrix} \begin{bmatrix} S_t \\ y_t \\ V_t \end{bmatrix} + \\ \mathrm{E}_t \begin{bmatrix} \Gamma_{ss}^{-1}\Phi_{sv} \\ \Phi_{yv} \\ \Gamma_{vv}^{-1}\Phi_{vv} \end{bmatrix} V_{t+1} + \begin{bmatrix} \Gamma_{ss}^{-1}\Psi_s \\ \Psi_y \\ 0 \end{bmatrix} \varepsilon_t = 0 \quad (35)$$

The task is to replace the large vectors S and V by vectors of lower dimension. This is possible without loss in accuracy if the following two conditions are satisfied.

1. There exists an $n_V \times n_v$ matrix $\bar{V}$ with $n_v < n_V \equiv dim(V)$ such that each possible value V_t in a solution of the model can be written as

$$V_t = \bar{V} v_t \quad (36)$$

[21] This is a slight deviation from the notation of Section 3.2, where V_t includes only the value vector.

The basis $\bar{V}$ spans a lower-dimensional space in which the value function lives. W.l.o.g. we can choose $\bar{V}$ as orthonormal so that $\bar{V}'\bar{V} = I$, and therefore $v_t = \bar{V}'V_t$. Section A.2 shows how to find such a $\bar{V}$ if it exists.

2. There exists an $n_s \times n_S$ matrix $\bar{S}$ with $n_s < n_S \equiv dim(S)$ and matrices $\hat{A}$, $\tilde{\Lambda}_{ys}$ and $\tilde{\Gamma}_{ys}$ such that

$$\bar{S}\Gamma_{ss}^{-1}\Lambda_{ss} = \hat{A}\bar{S}, \tag{37a}$$

$$\Lambda_{ys} = \tilde{\Lambda}_{ys}\bar{S}, \qquad \Gamma_{ys} = \tilde{\Gamma}_{ys}\bar{S} \tag{37b}$$

(37b) is satisfied if the rows of Λ_{ys} and Γ_{ys} are spanned by the rows of $\bar{S}$. We then define

$$s_t = \bar{S}S_t \tag{38}$$

The vector s_t should be interpreted as the statistics of the cross-sectional distribution that are necessary to compute the solution. These statistics are linear functions of the distribution. Section A.3 shows how to find such an $\bar{S}$ if it exists.

To write the model in reduced form, we premultiply the first block of equations in (35) by $\bar{S}$, and the third block by $\bar{V}'$. Using (36)–(38), the equation system (35) becomes

$$\begin{bmatrix} \hat{A} \\ \tilde{\Lambda}_{ys} \\ 0 \end{bmatrix} s_{t-1} + \begin{bmatrix} I & \bar{S}\Gamma_{ss}^{-1}\Gamma_{sy} & \bar{S}\Gamma_{ss}^{-1}\Gamma_{sv}\bar{V} \\ \tilde{\Gamma}_{ys} & \Gamma_{yy} & \Gamma_{yv}\bar{V} \\ 0 & \bar{V}'\Gamma_{vv}^{-1}\Gamma_{vy} & I \end{bmatrix} \begin{bmatrix} s_t \\ y_t \\ v_t \end{bmatrix} + \\ \mathrm{E}_t \begin{bmatrix} \bar{S}\Gamma_{ss}^{-1}\Phi_{sv}\bar{V} \\ \Phi_{yv}\bar{V} \\ \bar{V}'\Gamma_{vv}^{-1}\Phi_{vv}\bar{V} \end{bmatrix} v_{t+1} + \begin{bmatrix} \bar{S}\Gamma_{ss}^{-1}\Psi_s \\ \Psi_y \\ 0 \end{bmatrix} \varepsilon_t = 0 \tag{39}$$

It turns out that in all our example models substantial value and state reduction is possible such that (36) and (37) are satisfied with machine precision. In all our examples, the dimension of the reduced model (39) is small enough that the model can be solved for (s_t, y_t, v_t) by standard methods such as Sims (2001). The degree of model reduction is reported in Section A.5.

Notice the asymmetry between state reduction and value function reduction. For the value function, we require that $\bar{V}$ spans the space in which the value function "lives" in equilibrium. In contrast, we do not assume to know the space in which S_t lives. In

particular, $\bar{S}'$ does not span the space of realizations of S_t. What is required for state reduction is a set of statistics s_t that provides sufficient information about the state to solve the model. This is ensured by the commutation property (37). The realizations of S_t in a simulation also depend on the (arbitrary) initial state. Section A.4 shows how to recover S_t from s_t in a simulation.

A.2 Loss-less value function reduction

To find a lower-dimensional basis $\bar{V}$ of the value function space,[22] write the third line of equations in (35) as

$$V_t = \mathrm{E}_t \left(\tilde{\Gamma}_{vy} y_t + \tilde{\Phi}_{vv} V_{t+1} \right) \tag{40}$$

where $\tilde{\Gamma}_{vy} = -\Gamma_{vv}^{-1}\Gamma_{vy}$ and $\tilde{\Phi}_{vv} = -\Gamma_{vv}^{-1}\Phi_{vv}$. Iterating forward gives

$$\begin{aligned} V_t &= \mathrm{E}_t \left[\tilde{\Gamma}_{vy} y_t + \tilde{\Phi}_{vv}\tilde{\Gamma}_{vy} y_{t+1} + \tilde{\Phi}_{vv}^2 \tilde{\Gamma}_{vy} y_{t+2} + \ldots \right] \\ &= \sum_{i=0}^{\infty} \tilde{\Phi}_{vv}^i \tilde{\Gamma}_{vy} \, \mathrm{E}_t \, y_{t+i} \end{aligned} \tag{41}$$

At this stage, the terms $\mathrm{E}_t \, y_{t+i}$ are unknown, but (41) implies that V_t is spanned by the columns of the matrix $Q \equiv \left[\tilde{\Gamma}_{vy}, \tilde{\Phi}_{vv}\tilde{\Gamma}_{vy}, \tilde{\Phi}_{vv}^2\tilde{\Gamma}_{vy}, \ldots, \tilde{\Phi}_{vv}^N\tilde{\Gamma}_{vy}, \right]$ where N is chosen such that Q has full rank, and further terms do not increase the space. The matrix $\bar{V}$ is then given by an orthonormal basis of Q. This transformation is only useful if the rank of Q is substantially smaller than the dimension of V. An essential condition for this is that $\tilde{\Gamma}_{vy}$ has low rank, which is the case if the number of aggregate variables y affecting the decision problem of agents directly is small.

A.3 Loss-less state aggregation

As stated at the end of Section A.1, the task is to find a selection matrix $\bar{S}$ such that (37) is satisfied. We start with the following

Lemma 1. *Given an $n \times n$ matrix A and an $m \times n$ matrix C with $m \leq n$, define the*

[22] Reiter (2010a) proposes an iterative algorithm to determine $\bar{V}$, for which there is no convergence proof. The procedure described here avoids this problem and is faster.

$(m \cdot n) \times n$ matrix Q (in system theory called "observability matrix") as

$$Q \equiv \begin{bmatrix} C \\ CA \\ CA^2 \\ \cdots \\ CA^{n-1} \end{bmatrix} \tag{42}$$

Denote by $k \leq n$ the rank of Q. Then there are a $k \times n$ matrix K and a $k \times k$ matrix $\hat{A}$ such that

i) $KK' = I_k$

ii) $KA = \hat{A}K$

iii) The rows of C are spanned by the rows of K.

Proof. The singular value decomposition of Q can be written as

$$Q = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1' \\ V_2' \end{bmatrix} = U_1 S V_1' \tag{43}$$

$$S \equiv diag(\sigma_1, \ldots, \sigma_k) \tag{44}$$

where U_1 has dimension $m \cdot n \times k$, V_1 hs dimension $n \times k$, and $U_1'U_1 = V_1'V_1 = I_k$. Setting $K = V_1'$, then i) follows immediately from the properties of V_1, and ii) follows from 1) by setting $\hat{A} \equiv KAK'$. To get iii), notice that

$$C = \begin{bmatrix} I & 0 & \ldots & 0 \end{bmatrix} Q = \left(\begin{bmatrix} I & 0 & \ldots & 0 \end{bmatrix} U_1 S \right) K \tag{45}$$

□

To satisfy (37), we apply Lemma 1 setting A as $\Gamma_{ss}^{-1}\Lambda_{ss}$ and C as an arbitrary basis of $\begin{bmatrix} \Lambda_{ys}' & \Gamma_{ys}' \end{bmatrix}'$. $\bar{S}$ is then chosen as K of Lemma 1. State reduction is achieved if the rank of Q is smaller than the dimension of Λ_{ss}, Again, this requires that that the vector y is small so that Λ_{ys} and Γ_{ys} have small rank.

This construction of $\bar{S}$ has an intuitive interpretation. Assume that we want to predict the endogenous variables y_{t+i} for $i = 0, 1, \ldots$ in a linear model where y is related to the states S by $y_t = CS_t$ and the state transition equation is

$$S_t = AS_{t-1} + B\varepsilon_t, \qquad \mathrm{E}_{t-1}\, \varepsilon_t = 0 \tag{46}$$

Then

$$\mathrm{E}_t\, y_{t+i} = CA^i S_t \tag{47}$$

and the matrix Q in (42) expresses conditional expectations of future y's as a linear function of current states. A basis for Q then gives the linear combination of the states that contains all the relevant information about the expected y's. This procedures is therefore called the "Conditional Expectations Approach" in Reiter (2010a).

Notice that we do not know the system dynamics A before solving the model, so that the matrix $\bar{S}$ in (37) does not appear to capture the conditional expectations of the model solution. However, it follows from the commutation property (37a) that it is sufficient to use the $\Gamma_{ss}^{-1}\Lambda_{ss}$ instead of A for model reduction, so that $\bar{S}$ does in fact contain the relevant information about the model solution.

A.4 Simulating the linearized model

Simulating the reduced model (39) we obtain time series for the reduced variables (s_t, y_t, v_t). The full value vector V_t is then given by (36) as $V_t = \bar{V} v_t$. However, the full state vector S_t is not a function of (s_t, y_t, v_t), but is path-dependent and can be recovered only as part of a simulation of the model. To start the simulation, an initial state vector S_0 must be given, for example the deterministic steady state. The full state vector S_t is computed in step t of the simulation as follows:

1. From s_t and y_t, the model solution determines $\mathrm{E}_t\, v_{t+1}$, which gives $\mathrm{E}_t\, V_{t+1} = \bar{V}\, \mathrm{E}_t\, v_{t+1}$.

2. S_t is obtained from the first block of equations in (35) as

$$S_t = -\Gamma_{ss}^{-1}\left[\Lambda_{ss} S_{t-1} + \Gamma_{sy} y_t + \Gamma_{sv}\bar{V} v_t + \Phi_{sv}\bar{V}\, \mathrm{E}_t\, v_{t+1} + \Psi_s \varepsilon_t\right]$$

A.5 Model reduction in the example models

Table 5 reports the degree of reduction that is achieved by the loss-less reduction when applied to the models of Section 2.

Table 5: State reduction

Model	n_D	n_m	n_V	n_v
Divisible Labor	1400	188	350	97
Indivisible Labor	8500	316	1700	136
OLG model	16800	244	3360	183

Here, $n_D = n_\kappa \cdot n_\zeta$ denotes the size of the discrete grid of individual states, n_m the number of statistics describing the distribution. In the largest model, the number of variables is reduced from about 20,000 to about 400. Even further reductions would be possible, with minimal changes in results, by applying a less strict criterion for the rank of the Q-matrix in Sections A.2 and A.3. Models with a few hundred variables can be easily solved by QZ-decomposition. For more examples, cf. Reiter (2010a).

References

Ahn, S., G. Kaplan, B. Moll, T. Winberry, and C. Wolf (2018). When Inequality Matters for Macro and Macro Matters for Inequality. *NBER Macroeconomics Annual 32*(1), 1–75.

Antoulas, A. C. (2005). *Approximation of Large-Scale Dynamical Systems*. SIAM.

Auclert, A., B. Bardóczy, M. Rognlie, and L. Straub (2021, September). Using the Sequence-Space Jacobian to Solve and Estimate Heterogeneous-Agent Models. *Econometrica 89*(5), 2375–2408.

Bhandari, A., D. Evans, M. Golosov, and T. J. Sargent (2021). Inequality, Business Cycles, and Monetary-Fiscal Policy. *Econometrica 89*(6), 2559–2599.

Bilal, A. (2023). Solving heterogeneous agent models with the master equation. Manuscript,Harvard.

Boppart, T., P. Krusell, and K. Mitman (2018). Exploiting MIT shocks in heterogeneous-agent economies: the impulse response as a numerical derivative. *Journal of Economic Dynamics and Control 89*(C), 68–92.

Chang, Y. and S.-B. Kim (2007). Heterogeneity and aggregation: Implications for labor-market fluctuations. *American Economic Review 97*(5), 1939–1956.

Childers, D. (2018). Solution of Rational Expectations Models with Function-Valued States. Working Paper.

Den Haan, W. J. (1997). Solving dynamic models with aggregate shocks and heterogenous agents. *Macroeconomic Dynamics 1*, 355–86.

Dotsey, M., R. G. King, and A. L. Wolman (1999). State-dependent pricing and the general equilibrium dynamics of money and output. *The Quar- terly Journal of Economics 114*(2), 655–690.

Fernandez-Villaverde, J. and J. F. Rubio-Ramirez (2006). Solving DSGE models with perturbation methods and a change of variables. *Journal of Economic Dynamics and Control 30*(12), 2509–2531.

Gornemann, N., K. Kuester, and M. Nakajima (2021, May). Doves for the Rich, Hawks for the Poor? Distributional Consequences of Systematic Monetary Policy. ECONtribute Discussion Papers Series 089, University of Bonn and University of Cologne, Germany.

Grand, F. L. and X. Ragot (2022). Optimal policies with heterogeneous agents: Truncation and transitions. Manuscript.

Griewank, A. and A. Walther (2008). *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation, Second Edition*. SIAM e-books. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104).

Higham, N. (2002). *Accuracy and Stability of Numerical Algorithms* (2nd ed.). SIAM.

Jin, H.-H. and K. L. Judd (2002). Perturbation methods for general dynamic stochastic models. Stanford University.

Judd, K. L. (1998). *Numerical Methods in Economics*. Cambridge and London: MIT Press.

Judd, K. L. (2002). Perturbation methods with nonlinear changes of variables. Manuscript,Hoover Institution.

Krusell, P. and A. A. Smith (1998). Income and wealth heterogeneity in the macroeconomy. *Journal of Political Economy 106*(5), 867–96.

Kubler, F. and S. Scheidegger (2021). Uniformly self-justified equilibria: Existence and computation. Available at SSRN: https://ssrn.com/abstract=3995209.

McKay, A. and R. Reis (2016, January). The Role of Automatic Stabilizers in the U.S. Business Cycle. *Econometrica 84*, 141–194.

Mertens, T. M. and K. L. Judd (2017). Solving an incomplete markets model with a large cross-section of agents. ?

Press, W., B. Flannery, S. Teukolsky, and W. Vetterling (1986). *Numerical Recipes*. Cambridge University Press.

Reiter, M. (2009a). Solving heterogenous agent models by projection and perturbation. *Journal of Economic Dynamics and Control 33*(3), 649–665.

Reiter, M. (2009b). Solving heterogenous agent models by projection and perturbation. *Journal of Economic Dynamics and Control 33*(3), 649–665.

Reiter, M. (2010a). Approximate and almost-exact aggregation in dynamic stochastic heterogeneous-agent models. IHS Working Paper 258.

Reiter, M. (2010b). Solving the incomplete markets model with aggregate uncertainty by backward induction. *Journal of Economic Dynamics and Control 34*(1), 28–35.

Reiter, M., T. Sveen, and L. Weinke (2013). Lumpy investment and the monetary transmission mechanism. *Journal of Monetary Economics 60*(7), 821–834.

Rendahl, P. (2017). Linear time iteration.

Schmitt-Grohé, S. and M. Uribe (2004). Solving dynamic general equilibrium models using a second-order approximation to the policy function. *Journal of Economic Dynamics and Control 28*, 755–75.

Sims, C. A. (2001). Solving linear rational expectations models. *Computational Economics 20*(1-2), 1–20.

Takahashi, S. (2014, April). Heterogeneity and Aggregation: Implications for Labor-Market Fluctuations: Comment. *American Economic Review 104*(4), 1446–60.

Winberry, T. (2018). A toolbox for solving and estimating heterogeneous agent macro models. Quantitative Economics, forthcoming.

Young, E. R. (2010). Solving the incomplete markets model with aggregate uncertainty using the Krusell-Smith algorithm and non-stochastic simulations. *Journal of Economic Dynamics and Control 34*(1), 36–41.