Swapping the Nested Fixed Point Algorithm: A Class of Estimators for Discrete Markov Decision Models, Econometrica, 2002

Wenzhi Wang *

June 11, 2023

1. Introduction

We build on Hotz and Miller's work to show that the original fixed point problem in "value function space" can be reformulated as a fixed point problem in "probability space." That is, for any value of the structural parameters θ we prove that the vector of conditional choice probabilities P_θ associated with the solution of the dynamic programming problem can be obtained as the unique fixed point of a mapping in probability space: $P_\theta = \Psi_\theta(P_\theta)$. Moreover, at the fixed point P_θ the Jacobian matrix of Ψ_θ is always zero: $\partial\Psi_\theta(P_\theta)/\partial P = 0$. The policy iteration mapping $\Psi_\theta(P)$ is the cornerstone of our estimation procedure and the "zero Jacobian property" is the key to its finite sample and asymptotic properties. In our nested pseudo-likelihood algorithm (NPL) the inner algorithm maximizes in θ a pseudo-likelihood function based on choice probabilities $\Psi_\theta(P)$ where P is an estimate of choice probabilities. The outer algorithm is a fixed point algorithm that computes $\Psi_\theta(P)$ at the current parameter estimates to update the estimate of P. When the NPL algorithm is initialized with consistent nonparametric estimates of conditional choice probabilities, successive iterations return a sequence of estimators of the structural parameters that we call K-stage policy iteration (PI) estimators.

2. Discrete Markov Decision Processes

2.1. Definitions and Assumptions

There are two types of variables: the vector of state variables s and a control variable a that belongs to a finite set of mutually exclusive choice alternatives $A = \{1, 2, \ldots, J\}$. Time is discrete and it is indexed by t. At each period t an agent observes s_t and chooses a_t in order to maximize the

*This note is written during my Mphil study at the University of Oxford.

expected sum of current and future discounted utilities. Future values of some state variables are uncertain for the agent. His beliefs about uncertain future states can be represented by a Markov transition probability $p(s_{t+1} \mid s_t, a_t)$. The time horizon of the decision problem is infinite. Utility is time separable and $u(s_t, a_t)$ represents the one-period utility function. The parameter $\beta \in (0,1)$ is the time discount factor.

Under certain regularity conditions, Blackwell's theorem establishes the following properties. First, there exists a stationary, Markovian optimal decision rule $\delta(s_t)$: the decision at period t is the same as the decision at period $t+j$ if $s_t = s_{t+j}$. Therefore, we omit the subindex t for the rest of this section and we use s' to denote the vector of next period's state variables. Second, $\delta(s) = \text{argmax}_{a \in A} \{u(s,a) + \beta \int V(s') p(ds' \mid s, a)\}$, where the value function $V(\cdot)$ is the unique solution of the Bellman equation:

$$V(s) = \max_{a \in A} \left\{ u(s,a) + \beta \int V(s') p(ds' \mid s, a) \right\}. \tag{1}$$

We assume that the researcher knows the utility and the transition probability functions up to a vector of parameters θ. From an econometric point of view, we distinguish two types of state variables: $s = (x, \epsilon)$. The subvector x groups variables that are observed by both the agent and the researcher, whereas the subvector ϵ includes those state variables that are observed only by the agent. Given data on observable state variables and the actual choices made by agents, our goal is to obtain an estimate of θ.

Assumption 1. (Additivity) The one period utility function is additively separable in the observable and unobservable components: $u(s,a) = u(x,a) + \epsilon(a)$, where $\epsilon(a)$ is the ath component of the $J \times 1$ vector ϵ. The support of $\epsilon(a)$ is the real line for all a.

Assumption 2. (Conditional Independence) The transition probability of the state variables factors as $p(x', \epsilon' \mid x, a, \epsilon) = g(\epsilon' \mid x') f(x' \mid x, a)$, where $g(\cdot)$ has finite first moments and is continuous and twice differentiable in ϵ'.

Assumption 3. (Finite Domain of Observable State Variables) $x \in X = (x^1, \ldots, x^M)$.

We can exploit Assumptions 1-2 to obtain versions of the value functions and the Bellman operator in which the unobservables are integrated out. These versions will prove more useful than equation (1) in the analysis of the estimation problem. Let $V_\sigma(x)$ denote the expectation of the value function conditional on the state variables x (called integrated value function or ex ante value function): $V_\sigma(x) \equiv \int V(x,\epsilon) g(d\epsilon \mid x)$, where σ represents parameters that characterize the distribution of the ϵ's. Under Assumptions 1-3, $V_\sigma(x)$ solves the smooth Bellman equation:

$$V_\sigma(x) = \int \max_{a \in A} \left[u(x,a) + \epsilon(a) + \beta \sum_{x'} V_\sigma(x') f(x' \mid x, a) \right] g(d\epsilon \mid x). \tag{2}$$

The smoothed Bellman operator $\Gamma_\sigma(\cdot)$ defined by the RHS of this functional equation is a contraction mapping. It follows that V_σ is its unique fixed point: $V_\sigma = \Gamma_\sigma(V_\sigma)$.

The conditional choice probability $P(a \mid x)$ is the probability that alternative a is the optimal choice given the vector of observable state variables x:

$$P(a \mid x) = \int \mathbb{1}\left(a = \underset{j \in A}{\operatorname{argmax}} \left[v\left(x, j\right) + \epsilon\left(j\right)\right]\right) g(d\epsilon \mid x) \tag{3}$$

where $v(x,a)$ is the choice-specific value function $u(x,a) + \beta \sum_{x'} V_\sigma\left(x'\right) f\left(x' \mid x, a\right)$.

2.2. From Conditional Choice Probabilities to Value Functions

First, it can be shown that the choice probabilities conditional on any value of x are uniquely determined by the vector of normalized value functions or utility differencces $(\widetilde{v}(x,a) : a > 1)$, where $\widetilde{v}(x,a) \equiv v(x,a) - v(x,1)$. That is, there exists a vector mapping $Q_x(\cdot)$ such that

$$(P(a \mid x) : a > 1) = Q_x((\widetilde{v}(x,a) : a > 1)),$$

where, without loss of generality, we exclude the probability of alternative one. For instance, if unobservables have independent-across-alternatives extreme value distributions, the j-th component of this function takes the well known logistic form:

$$Q_x^j(\widetilde{v}(x,a)) = \frac{\exp\left(\widetilde{v}(x,j)/\sigma\right)}{1 + \sum_{a=2}^{J} \exp\left(\widetilde{v}(x,a)/\sigma\right)}.$$

A general representation of the mapping $Q_x(\cdot)$ can be obtained from the social surplus function:

$$S((v(x,a), a \in A), x) = \int \max_{a \in A} \left[v(x,a) + \epsilon(a)\right] g(d\epsilon \mid x). \tag{4}$$

The social surplus function computes expected utility, in the multinomial choice-random utility framework, as a function of the set of choice-specific utilities. Clearly, $V_\sigma(x) = S((v(x,a), a \in A), x)$. The Williams-Daly-Zachary (WDZ) theorem establishes the following properties of the social surplus function[1]: (i) it is strictly convex; (ii) it is additive: for any scalar $\alpha, S(\alpha + \{v(x,a)\}, x) = \alpha + S(\{v(x,a)\}, x)$; and (iii) its gradient is equal to the vector of conditional choice probabilities: $P(a \mid x) = \partial S(\{v(x,a)\}, x)/\partial v(x,a)$. Using properties (ii) and (iii) we obtain the representation $Q_x^j(\{\widetilde{v}(x,a) : a > 1\}) = \partial S([0, \{\widetilde{v}(x,a) : a > 1\}], x)/\partial v(x,j)$

Second, Hotz-Miller's invertibility proposition states that continuous differentiability and the strict convexity of the social surplus function imply that, for every x, $Q(\cdot)$ is invertible.

Third, Hotz and Miller also showed how invertibility can be exploited to obtain alternative representations of the value function in terms of choice probabilities. We now restate one such represen-

[1] See Rust (1994b, Structural Estimation of Markov Decision Processes, Theorem 3.1, and references therein). The strict convexity of the social surplus function is actually a strengthening of the WDZ theorem which follows from the unbounded support and continuity of the distribution of unobservables.

tation in our framework. First, notice that the Bellman equation (2) can be rewritten as

$$V_\sigma(x) = \sum_{a\in A} P(a \mid x) \left\{ u(x,a) + \mathbb{E}\left[\epsilon(a) \mid x, a\right] + \beta \sum_{x'} f\left(x' \mid x, a\right) V_\sigma\left(x'\right) \right\} \tag{5}$$

where $\mathbb{E}\left[\epsilon(a) \mid x, a\right]$ is the expectation of the unobservable $\epsilon(a)$ conditional on the optimal choice of alternative a:

$$\mathbb{E}\left[\epsilon(a) \mid x, a\right] = \left[P(a \mid x)\right]^{-1} \int \epsilon(a) \mathbb{1}\left(\widetilde{v}(x,a) + \epsilon(a) \geq \widetilde{v}(x,j) + \epsilon(j), \forall\, j \in A\right) g(d\epsilon \mid x). \tag{6}$$

Clearly, the conditional expectations $\mathbb{E}\left[\epsilon(a) \mid \cdot\right]$ are functions of the utility differences $\widetilde{v}(x,a)$. Since the mapping Q_x from utility differences into choice probabilities is invertible, it follows that these conditional expectations can be expressed as functions of the choice probabilities. We denote these functions by $e_x\left(a, \left\{P\left(j \mid x\right)\right\}\right)$. In the case of the extreme value unobservables they have the closed form $e_x\left(a, \left\{P\left(j \mid x\right)\right\}\right) = \gamma - \ln\left(P(a \mid x)\right)$ where γ is Euler's constant. Let's substitute these functions into (5) and stack M equations for each possible value of x. In compact matrix notation we get

$$V_\sigma = \sum_{a\in A} P(a) * \left[u(a) + e(a,P) + \beta F(a) V_\sigma\right] \tag{7}$$

where $*$ is the element-by-element product; V_σ is the $M \times 1$ vector describing the value function $V_\sigma(x)$; P is the $M(J-1) \times 1$ vector of conditional choice probabilities, alternative one excluded; $P(a), u(a)$, and $e(a,P)$ are $M \times 1$ vectors that stack the corresponding elements at all states for alternative a; and $F(a)$ is the $M \times M$ matrix of conditional transition probabilities $f\left(x' \mid x, a\right)$. The system of fixed point equations can be solved for the value function to obtain V_σ as a function of P:

$$V_\sigma = \phi(P) = \left(I_M - \beta F^U(P)\right)^{-1} \left\{ \sum_{a\in A} P(a) * \left[u(a) + e(a,P)\right] \right\} \tag{8}$$

where $F^U(P) = \sum_{a\in A} P(a) * F(a)$ is the $M \times M$ matrix of unconditional transition probabilities induced by P.

2.3. The Fixed Point Problem in Probability Space

In the preceding subsections we defined conditional choice probabilities in terms of the value function in equation (**??**) and we showed that the value function can be written in terms of conditional choice probabilities in equation (8). Substituting equation (8) into equation (**??**) for all choices and states we get:

$$P = \Psi(P) \equiv \Lambda\left(\phi(P)\right). \tag{9}$$

$\phi(\cdot)$ is a policy valuation operator that maps an $M(J-1) \times 1$ vector of conditional choice probabilities into a $M \times 1$ vector in value function spacing using Hotz and Miller's representation. $\Lambda(\cdot)$ is a policy improvement operator that maps an $M \times 1$ vector in value function space into a $M(J-1) \times 1$ vector of conditional choice probabilities. It stacks all choice probabilities associated with the value function, as defined in equation (**??**).

The policy iteration operator Ψ can be evaluated at any vector of conditional choice probabilities, optimal or not. For an arbitrary vector of choice probabilities P^0, the valuation operator $\phi(P^0)$ returns the value function corresponding to the arbitrary behavior represented by P^0. For an arbitrary vector of values, say V_σ^0, the policy improvement mapping $\Lambda(V_\sigma^0)$ returns the optimizing agent's choice probabilities under the assumption that expected utilities as of next period are given by the vector V_σ^0. Thus the composite mapping $\Psi(P^0)$ should be interpreted as giving the current optimal choice probabilities of an agent whose future behavior will be to randomize over alternatives according to P^0.

Notice that by the definitions in equations (**??**) and (8) we have that $V_\sigma = \phi(P)$ and $P = \Lambda(V_s)$; therefore, it is clear that the set of optimal choice probabilities P is a fixed point of Ψ. Thus, the original fixed point problem in "value space" can be reformulated as a fixed point problem in "probability space." The following prepositions establish the relationship between the two fixed point problems and an important property of the policy iteration operator.

Proposition 1. Under Assumptions 1-3:

1. Ψ has a unique fixed point P.
2. The sequence $P^K = \Psi(P^{K-1}), K = 1, \ldots, \infty$, converges to P for any P^0.
3. Equivalence of Ψ and Newton iterations: For any P^0, consider the pair of linked sequences $\{V_s^K, P^K\}$ defined by $V_\sigma^K = \phi(P^K), P^{K+1} = \Lambda(V_\sigma^K)$. Clearly, $P^K = \Psi(P^{K-1})$. Then, $\{V_\sigma^K\}$ is the sequence of Newton iterations converging to the unique solution of the Bellman equation (2).

Proposition 2. Under Assumptions 1-3, the Jacobian matrices of $\phi(\cdot)$ and $\Psi(\cdot)$ are zero at the fixed point P.

Proposition 2 establishes that at the fixed point it is not possible to increase expected utility by changing choice probabilities; that is, the optimal choice probabilities maximize the valuation operator locally. As a consequence, the Jacobian of the policy iteration operator is zero. This result is the key to the properties of the algorithm and the sequential estimators we propose in this paper.

3. Maximum Likelihood Estimation and Nested Algorithms

Let θ_u, θ_g, and θ_f be the vectors of unknown parameters in the utility function u, the density of unobservable g, and the conditional transition probability function f, respectively. That is, $\theta =\equiv (\theta_u, \theta_g, \theta_f)$. In order to guarantee the existence, consistency, and asymptotic normality of the ML estimator, we impose smoothness of the primitives with respect to θ.

Assumption 4. $u_{\theta_u}(x,a), g_{\theta_g}(\epsilon \mid x)$, and $f_{\theta_f}(x' \mid x, a)$ are continuus and twice differentiable with respect to θ.

Suppose our data set consists of a cross-section of observations from a random sample of individuals $\{x_i, a_i, x_i' : i = 1, \ldots, n\}$. Under Assumption 2, the log-likelihood function of the model can be

decomposed into conditional choice probability and transition probability terms as follows:

$$l(\theta) = l_1(\theta) + l_2\left(\theta_f\right) = \sum_{i=1}^{n} \ln P_\theta\left(a_i \mid x_i\right) + \sum_{i=1}^{n} \ln f_{\theta_f}\left(x_i' \mid x_i, a_i\right). \tag{10}$$

Consistent estimates of the conditional transition probability parameters θ_f can be obtained from transition data without having to solve the Markov decision model. In the rest of the paper we focus on the estimation of $\alpha \equiv \left(\theta_u, \theta_g\right)$ given initial consistent estimates of θ_f obtained from likelihood $l_2\left(\theta_f\right)$.

Let α^* denote the true value of α hereafter. The MLE of α^* can be computed using Rust's well known nested fixed point algorithm (NFXP). In this procedure, an 'inner' fixed point algorithm computes the conditional choice probabilities $P_\theta = \Psi(P_\theta)$ and their derivatives for given parameter values. The 'outer algorithm' feeds on this solution and maximizes the likelihood with respect to α using the BHHH method. We propose an alternative nested procedure:

Nested Pseudo Likelihood Algorithm (NPL):

Let $\hat{\theta}_f$ be an estimate of θ_f. Start with an initial guess for the conditional choice probabilities, $P^0 \in [0,1]^{MJ}$. At iteration $K \geq 1$, applying the following steps:

Step 1: Obtain a new pseudo-likelihood estimat of α, α^K, as

$$\alpha^K = \arg\max_{\alpha \in \Theta} \sum_{i=1}^{n} \ln \Psi_{\left(\alpha, \hat{\theta}_f\right)}\left(P^{K-1}\right)\left(a_i \mid x_i\right) \tag{11}$$

where $\Psi_\theta(P)(a \mid x)$ is the element (a,x) of $\Psi_\theta(P)$.

Step 2: Update P using the 'arg max' from step 1, i.e.,

$$P^K = \Psi_{\left(\alpha^K, \hat{\theta}_f\right)}\left(P^{K-1}\right). \tag{12}$$

Iterate in K until convergence in P (and α) is reached.

In our nested procedure we swap the order of the two algorithms. That is, the outer algorithm (step 2) iterates on Ψ to solve the fixed point problem, and the inner algorithm (step 1) maximizes a pseudo-likelihood function. The NFXP algorithm always converges to a root of the likelihood equations. We show that NPL satisfies a weaker version of the same property.

Proposition 3. (Equivalence of NFXP and NPL) Suppose the pseudo-likelihood maximization problems in (11) have unique interior solutions for any sample and any value of P. Then, if NPL converges, it does so to a root of the likelihood equations.

Example 1. Consider a class of models where: (i) the unobservables ϵ's have independent across alternatives, extreme value distributions; and (ii) there is multiplicative separability between x and θ_u in the utilities, i.e., $u_{\theta_u}(x,a) = h(x,a)'\eta(\theta_u)$ where $h(x,a)$ and $\eta(\theta_u)$ are known vector-valued functions with dimension p. Define $H(a)$ as the $M \times p$ matrix with rows $h(x,a)$ for each value of x.

In this case, the policy/iteration operator is

$$\Psi_{(\alpha,\theta_f)}(P)(a \mid x) = \frac{\exp\left\{\widetilde{h}_{(\theta_f,P)}(x,a)'\eta(\theta_u) + \widetilde{e}_{(\theta_f,P)}(x,a)\right\}}{\sum_{j=1}^{J}\exp\left\{\widetilde{h}_{(\theta_f,P)}(x,j)'\eta(\theta_u) + \widetilde{e}_{(\theta_f,P)}(x,j)\right\}}$$

where $\widetilde{h}_{(\theta_f,P)}(x,a)'$ is a row of the matrix

$$H(a) + \beta F(a)\left(I_M - \beta F^U_{\theta_f}(P)\right)^{-1}\sum_{j=1}^{J} P(j) * H(j),$$

and $\widetilde{e}_{(\theta_f,P)}(x,a)$ is an element of the vector

$$\beta F(a)\left(I_M - \beta F^U_{\theta_f}(P)\right)^{-1}\sum_{j=1}^{J} P(j) * [\gamma - \ln P(j)].$$

This model has several features that make the use of NPL specially advantageous. First and most important, it is very convenient that the extreme value assumption gives closed-form conditional expectation functions $e\big(j,P\big) = \gamma - \ln\big(P\big(j\big)\big)$. In general computing $e(\cdot)$ would involve first inverting $Q_x(\cdot)$ mappings to obtain utility differences and then solving the multiple integration problem in (6) given utility differences. For distributions other than the extreme value this may be a serious computational problem. Second, the extreme value assumption also implies that integration over unobservables during pseudo-likelihood estimation has a simple (logistic) closed form. Third, since $\left\{\widetilde{h}_{(\theta_f,P)}(x,a)\right\}$ and $\left\{\widetilde{e}_{(\theta_f,P)}(x,a)\right\}$ do not depend on α, they are fixed over a pseudo-likelihood estimation. To obtain $\Psi_{(\alpha,\theta_f)(P)(a|x)}$ for different values of α we do not have to repeat the inversion and multiplication of large matrices that is required for policy valuation. And fourth, the pseudo-likelihood function is globally concave in $\eta(\theta_u)$, which guarantees convergence of the hill-climbing pseudo-likelihood iterations for any initial value of θ_u. In contrast, to compute the probabilities $P_{\alpha,\theta_f}(a \mid x)$ that enter the likelihood function we have to invert and multiply $M \times M$ matrices repeatedly in policy iteration. Furthermore, the likelihood function is not globally concave in α, nor in a transformation of α. Therefore, convergence of NFXP's outer BHHH algorithm may require the use of optimal steps with a significant increase of the computational cost of estimation.

Finally, note that Proposition 2 is crucial to obtain equivalence of NFXP and NPL. Based on this, it is straightforward to see that the equivalence result also holds for full likelihood versions of NFXP and NPL. Also note that finite horizon models are covered in our infinite horizon framework if the decision period t is included among the observable state variables.

4. Sequential Policy Iteration Estimators

Let $\widehat{\theta}_f$ denote a consistent estimator of conditional transition probability parameters, and let P^0 be a consistent, nonparametric estimator of the true conditional choice probabilities P^*. Consider

using P^0 as an initial guess in our NPL algorithm. Performing one, two, and in general K iterations of the NPL algorithm yields a sequence $\{\widehat{\alpha}^1, \widehat{\alpha}^2, \ldots, \widehat{\alpha}^K\}$ of statistics that can be used as estimators of α^*. We call them *sequential policy iteration (PI) estimators*. Thus, for $K \geq 1$, the K-stage PI estimator is defined as:

$$\widehat{\alpha}^K = \arg\max_{\alpha \in \Theta} \sum_{i=1}^{n} \ln \Psi_{(\alpha, \tilde{\theta}_f)}\left(P^{K-1}\right)(a_i \mid x_i) \tag{13}$$

where $P^K = \Psi_{(\alpha, \tilde{\theta}_f)}\left(P^{K-1}\right)$, and P^0 is a consistent, nonparametric estimator of the true conditional choice probabilities P^*.

In this section, we study the asymptotic statistical properties of this sequence of estimators. The main result is Proposition 4, which shows that for any value of K the PI estimators are consistent and asymptotically equivalent to the partial MLE of α^*.

Proposition 4. Left for future study!