Quantitative Economics 10 (2019), 43–65 1759-7331/20190043

Strong convergence and dynamic economic models

Robert L. Bray
Operations Department, Kellogg School of Management at Northwestern University

Morton and Wecker (1977) stated that the value iteration algorithm solves a dynamic program's policy function faster than its value function when the limiting Markov chain is ergodic. I show that their proof is incomplete, and provide a new proof of this classic result. I use this result to accelerate the estimation of Markov decision processes and the solution of Markov perfect equilibria.

Keywords. Markov decision process, Markov perfect equilibrium, strong convergence, relative value iteration, dynamic discrete choice, nested fixed point, nested pseudo-likelihood.

JEL classification. C01, C13, C15, C61, C63, C65.

1. Introduction

I present a simple refinement to speed up the estimation of ergodic Markov decision processes. I exploit three facts:

1. The empirical likelihood of a dynamic model depends only on the dynamic program's policy function, not its value function.

2. The policy function depends only on the value function's relative differences, not its absolute level.

3. The value function's relative differences converge faster than its level under repeated Bellman contractions when the underlying stochastic process is ergodic. This is called strong convergence.

Morton and Wecker (1977) discovered this strong convergence property, but their proof is wanting. At one point, they implicitly replace a close-to-optimal policy with the optimal policy, which is unwarranted. I provide a new proof, based on a straightforward envelope theorem argument.

2. Markov decision process

I consider a dynamic program with discrete time periods, an infinite planning horizon, a finite state space, and an uncountable, compact action space. Taking a specific action in a specific state yields a specific utility. The goal is to determine the actions that yield the

Robert L. Bray: r-bray@kellogg.northwestern.edu
I would like to thank Victor Aguirregabiria, Achal Bassamboo, Thomas Bray, Seyed Emadi, Soheil Ghili, Arvind Magesan, John Rust, Dennis Zhang, and three anonymous referees for their helpful comments.

 https://doi.org/10.3982/QE833

maximum expected discounted utility. The following list defines the Markov decision process:

1. ι is the length-m vector of ones.
2. δ_i is the length-m unit vector indicating the ith position.
3. $\mathbb{a} \subset \mathbb{R}^\ell$ is the action space.
4. $\mathbb{x} \equiv \{x_1, \ldots, x_m\}$ is the state space.
5. $\beta \in [0, 1)$ is the discount factor.
6. $U : \mathbb{x} \to \mathbb{a}$ is a generic policy function that maps states to actions. Policy U specifies taking action $U(x_i)$ in state x_i.
7. $\mathbb{U} \equiv \mathbb{a}^{\mathbb{x}}$ is the set of permissible policy functions: if $U \in \mathbb{U}$ and $x_i \in \mathbb{x}$ then $U(x_i) \in \mathbb{a}$.
8. $\pi : \mathbb{U} \to \mathbb{R}^m$ is a continuous and uniformly bounded function that maps policy functions to length-n flow utility vectors: the ith element of $\pi(U)$ is the flow utility received from taking action $U(x_i)$ in state x_i.
9. $Q : \mathbb{U} \to \mathbb{R}^{m \times m}$ is a continuous function that maps policy functions to $m \times m$ stochastic matrices: the ijth element of $Q(U)$ is the probability of transitioning from state x_i to state x_j under action $U(x_i)$.
10. $V \in \mathbb{R}^m$ is a vector that characterizes a generic value function: the ith element of V denotes the expected discounted flow utility from state x_i.
11. $T_U : \mathbb{R}^m \to \mathbb{R}^m$ is the Bellman contraction operator:

(a) $T_U V \equiv \pi(U) + \beta Q(U)V$ characterizes the value of following policy U this period, given value function V next period;

(b) $T_U^n V \equiv T_U(T_U^{n-1}V) = (\sum_{t=0}^{n-1} \beta^t Q(U)^t \pi(U)) + \beta^n Q(U)^n V$ characterizes the value of following policy U for n periods, given value function V thereafter; and

(c) $T_U^\infty \equiv \lim_{n\to\infty} T_U^n V = \sum_{t=0}^{\infty} \beta^t Q(U)^t \pi(U) = (I - \beta Q(U))^{-1}\pi(U)$ characterizes the value of following policy U forever.

12. $\mathscr{U} : \mathbb{R}^m \to \mathbb{U}$ is the policy update function: $\mathscr{U}(V) \equiv \arg\max_{U\in\mathbb{U}} \iota' T_U V$.[1]
13. $T : \mathbb{R}^m \to \mathbb{R}^m$ is the value iteration operator:

(a) $TV \equiv T_{\mathscr{U}(V)}V$ characterizes the maximum value this period, given value function V next period; and

(b) $T^n V \equiv T(T^{n-1}V)$ characterizes the maximum value this period, given value function V in n periods.

14. $V^* \in \mathbb{R}^m$ is the optimal value function, implicitly defined as the unique fixed-point solution to Bellman's equation: $V^* \equiv TV^*$. The ith element of V^* is the maximum expected discounted flow utility from state x_i.

[1] To simplify the notation, I assume that each value function corresponds to a unique optimal policy. Note that maximizing the sum of the value function maximizes each element of the value function.

15. $U^* \in \mathbb{U}$ is the optimal policy function, defined as the policy corresponding to the optimal value function: $U^* \equiv \mathscr{U}(V^*)$.

16. $\|\cdot\|_1$ is the ℓ_1 norm: $\|x\| \equiv \sum_{i=1}^m |\delta_i' x|$.

17. $\|\cdot\|$ is the ℓ_∞ norm: $\|x\| \equiv \max_{i=1}^m |\delta_i' x|$.

18. $|||\cdot|||$ is the matrix norm induced by the ℓ_∞ norm: $|||A||| \equiv \sup\{\|Ax\| : \|x\| = 1\} = \max_{i=1}^m \|\delta_i' A\|_1$.

19. $\mathrm{Cond}(A) \equiv |||A|||\,|||A^{-1}|||$ is the condition number of matrix A (see Judd (1998, p. 67)).

20. $\lambda(Q)$ is the second-largest eigenvalue modulus of matrix Q.

21. $\psi(Q) \equiv \lim_{n\to\infty} \delta_1' Q^n$ is the stationary distribution associated with stochastic matrix Q.

22. $\Delta \equiv I - \iota\delta_1'$ is the $m \times m$ difference operator. Pre-multiplying a length-m vector by Δ subtracts the first element from every element: $\Delta[x_1, \ldots, x_m]' = [x_1 - x_1, \ldots, x_m - x_1]'$.

3. Relative value iteration

3.1 *Algorithm*

The following proposition establishes that the policy function only depends on the relative value function, ΔV.

Proposition 1 (White (1963)). *Differencing the value function does not affect the corresponding policy function*: $\mathscr{U}(\Delta V) = \mathscr{U}(V)$.

This result is intuitive: Changing the value function from V to ΔV is equivalent to reducing next period's flow utility by $\delta_1' V$. This utility loss is independent of this period's action, and thus does not affect this period's action.

The relative value iteration algorithm exploits Proposition 1. It proceeds as follows:[2]

1. `Initialize` $n := 0 \in \mathbb{R}$ `and` $V_0 := 0 \in \mathbb{R}^m$.
2. `Increment` n.
3. `Set` $V_n := \Delta T V_{n-1} = (\Delta T)^n V_0 = \Delta T^n V_0$.
4. `If` $\|V_n - V_{n-1}\| \geq \varepsilon(1-\beta)/(2\beta)$ `go to 2; otherwise go to 5`.
5. `Return` $\mathscr{U}(V_n)$.

The ε in step 4 specifies the convergence tolerance. It is usually impossible to calculate U^* exactly, so we must make do with an ε-optimal policy—a policy that yields within ε of the optimal value when followed forever (Puterman (2005)). The following proposition establishes that relative value iteration yields an ε-optimal policy.

Proposition 2 (Bray (2018)). *The relative value iteration algorithm always returns an ε-optimal policy after a finite number of iterations.*

[2]The equivalence of $(\Delta T)^n V_0$ and $\Delta T^n V_0$ follows from Lemma 4 in the Appendix.

3.2 *Strong convergence*

The only difference between the relative value iteration algorithm and the traditional value iteration algorithm is the presence of Δ in step 3: whereas traditional value iteration sets $V_n := T^n V_0$, relative value iteration sets $V_n := \Delta T^n V_0$. This subtle change reduces solution times, as the following proposition indicates.

PROPOSITION 3 (Morton and Wecker (1977)). *Whereas $\|T^n V - V^*\|$ is $O(\beta^n)$ as $n \to \infty$, $\|\Delta T^n V - \Delta V^*\|$ is $O(\beta^n \gamma^n)$ for all $\gamma > \lambda(Q(U^*))$. Thus, if the Markov chain is ergodic under policy U^* then $\lambda(Q(U^*)) < 1$ and relative value iteration converges strictly faster than traditional value iteration.*

I prove Proposition 3 in the Appendix. Morton and Wecker (1977) first articulated this result, but they did not comprehensively prove it. Specifically, they failed to prove their seventh theorem, which they claimed to be "A slight modification of Theorem 2." But this is not true. Their second theorem requires the set $\{Q(\mathscr{U}(T^n V)) : n \geq N\}$ to be "strongly ergodic of order λ" for some finite N; their seventh theorem invokes this result, but with $N = \infty$, which is not allowed. Setting $N = \infty$ quashes a crucial aspect of the problem: the policy function's restlessness. Indeed, the difference between $N < \infty$ and $N = \infty$ is the difference between a dynamic policy, which changes after every Bellman contraction, and a static policy, which always equals U^*. Setting $N = \infty$ implicitly replaces value iteration operator T with Bellman contraction operator T_{U^*}. But the analysis is trivial under T_{U^*} because $\Delta T_{U^*}^n V - \Delta V^* = \beta^n \Delta Q(U^*)^n (V - V^*)$ and $\Delta Q(U^*)^n$ is $O(\lambda(Q(U^*))^n)$. Unfortunately, T is not as tractable; for example,

$$
\begin{aligned}
T^1 V &= T_{\mathscr{U}(V)} V, \\
T^2 V &= T_{\mathscr{U}(T_{\mathscr{U}(V)} V)} T_{\mathscr{U}(V)} V, \\
T^3 V &= T_{\mathscr{U}(T_{\mathscr{U}(T_{\mathscr{U}(V)} V)} T_{\mathscr{U}(V)} V)} T_{\mathscr{U}(T_{\mathscr{U}(V)} V)} T_{\mathscr{U}(V)} V, \quad \text{and} \\
T^4 V &= T_{\mathscr{U}(T_{\mathscr{U}(T_{\mathscr{U}(T_{\mathscr{U}(V)} V)} T_{\mathscr{U}(V)} V)} T_{\mathscr{U}(T_{\mathscr{U}(V)} V)} T_{\mathscr{U}(V)} V)} \\
&\qquad \times T_{\mathscr{U}(T_{\mathscr{U}(T_{\mathscr{U}(V)} V)} T_{\mathscr{U}(V)} V)} T_{\mathscr{U}(T_{\mathscr{U}(V)} V)} T_{\mathscr{U}(V)} V.
\end{aligned}
$$

In general, V influences $T^n V$ in 2^n distinct ways—every Bellman contraction *doubles* the number of communication channels between the terminal value and the current value. I must show that the latter dissociates from the former despite this pathway doubling. Morton and Wecker overlooked this complication.

I prove Proposition 3 in a new way:

1. I use the envelope theorem to establish that

$$
\begin{aligned}
\frac{\partial}{\partial V} T^t V &= \frac{\partial}{\partial V} T_{p_t} \cdots T_{p_1} V |_{p_1 = \mathscr{U}(T^0 V), \ldots, p_t = \mathscr{U}(T^{t-1} V)} \\
&= \beta^t Q\big(\mathscr{U}\big(T^{t-1} V\big)\big) \cdots Q\big(\mathscr{U}\big(T^0 V\big)\big).
\end{aligned}
$$

This identity indicates that only one of the 2^t avenues by which V can affect $T^t V$ matters when V is near V^*.

2. I use the binomial theorem to establish that

$$\left\|\left\|Q\left(\mathscr{U}\left(T^{s+t-1}V\right)\right)\cdots Q\left(\mathscr{U}\left(T^{s}V\right)\right)-Q\left(U^{*}\right)^{t}\right\|\right\|<\varepsilon,$$

for sufficiently large s.

3. I use the Jordan normal form of $Q(U^*)$ to establish that $\|\|\Delta Q(U^*)^t\|\| \leq (\lambda(Q(U^*)) + \varepsilon)^t$, for sufficiently large t.

4. I use points 1, 2, and 3 to establish that

$$\left\|\Delta T^{s+2t}V-\Delta T^{s+t}V\right\|\leq\beta^{t}\left(\lambda\left(Q\left(U^{*}\right)\right)+\varepsilon\right)^{t}\left\|\Delta T^{s+t}V-\Delta T^{s}V\right\|,$$

for sufficiently large s and t.

5. I use point 4 to establish that $\|\Delta T^{s+jt}V - \Delta V^*\|$ is $O(\beta^{jt}(\lambda(Q(U^*)) + \varepsilon)^{jt})$ as $j \to \infty$, for sufficiently large s and t.

This proof formalizes a simple intuition: the relative value function converges faster because it depends on fewer utilities. Whereas the total value function depends on all payoffs not discounted to irrelevance, the relative value function depends only on the payoffs received before the state variables revert back to their limiting distribution; the payoffs received thereafter contribute to the value of each state evenly, and thus wash out upon differencing. Accordingly, the relative value function converges not at the rate of discounting, but at the rate of discounting times the rate at which the state variables revert to their stationary distribution. And the rate at which the state variables revert to their stationary distribution is $\lambda(Q(U^*))$.

I will illustrate with Bray et al.'s (2018) empirical inventory model. In the model, a supermarket manager controls a product's inventory levels by placing daily orders at a supplying distribution center. The dynamic program has three state variables: the inventory at the store, the inventory at the distribution center, and the expected demand. Each decision period lasts one day, so the discount factor is $\beta = 0.9997$ (which corresponds to an annual discount factor of $0.9997^{365} = 0.896$). Table 1 reports the quantiles of Bray et al.'s (2018) state transition matrix spectral sub-radii. The median spectral sub-radius is $\lambda(Q(U^*)) = 0.9775$; the corresponding product is a 250 ml bottle of Lulu brand cashew milk. I will focus henceforth on this median product.

The convergence rate of the value function under value iteration depends on one factor: the rate at which future utilities are discounted, β. In general, scaling the value function error by ε requires approximately $\log(\varepsilon)/\log(\beta)$ value iteration steps. Thus, scaling the value function error by 10^{-3} requires roughly $\log(10^{-3})/\log(0.9997) \approx 23{,}000$ Bellman contractions. These Bellman contractions compute the expected value of the next $23{,}000/365 \approx 63$ years' worth of utilities.

The convergence rate of the relative value function under relative value iteration depends on two factors: the rate at which future utilities are discounted, β, and the rate at which the state variables regress to their stationary distribution, $\lambda(Q(U^*))$. In general, scaling the relative value function error by ε requires approximately $\log(\varepsilon)/\log(\beta\lambda(Q(U^*)))$ relative value iteration steps. Thus, scaling the relative value function

TABLE 1. Bray et al. (2018) estimated 246 grocery store inventory dynamic programs. I calculate each dynamic program's state transition matrix spectral sub-radius and tabulate their three quartiles (0.25, 0.5, and 0.75), two extreme deciles (0.1 and 0.9), and minimum and maximum (0 and 1), by product group. For example, the minimum $\lambda(Q(U^*))$ across detergents is 0.9018, and the median $\lambda(Q(U^*))$ across all products is 0.9775. Although some of the statistics round up to one, all of the spectral sub-radii are less than one.

	0	0.1	0.25	0.5	0.75	0.9	1
Detergent	0.9018	0.9343	0.9694	0.9814	0.9878	0.9961	1.0000
Drinks	0.9361	0.9579	0.9693	0.9879	0.9944	0.9973	0.9998
Oil/Vinegar	0.9165	0.9204	0.9305	0.9593	0.9887	0.9928	0.9964
Oral Care	0.9198	0.9225	0.9353	0.9648	0.9725	0.9991	1.0000
Shampoo	0.8996	0.9133	0.9467	0.9738	0.9820	0.9984	0.9994
Tissues	0.9381	0.9383	0.9408	0.9473	0.9562	0.9762	0.9864
Toilet Paper	0.9411	0.9512	0.9569	0.9702	0.9945	0.9992	0.9999
Total	0.8996	0.9313	0.9573	0.9775	0.9898	0.9973	1.0000

error by 10^{-3} requires roughly $\log(10^{-3})/\log(0.9997 \cdot 0.9775) \approx 300$ Bellman contractions. These Bellman contractions compute the expected value of the next $300/365 \approx 0.82$ years' worth of utilities.

We can disregard utilities received thereafter because they are, essentially, independent of the current state variables—the system "forgets" the current state after 0.82 years, making all subsequent utilities moot. Figure 1 illustrates, plotting the distribution of the store's day-t inventory from two initial conditions: in the first, the three state variables equal their first quartiles on day 0; and in the second, the three state variables equal their third quartiles on day 0. The distributions coincide by day 256; thus, the first 256 utilities account for basically all the difference between the initial conditions' valuations. And factoring these 256 utilities requires only 256 Bellman contractions.

4. RELATIVE POLICY ITERATION

4.1 *Algorithm*

As the strong convergence analog of value iteration is relative value iteration, the strong convergence analog of policy iteration is relative policy iteration. The relative policy iteration algorithm proceeds as follows:

1. `Initialize` $n := 0 \in \mathbb{R}$ `and` $V_0 := 0 \in \mathbb{R}^m$.
2. `Increment` n.
3. `Set` $U_n := \mathscr{U}(V_{n-1})$.
4. `Set` $V_n := \Delta T_{U_n}^{\infty}$.
5. `If` $\|V_n - V_{n-1}\| \geq \varepsilon(1-\beta)/(2\beta)$ `go to 2; otherwise go to 6.`
6. `Return` $\mathscr{U}(V_n)$.

The following proposition establishes that this algorithm yields an ε-optimal policy.

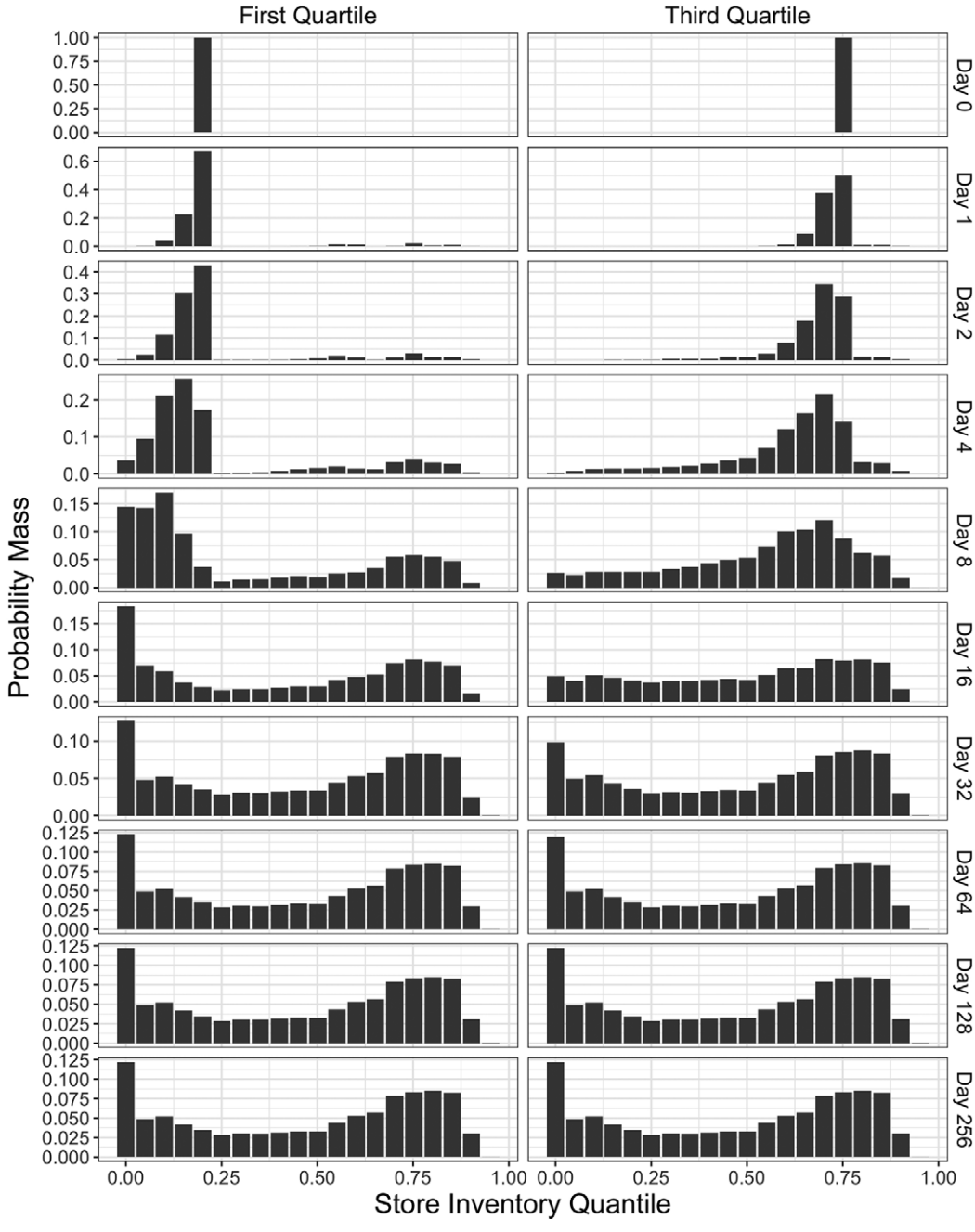

FIGURE 1. I plot the distribution of store inventories in Bray et al.'s (2018) Lulu brand cashew milk dynamic program. Specifically, I depict the day t distribution for $t \in \{0, 1, 2, 4, 8, 16, 32, 64, 128, 256\}$, given that all day-0 state variables equal their first quartiles (for the left panels) or their third quartiles (for the right panels). The distributions are essentially the same by day 256.

PROPOSITION 4 (Bray (2018)). *The relative policy iteration algorithm always returns an ε-optimal policy after a finite number of iterations.*

The only difference between relative policy iteration and traditional policy iteration is the presence of Δ in step 4: whereas traditional policy iteration sets $V_n := T^\infty_{U_n}$, relative policy iteration sets $V_n := \Delta T^\infty_{U_n}$. This Δ expedites the computation, provided the underlying Markov chain is ergodic. However, while the Δ in relative value iteration reduces the algorithm iteration count without changing the algorithm iteration difficulty, the Δ in relative policy iteration reduces the algorithm iteration difficulty without changing the algorithm iteration count. It is easier to implement a relative policy iteration step than a traditional policy iteration step because it is easier to evaluate $\Delta T^\infty_{U_n}$ than it is to evaluate $T^\infty_{U_n}$. There are three ways to calculate $T^\infty_{U_n}$ and $\Delta T^\infty_{U_n}$, and the Δ operator helps with each.

4.2 *Iterative policy evaluation*

The first way to calculate T^∞_U and ΔT^∞_U is to evaluate $T^n_U V$ and $\Delta T^n_U V$ for large values of n. The following proposition establishes that $\Delta T^n_U V$ converges to ΔT^∞_U faster than $T^n_U V$ converges to T^∞_U (when the underlying Markov chain is ergodic).

PROPOSITION 5 (Morton (1971)). *Whereas $\|T^n_U V - T^\infty_U\|$ is $O(\beta^n)$ as $n \to \infty$, $\|\Delta T^n_U V - \Delta T^\infty_U\|$ is $O(\beta^n \gamma^n)$ for all $\gamma > \lambda(Q(U))$. Thus, if the Markov chain is ergodic under policy U then $\lambda(Q(U)) < 1$ and relative policy iteration's policy evaluation step converges strictly faster than traditional policy iteration's policy evaluation step.*

4.3 *Forward simulation*

The second way to calculate T^∞_U and ΔT^∞_U is to simulate the future utilities received under policy U. Define $\widehat{\sigma}^n_s(x_i)$ as the average discounted utility received by s independent sample paths simulated from state x_i for n periods under policy U. That is, set $\widehat{\sigma}^n_s(x_i) \equiv s^{-1} \sum_{j=0}^{s-1} \sum_{k=0}^{n-1} \beta^k \delta'_{\ell(j,k)} \pi(U)$, where $\ell(j,0) = i$ and $\ell(j,k)$ is a multinoulli random variable with probability simplex $\delta'_{\ell(j,k-1)} Q(U)$. And define $\widehat{\sigma}^n_s \equiv [\widehat{\sigma}^n_s(x_1), \ldots, \widehat{\sigma}^n_s(x_m)]'$ as a vector of such simulation estimates.

Like all estimators, the mean square error of $\delta'_i \widehat{\sigma}^n_s = \widehat{\sigma}^n_s(x_i)$ decomposes into bias and variance components:

$$\begin{aligned} \mathrm{MSE}\big(\delta'_i \widehat{\sigma}^n_s\big) &\equiv \mathrm{E}\big(\big(\delta'_i \widehat{\sigma}^n_s - \delta'_i T^\infty_U\big)^2\big) \\ &= \mathrm{Bias}\big(\delta'_i \widehat{\sigma}^n_s\big)^2 + \mathrm{Var}\big(\delta'_i \widehat{\sigma}^n_s\big), \\ \text{where} \quad \mathrm{Bias}\big(\delta'_i \widehat{\sigma}^n_s\big) &\equiv \mathrm{E}\big(\delta'_i \widehat{\sigma}^n_s\big) - \delta'_i T^\infty_U, \\ \text{and} \quad \mathrm{Var}\big(\delta'_i \widehat{\sigma}^n_s\big) &\equiv \mathrm{E}\big(\big(\delta'_i \widehat{\sigma}^n_s - \mathrm{E}\big(\delta'_i \widehat{\sigma}^n_s\big)\big)^2\big). \end{aligned}$$

Equivalent expressions hold for $\delta'_i \Delta \widehat{\sigma}^n_s = \widehat{\sigma}^n_s(x_i) - \widehat{\sigma}^n_s(x_1)$. And the following proposition establishes that $\mathrm{Bias}(\delta'_i \Delta \widehat{\sigma}^n_s)$ falls faster with n than $\mathrm{Bias}(\delta'_i \widehat{\sigma}^n_s)$.

PROPOSITION 6. *Whereas* $\mathrm{Bias}(\delta_i'\widehat{\sigma}_s^n)$ *is* $O(\beta^n)$ *as* $n \to \infty$, $\mathrm{Bias}(\delta_i'\Delta\widehat{\sigma}_s^n)$ *is* $O(\beta^n\gamma^n)$ *for all* $\gamma > \lambda(Q(U))$. *Thus, if the Markov chain is ergodic under policy* U *then* $\lambda(Q(U)) < 1$ *and the bias in the relative value function estimate vanishes strictly faster than the bias in the total value function estimate.*

This proposition establishes that we do not have to simulate as far into the future as we previously thought. To avoid bias in $\widehat{\sigma}_s^n$, economists have generally set n large enough so that $\beta^n < \varepsilon$ (Arcidiacono and Ellickson (2011, p. 383)). But bias in $\widehat{\sigma}_s^n$ is irrelevant; only bias in $\Delta\widehat{\sigma}_s^n$ is relevant. And guaranteeing negligible $\Delta\widehat{\sigma}_s^n$ bias only requires that n be large enough to satisfy $(\beta\lambda(Q(U)))^n < \varepsilon$. Shorting the simulation horizon in this fashion speeds up the computation by a factor of $\log(\beta\lambda(Q(U)))/\log(\beta) = 1 + \log(\lambda(Q(U)))/\log(\beta)$.

Truncating the simulation horizon makes the estimator not only faster but also more accurate, as the following propositions imply.

PROPOSITION 7. *There exists* $b < 1$ *such that if* $\beta \in [b, 1)$, $\lambda(Q(U)) < 1$, *and* $\psi(Q(U))'T_U^\infty \neq 0$ *then*

$$\lim_{n\to\infty} \beta^{-2n}\big(\mathrm{Bias}\big(\delta_i'\widehat{\sigma}_s^{n+1}\big)^2 - \mathrm{Bias}\big(\delta_i'\widehat{\sigma}_s^n\big)^2\big) = -\big(1-\beta^2\big)\big(\psi\big(Q(U)\big)'T_U^\infty\big)^2 < 0 \quad \textit{and}$$

$$\lim_{n\to\infty} \beta^{-2n}\big(\mathrm{Var}\big(\delta_i'\widehat{\sigma}_s^{n+1}\big) - \mathrm{Var}\big(\delta_i'\widehat{\sigma}_s^n\big)\big)$$
$$= s^{-1}\pi(U)'\big(\mathrm{diag}\big(\psi\big(Q(U)\big)\big) - \psi\big(Q(U)\big)\psi\big(Q(U)\big)'\big)$$
$$\times \big(2\big(I - \Delta Q(U)/\beta\big)^{-1} - I\big)\pi(U) > 0.$$

In this case, there exists $S > 0$ *such that for all* $s > S$,

$$\lim_{n\to\infty} \beta^{-2n}\big(\mathrm{MSE}\big(\delta_i'\widehat{\sigma}_s^{n+1}\big) - \mathrm{MSE}\big(\delta_i'\widehat{\sigma}_s^n\big)\big) < 0,$$

and the limiting accuracy of the $\widehat{\sigma}_s^n$ *estimator increases with the length of the simulation horizon.*

PROPOSITION 8. *There exists* $b < 1$ *such that if* $\beta \in [b, 1)$ *and* $\lambda(Q(U)) < 1$ *then*

$$\lim_{n\to\infty} \beta^{-2n}\big(\mathrm{Bias}\big(\delta_i'\Delta\widehat{\sigma}_s^{n+1}\big)^2 - \mathrm{Bias}\big(\delta_i'\Delta\widehat{\sigma}_s^n\big)^2\big) = 0, \quad \textit{and}$$

$$\lim_{n\to\infty} \beta^{-2n}\big(\mathrm{Var}\big(\delta_i'\Delta\widehat{\sigma}_s^{n+1}\big) - \mathrm{Var}\big(\delta_i'\Delta\widehat{\sigma}_s^n\big)\big)$$
$$= 2s^{-1}\pi(U)'\big(\mathrm{diag}\big(\psi\big(Q(U)\big)\big) - \psi\big(Q(U)\big)\psi\big(Q(U)\big)'\big)$$
$$\times \big(2\big(I - \Delta Q(U)/\beta\big)^{-1} - I\big)\pi(U) > 0.$$

In this case,

$$\lim_{n\to\infty} \beta^{-2n}\big(\mathrm{MSE}\big(\delta_i'\Delta\widehat{\sigma}_s^{n+1}\big) - \mathrm{MSE}\big(\delta_i'\Delta\widehat{\sigma}_s^n\big)\big) > 0,$$

and the limiting accuracy of the $\Delta\widehat{\sigma}_s^n$ *estimator decreases with the length of the simulation horizon.*

Proposition 7 states that increasing n increases $\text{Var}(\delta_i'\widehat{\sigma}_s^n)$ by an order of β^{2n}/s and decreases $\text{Bias}(\delta_i'\widehat{\sigma}_s^n)^2$ by an order of β^{2n}. Thus, for sufficiently large s, lengthening the simulation horizon decreases $\text{MSE}(\delta_i'\widehat{\sigma}_s^n)$. Proposition 8 establishes that Proposition 7 is misleading, for lengthening the simulation horizon asymptotically *increases* the error that matters, $\text{MSE}(\delta_i'\Delta\widehat{\sigma}_s^n)$. Increasing n increases $\text{Var}(\delta_i'\Delta\widehat{\sigma}_s^n)$ by an order of β^{2n} and decreases $\text{Bias}(\delta_i'\Delta\widehat{\sigma}_s^n)^2$ by an order of $\beta^{2n}\lambda(Q(U))^{2n}$ (see Proposition 6). Thus, increasing n beyond $\log(\varepsilon)/\log(\beta\lambda(Q(U)))$ yields a variance increase without a commensurate bias decrease. Simulating too far forward yields an estimator both slower and noisier.

4.4 *System of equations*

The third way to calculate $T_{U_n}^\infty$ and $\Delta T_{U_n}^\infty$ is to solve linear equations $(I-\beta Q(U_n))T_{U_n}^\infty = \pi(U_n)$ and $(I-\beta\Delta Q(U_n))(\Delta T_{U_n}^\infty) = \Delta\pi(U_n)$. The latter system remains well conditioned as the discount factor goes to one, but the former system does not. Thus, we can compute $\Delta T_{U_n}^\infty$ to a higher degree of precision than $T_{U_n}^\infty$ when discounting is negligible, as the following proposition establishes.

Proposition 9. *If the Markov chain is ergodic under policy U, then*:

1. $\text{Cond}(I-\beta Q(U))$ *is* $O(\frac{1}{1-\beta})$ *as* $\beta\to 1$. *This indicates that policy iteration's policy evaluation equations become ill-conditioned as the discount factor approaches unity.*

2. $\text{Cond}(I-\beta\Delta Q(U))$ *is* $O(1)$ *as* $\beta\to 1$. *This indicates that relative policy iteration's policy evaluation equations remain well conditioned as the discount factor approaches unity.*

5. Application: Estimating Markov decision processes

We can leverage Section 3 and Section's 4 strong convergence results when estimating Markov decision processes. For example, we can position Rust's (1987) dynamic discrete choice problem in Section's 2 framework by replacing its finite action space with an infinite continuum of choice probabilities (see Aguirregabiria and Mira (2010, p. 49)). Most dynamic program estimators use some version of value iteration or policy iteration; we can accelerate these estimators by retooling them with relative value and policy iterations.

5.1 *Nested fixed point*

The nested fixed point (NFXP) estimator solves a sequence of dynamic programs. Rust (2000, p. 18) claimed the algorithm "has to compute the fixed point $[V = TV^*]$ in order to evaluate the likelihood function." This is incorrect. The estimator's empirical likelihood function depends only on the policy function, which in turn depends only on the value function's relative differences. Hence, we can replace Rust's (2000) value and policy iteration steps with quicker relative value and relative policy iteration steps.

Since I first proposed this technique, Chen (2017) and Kasahara and Shimotsu (2018) have independently validated it. Kasahara and Shimotsu (2018, p. 46) reported that my

TABLE 2. I reprint the first dozen rows of Chen's (2017) seventh table. The figures report the number of minutes required to solve each electric utility's acid-rain-prevention dynamic program with traditional value iteration, relative value iteration, and Bray's (2018) endogenous value iteration (which is a generalization of relative value iteration).

	Traditional	Relative	Endogenous
American Electric Power	404	27	5.1
Atlantic City Electric	303	9	4.0
Carolina Power and Light	410	24	3.2
Central Hudson Gas and Electric	301	9	3.9
Central Illinois Light	300	12	3.9
Central Operating	301	19	3.9
Cincinnati Gas and Electric	360	20	3.4
Dairyland Power Coop	304	30	4.0
Dayton Power and Light	366	9	3.9
Detroit Edison	407	19	3.2
Duke Energy	414	41	3.7
Empire District Electric	277	10	3.9

refinement "leads to substantial computational gains, reducing the average computational time and the average number of iterations by factors of 17 and 9, respectively." And Chen (2017, p. 3) reported that my refinement "vastly reduce[ed] the computation burden" of NFXP applied to her acid-rain-mitigation model. For example, solving each of her dynamic programs required 10,701 minutes with value iteration, 661 minutes with relative value iteration, and 126 minutes with endogenous value iteration (see Table 2). Endogenous value iteration is a generalization of relative value iteration; it enjoys an even stronger rate of convergence when the most persistent state variable is exogenous (Bray (2018)).

5.2 *Conditional choice probability estimators*

Hotz and Miller's (1993) infinite-horizon estimator, Aguirregabiria and Mira's (2002, 2007) nested pseudo-likelihood (NPL) estimators, Pakes, Ostrovsky, and Berry's (2007) "simple" estimator, and Pesendorfer and Schmidt-Dengler's (2008) least-squares estimator all follow the conditional choice probability (CCP) approach: (i) pre-estimate the agent's policy function in reduced form; (ii) choose a set of model primitives; (iii) apply a policy iteration step to the pre-estimated policy function under the given model primitives; and (iv) use the updated policy function to evaluate the empirical likelihood (or moment conditions) associated with the given model primitives. We can streamline these CCP estimators by replacing their policy iteration steps with relative policy iteration steps.

For example, suppose the flow utility vector is linear in the structural parameters: $\pi(U) = M(U)\theta$, where $M(U)$ is an $m \times k$ matrix of reward statistics and θ is a length-k vector of primitives to estimate. Aguirregabiria and Mira (2010, p. 51) explained that, in this case, the CCP bottleneck is calculating the $m \times k$ matrix $\widetilde{T}_U^\infty$ that satisfies

$\widetilde{T}_U^\infty = M(U) + \beta Q(U)\widetilde{T}_U^\infty$. The authors recommended computing $\widetilde{T}_U^\infty$ "by successive approximations, iterating [the fixed-point equation] which is a contraction mapping"; in other words, they suggested approximating $\widetilde{T}_U^\infty$ with $\widetilde{T}_U^n 0$ evaluated under large n, where $\widetilde{T}_U^n M \equiv \widetilde{T}_U(T_U^{n-1}M)$ and $\widetilde{T}_U M \equiv M(U) + \beta Q(U)M$. Since $\widetilde{T}_U^\infty \theta = T_U^\infty$ and $(\widetilde{T}_U^n 0)\theta = T_U^n$, Proposition 5 indicates that Aguirregabiria and Mira's (2010) iterative scheme converges at linear rate β. In contrast, Proposition 5 indicates that $\Delta\widetilde{T}_U^n 0$ converges to $\Delta\widetilde{T}_U^\infty$ at linear rate $\beta\lambda(Q(U))$. And Proposition 1 establishes that $\Delta\widetilde{T}_U^\infty$ is all we need.

5.3 *Simulation estimators*

Hotz, Miller, Sanders, and Smith's (1994) and Bajari, Benkard, and Levin's (2007) infinite-horizon simulation estimators are similar to Section's 5.2 CCP estimators, except they evaluate policies with forward simulation. Currently, these estimators simulate "far enough into the future so that the discounting renders future terms past this point irrelevant" (Arcidiacono and Ellickson (2011, p. 381)). It is generally accepted that the "main drawback of this particular [estimation scheme] is that when β is close to 1, many periods must be included to ensure the properties of the estimator are not unduly affected by the finite-horizon approximation" (Hotz et al. (1994, p. 277)). However, Section's 4.3 results indicate that we can obviate much of this work. We do not have to simulate the process until the flow utilities are discounted to oblivion; we only have to simulate the process until the state space scrambles.

Proposition 8 implies that truncating the simulation horizon will make Hotz et al.'s (1994) and Bajari, Benkard, and Levin's (2007) estimators both faster and more accurate. Simulating the utilities received after the state variables reach their stationary distribution increases the estimators' variance (since the random draws have positive standard deviation) without decreasing the estimators' bias (since the random draws have basically zero mean). Needlessly increasing the simulation horizon needlessly increases the simulation error.

I now demonstrate with a Monte Carlo simulation study. I generate 600 Rust-style dynamic discrete choice programs, each with 1000 discrete states and three discrete actions per state. Taking action a in state x in period t yields flow utility $u_1(a,x)\theta_1 + u_2(a,x)\theta_2 + e_t(a)$, where $e_t(a)$ is an independent standard Gumbel random variable that realizes in period t. The probability of transitioning from state x to state x', given action a, is $q(x'|x,a)$. I set probability vector $[q(x_1|x,a),\ldots,q(x_{1000}|x,a)]$ to an independent symmetric Dirichlet random variable with concentration parameter one; I set scalars $u_1(a,x)$, $u_2(a,x)$, θ_1, and θ_2 to independent standard normal random variables; and I set discount factor β to 0.99.

Following convention, I characterize the agent's optimal policy with CCPs $\{p(a|x)\}$, where $p(a|x)$ is the probability of choosing action a in state x, unconditional on the Gumbel shocks. I calculate these CCPs with relative value iteration.

The goal is to reverse engineer θ_1 and θ_2 from $\{u_i(a,x)\}$, $\{p(a|x)\}$, $\{q(x'|x,a)\}$, and β. I use Hotz et al.'s (1994) estimator, deploying s samples paths from state x_1 action a_1, s samples paths from state x_1 action a_2, and s samples paths from state x_1 action a_3. I then set $\widehat{\theta}_1$ and $\widehat{\theta}_2$ to the utility parameters that (i) equate $\log(p(a_2|x_1)) - \log(p(a_1|x_1))$

with the simulation-average a_2 discounted utility minus the simulation-average a_1 discounted utility, and (ii) equate $\log(p(a_3|x_1)) - \log(p(a_1|x_1))$ with the simulation-average a_3 discounted utility minus the simulation-average a_1 discounted utility (see Aguirregabiria and Mira (2010, p. 53)). I estimate 200 dynamic programs with $s = 100$, 200 dynamic programs with $s = 1000$, and 200 dynamic programs with $s = 10{,}000$. For each, I use both the traditional simulation horizon, $n_1 \equiv \log(\varepsilon)/\log(\beta)$, and my shorter simulation horizon, $n_2 \equiv \log(\varepsilon)/\log(\beta\lambda(Q(U)))$, where $\varepsilon \equiv 10^{-6}$ and $Q(U)$ is the state transition matrix implied by $\{p(a|x)\}$ and $\{q(x'|x,a)\}$.

For each dynamic program, I calculate the estimation time under horizon n_1 divided by the estimation time under horizon n_2, and I calculate the estimation error under horizon n_1 divided by the estimation error under horizon n_2. Table 3 reports these ratios' geometric means. Using the traditional simulation horizon is between 229 and 257 times slower, and between 2.39 and 3.36 times less accurate.

I will close with two technical points. First, Zobel and Scherer (2005, p. 133) listed several upper bounds that are easier to compute than $\lambda(Q(U))$. And second, Arcidiacono and Miller (2011, p. 1834) provided a means to replace an agent's observed policy with one that is more conducive to estimation. We can use this technique to further shorten the simulation horizon: rather than set the sample path length to $\log(\varepsilon)/\log(\beta\lambda(Q(U)))$, where U is the observed policy, we can set the sample path length to $\log(\varepsilon)/\log(\beta\lambda(Q(\widetilde{U})))$, where $\widetilde{U}$ is any policy we desire.

6. Application: Calculating dynamic equilibria

In the literature on dynamic games, (i) the canonical equilibrium concept is Maskin and Tirole's (1988a, 1988b) Markov perfect equilibrium, (ii) the canonical application is Ericson and Pakes's (1995) market entry problem, and (iii) the canonical solution method is Pakes and McGuire's (1994) Gauss–Jacobi algorithm. Pakes and McGuire's algorithm is just value iteration run in parallel across agents: in iteration n, each agent implements a value iteration step, assuming the other agents follow their iteration $n-1$ policies.

Table 3. I estimate 600 dynamic discrete choice problems with Hotz et al.'s (1994) and Bajari, Benkard, and Levin's (2007) simulation estimators. I estimate 200 problems with $s = 100$ sample paths, 200 with $s = 1000$ sample paths, and 200 with $s = 10{,}000$ sample paths. I estimate each problem with both the traditional simulation horizon, $n_1 \equiv \log(\varepsilon)/\log(\beta)$, and my shorter simulation horizon, $n_2 \equiv \log(\varepsilon)/\log(\beta\lambda(Q(U)))$, where $\varepsilon \equiv 10^{-6}$ and $\beta \equiv 0.99$. I tabulate the geometric means of the estimation time ratios under n_1 and n_2 and the geometric means of the estimation error ratios under n_1 and n_2. I measure the estimation error with the Euclidean distance between $[\theta_1, \theta_2]$ and $[\widehat{\theta}_1, \widehat{\theta}_2]$.

	100	1000	10,000
$\frac{\text{Time}(n_1)}{\text{Time}(n_2)}$	229.47	235.13	257.22
	(1.90)	(1.83)	(1.84)
$\frac{\text{Error}(n_1)}{\text{Error}(n_2)}$	2.39	2.88	3.36
	(0.28)	(0.28)	(0.33)

To exploit strong convergence, I difference the value functions after each Bellman contraction, transforming the algorithm from a multiagent version of value iteration to a multiagent version of relative value iteration.

To demonstrate, I solve Doraszelski and Judd's (2012) discrete-time version of Ericson and Pakes's (1995) Markov perfect equilibrium with both Pakes and McGuire's (1994) traditional algorithm (which calculates value functions) and my strong convergence analog (which calculates relative value functions). Table 4 reports the number of Bellman contractions each algorithm implemented under various β values. Taking β to one breaks Pakes and McGuire's value-iteration-based algorithm but not my relative-value-iteration-based algorithm.[3]

My algorithm converges even faster when I reformulate the problem to better leverage strong convergence. Porteus (1975) established that a dynamic program with flow utility vector $\pi(U)$ and state transition matrix $Q(U)$ has the same optimal policy as a dynamic program with flow utility vector $\widetilde{\pi}(U) = (1 - q(U))^{-1}\pi(U)$ and state transition matrix $\widetilde{Q}(U) = (1 - q(U))^{-1}(Q(U) - q(U)/\beta I)$, where $q(U)$ is the smallest diagonal element of $Q(U)$. However, if $q(U) > 0$ then $\lambda(\widetilde{Q}(U)) < \lambda(Q(U))$, and the relative value function converges faster under the alternative problem. Table 4 illustrates that Porteus's (1975) refinement accelerates my algorithm's convergence rate by another 26%.

7. Conclusion

There's no downside to exploiting strong convergence. Simplicity is not sacrificed because deriving the relative value function from the value function requires only one line of code: `rel.value.fn = value.fn - value.fn[1]`. And information is not sacrificed because deriving the value function from the relative value function requires only one Bellman contraction: $V^* = \Delta V^* + (1 - \beta)^{-1}(T\Delta V^* - \Delta V^*)$.[4]

Table 4. I report the number of Bellman contractions required to calculate Doraszelski and Judd's (2012) discrete-time Markov perfect equilibrium under five discount factors: $\beta \in \{0.9, 0.99, 0.999, 0.9999, 1\}$. I solve the equilibrium with Pakes and McGuire's (1994) multiagent value iteration algorithm, an equivalent multiagent relative value iteration algorithm, and the multiagent relative value iteration algorithm with Porteus's (1975) accelerant.

	0.9	0.99	0.999	0.9999	1
Value Iteration	246	2430	24,448	245,845	∞
Relative Value Iteration	244	639	964	1001	1005
Accelerated Relative Value Iteration	221	483	701	725	726

[3]Relative value iteration can accommodate $\beta = 1$ when the Markov chain is ergodic (see Morton and Wecker (1977)). This property apparently extends to the multiagent case.

[4]Lemmas 1 and 4 in the Appendix imply $(I - \Delta)(T\Delta V^* - \Delta V^*) = T\Delta V^* - \Delta V^*$. And Proposition 1 implies $T^n\Delta V^* = \sum_{t=0}^{n-1} \beta^t Q(U^*)^t \pi(U^*) + \beta^n Q(U^*)^n \Delta V^*$. With this, Lemma 2 implies $(T^{n+1} - T^n)\Delta V^* = \beta^n Q(U^*)^n \pi(U^*) - \beta^n Q(U^*)^n (I - \beta Q(U^*))\Delta V^* = \beta^n Q(U^*)^n (T\Delta V^* - \Delta V^*) = \beta^n Q(U^*)^n (I - \Delta)(T\Delta V^* - \Delta V^*) = \beta^n (I - \Delta)(T\Delta V^* - \Delta V^*) = \beta^n (T\Delta V^* - \Delta V^*)$. And this implies $V^* = \Delta V^* + \sum_{n=0}^{\infty}(T^{n+1} - T^n)\Delta V^* = \Delta V^* + \sum_{n=0}^{\infty} \beta^n (T\Delta V^* - \Delta V^*) = \Delta V^* + (1 - \beta)^{-1}(T\Delta V^* - \Delta V^*)$.

The technique's raison d'être is the high-frequency problem with low persistence. For example, suppose the period length is one day and the system equilibrates within a year. The first assumption implies that the per-period discount factor is around $\beta = 0.9998$ (for a per-anum discount factor of $0.9998^{365} = 0.93$), which implies that value iteration requires around $\log(10^{-6})/\log(0.9998) \approx 69{,}000$ Bellman contractions. But the second assumption implies that relative value iteration requires only around 365 Bellman contractions.

Appendix: Proofs

Lemma 1. $\Delta^2 = \Delta$.

Proof. $(I-\Delta)^2 = \iota\delta_1'\iota\delta_1' = \iota\delta_1' = I - \Delta$. This implies the result. □

Lemma 2. $Q(U)^n(I-\Delta) = (I-\Delta)$ *for all* $n \in \mathbb{N}$.

Proof. The rows of a stochastic matrix sum to one. This implies $Q(U)\iota = \iota$, which implies $Q(U)(I-\Delta) = Q(U)\iota\delta_1' = \iota\delta_1' = (I-\Delta)$. By induction, this implies the result. □

Lemma 3. $\Delta Q(U)\Delta = \Delta Q(U)$.

Proof. Lemmas 1 and 2 imply $\Delta Q(U)(I-\Delta) = \Delta(I-\Delta) = \Delta - \Delta = 0$. □

Proposition 1 (White (1963)). *Differencing the value function does not affect the corresponding policy function*: $\mathscr{U}(\Delta V) = \mathscr{U}(V)$.

Proof. Lemma 2 implies

$$\begin{aligned}
\mathscr{U}(\Delta V) &= \operatorname*{arg\,max}_{U\in\mathbb{U}} \iota'(\pi(U) + \beta Q(U)\Delta V) \\
&= \operatorname*{arg\,max}_{U\in\mathbb{U}} \iota'(\pi(U) + \beta Q(U)\Delta V + \beta(I-\Delta)V) \\
&= \operatorname*{arg\,max}_{U\in\mathbb{U}} \iota'(\pi(U) + \beta Q(U)\Delta V + \beta Q(U)(I-\Delta)V) \\
&= \operatorname*{arg\,max}_{U\in\mathbb{U}} \iota'(\pi(U) + \beta Q(U)V).
\end{aligned}$$

□

Lemma 4. $\Delta T\Delta = \Delta T$.

Proof. Proposition 1 and Lemma 3 imply $\Delta T\Delta V = \Delta\pi(\mathscr{U}(\Delta V)) + \beta\Delta Q(\mathscr{U}(\Delta V))\Delta V = \Delta\pi(\mathscr{U}(V)) + \beta\Delta Q(\mathscr{U}(V))V = \Delta TV$. □

Proposition 2 (Bray (2018)). *The relative value iteration algorithm always returns an ε-optimal policy after a finite number of iterations.*

Proof. This is a special case of Bray's (2018) fourth proposition. □

LEMMA 5. *For all $\varepsilon > 0$, there exists $N(\varepsilon) > 0$ such that $\|\Delta Q(U^*)^t\| \leq (\lambda(Q(U^*)) + \varepsilon)^t$ for all $t > N(\varepsilon)$.*

PROOF. This follows from the Jordan normal form of $Q(U^*)$ and the fact that ι resides both in the null space of Δ and the eigenspace corresponding to $Q(U^*)$'s largest eigenvalue. □

LEMMA 6. *For all $t > 0$ and $\varepsilon > 0$, there exists $N(t, \varepsilon) > 0$ such that for all $s > N(t, \varepsilon)$*

$$\frac{\|\Delta T^{s+2t}V - \Delta T^{s+t}V\|}{\|\Delta T^{s+t}V - \Delta T^{s}V\|} \leq \beta^t \left\| \Delta \prod_{j=0}^{t-1} Q\big(\mathscr{U}\big(T^{s+j}V\big)\big) \right\| + \varepsilon.$$

PROOF. The envelope theorem implies

$$\begin{aligned}
\frac{\partial}{\partial V} TV &= \frac{\partial}{\partial V}\Big(\max_{U \in \mathbb{U}} \pi(U) + \beta Q(U)V\Big) \\
&= \frac{\partial}{\partial V}\big(\pi(U) + \beta Q(U)V\big)|_{U=\mathscr{U}(V)} \\
&= \beta Q\big(\mathscr{U}(V)\big).
\end{aligned}$$

By induction, this implies $\frac{\partial}{\partial V} T^t V = \beta^t \prod_{j=0}^{t-1} Q(\mathscr{U}(T^j V))$. This, in turn, implies

$$T^t T^{s+t} V = T^t T^s V + \beta^t \left(\prod_{j=0}^{t-1} Q\big(\mathscr{U}\big(T^j T^s V\big)\big) \right) \big(T^{s+t}V - T^s V\big) + o\big(T^{s+t}V - T^s V\big).$$

With this, Lemma 3 implies

$$\begin{aligned}
\Delta T^t T^{s+t} V = \Delta T^t T^s V &+ \beta^t \Delta \left(\prod_{j=0}^{t-1} Q\big(\mathscr{U}\big(T^j T^s V\big)\big) \right) \big(\Delta T^{s+t}V - \Delta T^s V\big) \\
&+ o\big(T^{s+t}V - T^s V\big).
\end{aligned}$$

Since $\lim_{s\to\infty} T^{s+t}V - T^s V = V^* - V^* = 0$, this implies the result. □

LEMMA 7. *For all $t > 0$ and $\varepsilon > 0$, there exists $N(t, \varepsilon) > 0$ such that $\|Q(U^*)^t - \prod_{j=0}^{t-1} Q(\mathscr{U}(T^{s+j}V))\| \leq \varepsilon$ for all $s > N(t, \varepsilon)$.*

PROOF. Choose $\xi > 0$ small enough so that $t \sum_{j=1}^{t-1} \binom{t-1}{j} \|Q(U^*)\|^{t-1-j} \xi^{j+1} < \varepsilon$. And choose $N(t, \varepsilon)$ large enough so that $\|Q(\mathscr{U}(T^s V)) - Q(U^*)\| \leq \xi$ for all $s > N(t, \varepsilon)$ (doing so is possible because $\lim_{s\to\infty} \mathscr{U}(T^s V) = U^*$ and function Q is continuous). Now, choosing $s > N(t, \varepsilon)$ yields

$$\begin{aligned}
&\left\| Q\big(U^*\big)^t - \prod_{j=0}^{t-1} Q\big(\mathscr{U}\big(T^{s+j}V\big)\big) \right\| \\
&= \left\| \sum_{i=0}^{t-1} \prod_{j=0}^{t-1} \big(\mathbb{1}(i \neq j) Q\big(U^*\big) + \big(Q\big(\mathscr{U}\big(T^{s+j}V\big)\big) - Q\big(U^*\big)\big)\big) \right\|
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{i=0}^{t-1}\prod_{j=0}^{t-1}\bigl(\mathbb{1}(i\neq j)\bigl|\!\bigl|\!\bigl|Q\bigl(U^*\bigr)\bigr|\!\bigr|\!\bigr| + \bigl|\!\bigl|\!\bigl|Q\bigl(\mathscr{U}\bigl(T^{s+j}V\bigr)\bigr) - Q\bigl(U^*\bigr)\bigr|\!\bigr|\!\bigr|\bigr)\\
&\leq \sum_{i=0}^{t-1}\prod_{j=0}^{t-1}\bigl(\mathbb{1}(i\neq j)\bigl|\!\bigl|\!\bigl|Q\bigl(U^*\bigr)\bigr|\!\bigr|\!\bigr| + \xi\bigr)\\
&= t\xi\bigl(\bigl|\!\bigl|\!\bigl|Q\bigl(U^*\bigr)\bigr|\!\bigr|\!\bigr| + \xi\bigr)^{t-1}\\
&= t\sum_{j=1}^{t-1}\binom{t-1}{j}\bigl|\!\bigl|\!\bigl|Q\bigl(U^*\bigr)\bigr|\!\bigr|\!\bigr|^{t-1-j}\xi^{j+1} \leq \varepsilon.
\end{aligned}
$$

□

Proposition 3 (Morton and Wecker (1977)). *Whereas $\|T^nV - V^*\|$ is $O(\beta^n)$ as $n\to\infty$, $\|\Delta T^nV - \Delta V^*\|$ is $O(\beta^n\gamma^n)$ for all $\gamma > \lambda(Q(U^*))$. Thus, if the Markov chain is ergodic under policy U^* then $\lambda(Q(U^*)) < 1$ and relative value iteration converges strictly faster than traditional value iteration.*

Proof. Let $N_5(\varepsilon)$, $N_6(t,\varepsilon)$, and $N_7(t,\varepsilon)$ represent the thresholds characterized in Lemmas 5, 6, and 7. Choose ε, t, and s such that $\varepsilon \in (0, 1-\beta\lambda(Q(U^*)))$, $t > N_5(\varepsilon/2)$, and $s > \max(N_6(t,\xi(\varepsilon)), N_7(t,\xi(\varepsilon)))$, where $\xi(\varepsilon) = (1+2\beta^t)^{-1}(\beta^t(\lambda(Q(U^*)) + \varepsilon)^t - \beta^t(\lambda(Q(U^*)) + \varepsilon/2)^t)$. With this, Lemmas 5, 6, and 7 imply

$$
\begin{aligned}
&\frac{\bigl\|\Delta T^{s+(i+2)t}V - \Delta T^{s+(i+1)t}V\bigr\|}{\bigl\|\Delta T^{s+(i+1)t}V - \Delta T^{s+it}V\bigr\|}\\
&\leq \beta^t\left|\!\left|\!\left|\Delta\prod_{j=0}^{t-1}Q\bigl(\mathscr{U}\bigl(T^{s+it+j}V\bigr)\bigr)\right|\!\right|\!\right| + \xi(\varepsilon)\\
&\leq \beta^t\bigl|\!\bigl|\!\bigl|\Delta Q\bigl(U^*\bigr)^t\bigr|\!\bigr|\!\bigr| + \beta^t\bigl|\!\bigl|\!\bigl|\Delta\bigr|\!\bigr|\!\bigr|\left|\!\left|\!\left|\left(\prod_{j=0}^{t-1}Q\bigl(\mathscr{U}\bigl(T^{s+it+j}V\bigr)\bigr)\right) - Q\bigl(U^*\bigr)^t\right|\!\right|\!\right| + \xi(\varepsilon)\\
&\leq \beta^t\bigl(\lambda\bigl(Q\bigl(U^*\bigr)\bigr) + \varepsilon/2\bigr)^t + 2\beta^t\xi(\varepsilon) + \xi(\varepsilon)\\
&= \beta^t\bigl(\lambda\bigl(Q\bigl(U^*\bigr)\bigr) + \varepsilon\bigr)^t.
\end{aligned}
$$

By induction, this implies

$$
\frac{\bigl\|\Delta T^{s+(i+1)t}V - \Delta T^{s+it}V\bigr\|}{\bigl\|\Delta T^{s+t}V - \Delta T^{s}V\bigr\|} \leq \beta^{it}\bigl(\lambda\bigl(Q\bigl(U^*\bigr)\bigr) + \varepsilon\bigr)^{it},
$$

which implies

$$
\begin{aligned}
\frac{\bigl\|V^* - \Delta T^{s+jt}V\bigr\|}{\bigl\|\Delta T^{s+t}V - \Delta T^{s}V\bigr\|} &= \frac{\left\|\sum_{i=j}^{\infty}\Delta T^{s+(i+1)t}V - \Delta T^{s+it}V\right\|}{\bigl\|\Delta T^{s+t}V - \Delta T^{s}V\bigr\|}\\
&\leq \sum_{i=j}^{\infty}\frac{\bigl\|\Delta T^{s+(i+1)t}V - \Delta T^{s+it}V\bigr\|}{\bigl\|\Delta T^{s+t}V - \Delta T^{s}V\bigr\|}
\end{aligned}
$$

$$\leq \sum_{i=j}^{\infty} \beta^{it}\big(\lambda\big(Q\big(U^*\big)\big)+\varepsilon\big)^{it}$$

$$= \frac{\beta^{jt}\big(\lambda\big(Q\big(U^*\big)\big)+\varepsilon\big)^{jt}}{1-\beta\big(\lambda\big(Q\big(U^*\big)\big)+\varepsilon\big)},$$

which implies $\|\Delta T^{s+jt}V - \Delta V^*\|$ is $O(\beta^{jt}(\lambda(Q(U^*))+\varepsilon)^{jt})$ as $j\to\infty$, which implies $\|\Delta T^n V - \Delta V^*\|$ is $O(\beta^n(\lambda(Q(U^*))+\varepsilon)^n)$ as $n\to\infty$. □

Proposition 4 (Bray (2018)). *The relative policy iteration algorithm always returns an ε-optimal policy after a finite number of iterations.*

Proof. This is a special case of Bray's (2018) fourth proposition. □

Proposition 5 (Morton (1971)). *Whereas $\|T_U^n V - T_U^\infty\|$ is $O(\beta^n)$ as $n\to\infty$, $\|\Delta T_U^n V - \Delta T_U^\infty\|$ is $O(\beta^n\gamma^n)$ for all $\gamma > \lambda(Q(U))$. Thus, if the Markov chain is ergodic under policy U then $\lambda(Q(U))<1$ and relative policy iteration's policy evaluation step converges strictly faster than traditional policy iteration's policy evaluation step.*

Proof. Consider an auxiliary problem in which $\mathbb{U}=\{U\}$. In this case $\mathscr{U}(V)=U$, and thus $T_U V = T_{\mathscr{U}(V)}V = TV$, and thus $T_U^t V = T^t V$. With this, Proposition 3 implies the result. □

Proposition 6. *Whereas* $\text{Bias}(\delta_i'\widehat{\sigma}_s^n)$ *is* $O(\beta^n)$ *as* $n\to\infty$, $\text{Bias}(\delta_i'\Delta\widehat{\sigma}_s^n)$ *is* $O(\beta^n\gamma^n)$ *for all* $\gamma > \lambda(Q(U))$. *Thus, if the Markov chain is ergodic under policy U then $\lambda(Q(U))<1$ and the bias in the relative value function estimate vanishes strictly faster than the bias in the total value function estimate.*

Proof. Proposition 5 implies the result, since

$$\begin{aligned}\text{Bias}\big(\delta_i'\widehat{\sigma}_s^n\big) &= \left(\sum_{k=0}^{n-1}\beta^k \text{E}(\delta_{\ell(0,k)})'\pi(U)\right) - \delta_i' T_U^\infty \\ &= \left(\sum_{k=0}^{n-1}\beta^k \delta_i' Q(U)^k \pi(U)\right) - \delta_i' T_U^\infty \\ &= \delta_i'\big(T_U^n 0 - T_U^\infty\big),\end{aligned}$$

and, similarly, $\text{Bias}(\delta_i'\Delta\widehat{\sigma}_s^n) = \delta_i'(\Delta T_U^n 0 - \Delta T_U^\infty)$. □

Proposition 7. *There exists $b<1$ such that if $\beta\in[b,1)$, $\lambda(Q(U))<1$, and $\psi(Q(U))'T_U^\infty \neq 0$ then*

$$\lim_{n\to\infty}\beta^{-2n}\big(\text{Bias}\big(\delta_i'\widehat{\sigma}_s^{n+1}\big)^2 - \text{Bias}\big(\delta_i'\widehat{\sigma}_s^{n}\big)^2\big) = -\big(1-\beta^2\big)\big(\psi\big(Q(U)\big)'T_U^\infty\big)^2 < 0 \quad \textit{and}$$

$$\lim_{n\to\infty}\beta^{-2n}\big(\text{Var}\big(\delta_i'\widehat{\sigma}_s^{n+1}\big) - \text{Var}\big(\delta_i'\widehat{\sigma}_s^{n}\big)\big)$$

$$= s^{-1}\pi(U)'\big(\operatorname{diag}\big(\psi\big(Q(U)\big)\big) - \psi\big(Q(U)\big)\psi\big(Q(U)\big)'\big)$$
$$\times \big(2\big(I - \Delta Q(U)/\beta\big)^{-1} - I\big)\pi(U) > 0.$$

In this case, there exists $S > 0$ such that for all $s > S$,

$$\lim_{n\to\infty} \beta^{-2n}\big(\operatorname{MSE}\big(\delta_i'\widehat{\sigma}_s^{n+1}\big) - \operatorname{MSE}\big(\delta_i'\widehat{\sigma}_s^{n}\big)\big) < 0,$$

and the limiting accuracy of the $\widehat{\sigma}_s^n$ estimator increases with the length of the simulation horizon.

Proof. First, the proof of Proposition 6 establishes that $\operatorname{Bias}(\delta_i'\widehat{\sigma}_s^n) = \delta_i'T_U^n 0 - \delta_i'T_U^\infty = -\beta^n\delta_i'Q(U)^nT_U^\infty$, which implies

$$\lim_{n\to\infty} \beta^{-2n}\big(\operatorname{Bias}\big(\widehat{\sigma}_s^{n+1}\big)^2 - \operatorname{Bias}\big(\widehat{\sigma}_s^{n}\big)^2\big) = \lim_{n\to\infty} \beta^2\big(\delta_i'Q(U)^{n+1}T_U^\infty\big)^2 - \big(\delta_i'Q(U)^nT_U^\infty\big)^2$$
$$= -\big(1-\beta^2\big)\big(\psi(U)'T_U^\infty\big)^2.$$

Second, direct manipulation yields

$$\begin{aligned}
\operatorname{Cov}(\delta_{\ell(n-t)}, \delta_{\ell(n)}) &= \operatorname{E}\big(\delta_{\ell(n-t)}\delta_{\ell(n)}'\big) - \operatorname{E}(\delta_{\ell(n-t)})\operatorname{E}\big(\delta_{\ell(n)}'\big)\\
&= \left(\sum_{j=1}^m\sum_{k=1}^m \big(\delta_i'Q(U)^{n-t}\delta_j\big)\big(\delta_j'Q(U)^t\delta_k\big)\delta_j\delta_k'\right) - Q(U)^{n-t'}\delta_i\delta_i'Q(U)^n\\
&= \left(\sum_{j=1}^m\sum_{k=1}^m \delta_j\delta_i'Q(U)^{n-t}\delta_j\delta_j'Q(U)^t\delta_k\delta_k'\right) - Q(U)^{n-t'}\delta_i\delta_i'Q(U)^n\\
&= \left(\sum_{j=1}^m \delta_j\delta_i'Q(U)^{n-t}\delta_j\delta_j'Q(U)^t\right) - Q(U)^{n-t'}\delta_i\delta_i'Q(U)^n\\
&= \left(\left(\sum_{j=1}^m \delta_j\delta_i'Q(U)^{n-t}\delta_j\delta_j'\right) - Q(U)^{n-t'}\delta_i\delta_i'Q(U)^{n-t}\right)Q(U)^t\\
&= \big(\operatorname{diag}\big(\delta_i'Q(U)^{n-t}\big) - Q(U)^{n-t'}\delta_i\delta_i'Q(U)^{n-t}\big)Q(U)^t.
\end{aligned}$$

Third, Lemma 2 implies

$$\begin{aligned}
&\big(\operatorname{diag}\big(\delta_i'Q(U)^{n-t}\big) - Q(U)^{n-t'}\delta_i\delta_i'Q(U)^{n-t}\big)(I-\Delta)\\
&\quad= \operatorname{diag}\big(\delta_i'Q(U)^{n-t}\big)(I-\Delta) - Q(U)^{n-t'}\delta_i\delta_i'(I-\Delta)\\
&\quad= \operatorname{diag}\big(\delta_i'Q(U)^{n-t}\big)\iota\delta_1' - Q(U)^{n-t'}\delta_i\delta_i'\iota\delta_1'\\
&\quad= Q(U)^{n-t'}\delta_i\delta_1' - Q(U)^{n-t'}\delta_i\delta_1'\\
&\quad= 0.
\end{aligned}$$

Fourth, the second and third points imply that

$$\operatorname{Cov}(\delta_{\ell(n-t)}, \delta_{\ell(n)}) = \bigl(\operatorname{diag}\bigl(\delta_i' Q(U)^{n-t}\bigr) - Q(U)^{n-t'}\delta_i\delta_i' Q(U)^{n-t}\bigr)\Delta Q(U)^t.$$

Fifth, set $b > \lambda(Q(U))$, so that $\beta > \lambda(Q(U))$. With this, Lemmas 3 and 5 imply that $(I - \Delta Q(U)/\beta)^{-1} = \sum_{t=0}^{\infty}(\Delta Q(U)/\beta)^t = \sum_{t=0}^{\infty}\Delta Q(U)^t/\beta^t$ exists, which implies

$$\begin{aligned}
&\lim_{n\to\infty}\sum_{t=0}^{n}\operatorname{Cov}(\delta_{\ell(n-t)}, \delta_{\ell(n)})/\beta^t \\
&\quad= \lim_{n\to\infty}\sum_{t=0}^{n}\bigl(\operatorname{diag}\bigl(\delta_i' Q(U)^{n-t}\bigr) - Q(U)^{n-t'}\delta_i\delta_i' Q(U)^{n-t}\bigr)\Delta Q(U)^t/\beta^t \\
&\quad= \bigl(\operatorname{diag}\bigl(\psi(U)\bigr) - \psi(U)\psi(U)'\bigr)\sum_{t=0}^{\infty}\Delta Q(U)^t/\beta^t \\
&\quad= \bigl(\operatorname{diag}\bigl(\psi(U)\bigr) - \psi(U)\psi(U)'\bigr)\bigl(I - \Delta Q(U)/\beta\bigr)^{-1}.
\end{aligned}$$

Sixth, this implies

$$\begin{aligned}
&\lim_{n\to\infty}\beta^{-2n}\bigl(\operatorname{Var}\bigl(\delta_i'\widehat{\sigma}_s^{n+1}\bigr) - \operatorname{Var}\bigl(\delta_i'\widehat{\sigma}_s^{n}\bigr)\bigr) \\
&\quad= \lim_{n\to\infty}\beta^{-2n}s^{-1}\left(\operatorname{Var}\left(\sum_{k=0}^{n}\beta^k\delta_{\ell(0,k)}'\pi(U)\right) - \operatorname{Var}\left(\sum_{k=0}^{n-1}\beta^k\delta_{\ell(0,k)}'\pi(U)\right)\right) \\
&\quad= s^{-1}\lim_{n\to\infty}\pi(U)'\left(-\operatorname{Cov}(\delta_{\ell(n)}, \delta_{\ell(n)}) + 2\sum_{t=0}^{n}\beta^{-t}\operatorname{Cov}(\delta_{\ell(n)}, \delta_{\ell(n-t)})\right)\pi(U) \\
&\quad= s^{-1}\pi(U)'\bigl(\operatorname{diag}\bigl(\psi(U)\bigr) - \psi(U)\psi(U)'\bigr)\bigl(2\bigl(I - \Delta Q(U)/\beta\bigr)^{-1} - I\bigr)\pi(U).
\end{aligned}$$

Finally, I establish that this quantity is positive for β sufficiently close to one. Lemmas 3 and 5 imply that $(I - \Delta Q(U)/\beta)^{-1}$ is differentiable in β at $\beta = 1$. So it suffices to consider the $\beta = 1$ case. Clearly, $\lim_{n\to\infty}\operatorname{Var}(\delta_i'\widehat{\sigma}_s^n) = \infty$ when $\beta = 1$, and thus $\lim_{n\to\infty}\operatorname{Var}(\delta_i'\widehat{\sigma}_s^{n+1}) - \operatorname{Var}(\delta_i'\widehat{\sigma}_s^n) > 0$ when $\beta = 1$. □

PROPOSITION 8. *There exists $b < 1$ such that if $\beta \in [b, 1)$ and $\lambda(Q(U)) < 1$ then*

$$\begin{aligned}
&\lim_{n\to\infty}\beta^{-2n}\bigl(\operatorname{Bias}\bigl(\delta_i'\Delta\widehat{\sigma}_s^{n+1}\bigr)^2 - \operatorname{Bias}\bigl(\delta_i'\Delta\widehat{\sigma}_s^{n}\bigr)^2\bigr) = 0, \quad \textit{and} \\
&\lim_{n\to\infty}\beta^{-2n}\bigl(\operatorname{Var}\bigl(\delta_i'\Delta\widehat{\sigma}_s^{n+1}\bigr) - \operatorname{Var}\bigl(\delta_i'\Delta\widehat{\sigma}_s^{n}\bigr)\bigr) \\
&\quad= 2s^{-1}\pi(U)'\bigl(\operatorname{diag}\bigl(\psi\bigl(Q(U)\bigr)\bigr) - \psi\bigl(Q(U)\bigr)\psi\bigl(Q(U)\bigr)'\bigr) \\
&\qquad\times\bigl(2\bigl(I - \Delta Q(U)/\beta\bigr)^{-1} - I\bigr)\pi(U) > 0.
\end{aligned}$$

In this case,

$$\lim_{n\to\infty}\beta^{-2n}\bigl(\operatorname{MSE}\bigl(\delta_i'\Delta\widehat{\sigma}_s^{n+1}\bigr) - \operatorname{MSE}\bigl(\delta_i'\Delta\widehat{\sigma}_s^{n}\bigr)\bigr) > 0,$$

and the limiting accuracy of the $\Delta\widehat{\sigma}_s^n$ estimator decreases with the length of the simulation horizon.

PROOF. This follows from Propositions 6 and 7. The variance term increases by a factor of two because $\delta_i'\Delta\widehat{\sigma}_s^n$ depends on both the sample paths deployed from state x_i and the sample paths deployed from state x_1. □

PROPOSITION 9. *If the Markov chain is ergodic under policy U, then*:

1. $\mathrm{Cond}(I-\beta Q(U))$ *is* $O(\frac{1}{1-\beta})$ *as* $\beta\to 1$. *This indicates that policy iteration's policy evaluation equations become ill-conditioned as the discount factor approaches unity.*

2. $\mathrm{Cond}(I-\beta\Delta Q(U))$ *is* $O(1)$ *as* $\beta\to 1$. *This indicates that relative policy iteration's policy evaluation equations remain well conditioned as the discount factor approaches unity.*

PROOF. With the $(I-\beta Q(U))^{-1}=\sum_{t=0}^{\infty}\beta^t Q(U)^t$ identity, it is straightforward to show that (i) $\|\!|I-\beta Q(U)|\!\| = 1+\beta-2\beta\min_{i=1}^m \delta_i' Q(U)\delta_i$, (ii) $\|\!|(I-\beta Q(U))^{-1}|\!\| = (1-\beta)^{-1}$, and (iii) $\|\!|I-\beta\Delta Q(U)|\!\| \le 1+\beta\max_{i=1}^m \|(\delta_i-\delta_1)'Q(U)\|_1$. Bounding $\lim_{\beta\to 1}\|\!|(I-\beta\Delta Q(U))^{-1}|\!\|$ is more difficult. To do so, choose $\varepsilon>0$ such that $\varepsilon\le(1-\lambda(Q(U)))/2$. And choose $n>N(\varepsilon)$, where $N(\varepsilon)$ is defined in Lemma 5. With this, Lemmas 3 and 5 imply

$$
\begin{aligned}
\left\|\!\left|\left(I-\beta\Delta Q(U)\right)^{-1}\right|\!\right\| &= \left\|\!\left|\sum_{t=0}^{\infty}\left(\beta\Delta Q(U)\right)^t\right|\!\right\| \\
&= \left\|\!\left|\sum_{t=0}^{\infty}\beta^t\Delta Q(U)^t\right|\!\right\| \\
&\le \sum_{t=0}^{\infty}\left\|\!\left|\beta^t\Delta Q(U)^t\right|\!\right\| \\
&\le \sum_{t=0}^{n}\left\|\!\left|\beta^t\Delta Q(U)^t\right|\!\right\| + \sum_{t=n}^{\infty}\left(\beta\lambda\left(Q(U)\right)+\varepsilon\right)^t \\
&\le \sum_{t=0}^{n}\left\|\!\left|\beta^t\Delta Q(U)^t\right|\!\right\| + \left(1-\beta\lambda\left(Q(U)\right)-\varepsilon\right)^{-1} \\
&\le \sum_{t=0}^{n}\left\|\!\left|\Delta Q(U)^t\right|\!\right\| + \left(\left(1-\lambda\left(Q(U)\right)\right)/2\right)^{-1},
\end{aligned}
$$

which is independent of β. □

REFERENCES

Aguirregabiria, V. and P. Mira (2002), "Swapping the nested fixed point algorithm: A class of estimators for discrete Markov decision models." *Econometrica*, 70 (4), 1519–1543. [53]

Aguirregabiria, V. and P. Mira (2007), "Sequential estimation of dynamic discrete games." *Econometrica*, 75 (1), 1–53. [53]

Aguirregabiria, V. and P. Mira (2010), "Dynamic discrete choice structural models: A survey." *Journal of Econometrics*, 156 (1), 38–67. [52, 53, 54, 55]

Arcidiacono, P. and P. B. Ellickson (2011), "Practical methods for estimation of dynamic discrete choice models." *Annual Review of Economics*, 3 (1), 363–394. [51, 54]

Arcidiacono, P. and R. Miller (2011), "Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity." *Econometrica*, 79 (6), 1823–1867. [55]

Bajari, P., L. Benkard, and J. Levin (2007), "Estimating dynamic models of imperfect competition." *Econometrica*, 75 (5), 1331–1370. [54, 55]

Bray, R. L. (2018), "Markov decision processes with exogenous variables." Under Review at *Management Science*, 1–28. [45, 50, 53, 57, 60]

Bray, R. L., Y. Yao, Y. Duan, and J. Huo (2018), "Ration gaming and the bullwhip effect." *Operations Research*, forthcoming, 1–30. [47, 48, 49]

Chen, C. (2017), "Slow focus: Belief evolution in the U.S. acid rain program." Working Paper, 1–49. [52, 53]

Doraszelski, U. and K. L. Judd (2012), "Avoiding the curse of dimensionality in dynamic stochastic games." *Quantitative Economics*, 3 (1), 53–93. [56]

Ericson, R. and A. Pakes (1995), "Markov-perfect industry dynamics: A framework for empirical work." *The Review of Economic Studies*, 62 (1), 53–82. [55, 56]

Hotz, V. J. and R. A. Miller (1993), "Conditional choice probabilities and the estimation of dynamic models." *The Review of Economic Studies*, 60 (3), 497–529. [53]

Hotz, V. J., R. A. Miller, S. Sanders, and J. Smith (1994), "A simulation estimator for dynamic models of discrete choice." *The Review of Economic Studies*, 61 (2), 265–289. [54, 55]

Judd, K. L. (1998), *Numerical Methods in Economics*. MIT Press, Cambridge, MA. [45]

Kasahara, H. and K. Shimotsu (2018), "Estimation of discrete choice dynamic programming models." *The Journal of Japanese Economic Association*, 69 (1), 28–58. [52]

Maskin, E. and J. Tirole (1988a), "A theory of dynamic oligopoly, I: Overview and quantity competition with large fixed costs." *Econometrica*, 56 (3), 549–569. [55]

Maskin, E. and J. Tirole (1988b), "A theory of dynamic oligopoly, II: Price competition, kinked demand curves, and Edgeworth cycles." *Econometrica*, 56 (3), 571–599. [55]

Morton, T. E. (1971), "On the asymptotic convergence rate of cost differences for Markovian decision processes." *Operations Research*, 19 (1), 244–248. [50, 60]

Morton, T. E. and W. E. Wecker (1977), "Discounting, ergodicity and convergence for Markov decision processes." *Management Science*, 23 (8), 890–900. [43, 46, 56, 59]

Pakes, A. and P. McGuire (1994), "Computing Markov-perfect Nash equilibria: Numerical implications of a dynamic differentiated product model." *The RAND Journal of Economics*, 25 (4), 555–589. [55, 56]

Pakes, A., M. Ostrovsky, and S. Berry (2007), "Simple estimators for the parameters of discrete dynamic games (with entry/exit examples)." *The RAND Journal of Economics*, 38 (2), 373–399. [53]

Pesendorfer, M. and P. Schmidt-Dengler (2008), "Asymptotic least squares estimators for dynamic games." *The Review of Economic Studies*, 75 (3), 901–928. [53]

Porteus, E. L. (1975), "Bounds and transformations for discounted finite Markov decision chains." *Operations Research*, 23 (4), 761–784. [56]

Puterman, M. L. (2005), *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, Hoboken, NJ. [45]

Rust, J. (1987), "Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher." *Econometrica*, 55 (5), 999–1033. [52]

Rust, J. (2000), "Nested fixed point algorithm documentation manual." Unpublished Manuscript (6), 1–43. [52]

White, D. J. (1963), "Dynamic programming, Markov chains, and the method of successive approximations." *Journal of Mathematical Analysis and Applications*, 6 (3), 373–376. [45, 57]

Zobel, C. W. and W. T. Scherer (2005), "An empirical study of policy convergence in Markov decision process value iteration." *Computers and Operations Research*, 32 (1), 127–142. [55]

Co-editor Karl Schmedders handled this manuscript.

Manuscript received 31 January, 2017; final version accepted 3 April, 2018; available online 5 April, 2018.