Polytime Computation of Strong- and n-Present-Value Optimal Policies in Markov Decision Chains

by

Michael O'Sullivan

and

Arthur F. Veinott, Jr.

Introduction

- Markov Decision Chain: $S < \infty$ states s, finite actions $a \in A_s$, stationary policy $\delta = (\delta^s)$, reward r_δ, Submarkov matrix P_δ, V^ρ_δ is present value of income with interest rate $\rho > 0$ using δ
- Optimality Concepts: Strengths and Weaknesses
  - ♦ Maximum Present Value (MPV): $V^\rho \equiv \max_\delta V^\rho_\delta$

    How small should ρ be if the distant future is relevant?

    MPV is undefined if $\rho = 0$.

    What if the problem does not involve interest?
  - ♦ Maximum Reward Rate (MRR): $\max_\delta \lim_{\rho \downarrow 0} \rho V^\rho_\delta$

    If δ has MRR, then $V^\rho_\delta - V^\rho = O(1)$.

    MRR does not reflect transient effects.

    The rewards earned in any finite horizon are irrelevant.

    In transient systems, all rewards are irrelevant.

    MRR policies are not nearly SPV (see below) optimal.
  - ♦ Limiting Present Value [Bl62]: $V^\rho_\delta - V^\rho = o(1)$

    LPV is stronger than MRR.

    LPV reflects transient effects.

    LPV policies are nearly SPV optimal.

LPV is the right way to define MPV when $\rho = 0$.

In transient systems, LPV is equivalent to MV.

- n-Present Value [Ve69]: $V_\delta^\rho - V^\rho = o(\rho^n)$, $n \quad -1$

  n-PV polices are MRR if $n = -1$ and LPV if $n = 0$.

  n-PV policies are nearly SPV for $n \quad 0$.

  The quality of n-PV policies rises with n.

  The set of n-PV policies diminishes with n.

  Computation: Find n-PV Opt given $(n-1)$-PV Opt

- Strong Present Value [Bl62]: $V_\delta^\rho - V^\rho = 0$, $\rho^* > \rho > 0$

  Such policies are robust under small changes in ρ.

  The sets of n-PV & SPV policies coincide for $n \quad S$.

  SPV policies are MRR and SPV for all $n \geq -1$.

  But SPV policies can be more expensive to compute.

- Extensions
  - Immigration [RV92]
  - Perturbation of Interest Rates [Ve69]
  - Long Time Horizons: n-Cesàro-overtaking optimality
  - Continuous Time Parameter
  - Markov Population Decision Chains [RV92]

Background [MV69, Ve69, Ve74]

- Laurent Expansion of $V^\rho_\delta = \sum_{-1}^{\infty} \rho^n v^n_\delta$
- δ is n-Optimal if $V^n_\delta \succeq V^n_\gamma$ all γ, $V^n_\gamma \equiv \left(v^{-1}_\gamma, \ldots, v^n_\gamma\right)$
- $\Delta_n \equiv$ Set of (stationary) n-Optimal Policies
  - ♦ $\delta \in \Delta_{-1} \Leftrightarrow \delta$ Maximum Reward Rate (MRR)
  - ♦ $\delta \in \Delta_0 \quad \Leftrightarrow \delta$ Present-Value Optimal $\rho = 0$
- Selectivity Increases with n: $\Delta_n \supseteq \Delta_{n+1}$
- S-Optimal $\Leftrightarrow$ Strong Present-Value Optimal
- Comparison $\qquad g^n_{\gamma\delta} \equiv r^n_\gamma + Q_\gamma v^n_\delta - v^{n-1}_\delta$, $n \quad -1$

$$Q_\gamma \equiv P_\gamma - I,\ r^0_\gamma \equiv r_\gamma,\ r^n_\gamma \equiv 0 \text{ all } n \neq 0$$

$$r_\gamma + P_\gamma V^\rho_\delta - (1+\rho)V^\rho_\delta = r_\gamma + Q_\gamma V^\rho_\delta - \rho V^\rho_\delta = \sum_{-1}^{\infty} \rho^n g^n_{\gamma\delta}$$

- $r^n_\delta + Q_\delta v^n = v^{n-1}$, $n \leq m+1 \ \Rightarrow V^m = V^m_\delta$
- $\delta \in \Delta_n \Rightarrow G^n_{\gamma\delta} \equiv (g^{-1}_{\gamma\delta}, \ldots, g^n_{\gamma\delta}) \preceq 0$ all γ
- $\delta \in \Delta_n \Leftarrow G^{n+1}_{\gamma\delta} \preceq 0$ all $\gamma \qquad$ n-Optimality Condition
- Policy-Improvement Method
- $\mathcal{E}^n_\delta \equiv \{\gamma : G^n_{\gamma\delta} = 0\}$, $n \geq -1$
- $\delta \in \Delta_{n-1} \Rightarrow \mathcal{E}^n_\delta \subseteq \Delta_{n-1} \subseteq \mathcal{E}^{n-1}_\delta$

Maximum Transient Value (MTV)
[Ve69], [EV72]

- State s is Transient or Recurrent under δ according as s^{th} Column of P_δ^* is 0 or Not.

- δ is Transient or Recurrent according as P_δ^* is 0 or Not.

- If δ is Transient, its Value is $V_\delta = \sum_{i=1}^{\infty} P_\delta^{i-1} r_\delta$.

- δ has MTV if $V_\delta \quad V_\gamma$, all Transient γ.

- Characterization of MTV. If δ Transient, δ has MTV $\Leftrightarrow$ $V = V_\delta$ is Least Fixed (resp., Excessive) Point of $\mathcal{R}$,

$$\mathcal{R}V \equiv \max_\gamma \, (r_\gamma + P_\gamma V), \text{ all } V \in \Re^S.$$

- Dual Linear Program to Find MTV. Choose V to Minimize $1V$ Subject to $V \quad r_\gamma + P_\gamma V$, all γ. Each Optimal Basis Corresponds to MTV δ.

Maximum Reward-Rate (MRR)

[Ho60], [Ba61], [Bl62], [DF68], [HK78]

- Policy Improvement for $n = -1$
- Dual Linear Programming

Given $w \gg 0$, find v^{-1} and v^{0} that

$$\begin{array}{lll} \text{minimize} & w v^{-1} & \\ \text{subject to} & v^{-1} - Q_\gamma v^{0} & r_\gamma \\ & -Q_\gamma v^{-1} & 0 \end{array}$$

for all $\gamma \in \Delta$.

This linear program has optimal solution $(v^{-1} \;\; v^{0})$; $v^{-1} = \max_\gamma v_\gamma^{-1}$. There is a procedure to find a δ such that $v^{-1} = v_\delta^{-1}$.

Each such δ has MRR.

Sequential Decomposition Method (OV)

For finding an n-optimal policy

- Given δ is $(n-1)$-Optimal
  - (a) Find ζ Satisfying $(n-1)$-Optimality Conditions.
  - (b) Find η that is n-Optimal on States Recurrent under some n-Optimal Policy.
  - (c) Find θ that is n-Optimal.

Theorem. Sequential Decomposition.

(a) Given $(n-1)$-Optimal δ,

- Solve MTV Problem: Find Least Solution v^n of

$$v^n = \max_{\gamma \in \mathcal{E}_\delta^{n-1}} (r_\gamma^n - v_\delta^{n-1} + P_\gamma v^n) \vee v_\delta^n,$$

so $v^n = V_\kappa$ has MTV in this System $(\mathcal{E}_\delta^{n-1} \supseteq \Delta_{n-1})$.

- Set $\zeta^s = \kappa^s$ if $V_{\kappa s} > v_{\delta s}^n$ & $\zeta^s = \delta^s$ if $V_{\kappa s} = v_{\delta s}^n$.
- $v_\zeta^n = V_\kappa$ & ζ Satisfies $(n-1)$-Optimality Conditions.

(b) Given ζ Satisfying $(n-1)$-Optimality Conditions,

- $$\max_{\gamma \in \mathcal{E}_\zeta^n} v_\gamma^n = v_\zeta^n + \max_{\gamma \in \mathcal{E}_\zeta^n} P_\gamma^* (r_\gamma^{n+1} - v_\zeta^n).$$

- Each η solving above MRR Problem is $(\mathcal{E}_\zeta^n \subseteq \Delta_{n-1})$
  - $(n-1)$-Optimal and
  - n-Optimal on States Recurrent under some

  n-Optimal Policy.

(c) Given η Satisfying Last Two Conditions,

- Solve MTV Prob: Find Least Solution v^n of

$$v^n = \max_{\gamma \in \mathcal{E}_\eta^{n-1}} (r_\gamma^n - v_\eta^{n-1} + P_\gamma v^n) \vee v_\eta^n,$$

so $v^n = V_\kappa$ is MTV in this System $(\mathcal{E}_\eta^{n-1} \supseteq \Delta_{n-1})$.

- Set $\theta^s = \kappa^s$ if $V_{\kappa s} > v_{\eta s}^n$ & $\theta^s = \eta^s$ if $V_{\kappa s} = v_{\eta s}^n$.
- $v_\theta^n = V_\kappa$ and θ is n-Optimal.

Remarks: n-Optimality-Problem

- $2n+3$ MTV and $n+2$ MRR Subproblems
  - ♦ Each Solvable in Polytime by LP in Original Data

    To see why, consider Theorem 1a. The inputs to the MTV problem there are $v_\delta^{n-1}$and v_δ^n. They can be found in polytime by solving $r_\delta^m + Q_\delta v^m = v^{m-1}$, $m \le n{+}1 \;\Rightarrow V^n = V_\delta^n$.
  - ♦ Each Solvable by Policy Improvement
  - ♦ Subproblem Solutions Distinct
  - ♦ Subproblems Independent of Solution Method
  - ♦ State-Classification Refinements
  - ♦ More Efficient Algorithms for Special Structures
- n-Optimality-Problem Solvable in Polytime
- n-Optimality-Conditions Problem

Unique Transition Systems [OV]

- Unique Transition System: One in which Action is Next (Non-Stopped) State Visited

- Includes Standard Deterministic Systems

- Finding $(n-2)$-Optimal Policy is Solvable in Strongly Polynomial Time, viz.,
  - ♦ $O(nS[A \wedge S^2] + A)$ Time — Undiscounted
  - ♦ $O(nS^2[A \wedge S^2] + A)$ Time — Discounted
  - ♦ $O(nS^2 A \log S)$ Time — General

($A = \#$ state-action pairs)