Automatica 153 (2023) 111024

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Brief paper

Interval dominance based structural results for Markov decision process☆

Vikram Krishnamurthy

School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, 14853, USA

ARTICLE INFO

Article history:
Received 8 April 2022
Received in revised form 15 December 2022
Accepted 18 February 2023
Available online xxxx

Keywords:
MDP
Interval Dominance
Monotone policy
Supermodularity
TP3 stochastic matrices
Reinforcement learning

ABSTRACT

Structural results impose sufficient conditions on the model parameters of a Markov decision process (MDP) so that the optimal policy is an increasing function of the underlying state. The classical assumptions for MDP structural results require supermodularity of the rewards and transition probabilities. However, supermodularity does not hold in many applications. This paper uses a sufficient condition for interval dominance (called $\mathcal{I}$) proposed in the micro-economics literature, to obtain structural results for MDPs under more general conditions. We present several MDP examples where supermodularity does not hold, yet $\mathcal{I}$ holds, and so the optimal policy is monotone; these include sigmoidal rewards (arising in prospect theory for human decision making), bi-diagonal and perturbed bi-diagonal transition matrices (in optimal allocation problems). We also consider MDPs with TP3 transition matrices and concave value functions. Finally, reinforcement learning algorithms that exploit the differential sparse structure of the optimal monotone policy are discussed.

1. Introduction

Markov decision processes (MDPs) are controlled Markov chains. Brute force numerical solution to compute the optimal policy of an MDP with a large state and action space is expensive and yields little insight into the structure of the controller. *Structural results* for MDPs are widely studied in stochastic control, operations research and economics (Amir, 2005; Heyman & Sobel, 1984; Puterman, 1994; Topkis, 1998). They impose sufficient conditions on the parameters of an MDP model so that there exists an optimal policy $\mu^*(x)$ that is increasing[1] in the state x, denoted as $\mu^*(x) \uparrow x$. Such monotone optimal policies are useful as they yield insight into the structure of the optimal controller of the MDP. Put simply, they provide a mathematical justification for rule of thumb heuristics such as choose a "larger" control action for a "larger" state. Also, since monotone optimal policies are differentially sparse (see Section 5), optimization algorithms and reinforcement learning algorithms that exploit this sparsity can solve the MDP efficiently (Krishnamurthy, 2016; Mattila, Rojas, Krishnamurthy, & Wahlberg, 2017).

The classical assumption (Heyman & Sobel, 1984; Puterman, 1994) for the existence of a monotone policy in a MDP relies on supermodularity (Liu, Chong, Pezeshki, & Zhang, 2020; Topkis, 1998). By imposing supermodularity conditions on the rewards and transition probabilities of the MDP, the classical proof shows that the Q function in Bellman's dynamic programming equation is supermodular. (These conditions are reviewed in Section 2.) With $\mathcal{X} = \{1, \ldots, X\}$, $\mathcal{A} = \{1, \ldots, A\}$ denoting a finite state space and action space, recall (Topkis, 1998) that $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is supermodular[2] if it has increasing differences:

$$\phi(\bar{x}, \bar{a}) - \phi(\bar{x}, a) \geq \phi(x, \bar{a}) - \phi(x, a), \quad \bar{x} > x,\ \bar{a} > a. \tag{1}$$

Then the well known Topkis' theorem (Topkis, 1998) states that supermodularity is a sufficient condition for

$$a^*(x) \in \operatorname*{arg\,max}_{a \in \mathcal{A}} \phi(x, a) \uparrow x. \tag{2}$$

So if it can be shown for an MDP that its Q function is supermodular, then Topkis theorem implies that there exists an optimal policy that is monotone: $\mu^*(x) \in \arg\max_{a \in \mathcal{A}} Q(x, a) \uparrow x$.

However, supermodularity is a restrictive sufficient condition for the existence of a monotone optimal policy; it imposes conditions on the rewards and transition probabilities that may not hold in many cases.

☆ This research was partially supported by U.S. Army Research Office grant W911NF-21-1-0093, Air Force Office of Scientific Research, USA grant FA9550-22-1-0016 and National Science Foundation, USA grant CCF-2112457. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Valery Ugrinovskii under the direction of Editor Ian R. Petersen.

E-mail address: vikramk@cornell.edu.

[1] We use increasing in the weak sense to mean non-decreasing.

[2] More generally supermodularity applies to lattices with a partial order (Topkis, 1998). In our simple setup of (1), Puterman (1994) uses the terminology 'superadditive'.

https://doi.org/10.1016/j.automatica.2023.111024
0005-1098/© 2023 Elsevier Ltd. All rights reserved.

Recently, Quah and Strulovici (2009) introduced the *Interval Dominance* condition which is necessary and sufficient for (2) to hold. For the purposes of our paper, Quah and Strulovici (2009, Proposition 3) gives the following useful sufficient condition[3] for $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ to satisfy interval dominance:

$$\phi(\bar{x}, a+1) - \phi(\bar{x}, a) \geq \alpha_{x,\bar{x},a}\big[\phi(x, a+1) - \phi(x, a)\big], \quad \bar{x} > x \tag{3}$$

where the scalar valued function $\alpha_{x,\bar{x},a} > 0$ (strictly non-negative) is increasing in a. We symbolically denote (3) as the condition $(\phi, \alpha) \in \mathcal{I}$. Comparing supermodularity (1) with $\mathcal{I}$ in (3), we see that supermodularity is a special case of $\mathcal{I}$ when $\alpha_{x,\bar{x},a} = 1$. An important property of $\mathcal{I}$ is that it compares adjacent actions a and $a+1$. A more restrictive condition would be to replace $a+1$ with any action $\bar{a} > a$ in (3). However, this stronger condition (which in analogy to (1) can be called α-supermodularity) is highly restrictive and does not hold for MDP examples considered below.

Main Results. This paper shows how $\mathcal{I}$ in (3) applies to obtain structural results for MDPs under more general conditions than the classical supermodularity conditions. Theorems 1 and 2 are our main results. To avoid technicalities we consider finite state, finite action MDPs which are either finite horizon or discounted reward infinite horizon. We present several MDP examples where the Q functions satisfy $\mathcal{I}$ but not supermodularity, and the optimal policy is monotone. One important class comprises MDPs with sigmoidal and concave rewards; since a sigmoidal function comprises convex and concave segments, supermodularity rarely holds. Such sigmoidal rewards arise in prospect theory (behavioral economics) based models for human decision making (Kahneman & Tversky, 1979). A second important class of examples we will consider involves perturbed bi-diagonal transition matrices for which the standard supermodularity assumptions do not hold. Bi-diagonal transition matrices arise in optimal allocation with penalty costs (Derman, Lieberman, & Ross, 1976; Ross, 1983). The result in the Appendix complements this classical result for possibly non-submodular costs. Finally, a third class of examples comprises MDPs with integer concave value functions. Theorem 2 and Corollary 5 impose TP3 (totally positive of order 3) assumptions along with $\mathcal{I}$ to show that the optimal policy is monotone. An extension of the classical TP3 result of Karlin (1968, pg 23) is proved to characterize the $\mathcal{I}$ condition for MDPs with bi-diagonal and tri-diagonal transition matrices. Such MDPs model controlled random walks (Puterman, 1994) and arise in the control of queuing and manufacturing systems.

2. Background. Supermodularity based results

An infinite horizon discounted reward MDP model is the tuple $(\mathcal{X}, \mathcal{A}, (P(a), r(a), a \in \mathcal{A}), \rho)$. Here $\mathcal{X} = \{1, \ldots, X\}$ denotes the finite state space, and we will denote $x_k \in \mathcal{X}$ as the state at time $k = 0, 1, \ldots$. Also $\mathcal{A} = \{1, \ldots, A\}$ is the action space, and we will denote $a_k \in \mathcal{A}$ as the action chosen at time k. $P(a)$ are $X \times X$ stochastic matrices with elements $P_{ij}(a) = \mathbb{P}(x_{k+1} = j | x_k = i, a_k = a)$, $r(a)$ are X dimensional reward vectors with elements denoted $r(x, a)$, and $\rho \in (0, 1)$ is the discount factor.

The action at each time k is chosen as $a_k = \mu(x_k)$ where μ denotes a stationary policy $\mu : \mathcal{X} \to \mathcal{A}$. The optimal stationary policy $\mu^* : \mathcal{X} \to \mathcal{A}$ is the maximizer of the infinite horizon discounted reward J_μ:

$$\begin{aligned} &\mu^*(x) \in \arg\max_{\mu} J_\mu(x), \\ &J_\mu(x) = \mathbb{E}_\mu\{\sum_{k=0}^{\infty} \rho^k r(x_k, a_k) \mid x_0 = x\} \end{aligned} \tag{4}$$

The optimal stationary policy μ^* satisfies Bellman's dynamic programming equation

$$\begin{aligned} &\mu^*(x) \in \arg\max_{a \in \mathcal{A}}\{Q(x, a)\}, \;\; V(x) = \max_{a \in \mathcal{A}}\{Q(x, a)\}, \\ &Q(x, a) = r(x, a) + \rho \sum_{j=1}^{X} P_{xj}(a) V(j) \end{aligned} \tag{5}$$

An MDP with finite horizon N is the tuple

$(\mathcal{X}, \mathcal{A}, (P(a), r(a), a \in \mathcal{A}), \tau)$ where τ is the X-dimensional terminal reward vector. (In general $P(a)$ and $r(a)$ can depend on time k; for notational convenience we suppress this time dependency.) The optimal policy sequence $\mu_0, \ldots, \mu_{N-1}$ is given by Bellman's recursion: $V_N(x) = \tau_x$, $x \in \mathcal{X}$, and for $k = 0, \ldots, N$,

$$\begin{aligned} &\mu_k^*(x) \in \arg\max_{a \in \mathcal{A}}\{Q_k(x, a)\}, \;\; V_k(x) = \max_{a \in \mathcal{A}}\{Q_k(x, a)\} \\ &Q_k(x, a) = r(x, a) + \sum_{j=1}^{X} P_{xj}(a) V_{k+1}(j) \end{aligned} \tag{6}$$

Monotone policies using supermodularity

The classical supermodularity assumptions for an MDP are:

(A1) Rewards $r(x, a) \uparrow x$ for each a.
(A2) $P_x(a) \leq_s P_{x+1}(a)$ for each x, a, where $P_x(a)$ is the xth row of matrix $P(a)$.[4]
(A3) $r(x, a)$ is supermodular in (x, a).
(A4) $\sum_{j \geq l} P_{xj}(a)$ is supermodular in x, a for each $l \in \mathcal{X}$.
(A5) The terminal reward $\tau_x \uparrow x$.

The following textbook result establishes Q_k and Q are supermodular; so the optimal policy is monotone:

Proposition 1 (*Heyman & Sobel, 1984; Puterman, 1994*). *(i) For a discounted reward MDP, under (A1)–(A4), the optimal policy $\mu^*(x)$ in (5)[5] $\uparrow x$.*

(ii) For a finite horizon MDP, under (A1)–(A5), the optimal policy sequence $\mu_k^*(x)$, $k = 0, \ldots, N-1$, satisfying (6) $\uparrow x$.

3. MDP structural results using interval dominance

The supermodular conditions (A3), (A4) on the rewards and transition probabilities, are restrictive. We relax these with the interval dominance condition $\mathcal{I}$ defined in (3) as follows:

(A6) For $\beta_{x,\bar{x},a} > 0$ and $\uparrow a$, the rewards satisfy

$$r(\bar{x}, a+1) - r(\bar{x}, a) \geq \beta_{x,\bar{x},a}\left[r(x, a+1) - r(x, a)\right], \; \bar{x} > x$$

[3] If $\alpha_{x,\bar{x},a}$ is a fixed constant independent of $x, \bar{x}, a$, then (3) is sufficient for the single crossing property (Milgrom & Shannon, 1994), namely, RHS of (1) ≥ 0 implies LHS of (1) ≥ 0. Supermodularity implies single crossing which in turn implies interval dominance; see also Amir (2005) for a tutorial exposition. The condition (3) is sufficient for interval dominance and is the main condition that we will use.

[4] $\leq_s$ denotes first order stochastic dominance, namely, $\sum_{j=l}^{X} P_{x,j}(a) \leq \sum_{j=l}^{X} P_{x+1,j}(a)$, $l \in \mathcal{X}$.

[5] More precisely, there exists a version of the optimal policy that is nondecreasing in x. (4) uses the notation $\in$ since the optimal policy is not necessarily unique.

(A7) With $\bar{x} > x$ and $\alpha_{x,\bar{x},a} > 0 \uparrow a$, the transition probabilities satisfy (recall $\geq_s$ denotes first order dominance)

$$\frac{P_{\bar{x}}(a+1) + \alpha_{x,\bar{x},a}\, P_x(a)}{1+\alpha_{x,\bar{x},a}} \geq_s \frac{P_{\bar{x}}(a) + \alpha_{x,\bar{x},a}\, P_x(a+1)}{1+\alpha_{x,\bar{x},a}}.$$

(A8) There exist $\alpha_{x,\bar{x},a} = \beta_{x,\bar{x},a}$ for which (A6), (A7) hold.

Remark. If $\alpha_a = \beta_a = 1$, then (A6) and (A7) are equivalent to supermodularity conditions (A3) and (A4). Note that (A8) is sufficient for the sum of two $\mathcal{I}$ functions to be $\mathcal{I}$.

(A6) and (A7) compare adjacent actions $a+1$ and a. A more restrictive condition is to replace $a+1$ with any action $\bar{a} > a$ in (A6) and (A7). This stronger condition does not hold in the MDP examples below. This is the reason why the $\mathcal{I}$ condition is useful.

Main Result. The following is our main result.

Theorem 1. *(i) For a discounted reward MDP, under (A1), (A2), (A6), (A7), (A8), there exists an optimal stationary policy $\mu^*(x)$ satisfying (5) which is $\uparrow x$.*
(ii) For a finite horizon MDP, under (A1), (A2), (A5), (A6), (A7), (A8), there exists an optimal policy sequence $\mu_k^*(x)$, $k = 0, \ldots, N$ satisfying (6) which is $\uparrow x$.

Remark. Theorem 1 also holds for average reward MDPs that are unichain (Puterman, 1994) so that a stationary optimal policy exists. This is because our proof uses the value iteration algorithm, and for average reward problems, the same ideas directly apply to the relative value iteration algorithm.

Proof. The standard textbook proof (Puterman, 1994) shows via induction that for the finite horizon case, (A1), (A2), (A5) imply that $Q_k(x,a) \uparrow x$ for each $a \in \mathcal{A}$, and therefore $V_k(x) \uparrow x$. The induction step also constitutes the value iteration algorithm for the infinite horizon case, and shows that $Q(x,a)$ and $V(x) \uparrow x$.

Next, since $V(x) \uparrow x$, (A7) implies that for $\bar{x} > x$,

$$\sum_{j=1}^{X} \big[P_{\bar{x},j}(a+1) - P_{\bar{x},j}(a)\big]V(j)$$
$$\geq \alpha_{x,\bar{x},a}\Big(\sum_{j=1}^{X}\big[P_{x,j}(a+1) - P_{x,j}(a)\big]V(j)\Big) \tag{7}$$

Assumption (A6) implies the rewards satisfy $\mathcal{I}$. Finally, (A8) implies for $\bar{x} > x$,

$$r(\bar{x}, a+1) - r(\bar{x}, a) + \sum_{j=1}^{X}\big[P_{\bar{x},j}(a+1) - P_{\bar{x},j}(a)\big]V(j)$$
$$\geq \gamma_{x,\bar{x},a}\left[r(x, a+1) - r(x,a) + \sum_{j=1}^{X}\big[P_{x,j}(a+1) - P_{x,j}(a)\big]V(j)\right]$$

for $\gamma = \alpha = \beta$. Thus $(Q, \gamma) \in \mathcal{I}$ implying that (2) holds. □

3.1. Example 1. MDPs with interval dominant rewards

Our first example considers MDPs with sigmoidal and concave[6] rewards specified in Example (i) below. Let us give some visual intuition. Supermodularity is difficult to ensure since a sigmoidal reward comprises a convex segment followed by a concave segment. In Fig. 1(a), reward $r(x,1)$ is sigmoidal, while $r(x,2)$ and $r(x,3)$ are concave in x. Since concave reward $r(x,3)$ intersects sigmoidal reward $r(x,1)$ multiple times, the single crossing condition and therefore supermodularity (A3) does not hold. More directly, $r(x,3) - r(x,1)$ is not increasing and so not supermodular. But condition $\mathcal{I}$ (A6) holds. Specifically, $r(x,2) - r(x,1)$ is single crossing, and $r(x,3) - r(x,2)$ is single crossing. Note that $\mathcal{I}$ does not require $r(x,3) - r(x,1)$ to be single crossing.

Consider a discounted reward MDP. Assume:

(Ex.1) For each pair of actions $a, a+1$, assume there is state x_a^* such that $r(x,a+1) \leq r(x,a)$, $P_x(a+1) \leq_s P_x(a)$ for $x \leq x_a^*$. Also $r(x,a+1) \geq r(x,a)$, $P_x(a+1) \geq_s P_x(a)$ for $x \geq x_a^*$.

Corollary 1. *Consider a discounted reward MDP. Assume (A1), (A2), (Ex.1). Then Theorem 1 holds.*

Compared to Proposition 1, Corollary 1 does not impose supermodularity conditions on the rewards or transition probabilities. (Ex.1) is weaker than the single crossing condition.

Proof. We verify that condition (A6), (A7), (A8) of Theorem 1 hold:

First consider $x < \bar{x} \leq x_a^*$. Since $r(x,a) \geq r(x,a+1)$, and $r(\bar{x},a) \geq r(\bar{x},a+1)$, (A6) holds for all $\beta \in [\beta^*_{x,\bar{x},a}, \infty)$ for some $\beta^*_{x,\bar{x},a} > 0$. Also $P_x(a+1) \leq_s P_x(a)$ implies (A7) holds for all $\alpha_{x,\bar{x},a} \in [\alpha^*_{x,\bar{x},a}, \infty)$ for some $\alpha^*_{x,\bar{x},a} > 0$. So we can choose $\alpha = \beta = \max_a\{\alpha^*_{x,\bar{x},a}, \beta^*_{x,\bar{x},a}\}$ independent of a so that (A8) holds.

Next consider $\bar{x} > x \geq x_a^*$. Then (A6) holds for all $\beta \in (0, \beta^*_{x,\bar{x},a}]$ for some $\beta^*_{x,\bar{x},a} > 0$. Also $P_x(a+1) \geq_s P_x(a)$ implies (A7) holds for all $\alpha \in (0, \alpha^*_{x,\bar{x},a}]$ for some $\alpha^*_{x,\bar{x},a} > 0$. Therefore, we can choose $\alpha = \beta = \min_a\{\alpha^*_{x,\bar{x},a}, \beta^*_{x,\bar{x},a}\}$ independent of a so that (A8) holds. Finally, for $x \leq x_a^*$ and $\bar{x} > x_a^*$, (A6) and (A7) hold for all $\alpha, \beta > 0$. So Theorem 1 applies and $\mu^*(x) \uparrow x$. □

Example (i). Sigmoidal[7] and Concave Rewards

The following MDP parameters satisfy Corollary 1: $X = 201$, $A = 3$. The action dependent transition matrices are $P_i(1) = P_{i-1}(1) + \mu(e_X - e_1)$, $\mu = \frac{0.004}{X}$, $\epsilon = \frac{0.05}{X}$,

$$P_i(a+1) = \begin{cases} P_i(a) - \epsilon(e_X - e_1), & i \leq 50, \\ P_i(a) + \epsilon(e_X - e_1), & i > 50 \end{cases}$$

Here e_i denotes the unit X-dimension row vector with 1 in the ith position.

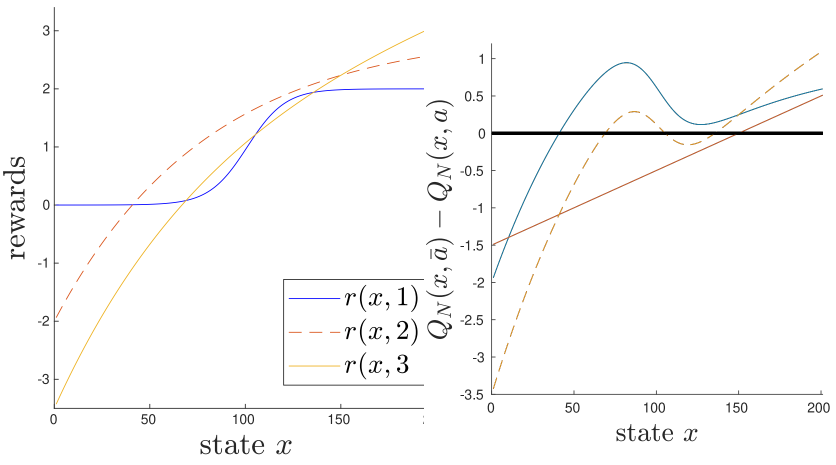
With $\theta = [2, X-1, 20, 5, 80, -2, 5, 80, -3.5, 0.01]$,

$$\begin{aligned} r(x,1) &= \frac{\theta_1}{1+\exp(\frac{x-\theta_2}{\theta_3})}\ \text{(sigmoidal)}\ , \\ r(x,2) &= \theta_4(1-\exp(-\frac{x}{\theta_5})) + \theta_6\ \text{(concave)}\ , \\ r(x,3) &= \theta_7(1-\exp(-\frac{x}{\theta_8})) + \theta_9 + \theta_{10}\,x\ \text{(concave)} \end{aligned} \tag{8}$$

Fig. 1(b) shows the non-supermodular Q_N for $N = 100$, $\rho = 0.9$. $Q_N(x,3) - Q_N(x,1)$ (broken line) intersects the horizontal axis three times; so single crossing does not hold. $Q_N(x,2) - Q_N(x,1)$ (blue line) is non-monotone (non-supermodular). Statement 1, Corollary 1 applies; so the optimal policy is monotone.

[6] Throughout this paper convex (concave) means integer convexity (concavity). Since $x \in \{1, \ldots, X\}$, integer convex ϕ means $\phi(x+1) - \phi(x) \geq \phi(x) - \phi(x-1)$. We do not consider higher dimensional discrete convexity such as multimodularity; see Section 5.

[7] Sigmoidal rewards/costs are ubiquitous. They arise in logistic regression, prospect theory in behavioral economics, and wireless communications.



(a) Rewards

(b) Q-function for MDP

Fig. 1. Interval Dominant Rewards that are not single crossing and so not supermodular. If supermodularity holds then the curves would be increasing with x. Yet $\mathcal{I}$ holds by Corollary 1 and the optimal policy is monotone; see Example (i). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Example (ii). Prospect theory based rewards

In prospect theory (Kahneman & Tversky, 1979), an agent (human decision maker) a has utility $r(x,a)$ that is asymmetric sigmoidal in x. This asymmetry reflects a human decision maker's risk seeking behavior (larger slope) for losses and risk averse behavior (smaller slope) for gains. With X an even integer, the prospect theory rewards are asymmetric sigmoidals:

$$r(x,a) = \frac{2(\mu\,(x-1))^{\theta(a)}}{1+(\mu\,(x-1))^{\theta(a)}} - 1, \quad \mu = 2/(X-2),\ \theta(a) > 1 \quad (9)$$

so they cross zero at $x = X/2$. The shape parameter $\theta(a)$ determines the slope of the reward curve $r(x,a)$.

Suppose the agents (investment managers) range from $a = 1$ (cautious) to $a = A$ (aggressive); so the shape parameter $\theta(a) \uparrow a$. The value of an investment evolves according to Markov chain x_k with transition probabilities $P(a_k)$ based on agent a_k. Since agent $a+1$ is more aggressive (risk seeking) than agent a in losses and gains, it incurs higher volatility. So the xth row of $P(a)$ and $P(a+1)$ satisfy

$$P_x(a+1) \leq_s P_x(a),\ x < \frac{X}{2};\ P_x(a+1) \geq_s P_x(a),\ x \geq \frac{X}{2} \quad (10)$$

The aim is to choose the optimal agent a_k at each time k to maximize the discounted infinite horizon reward. Since $r(x,a)$ is single crossing but not supermodular, (A3) does not apply.

Corollary 2. *Consider a discounted reward MDP with $r(x,a)$ specified by (9) and $\theta(a) \uparrow a$. Assume (A2), (10) hold. Then Theorem 1 holds.*

The proof follows from Corollary 1 with $x^*_a = X/2$.

3.2. Example 2. Interval dominant transition probabilities

Corollary 3. *Consider the discounted reward MDP with $r(x,a) = \phi(x)$ where $\phi \uparrow x$ and non-negative. Suppose the ith row of transition matrix $P(a)$ is*

$$P_i(a) = p + \Delta_{i,a}\,(e_X - e_1) \quad (11)$$

Here e_i denotes the unit X-dimension row vector with 1 in the ith position. p is an arbitrary X-dimensional probability row vector. Also $\Delta_{1,a} = 0$, $\Delta_{i,a} \in [0,1]$ are $\uparrow i$, and satisfy $\mathcal{I}$ (3). (Also, $\Delta_{i,a} \leq \min\{p_1, 1-p_X\}$ to ensure $P(a)$ is valid transition matrix.) Then Theorem 1 holds.

Compared to supermodularity (A4) of the transition probabilities, Corollary 3 imposes weaker conditions: Δ satisfy $\mathcal{I}$ (3) and p can be any probability vector. Since Δ only needs to satisfy $\mathcal{I}$ (suitably scaled and shifted to ensure valid probabilities), (11) offers considerable flexibility in choice of the transition matrices.

Proof. Reward $r(x,a) = \phi(x)$ satisfies (A1), (A6) for all $\beta_{x,\bar{x},a} > 0$. Also $\Delta_{x,a} \uparrow x$ implies (A2) holds. Next let us verify (A7). Using (11), we need to verify

$$(\Delta_{\bar{x},a+1} - \Delta_{\bar{x},a}) \sum_{j\geq l}(e_X - e_1)'e_j$$
$$\geq \alpha_{x,\bar{x},a}\Big[(\Delta_{x,a+1} - \Delta_{x,a}) \sum_{j\geq l}(e_X - e_1)'e_j\Big] \quad (12)$$

where $\alpha_{x,\bar{x},a} > 0 \uparrow a$. Since $\sum_{j\geq l}(e_X - e_1)'e_j \geq 0$, clearly $\Delta_{i,a}$ satisfying (3) for some $\alpha_{x,\bar{x},a} > 0 \uparrow a$ is a sufficient condition for (12) to hold. Since $\beta_{x,\bar{x},a} > 0$ is unrestricted, we can choose $\beta_{x,\bar{x},a} = \alpha_{x,\bar{x},a}$. Hence (A8) holds. Thus Theorem 1 holds. □

Example. Suppose p is an arbitrary probability vector, and Δ is chosen as the rewards (8) suitably scaled and shifted. Then the transition matrices inherit the sigmoidal and concave structures of Section 3.1.

3.3. Example 3. Perturbed bi-diagonal transition matrices

This section illustrates the $\mathcal{I}$ condition in MDPs with perturbed bi-diagonal transition matrices. The Appendix discusses a finite horizon MDP example in optimal allocation problems with penalty costs (Derman et al., 1976; Ross, 1983). It also has applications in wireless transmission control (Ngo & Krishnamurthy, 2010).

Consider an infinite horizon discounted reward MDP with $P^\epsilon(a), a \in \mathcal{A}$ specified by parameter $p_a \in [0,1]$ are $P^\epsilon_{X,X-1}(a) = p_a$, $P^\epsilon_{X,X}(a) = 1 - p_a$

$$\begin{aligned} &P^\epsilon_{11}(a) = 1 - (A-a)\epsilon, \quad P^\epsilon_{1,X}(a) = (A-a)\epsilon, \\ &P^\epsilon_{ii}(a) = 1 - p_a - (A-a)\epsilon, \quad P^\epsilon_{i,i-1}(a) = p_a, \\ &P^\epsilon_{i,X}(a) = (A-a)\epsilon, \quad i = 2,\dots,X-1 \end{aligned} \quad (13)$$

where $\epsilon \ll 1$ is a small positive real. We assume that $p_a \uparrow a$. When $\epsilon = 0$, $P^\epsilon(a)$ are bi-diagonal transition matrices; so ϵ can

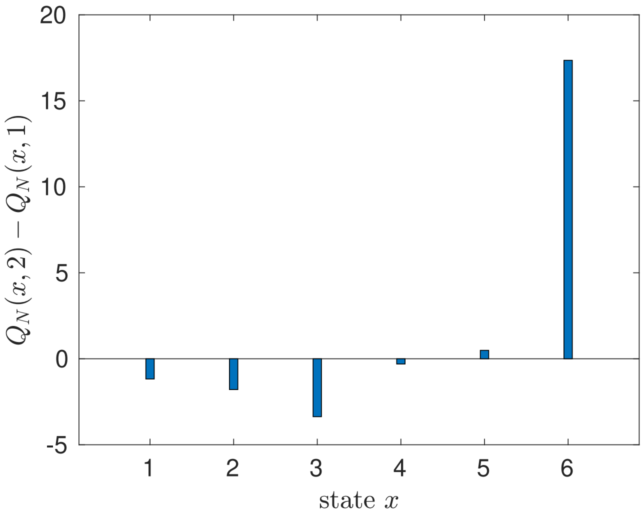


Fig. 2. The Q-function is not-supermodular for an MDP with perturbed bi-diagonal matrices; yet the optimal policy $\mu^*(x)$ is increasing in state x by Corollary 4.

be viewed as a perturbation probability of a bi-diagonal transition matrix.

Supermodularity (A4) of the transition matrices (13) holds if $\epsilon \geq p_{a+1} - p_a$. In this section we assume ϵ is a small parameter with $\epsilon \leq \min_a p_{a+1} - p_a$, so that (A4) does not hold. Therefore, textbook Proposition 1 does not hold. We show how the $\mathcal{I}$ condition and Theorem 1 apply.

Remark. In our result below, to show condition $\mathcal{I}$ holds, we choose $\alpha_a = \beta_a = (p_{a+1} - p_a)/\epsilon = \gamma_a$. If p_a is differentiable wrt a, then as $\epsilon \to 0$, i.e., for an MDP with bi-diagonal transition matrices, this can be interpreted as choosing $\alpha_a = \beta_a = dp_a/da$.

Corollary 4. *Consider a discounted cost MDP with transition probabilities (13). Assume $p_a \uparrow a$ and $p_{a+1} - p_a = \gamma_a \epsilon$ for some positive real number γ_a increasing in a. Assume (A1) and that*

$$r(i+1, a+1) - r(i+1, a) \geq \beta_a \left[r(i, a+1) - r(i, a) \right] \tag{14}$$

for some $\beta_a \uparrow a$ with $\beta_a \geq \gamma_a$. Then $\mu^*(x) \uparrow x$.

Proof. We verify that the assumptions in Theorem 1 hold. (A1) holds by assumption. From the structure of $P^\epsilon(a)$ in (13) it is clear that (A2) holds. Considering actions a and $a+1$, it is verified that (A7) holds for all $\alpha_a \geq (p_{a+1} - p_a)/\epsilon = \gamma_a$. Next by assumption (14), (A6) holds for $\beta_a \geq \gamma_a$. So we can choose $\alpha_a = \beta_a \geq \gamma_a$, and so (A8) holds. □

Example. $A = 2, X = 6$, $p_1 = 0.3$, $p_2 = p_1 + 20\epsilon$, $\epsilon = 10^{-3}$, $\rho = 0.9$, $N = 200$, $r' = \begin{bmatrix} 1 & 3.5 & 6 & 6 & 11 & 43 \\ 0 & 2 & 3 & 6 & 12 & 63 \end{bmatrix}$. Given the transition probabilities, we choose $\alpha \geq 20$. Also for the rewards, we choose $\beta = 20$ in (14). So Corollary 4 holds. Fig. 2 shows $Q_N(x,a)$ is not supermodular, yet the optimal policy is monotone with $\mu^*(i) = 1$ for $i \in \{1, 2, 3, 4\}$ and $\mu^*(i) = 2$ for $i \in \{5, 6\}$.

4. Example 4. MDPs with concave value functions

Theorem 1 used first order dominance and monotone costs to establish $\mathcal{I}$ and therefore monotone optimal policies. In comparison, this section extends Theorem 1 to MDPs where the value function is concave. We use second order stochastic dominance and concave costs to establish $\mathcal{I}$ and therefore monotone optimal policies. The results below assume a TP3 transition matrix; see Karlin (1968) for the rich structure involving their diminishing variation property. For convenience we minimize costs instead of maximize rewards.

(C1) Costs $c(x,a)$ are $\uparrow x$ and concave in x for each a.
(C2) $P(a)$ is TP3 with $\sum_{j=1}^{X} j P_{ij}(a) \uparrow i$ and concave in i. Totally positive of order 3 means that each 3rd order minor of $P(a)$ is non-negative.
(C3) For $\beta_{x,\bar{x},a} > 0$ and $\uparrow a$, $c(\bar{x}, a+1) - c(\bar{x}, a) \geq \beta_{x,\bar{x},a} \left[c(x, a+1) - c(x,a) \right]$, $\quad \bar{x} > x$.
(C4) For $\alpha_{x,\bar{x},a} > 0$ and $\uparrow a$, $\frac{P_{\bar{x}}(a+1) + \alpha_{x,\bar{x},a} P_x(a)}{1+\alpha_{x,\bar{x},a}} >_2 \frac{P_{\bar{x}}(a) + \alpha_{x,\bar{x},a} P_x(a+1)}{1+\alpha_{x,\bar{x},a}}$, $\bar{x} > x$ where $>_2$ denotes second order stochastic dominance.[8]
(C5) Terminal cost $\tau_x \uparrow x$ and concave in x.

Remarks. (i) As shown in the proof, (C1) (concavity), (C2), (C5) imply the value function is concave and increasing. These together with (C3), (C4) and (A8) imply $\mathcal{I}$ holds and so the optimal policy is monotone.

(ii) (C2) generalizes the assumption that $\sum_j j P_{ij}$ is linear increasing in i. The classical result in Karlin (1968, pg 23) states: Suppose P is a TP3 transition matrix and $\sum_j j P_{ij}$ is linear increasing in i. If vector V is concave, then vector PV is concave. However, for bi-diagonal and tri-diagonal transition matrices, $\sum_j j P_{ij}$ is concave (or convex) and not linear in i (see examples below). This is why we introduced (C2). Since the classical result requires $\sum_j j P_{ij}$ being linear in i, it no longer applies. So we will prove a generalization that handles the case where $\sum_j j P_{ij}$ is concave in i (see Lemma 1).

Theorem 2. *(i) For a discounted cost MDP under (C1)–(C4), (A8), optimal policy $\mu^*(x) \downarrow x$.*
(ii) For a finite horizon MDP, under (C1)–(C5), (A8), optimal policy sequence $\mu_k^*(x) \downarrow x$, $k = 0, \ldots, N$.

Corollary 5. *Consider the modified assumptions: (C1): increasing replaced by decreasing; (C2) concave replaced with convex; (C3): inequality involving costs reversed; (C4): $>_2$ replaced by convex dominance[9] $>_c$; (C5): increasing replaced by decreasing. Under these assumptions and (A8), Theorem 2 holds with the modification $\mu^*(x)$ and $\mu_k^*(x) \uparrow x$.*

Proof of Theorem 2. We prove statement (ii). The proof of statement (i) is similar and omitted.

First we show by induction that $V_k(i) \uparrow i$ for $k = N, \ldots, 1$. By (C5), $V_N(i) = \tau_i \uparrow i$. Assume $V_{k+1}(i) \uparrow i$. TP3 (C2) implies TP2 which preserves monotone functions (Karlin, 1968, pg 23; Lehmann & Casella, 1998), namely, $\sum_j P_{ij}(a) V_{k+1}(j) \uparrow i$. This together with (C1) implies $Q_k(i,a) \uparrow i$. Thus $V_k(i) = \min_a Q_k(i,a) \uparrow i$.

Next we show by induction that $V_k(i)$ is concave in i. By (C5), $V_N = \tau$ is concave. Assume V_{k+1} is concave. Then (C2) implies $\sum_j P_{ij}(a) V_{k+1}(j)$ is concave in i (see Lemma 1). Since $c(i,a)$ is concave by (C1), it follows that $Q_k(i,a) = c(i,a) + \sum_j P_{ij}(a) V_{k+1}(j)$ is concave in i. Since concavity is preserved by minimization, $V_k(i) = \min_a Q_k(i,a)$ is concave. Finally, $V_k(i)$ increasing and

[8] If p, q are probability vectors, then $p >_2 q$ if $\sum_{l \leq m} \sum_{j \leq l} p_j \leq \sum_{l \leq m} \sum_{j \leq l} q_j$ for each m. Equivalently, $p >_2 q$ iff $f'p \geq f'q$ for vector f increasing and concave. Recall $'$ denotes transpose.

[9] If p, q are probability vectors, then $p >_c q$ if $\sum_{l \geq m} \sum_{j \geq l} p_j \geq \sum_{l \geq m} \sum_{j \geq l} q_j$ for each m. Equivalently, $p >_c q$ iff $f'p \geq f'q$ for f increasing and convex.

concave in i and (C4) implies (7) holds for all $\alpha_{x,\bar{x},a} \geq 1$. Then with (C3), (A8), the proof is identical to Theorem 1. □

The following lemma used in the proof of Theorem 2 extends the result in Karlin (1968, pg 23).

Lemma 1. *Suppose P satisfies (C2). If V is concave and increasing, then $P\,V$ is concave and increasing.*

Proof. First TP3 preserves monotonicity, so PV is increasing. Next, since V is concave and increasing, then for any $a > 0$ and $b \in \mathbb{R}$, $V(j) - (aj + b)$ has two or fewer sign changes in the order $-, +, -$ as j increases from 1 to X. Let $\phi_i(a, b) = \sum_j P_{ij}(aj + b)$. Since P is TP3, the diminishing variation property of TP3 implies $\sum_j P_{ij}\,V_j - \phi_i(a, b)$ also has two or fewer sign changes in the order $-, +, -$ as i increases from 1 to X. Assume two sign changes occur; then for some $i_1 < i_2$, $\sum_j P_{ij}\,V_j \geq \phi_i(a, b)$ for $i_1 \leq i \leq i_2$. Since $\phi_i(a, b)$ is integer concave in i by (C2), it lies above the line segment L_i that connects (i_1, ϕ_{i_1}) to (i_2, ϕ_{i_2}). So $\sum_j P_{ij}\,V_j \geq \phi_i(a, b) \geq L_i$, $i_1 \leq i \leq i_2$ Finally, for arbitrary $i_1 < i_2 \in \{1, \ldots, X\}$, we can choose $a = \frac{\sum_j V_j(P_{i_2,j} - P_{i_1,j})}{\sum_j j(P_{i_2,j} - P_{i_1,j})}$ and $b = \sum_j P_{i_1,j} V_j - a \sum_j j\,P_{i_1,j}$ so that $\sum_j P_{ij}\,V_j = \phi_i(a, b) = L_i$ at $i = i_1, i_2$. Clearly, $\sum_j P_{ij}\,V_j \geq L_i$ for arbitrary $i_1 \leq i \leq i_2$ and $\sum_j P_{ij}\,V_j = L_i$ for $i = i_1, i_2$ implies $\sum_j P_{ij}\,V(j)$ is concave. □

Example (i). Bi-diagonal transition matrices

Theorem 2 applies to bi-diagonal transition matrices with possibly non-supermodular costs; this is in contrast to Section 3.3 where we considered perturbed bi-diagonal matrices. Consider an MDP with bi-diagonal transition matrices $P_{i.i}(a) = 1 - p_a,\ P_{1,i+1} = p_a,\ P_{X,X}(a) = 1,\quad a \in \{1, \ldots, A\}$. Then $\sum_j jP_{ij}(a) = i + p_a$ for $i < X$ and X for $i = X$; so $\sum_j jP_{ij}(a)$ is increasing and concave in i ((C2) holds). Assume $p_a \downarrow a$. Then (C4) is equivalent to $\sum_{l \leq m} \sum_{j \leq l} P_{\bar{x},j}(a + 1) - P_{\bar{x},j}(a) \leq \alpha_{x,\bar{x},a}(\sum_{l \leq m} \sum_{j \leq l} P_{x,j}(a + 1) - P_{x,j}(a))$. Since $p_a \geq p_{a+1}$, it follows that (C4) holds for all $\alpha_{x,\bar{x},a} \geq 1$. If (C1), (C3) hold for some $\beta_{x,\bar{x},a} > 1$, then Theorem 2 holds.
Numerical example. Consider a discounted cost MDP with $A = 2$, $X = 50$, $p_1 = 0.8$, $p_2 = 0.7$, $\rho = 0.95$, $N = 200$, $c(x, 1) = \theta_1 x^2 + \theta_2 x + \theta_3$, $\quad c(x, 2) = \theta_4\big(1 - \exp(\theta_5 x + \theta_6)\big)$, $\theta = [-0.01, 1, 8.8, 25, -0.1, -0.4]$. It can be verified that the cost is not supermodular, but the conditions of Theorem 2 are satisfied. So the value function is concave and optimal policy is decreasing. Fig. 3 shows $Q_N(x, a)$ is not submodular.

Example (ii). Tri-diagonal transition matrices

Corollary 5 applies to MDPs with tri-diagonal transition matrices where $P_{i-1,i}(a) = p_a,\ P_{i+1,i} = q_a,\ P_{ii} = 1 - p_a - q_a,\ P_{11}(a) = 1,\ \ P_{X-1,X} = 1 - s_a, P_{X,X} = s_a$. If $P(a)$ is TP3 and $q_a < p_a$, $s_a > 1 + q_a - p_a$ hold, then $\sum_i jP_{ij}(a)$ is increasing and convex in i; so modified (C2) holds. Also, if $q_a \uparrow a$, $p_a \downarrow a$, $q_{a+1} - q_a \geq p_{a+1} - p_a$, $s_{a+1} - s_a > q_{a+1} - q_a + p_a - p_{a+1}$, then convex dominance (modified (C4)) holds for all $\alpha \in (0, 1]$. If the costs are chosen so modified (C1), and modified (C3) hold for $\beta_{x,\bar{x},a} \leq 1$, then Corollary 5 holds and the optimal policy is monotone.
Numerical example. Consider a discounted cost MDP with $A = 2$, $X = 35$, tri-diagonal transition matrices with $p_1 = 0.2, p_2 = 0.1, q_1 = 0.05, q_2 = 0.1, s_1 = 0.95, s_2 = 1$. Also $\rho = 0.95$, $N = 200$, $c(x, 1) = -(\theta_1 + \theta_2\,x^3)$, $\quad c(x, 2) = -(\theta_3 + \theta_4 x^3)$ where $\theta = [15, 0.3/4^3, 1, 3/4^3]$. The cost $c(x, a)$ is not submodular but Corollary 5 holds. Fig. 4 shows the non-submodular $Q_N(x, a)$.

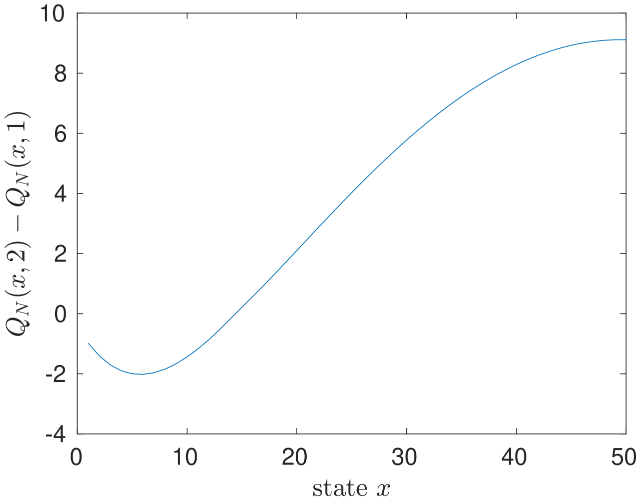


Fig. 3. Non-submodular Q function for MDP with bi-diagonal transition matrix that satisfies the assumptions of Theorem 2.

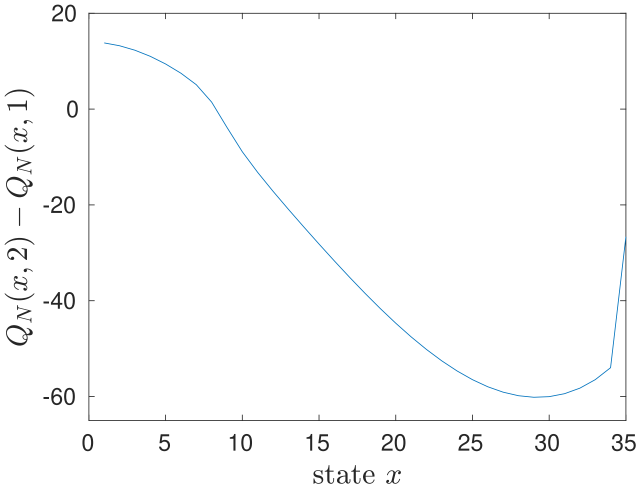


Fig. 4. Non-submodular Q function for MDP with tri-diagonal transition matrix that satisfies Corollary 5.

5. Summary and discussion

The classical structural result for MDPs uses supermodularity to establish the existence of monotone optimal policies. This paper proposes a more general condition, which we call the $\mathcal{I}$ condition, that was developed in the micro-economics literature. We presented several examples of MDPs which satisfy $\mathcal{I}$ including sigmoidal costs, and bi-diagonal/perturbed bi-diagonal transition matrices. The structural results in Section 3, namely, Theorem 1, Corollaries 1, 3, 4 and Theorem 3 used first order stochastic dominance to establish $\mathcal{I}$ for several examples. of MDPs. In comparison, Theorem 2 in Section 4 discussed examples of $\mathcal{I}$ in MDPs with concave value functions; we used TP3 assumptions and second order (convex) stochastic dominance to prove the existence of monotone optimal policies.
Discussion. Reinforcement Learning (RL) and Differential sparse Policies: Once the existence of a monotone optimal policy has been established, RL algorithms that exploit this structure can be constructed. Q-learning algorithms that exploit the $\mathcal{I}$ condition can be obtained by generalizing the supermodular Q-learning algorithms in Krishnamurthy (2016). The second approach is to develop policy search RL algorithms. In particular, when A is small

and X is large, then since $\mu^*(x) \uparrow x$, it is differentially sparse, that is $\mu^*(x+1) - \mu^*(x)$ is positive only at $A-1$ values of x, and zero for all other x. In Mattila et al. (2017), LASSO based methods are developed to exploit this sparsity and significantly accelerate search for $\mu^*(x)$; they build on the nearly-isotonic regression techniques in Tibshirani, Hoefling, and Tibshirani (2011). The idea is to add a rectified l_1-penalty $\sum_{x=1}^{X-1} |\mu^l(x) - \mu^l(x+1)|_+$ to the cost in the optimization problem (here μ^l is the estimate of the optimal policy at iteration l of the optimization algorithm). Intuitively, this modifies the cost surface to be more steep in the direction of monotone policies resulting in faster convergence of an iterative optimization algorithm.

Appendix. Optimal allocation with penalty cost

This appendix discusses a finite horizon penalty-cost MDP with perturbed bi-diagonal transition matrices (13). This has applications in optimal allocation problems with penalty costs (Derman et al., 1976; Ross, 1983) and wireless transmission control (Ngo & Krishnamurthy, 2010). We assume $\epsilon < p_{a+1} - p_a$; so as discussed in Section 3.3 supermodularity condition (A4) does not hold.

As in Example 4.2 in Derman et al. (1976) and Ross (1983, pg.8), we consider an N-horizon MDP model. There are N-stages to construct X components sequentially. If effort $c(x,a)$ is allocated then the component is constructed with successfully with probability p_a. Our transition matrices are specified by the perturbed bi-diagonal matrices (13). At the end of N stages, the penalty cost incurred is τ_i if we are i components short, where $i = \{1,\dots,X\}$, with $\tau_1 = 0$. Note that Ross (1983) considers a continuous action space $\mathcal{A} = [0,A]$, $c(x,a) = a$ where $a \in \mathcal{A}$ and bi-diagonal matrices ($\epsilon = 0$). Below we show how the $\mathcal{I}$ condition applies to non-supermodular cost structures with perturbed bi-diagonal matrices. Such cases cannot be handled by the convexity based supermodularity in Ross (1983).

We consider the discrete action space $\mathcal{A} = \{1,\dots,A\}$ corresponding to discretization of the continuous valued actions: $\bar{\mathcal{A}} = \{0, \epsilon, 2\epsilon, \dots, (A-1)\epsilon\}$. Recall ϵ are perturbation probabilities of the bi-diagonal transition matrices in (13). The costs and transition probability parameter p_a are

$$\text{Costs: } c(x,a)\epsilon, \quad p_{a+1} - p_a = \epsilon\,\gamma_a \;\; \gamma_a > 0. \tag{A.1}$$

We make the following assumptions.

(A9) $\gamma_a \geq 1$ and $\uparrow a$. (This is relaxed in remark below.)
(A10) Terminal cost τ_x convex and $\uparrow x$ with $\tau_1 = 0$. Cost $c(x,a) \downarrow x$. (More generally, $\bar{c}(x,a)$ in (A.2) $\downarrow x$.)

Main Result. We will work with the modified value function $W_k(x) = V_k^\epsilon(x) - \tau_x$. This is convenient since the terminal condition is $W_N(i) = 0$ for all i. The dynamic programming recursion (6) expressed in terms of $W_k(x)$ and minimizing the cumulative cost (rather than maximizing the cumulative reward) is $\mu_k^*(x) = \arg\min_a \bar{Q}_k(x,a)$, $W_k(x) = \min_a \bar{Q}_k(x,a)$, $k = 0,\dots,N-1$,

$$\begin{aligned}\bar{Q}_k(i,a) &= \bar{c}(i,a) + \big(1 - p_a - \epsilon(A-a)\big)\,W_{k+1}(i) \\ &\qquad + p_a W_{k+1}(i-1) \\ \bar{c}(i,a) &= \epsilon\,c(i,a) + p_a\,(\tau_{i-1} - \tau_i) + \epsilon\,(A-a)(\tau_X - \tau_i), \\ i &= 1,\dots,X-1 \\ \bar{Q}_k(X,a) &= \bar{c}(X,a) + p_a W_{k+1}(X-1) + (1-p_a)\,W_{k+1}(X), \\ \bar{c}(X,a) &= \epsilon\,c(X,a) + p_a(\tau_{X-1} - \tau_X)\end{aligned} \tag{A.2}$$

Theorem 3. *Consider the N-horizon MDP with costs and transition probabilities specified by (A.1), (13). Assume (A9) and (A10). Suppose $\min_a \gamma_a > 1$ and the costs satisfy*

$$\tau_{i+1} \geq \tau_X + \frac{\gamma_a^2\,(\tau_i - \tau_{i-1})}{\gamma_a - 1} + \frac{\Delta(i+1,a) - \gamma_a \Delta(i,a)}{\gamma_a - 1} \tag{A.3}$$

for $i = 2,\dots,X-1$ where $\Delta(i,a) = c(i,a+1) - c(i,a)$ and perturbation probabilities $\epsilon \in \big(0, \min_a(p_{a+1} - p_a)\big)$. Then optimal policy $\mu_k^*(i)$, $k = 1,\dots,N-1 \uparrow i$.

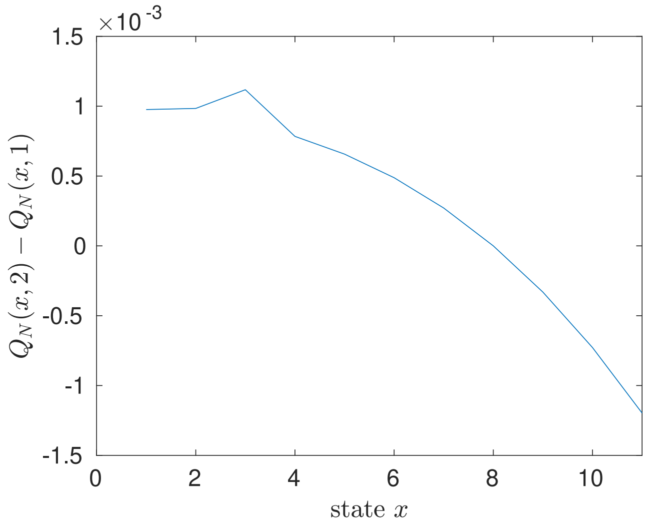


Fig. A.1. Non submodular Q function for optimal allocation.

Remarks. 1. Theorem 3 can be viewed as complementary result to the structural result in Derman et al. (1976) and Ross (1983). If we choose the same instantaneous cost as Ross (1983), namely $c(x,a) = f\,a$ for some constant f, then (A.3) becomes $\tau_{i+1} \geq \tau_X + \frac{\gamma_a^2\,(\tau_i - \tau_{i-1})}{\gamma_a - 1} - f$. But terminal costs satisfying this condition yield monotone policies that are degenerate, namely, $\mu_k^*(i) = 1$ for all i. So for $c(x,a) = f\,a$, the $\mathcal{I}$ condition does not yield a useful result. It is necessary to exploit convexity of the value function, as in Ross (1983), to obtain non-degenerate optimal policies. On the other hand, the $\mathcal{I}$ condition (A.3) allows for non-submodular costs and yields monotone policies (see examples below). For such cases, it is not clear how to extend the convexity based submodularity proof in Ross (1983) (which applies when $\epsilon = 0$) to the MDP (13) for arbitrary $\epsilon > 0$.

2. (A9) is equivalent to $p_a \uparrow a$ and convex. (A9) can be relaxed to $p_a \uparrow a$ by imposing stronger conditions on (A.3), see (A.4). The convexity (A10) of terminal costs implies $\bar{c}(i,a)$ in (A.2) is decreasing. Recall decreasing costs (A1) is used to show submodularity (and Theorem 1).

Examples. We chose the MDP parameters in (13), (A.1) as $X = 11$, $A = 2$, $\gamma_a = 1.2$, $\epsilon = 10^{-6}$, $\tau = [0, 1, 2, 4, 8, 15, 25, 40, 60, 90, 200]$. Fig. A.1 displays $Q_k(x,2) - Q_k(x,1)$ when $c(x,1) = 0$, $c(x,2) = \epsilon(f + 2.5\,x^4 I(x \leq 3) - (x+2)^3)$, $f = 10^3$. Notice $Q(x,a)$ is not submodular. But Theorem 3 holds; so $\mu_k^*(x) \uparrow x$.

Proof of Theorem 3. Using (A.2), the proof follows straightforwardly by verifying the assumptions in Theorem 1. □

Remark. Choosing $\alpha = \bar{\gamma} = \max_a \gamma_a$ in the proof, we obtain a stronger sufficient condition than (A.3):

$$\begin{aligned}\tau_{i+1} \geq \tau_X \frac{\bar{\gamma} - 1}{\gamma_a - 1} &+ \frac{\bar{\gamma}\,\gamma_a\,(\tau_i - \tau_{i-1})}{\gamma_a - 1} \\ &+ \frac{\Delta(i+1,a) - \bar{\gamma}\,\Delta(i,a)}{\gamma_a - 1} + \frac{(\gamma_a - \bar{\gamma})\tau_i}{\gamma_a - 1}\end{aligned} \tag{A.4}$$

Since $\alpha = \beta$ is a constant and not a dependent, (A9) is relaxed to $\gamma_a > 1$.

References

Amir, R. (2005). Supermodularity and complementarity in economics: An elementary survey. *Southern Economic Journal*, *71*(3), 636–660.

Derman, C., Lieberman, G. J., & Ross, S. M. (1976). Optimal system allocations with penalty cost. *Management Science*, *23*(4), 399–403.

Heyman, D. P., & Sobel, M. J. (1984). *Stochastic models in operations research, Vol. 2*. McGraw-Hill.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Karlin, S. (1968). *Total positivity, Vol. 1*. Stanford Univ..

Krishnamurthy, V. (2016). *Partially observed markov decision processes. From filtering to controlled sensing*. Cambridge University Press.

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (2nd ed.). Springer-Verlag.

Liu, Yajing, Chong, Edwin K. P., Pezeshki, Ali, & Zhang, Zhenliang (2020). Submodular optimization problems and greedy strategies: A survey. *Discrete Event Dynamic Systems*, *30*(3), 381–412.

Mattila, Robert, Rojas, Cristian R., Krishnamurthy, Vikram, & Wahlberg, Bo (2017). Computing monotone policies for Markov decision processes: a nearly-isotonic penalty approach. *IFAC-PapersOnLine*, *50*(1), 8429–8434.

Milgrom, P., & Shannon, C. (1994). Monotone comparative statics. *Econometrica*, *62*(1), 157–180.

Ngo, M. H., & Krishnamurthy, V. (2010). Monotonicity of constrained optimal transmission policies in correlated fading channels with ARQ. *IEEE Transactions on Signal Processing*, *58*(1), 438–451.

Puterman, M. (1994). *Markov decision processes*. John Wiley.

Quah, J., & Strulovici, B. (2009). Comparative statics, informativeness, and the interval dominance order. *Econometrica*, *77*(6), 1949–1992.

Ross, S. (1983). *Introduction to stochastic dynamic programming*. San Diego, California: Academic Press.

Tibshirani, Ryan J., Hoefling, Holger, & Tibshirani, Robert (2011). Nearly-isotonic regression. *Technometrics*, *53*(1), 54–61.

Topkis, D. M. (1998). *Supermodularity and complementarity*. Princeton University Press.

Vikram Krishnamurthy is a professor in the School of Electrical & Computer Engineering, Cornell University. From 2002–2016 he was a Professor and Canada Research Chair at the University of British Columbia, Canada. His research interests include statistical signal processing and stochastic control in social networks and adaptive sensing. He served as Distinguished Lecturer for the IEEE Signal Processing Society and Editor-in-Chief of the IEEE Journal on Selected Topics in Signal Processing. In 2013, he was awarded an Honorary Doctorate from KTH (Royal Institute of Technology), Sweden. He is author of the books *Partially Observed Markov Decision Processes* published by Cambridge University Press in 2016.