Finite Horizon Risk Sensitive MDP and Linear Programming

Atul Kumar, Veeraruna Kavitha and N. Hemachandra
IEOR, Indian Institute of Technology Bombay, India

***Abstract*— In the context of standard Markov decision processes (MDPs), the connection between Dynamic Program (DP) and Linear Program (LP) is well understood and is well established under sufficiently general conditions. LP based approach facilitates solving the constrained MDPs. Multiplicative or Risk sensitive MDPs, introduced to control the fluctuations/variations around the expected value, are relatively less studied objects. DP equations are considerably well understood even in the context of Risk MDPs, however the LP connection is not known. We consider a finite horizon risk MDP problem and establish the connections between the DP and LP approaches. We augment the state space with a suitable component, to obtain the optimal policies for constrained risk MDPs. We apply this results to a server selection problem in $Ber/M/K/K$ queues, with a constraint on the utilization of the fast server. We discuss some interesting structural properties of the risk optimal policies.**

I. Introduction

Markov Decision Process (MDP) is a mathematical framework used to solve the problem of sequential decision making in stochastic situations ([6], [3], [2], [4] etc). The aim of MDP is to find an optimal policy for the decision makers. A policy is a sequence of decisions one for each time slot, possibly depending upon the (current state or history of all states) state of the system. MDP considers a running cost at every time step, depending upon the state and the action taken at that time step, and obtains an optimal policy that optimizes the expected value of the sum (integral in case of continuous time problems) of the running costs over all the time slots under consideration. In case of finite horizon problems, it also considers a terminal cost.

There can be three varieties of MDP problems based on the time horizon for which the problem spans. It is finite time horizon problem if the sum cost is considered for a finite duration. In case of infinite time horizon problems, variants like discounted cost, average cost and total cost are considered. The focus of this paper is on finite horizon problems.

In many scenarios, the agents are interested not just in the average cost. But some agents would like to reduce the risk on most of the sample paths. Worst case analysis deals with an extreme case in this direction. While risk sensitive framework offers varying ranges of importance to sample paths and average value as controlled by a parameter. Depending upon the parameter, called risk parameter, it provides importance to higher moments of the sum cost. In all, while average cost/linear MDPs are concerned about first moment of the sum cost, risk sensitive MDPs incorporate higher moments of the cost, to control the variability/fluctuations about the expected value. The linear MPDs are also viewed as risk neutral MDPs.

The linear MDP is a well studied topic and many solutions approaches are known. Dynamic programming (DP), Linear programming (LP), Value iteration are some of them ([6], [3], [2], [4] etc). DP obtains the value function, the optimal cost to go till termination from any time and any state, using backward induction. Alternatively value functions can be obtained using a solution of an appropriate Linear Program (LP). The dual LP directly provides the optimal policy (e.g., [6], [2]). Relatively risk sensitive MDPs are studied to a limited extent. Nevertheless, dynamic programming approach can be applied even in the context of finite horizon risk MDP ([5]). To the best of our knowledge, the connection between risk sensitive MDPs and an appropriate linear program is not yet established and this is the main focus of the paper. This connection does not solve the dimensionality problem. Nevertheless, the advent of fast LP solvers makes this a very attractive alternative. Further and more importantly one can incorporate constraints in the MDP framework. Our LP connection thus provides computational methods to solve such constrained risk sensitive MDPs. Availability of resources can be captured as suitable constraints and hence solutions to constrained MDPs are important.

The work on risk sensitive control is vast and varied and we give a sample of some of the strands. The pioneering work is by Howard and Mathieson [7]. The backward recursion dynamic programming equations in finite horizon setting are of multiplicative type and algorithms to compute optimal polices in this model are known. In general, the optimal policies in infinite horizon setting tend to be non stationary, [8], etc. Some papers identify suitable sufficient conditions that ensure that the optimal polices are stationary and also develop algorithms to compute the same ([10]) or approximate optimal policies. Other papers also explore the relations between robust MDPs and risk sensitive MDPs [9]. We develop LP based algorithm that computes optimal risk sensitive polices for constrained risk MDPs.

Notation: The bold letters represent the vectors, e.g.,

$\mathbf{y} = \{y(t,x,a)\}_{t,x,a}$ represents a feasible vector of dual LP (9), given below. While $\mathbf{x}_n^t$ represents the vector $\mathbf{x}_n^t = [x_n, \cdots, x_t]$. The random variables are represented by capital letters, while their realization by the corresponding small letters. When required to specify the time index, a subscript of the time index is used. When not required it is avoided. For example, x represents a realization of random variable X_t for any t. If it required to represent a realization of the pair of random variables X_t, X_n, then we use x_t, x_n. The realizations for random variables of subsequent time slots, like X_t, X_{t+1}, are represented by (x, x').

II. Risk sensitive MDP framework

Risk sensitive MDP, as in the case of linear MDP, consists of a set $\mathcal{X}$ of all possible states, a set $\mathcal{A}$ of all possible actions and an immediate reward function

$$r_t : \mathcal{X} \times \mathcal{A} \to \mathcal{R} \text{ for each time slot } t.$$

The terminal cost r_T depends only upon the state $x \in \mathcal{X}$. The state, action spaces $\mathcal{X}$, $\mathcal{A}$ do not depend the time slot t, that is we consider the same set for all the time indices. It is further characterized by a transition function $p : \mathcal{X} \times \mathcal{A} \to \mathcal{X}$, which defines the action dependent state transitions. Here $p(x'|x,a)$ gives the probability of the state transition from x to x', when action a is chosen.

We consider a finite horizon problem and let $\{X_t\}_{t \le T}$, $\{A_t\}_{t \le T-1}$ respectively represent the trajectories of the state and the action processes. The terminal cost, cost in final time slot T, depends only upon the (final) state. A policy $\Pi^t = (\pi_t, \pi_{t+1} \cdots \pi_{T-1})$ is a sequence of state dependent and possibly randomized actions, given for time slots between t and $T-1$. Given a policy Π^t, the state, action pair evolve randomly over the time slots $t < n < T$, with transitions given by the following laws:

$$\begin{aligned} q_n^{\Pi^t}(x', a'|x, a) &= P(X_n = x', A_n = a'|X_{n-1} = x, A_{n-1} = a) \\ &= \pi_n(x', a') p(x'|x, a) \text{ where} \\ p(x'|x, a) &= P(X_n = x'|X_{n-1} = x, \, A_{n-1} = a) \text{ and} \\ \pi_n(x', a') &= P(A_n = a'|X_n = x'). \end{aligned} \quad (1)$$

The above evolution further depends upon the initial condition, i.e., the initialization of the stating point X_t. Let E^{x,Π^t} represent the expectation operator with initial condition $X_t = x$ and when the policy Π^t is used. Let E^{α,Π^t} represent the same expectation operator when the initial condition is distributed according to α, written as $X_t \sim \alpha$. Here $\alpha(x) = P(X_t = x)$. We are interested in optimizing the following risk sensitive objective:

$$\begin{aligned} \tilde{J}_t(\alpha, \Pi^t) &= \gamma^{-1} \log \left(J_t(\alpha, \Pi^t) \right) \text{ where} \\ J_t(\alpha, \Pi^t) &= E^{\alpha,\Pi^t} \left[e^{\gamma \sum_{n=t}^{T-1} r_n(X_n, A_n) + r_T(X_T)} \right]. \end{aligned} \quad (2)$$

The above equation represents the cost to go from time slot t to T under the policy Π^t, with $X_t \sim \alpha$. The value function, a function of (x,t), is defined as the optimal value of the above risk sensitive objective given the initial condition $X_t = x$:

$$V_t(x) := \min_{\Pi^t \in \mathcal{D}^t} \tilde{J}_t(x, \Pi^t) \text{ for any } x \in \mathcal{X}, \quad (3)$$

where $\mathcal{D}^t$ represents the space of policies Π^t.

Dynamic Programming

We are interested in the optimal policy $\Pi^{0*} = \Pi^*$ (we discard 0 in superscript when it starts from 0) that optimizes the risk cost $J_0(x, \Pi^0)$, or equivalently a policy that achieves the value function, i.e., a Π^* such that $\quad V_0(x) = \tilde{J}_0(x, \Pi^*)$ for all $x \in \mathcal{X}$.

Dynamic programming (DP) is a well known technique, that provides a solution to such control problems, and DP equations are given by backward induction as below for any $x \in \mathcal{X}$ (see [5]):

$$V_T(x) = r_T(x), \text{ and for any } 0 \le t \le T-1,$$

$$V_t(x) = \min_{a \in \mathcal{A}} \left\{ r_t(x,a) + \frac{1}{\gamma} \log \left[\sum_{x' \in \mathcal{X}} p(x'|x,a) e^{\gamma V_{t+1}(x')} \right] \right\}.$$

We consider the following translation of the value function, to simplify the above set of equations:

$$u_t(x) = e^{\gamma V_t(x)} \text{ for all } 0 \le t \le T, \text{ and } x \in \mathcal{X}.$$

Note by monotonicity and continuity u_t for any t is minimum value of J_t given in (2):

$$u_t(x) = \min_{\Pi^t} J_t(x, \Pi^t).$$

The DP equations can now be rewritten as:

$$\begin{aligned} u_T(x) &= e^{\gamma r_T(x)} \text{ for any } x \in \mathcal{X} \text{ and} & (5) \\ u_t(x) &= \min_a \left\{ e^{\gamma r_t(x,a)} \sum_{x' \in \mathcal{X}} p(x'|x,a) u_{t+1}(x') \right\} \\ &\quad \text{for any } 0 \le t \le T-1, \text{ and } x \in \mathcal{X}. & (6) \end{aligned}$$

For ease of notations, we absorb γ into the cost functions $\{r_t\}$. One needs to solve the above set of equations to obtain the value function:

$$\underline{\mathbf{u}}^* = \{u_t^*(x); t < T, x \in \mathcal{X}\},$$

and the optimizers in the minimization step will provide us the optimal policy ([6], [5] etc).

III. Linear Programming Formulation

The dynamic programming based approach suffers from the curse of dimension. As we increase the number of states and/or time epochs, the complexity of the problem increases significantly. This results in limited applicability of dynamic programming. In the context of linear MDPs, it is a well known fact that a DP problem can be reformulated as a Linear Program (LP), under considerable generality (see for e.g., [6] in the context of infinite horizon problems). However this conversion may

$$
\begin{aligned}
C_{t,\pi_t,a} &= \begin{bmatrix} \pi_t(1,a)e^{r_t(1,a)} & 0 & \cdot\;\cdot\;\cdot & 0 \\ 0 & \pi_t(2,a)e^{r_t(2,a)} & \cdot\;\cdot\;\cdot & 0 \\ \cdot & \cdot & \cdot\;\cdot\;\cdot & 0 \\ \cdot & \cdot & \cdot\;\cdot\;\cdot & 0 \\ \cdot & \cdot & \cdot\;\cdot\;\cdot & 0 \\ 0 & \cdot & \cdot\;\cdot\;\cdot & \pi_t(N,a)e^{r_t(N,a)} \end{bmatrix}, \\
P_a &= \begin{bmatrix} p(1|1,a) & p(2|1,a) & \cdot\;\cdot\;\cdot & p(N|1,a) \\ p(1|2,a) & p(2|2,a) & \cdot\;\cdot\;\cdot & p(N|2,a) \\ \cdot & \cdot & \cdot\;\cdot\;\cdot\;\cdot & \cdot \\ \cdot & \cdot & \cdot\;\cdot\;\cdot\;\cdot & \cdot \\ \cdot & \cdot & \cdot\;\cdot\;\cdot\;\cdot & \cdot \\ p(1|N,a) & p(2|N,a) & \cdot\;\cdot\;\cdot & p(N|N,a) \end{bmatrix}
\end{aligned} \tag{4}
$$

not solve the problem of dimension. But recent improvements in LP solvers makes it an attractive alternative. Further and more importantly the LP based approach extends easily and *provides solutions for constrained MDPs.*

In the coming sections, as in the case of linear MDP (see for e.g., [6]), we will obtain two relevant LPs (a primal and a dual LP). The solution of the primal LP will be the translated value function vector, $\underline{\mathbf{u}}^*$, which is the function value on the left hand side (LHS) of the DP equation (6), at the optimizer(s). On the other hand the solution of the dual LP will directly provide the optimal policy $\Pi^* \in \mathcal{D}$ of the control problem (3).

We begin with introducing some more notations. Let N be the number of elements of the state space and without loss of generality let $\mathcal{X} = \{1, \cdots, N\}$. Let $\mathbf{u}_t = [u_t(1) \cdots, u_t(N)]$ represent an N dimensional vector indexed by time t, indicative of the possible value function for different states at time t. And let the combined vector that includes the value function for all combinations of time slots and states be rewritten as below:

$$\underline{\mathbf{u}} = \{u_t(x); t < T, x \in \mathcal{X}\} = [\mathbf{u}_0, \mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_{T-1}].$$

Define the operator that operates on the combined vector $\underline{\mathbf{u}}$ by:

$$
\begin{aligned}
\mathcal{L}\underline{\mathbf{u}} &= [L_0\underline{\mathbf{u}}, L_1\underline{\mathbf{u}}, \cdots, L_{T-1}\underline{\mathbf{u}}] \text{ where} \\
L_t\underline{\mathbf{u}} &:= \inf_{\pi_t} \sum_a C_{t,\pi_t,a} P_a \mathbf{u}_{t+1}
\end{aligned} \tag{7}
$$

with the matrices $C_{t,\pi_t,a}, P_a$ are defined using (4), placed at the top of the page and $\mathbf{u}_T = \{u_T(x); x \in \mathcal{X}\}$ is given by equation (5). The above operator is constructed using the right hand side (RHS) of the DP equation (6). We now have the following theorems (whose proofs are in Appendix):

Theorem 1: Any vector $\underline{\mathbf{u}}$ with $\mathcal{L}\underline{\mathbf{u}} \geq \underline{\mathbf{u}}$ (the component wise inequality), satisfies: $\underline{\mathbf{u}}^* \geq \underline{\mathbf{u}}$.

Theorem 2: Any vector $\underline{\mathbf{u}}$ with $\mathcal{L}\underline{\mathbf{u}} \leq \underline{\mathbf{u}}$, satisfies: $\underline{\mathbf{u}}^* \leq \underline{\mathbf{u}}$.

Any vector $\underline{\mathbf{u}}$ that satisfies the constraint, $\mathcal{L}\underline{\mathbf{u}} \geq \underline{\mathbf{u}}$, i.e., when

$$u_t \leq \sum_a C_{t,\pi_t,a} P_a \mathbf{u}_{t+1} \text{ for all } t \text{ and } \pi_t \in \Pi, \tag{8}$$

by Theorem 1, is lower than the value function of risk MDP, $\underline{\mathbf{u}}^*$. It is trivial to check that $\underline{\mathbf{u}}^*$ also satisfies (8). Thus it is the greatest lower bound among all vectors that satisfy (8). Hence we have the following LP for primal. Following similar procedure as in ([6]), we chose a nonnegative vector $\alpha(x), x \in \mathcal{X}$ that satisfies $\sum_{x\in\mathcal{X}} \alpha(x) = 1$. Using this one can obtain an equivalent LP, whose solution equals the value function vector $\underline{\mathbf{u}}^*$.

Primal Linear Program

$$\max_{\left\{\{u_t(x)\}_{x\in\mathcal{X}, t\leq T-1}\right\}} \sum_{x\in\mathcal{X}} \alpha(x) u_0(x)$$

$$
\begin{aligned}
\text{subject to:} \qquad u_{T-1}(x) &\leq b_{x,a} \text{ for all } x, a, \\
u_t(x) - e^{r_t(x,a)} \sum_{x'\in\mathcal{X}} p(x'|x,a) u_{t+1}(x') &\leq 0 \\
&\text{for all } a, x \text{ and } t \leq T-2 \\
\text{with } b_{x,a} := e^{r_{T-1}(x,a)} \sum_{x'\in\mathcal{X}} p(x'|x,a) e^{r_T(x')}.
\end{aligned}
$$

The vector α can be interpreted as the distribution of initial state, X_0. The dual of the above LP, is given by:

Dual Linear Program

$$\min_{\mathbf{y}=\{y(t,x,a); t\leq T-1, x\in\mathcal{X}, a\in\mathcal{A}\}} \sum_a \sum_{x\in\mathcal{X}} b_{x,a}\, y(T-1,x,a)$$

subject to:

$$\sum_a y(0,x',a) = \alpha(x') \text{ for all } x' \in \mathcal{X}, \tag{9}$$

$$
\begin{aligned}
\sum_a y(t,x',a) &= \sum_a \sum_x e^{r_{t-1}(x,a)} p(x'|x,a) y(t-1,x,a) \\
&\text{for all } 1 \leq t \leq T-1 \text{ and } x' \in \mathcal{X}.
\end{aligned} \tag{10}
$$

Below we give a series of results connecting the dual LP (9) and the translated risk MDP (6). Some of the proofs and results of this section have similar structure as that given in [6]. However there are significant changes due to risk sensitive nature of the cost.

A. *Feasible region $\mathcal{F}$ and the set of risk policies $\mathcal{D}$:*

We say that a vector $\mathbf{y}$ is feasible if it satisfies the dual constraints (9), (10) and let $\mathcal{F}$ represent this feasible region. We first show a one to one correspondence between the two spaces, $\mathcal{F}$ and $\mathcal{D}$.

Theorem 3: (i) For any policy $\Pi \in \mathcal{D}$ of risk MDP, there exists a vector $\mathbf{y_\Pi}$ which satisfies all the constraints of dual LP (9), i.e., $\mathbf{y_\Pi} \in \mathcal{F}$. The feasible vector is given by the equation (see (1)):

$$\begin{aligned} y_\Pi(0,x_0,a_0) &= \alpha(x_0)\pi_0(x_0,a_0) \text{ for all } x_0 \in \mathcal{X}, a_0 \in \mathcal{A}, \\ y_\Pi(t,x_t,a_t) & \\ &= \sum_{\mathbf{a}_0^{t-1},\mathbf{s}_0^{t-1}} \alpha(x_0) e^{\sum_{n=0}^{t-1} r_n(x_n,a_n)}\, \Pi_{n=0}^t\, q_n^\Pi(x_n,a_n|x_{n-1},a_{n-1}) \\ &\qquad \text{for all } x_t \in \mathcal{X}, a_t \in \mathcal{A}, \text{ and } 1 \le t < T. \end{aligned} \tag{11}$$

In the above we define

$$q_0^\pi(x_0,a_0|x_{-1},a_{-1}) := \pi_0(x_0,a_0).$$

(ii) Given a vector $\mathbf{y} \in \mathcal{F}$, define a policy $\Pi_\mathbf{y}$ using the following rule:

$$\pi_{\mathbf{y},t}(x,a) := \frac{y(t,x,a)}{\sum_{a'} y(t,x,a')} \text{ for all } x \in \mathcal{X}, \text{ and } a \in \mathcal{A}. \tag{12}$$

The vector $\mathbf{y}_{\Pi_\mathbf{y}}$ defined by equation (11) of point (i) is again in feasible region and equals $\mathbf{y}$.

Proof: The proof is provided in Appendix. ■

B. *Expectation at optimal policy:*

To obtain the connection between risk MDP problem and the dual LP, one needs to study the connection between the risk sensitive cost for a given policy Π and the dual objective function at the feasible point $\mathbf{y}_\Pi$, defined using Π. We also require similar connecting between the feasible point $\mathbf{y}$ and the corresponding policy $\Pi_\mathbf{y}$. Further, we would like to solve constrained risk sensitive MDP problems (in section IV). The constraints usually bound the expected value of some function of the state, action random trajectories. In all, we require the expression for the expected value of a given function, in terms of the dual variable $\mathbf{y} \in \mathcal{F}$. As a first step, we have the following, with proof in Appendix:

Lemma 1: Let $X_0 \sim \alpha$. For any feasible point $\mathbf{y}$ of dual LP, integrable function f and $t < T$

$$\begin{aligned} \sum_{x,a} y(t,x,a) f(x,a) &= E^{\alpha,\Pi_\mathbf{y}}\left[\Psi_t^{-1} f(X_t,A_t)\right] \text{ with} \\ \Psi_t &:= e^{-\sum_{n=0}^{t-1} r_n(X_n,A_n)}. \end{aligned} \tag{13}$$

Further, for any integrable function f of the last two states X_{T-1}, X_T and the final action A_{T-1}, we have:

$$\begin{aligned} \sum_{x,a,x'} & y(T-1,x,a) p(x'|x,a) f(x,a,x') \\ &= E^{\alpha,\Pi_\mathbf{y}}\left[\Psi_{T-1}^{-1} f(X_{T-1},A_{T-1},X_T)\right]. \end{aligned} \tag{14}$$

■

We have the same result when we replace $\mathbf{y}$, $\Pi_\mathbf{y}$ with $\mathbf{y}_\Pi$, Π respectively, following exactly similar steps as in the previous theorem.

Lemma 2: For any policy $\Pi \in \mathcal{D}$ of risk MDP and for any integrable function f, we have:

$$\begin{aligned} \sum_{x,a,x'} & y_\Pi(T-1,x,a) p(x'|x,a) f(x,a,x') \\ &= E^{\alpha,\Pi}\left[\Psi_{T-1}^{-1} f(X_{T-1},A_{T-1},X_T)\right]. \end{aligned} \tag{15}$$

■

If we use the above theorem with the function,

$$f(x,a,x') = e^{r_{T-1}(x,a)} e^{r_T(x')},$$

we obtain the following for any $(\mathbf{y}, \Pi_\mathbf{y})$:

$$\begin{aligned} \sum_{x,a} y(T-1,x,a) & \sum_{x'} p(x'|x,a) e^{r_{T-1}(x,a)} e^{r_T(x')} \\ &= E^{\alpha,\Pi_\mathbf{y}}\left[e^{\sum_{n=0}^{T-1} r_n(X_n,A_n)} e^{r_T(X_T)}\right]. \end{aligned}$$

Note that the LHS is the dual objective (9) at point $\mathbf{y}$ and RHS is the risk cost at policy $\Pi_\mathbf{y}$. This is the basic element in proving the equivalence of optimal policies and optimal dual solutions, given below.

C. *Optimal policies and the dual solutions*

The following theorem shows the relation between the two optimizers (proof in Appendix).

Theorem 4: (a) If $\mathbf{y}^*$ is an optimal solution of the dual LP, then $\Pi_{\mathbf{y}^*}$ defined by (12) is an optimal policy for risk MDP.

(b) If Π^* is an optimal policy for risk MDP, then $\mathbf{y}_{\Pi^*}$ is an optimal solution of the dual LP. ■

IV. Constrained risk MDP

We now consider a constrained MDP problem, with an additional constraint as given below:

$$\min_\Pi J_0(\alpha,\Pi) \tag{16}$$

$$\text{Subject to: } \sum_t E^{\alpha,\Pi}\left[f_t(X_t,A_t)\right] \le B,$$

for some set of integrable functions $\{f_t\}$, initial distribution α and bound B. The equation (13) of Lemma 1 could have been useful in obtaining the expectation defining the constraint, but for the extra factor Ψ_t^{-1}, as seen from the RHS of the equation (13). We propose to add Ψ_t as additional state component to the original Markov chain $\{X_t\}$ to tackle this problem. We consider a two component Markov chain $\{(X_t, \Psi_t)\}$ and the corresponding probability transition matrix depends explicitly upon time index as below:

$$\tilde{p}_{t+1}(x', \psi'_{t+1} | x, \psi_t, a) = 1_{\{\psi'_{t+1} = \psi_t e^{-r_t(x,a)}\}} p(x'|x, a).$$

With the introduction of the new state component, for any dual LP feasible point $\mathbf{y}$ we have:

$$\sum_{x,\psi_t,a} y(t, x, \psi_t, a)\, \psi_t\, f(x, a) = E^{\alpha,\Pi_{\mathbf{y}}} \left[f(X_t, A_t) \right]. \quad (17)$$

Thus one can obtain optimal policy of constrained risk MDP (16) by considering an additional state component and by adding an extra constraint to the dual LP (9) as below:

$$\min \sum_a \sum_x e^{r_{T-1}(x,a)} \left[\sum_{x' \in \mathcal{X}} p(x'|x, a) e^{r_T(x')} \right] y(T-1, x, a) \quad (18)$$

subject to:

$$\begin{aligned}
y(t, x, a) &= \sum_{\psi_t} y(t, x, \psi_t, a) \text{ for all } t \\
\sum_a y(0, x, \psi_0, a) &= \alpha(x) 1_{\{\psi_0 = 1\}} \text{ for all } x, \psi_0 \\
\sum_a y(t, x', \psi'_t, a) &= \\
&\sum_{a,x,\psi_{t-1}} e^{r_{t-1}(x,a)} \tilde{p}(x', \psi'_t | x, \psi_{t-1}, a) y(t-1, x, \psi_{t-1}, a) \\
&\qquad \text{for all } 1 \leq t \leq T-1 \text{ and } x', \psi'_t \text{ and} \\
\sum_t \sum_{x,\psi_t,a} & y(t, x, \psi_t, a) \psi_t f_t(x, a) \leq B.
\end{aligned}$$

We would like to mention here that $\psi_0 = 1$ is always initialized to one, Ψ_1 can take at maximum $|\mathcal{X}| \times |\mathcal{A}|$ values while Ψ_t for any t can take at maximum $|\mathcal{X}|^t \times |\mathcal{A}|^t$ possible values. There will also be considerable deletions if the concerned mapping

$$(\mathbf{a}_0^t, \mathbf{x}_0^t) \mapsto e^{-\sum_{n=0}^t r_n(x_n, a_n)}$$

is not one-one. One needs to consider this time dependent state space while solving the dual LP given above and we omit the discussion of these obvious details.

V. Applications

In [11], we applied LP based approach to solve a constrained risk sensitive cost that arises naturally in the context of Delay Sensitive Networks (DTNs). The probability of message delivery failure with exponentially distributed contacts turns out to have a risk sensitive form. The direct solution to the power constrained problem works significantly superior in comparison with the solution obtained for a model with soft constraints.

In this paper we consider another example, which investigates the effect of risk sensitive cost on optimal policy. As the risk factor increases, the optimal policies are no more monotone.

A. *Queueing with losses*

We consider a queueing system with two possible service options. The fast service facility offers service at rate μ_1 and is expensive, while the service rate of the slower one is μ_0 with $\mu_0 < \mu_1$. The system can support at maximum N jobs in parallel and any job arrival that finds all the N servers busy, leaves the system without service. Aim is to utilize the fast service facility in an optimal manner which minimizes the total number of jobs lost in a given time horizon, while maintaining the utilization of the fast service facility within a given limit.

We consider a queueing system with Bernoulli arrivals. In every time slot (of unit duration), a customer arrives with probability δ and there is no arrival with probability $1-\delta$. The job demands are exponentially distributed with parameter μ_1 (parameter μ_0) when served by the fast (slow) server. Let X_t represent the number of customers in the system and let A_t be the indicator of the service type used in time slot t. The flag $A_t = 1$ implies faster service facility is used across all the servers, while $A_t = 0$ implies the use of slower service facility. A customer leaves the system after service completion, in one time slot with probability $1 - \Theta_{A_t}$ where

$$\Theta_a := e^{-\mu_a}, \ \mu_a := \mu_1 1_{\{a=1\}} + \mu_0 1_{\{a=0\}}.$$

Thus the transition probability matrix of this controlled Markov chain is given by

$$\begin{aligned}
& p(x'|x, a) \\
&= \begin{cases}
\Theta_a^N + \delta N \Theta_a^{N-1}(1 - \Theta_a) & \text{if } \ x' = x = N \\
\delta \Theta_a^x 1_{\{x<N\}} & \text{if } \ x' = x + 1 \\
(1-\delta)(1-\Theta_a)^x & \text{if } \ x' = 0 \\
\delta \binom{x}{x'-1} \Theta_a^{x'-1} (1-\Theta_a)^{x-x'+1} & \\
+(1-\delta) \binom{x}{x'} \Theta_a^{x'} (1-\Theta_a)^{x-x'} & \text{if } 0 < x' \leq x \\
0 & \text{else.}
\end{cases}
\end{aligned}$$

With G_t representing the flag indicating the arrival of a customer in time slot t, the total number of customers lost in a total of T time slots is given by:

$$\sum_{t=0}^T 1_{\{X_t = N\}} G_t$$

and we are interested in minimizing the corresponding risk sensitive cost for a given risk parameter γ

$$J(x, \Pi) = E^{x,\Pi} \left[e^{\gamma \sum_{t=0}^T 1_{\{X_t=N\}} G_t} \right].$$

Theorem 5: The required risk sensitive cost has a simpler form as below:

$$\begin{aligned}
J(x, \Pi) &= E^{x,\Pi} \left[e^{\beta \sum_{t=0}^T 1_{\{X_t=N\}}} \right] \text{ with} \\
\beta &= \ln\left(\delta e^\gamma + (1-\delta)\right). \quad (19)
\end{aligned}$$

Proof : Note that the arrivals in the time slots with $X_t = N$ are lost, because all the servers are busy. These arrivals does not change the number in the system in the next time slot X_{t+1} and hence are independent of the system evolution. By conditioning on the Markov chain trajectory $\{X_t\}_{t=0}^T$ and because of the independence just discussed above, we have:

$$J(x, \Pi) = E^{x,\Pi}\left[g_\delta^{\sum_{t=0}^T 1_{\{X_t=N\}}}\right] \text{ with } g_\delta := E\left[e^{\gamma G_t}\right].$$

Let $\beta := \ln(g_\delta)$, so that $e^\beta = g_\delta$. ■

We would like to optimize the above risk sensitive cost under the following constraint for a given utilization bound B:

$$E^{x,\Pi}\left[\sum_{t=0}^T 1_{\{A_t=1\}} X_t\right] \leq B.$$

Basically when fast facility is chosen as option in any time slot, X_t number of servers are using fast facility and hence the above constraint.

Numerical analysis

We obtain the optimal policy for the above queueing based control problem,

$$\min_\Pi E^{x,\Pi}\left[e^{\beta \sum_{t=0}^T 1_{\{X_t=N\}}}\right] \text{ such that}$$
$$E^{x,\Pi}\left[\sum_{t=0}^T 1_{\{A_t=1\}} X_t\right] \leq B,$$

by solving the corresponding LP (18), where β depends upon the risk parameter γ as given by Theorem 5. We did most of the coding in Matlab except for the LP part. We used AMPL to model the LP and Gurobi solver to solve the LP. The solution $\mathbf{y}^*$ of the LP provides the optimal policy $\Pi_{\mathbf{y}^*}$ as given by equation (12).

In Figure 1, we consider a system with 3 servers. We plot the optimal policy for two values of γ. The optimal policy with $x = 0$ (i.e., with no customers in the system) has no impact as the server(s) are not utilized. With both values of γ, the optimal policy with one customer in the system, i.e., with $x = 1$, is to switch off the fast serve facility at all time slots. But there is a big difference for the remaining two states $x = 2, 3$ and these policies are plotted in the figure. We plot the probability of fast service, as dictated by optimal policy, with states $x = 2$ and $x = 3$ across the time slots. When $\gamma = 0.001$, the optimal policies are threshold type. With $x = 2$ it switches off the fast facility at 5th time slot, while with $x = 3$ it switches off at the 2nd time slot. The risk cost is close to the linear cost with small values of risk factor γ, the optimal policies are well understood to be threshold type for linear control and this explains the figure for the case with $\gamma = 0.001$.

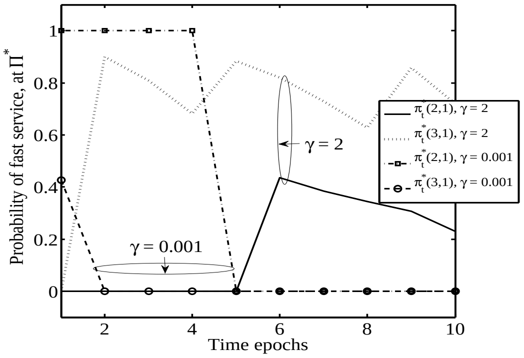

Fig. 1. Optimal policy with $T = 10$, $N = 3$, $\mu_1 = 0.3$, $\mu_0 = 0.1$ and $B = 2.5$. For $x = 1$ optimal policy always uses slow server.

With $\gamma = 2$, the risk optimal policies are no more threshold. In fact they are not even monotone as seen from the Figure 1. Further, the probability of fast service is higher at smaller states ($x = 2$) with small γ, while the opposite is true when $\gamma = 2$. With larger importance to risk cost, the policy is more cautious at the points of high risk, i.e., when $x = 3$. We estimate the average number of customers lost, at optimal policy, using Lemma 1 as below:

$$\begin{aligned} E^{x,\Pi^*}[N_{lost}] &= E^{x,\Pi^*}\left[\sum_t 1_{\{X_t=N\}} G_t\right] \\ &= \sum_t \sum_{x,a,\psi_t} y^*(t,x,a)\psi_t 1_{\{x=N\}}\delta. \end{aligned}$$

The average number lost equals 1.37 and 1.45 respectively at $\gamma = 0.01$ and 2. This is obvious because with high risk factor, the importance is drifted away from the average number lost.

We considered many more numerical examples and found similar characterization of the optimal policies.

CONCLUSIONS

We consider a finite horizon risk MDP problem and establish the connections between the DP and LP approaches. We show that the solution of the unconstrained risk MDP problem (3) can be obtained via the solution of any one of the two LPs, a primal and a dual. The primal solution provides the value function while the dual solution directly provides the risk optimal policy. It is not straightforward to extend the solution to the constrained risk MDP problem. We augment the state space with a suitable component, that at any time slot captures the effect of the risk cost until that slot. We propose a third LP using the augmented state space transitions, which provides the solution to the constrained risk MDP problem.

We apply the results so obtained, to study the server selection problem in the context of Bernoulli queues with losses. Our aim is to minimize the number of customers lost, i.e., returned without service. We consider minimizing the risk version of the cost and optimize it under a fast server utilization constraint. The optimal policy is a threshold policy when the risk factors are close to zero. It is well known that the risk MDP is close to the linear MDP with small risk factors and hence a threshold policy is anticipated. However, with large risk factors the risk optimal policy is no longer threshold type. The policies are not even monotone. Further we notice that the probability of choosing fast server is higher at larger states. With higher preference to risk cost, the policy emphasizes utilization of the fast server at high risk states, the larger states.

Thus the proposed LPs are useful in obtaining the solutions of the constrained/ unconstrained finite horizon risk MDPs.

Acknowledgements

The work towards this problem originated with Prof. Eitan Altman's remark about finding the connections between LPs and risk sensitive MDP problems.

Appendix: Proofs

The proofs of this section follow similar structure as given in [6]. However there are significant changes due to risk sensitive nature of the cost.

Proof of Theorem 1: Consider any vector $\underline{\mathbf{u}}$ satisfying $\underline{\mathbf{u}} \le \mathcal{L}\underline{\mathbf{u}}$. Consider any policy $\Pi' = [\pi_0', \pi_1', \cdots, \pi_{T-1}']$. By definition of the operator $\mathcal{L}$

$$\mathbf{u}_0 \quad \le \quad \inf_{\pi_0} \left\{ \sum_{a_0} C_{0,a_0,\pi_0} P_{a_0} \mathbf{u_1} \right\} \tag{20}$$

$$\le \sum_{a_0} C_{0,a_0,\pi_0'} P_{a_0} \mathbf{u_1}$$

$$\le \sum_{a_0} C_{0,a_0,\pi_0'} P_{a_0} \sum_{a_1} C_{1,a_1,\pi_1'} P_{a_1} \mathbf{u_2}$$

$$\vdots$$

$$\le \sum_{\mathbf{a}_0^{T-1}} C_{0,a_0,\pi_0'} P_{a_0} C_{1,a_1,\pi_1'} P_{a_1} \cdots C_{T-1,a_{T-1},\pi_{T-1}'} P_{a_{T-1}} \mathbf{u_T}$$

$$= \mathbf{J}_0(\Pi') \text{ with } \mathbf{J}_0(\Pi') := \begin{bmatrix} J_0(1,\Pi') \\ J_0(2,\Pi') \\ \vdots \\ J_0(N,\Pi') \end{bmatrix}. \tag{21}$$

This is true for any policy Π'. Thus

$$\mathbf{u}_0 \le \inf_{\Pi'} \mathbf{J}_0(\Pi') = \mathbf{u}_0^*.$$

Following exactly similar logic one can show for all $1 \le t < T$ that

$$\mathbf{u}_t \le \mathbf{u}_t^*. \qquad \blacksquare$$

Proof of Theorem 2:
Let us consider any vector $\underline{\mathbf{u}}$ which satisfies $\underline{\mathbf{u}} \ge \mathcal{L}\underline{\mathbf{u}}$. By definition of $\mathcal{L}$:

$$\mathbf{u}_0 \ge \inf_{\pi_0} \left\{ \sum_{a_0} C_{0,a_0,\pi_0} P_{a_0} \mathbf{u_1} \right\} \tag{22}$$

Consider any $\epsilon_0 \ge 0$, by definition of infimum there exists a policy π_0' such that:

$$\mathbf{u}_0 \quad \ge \sum_{a_0} C_{0,a_0,\pi_0'} P_{a_0} \mathbf{u_1} - \epsilon_0$$

By boundedness of the matrices (finite states and actions) involved and further choosing the policies π_1', π_2' etc., inductively we obtain the following for any increasing sequence of $\{\epsilon_i\}$ and with $\epsilon = \epsilon_{T-1}$:

$$\mathbf{u}_0 \ge \sum_{a_0} C_{0,a_0,\pi_0'} P_{a_0} \sum_{a_1} C_{1,a_1,\pi_1'} P_{a_1} \mathbf{u_2} - \epsilon_1$$

$$\vdots$$

$$\ge \sum_{\mathbf{a}_0^{T-1}} C_{0,a_0,\pi_0'} P_{a_0} C_{1,a_1,\pi_1'} P_{a_1} \cdots C_{T-1,a_{T-1},\pi_{T-1}'} P_{a_{T-1}} \mathbf{u_T} - \epsilon$$

Note in the above that, for example, ϵ_1 is chosen such that (for appropriate choice of ϵ_1'):

$$\epsilon_1 \ge \sum_{a_0} C_{0,a_0,\pi_0'} P_{a_0} \epsilon_1' + \epsilon_0.$$

Thus as in the proof of the previous theorem,

$$\begin{aligned} &\mathbf{u_0} \\ &\ge \sum_{\mathbf{a}_0^{T-1}} C_{0,a_0,\pi_0'} P_{a_0} C_{1,a_1,\pi_1'} P_{a_1} \cdots C_{T-1,a_{T-1},\pi_{T-1}'} P_{a_{T-1}} \mathbf{u_T} - \epsilon \\ &= \mathbf{J}_0(\Pi') - \epsilon \text{ with } \Pi' = [\pi_0', \pi_1', \cdots, \pi_{T-1}']. \end{aligned}$$

Thus

$$\mathbf{u}_0 \ge \mathbf{J}_0(\Pi') - \epsilon \ge \inf_{\Pi} \mathbf{J}_0(\Pi) - \epsilon = \mathbf{u}_0^* - \epsilon.$$

Thus for any $\epsilon > 0$ one can chose appropriate increasing sequence of $\{\epsilon_i\}$ such that

$$\mathbf{u}_0 \ge \mathbf{u}_0^* - \epsilon.$$

Since $\epsilon > 0$ is arbitrary, consider the limit $\epsilon \to 0$ and hence

$$\mathbf{u}_0 \ge \mathbf{u}_0^*.$$

Following exactly similar logic one can show for all $1 \le t < T$ that

$$\mathbf{u}_t \ge \mathbf{u}_t^*. \qquad \blacksquare$$

Proof of Theorem 3:
It is easy to see that the defined point $\mathbf{y}_\Pi$ satisfies the first constraint (9), as by definition for all x_0

$$\sum_{a_0} y_\Pi(0, x_0, a_0) = \sum_{a_0} \alpha(x_0) \pi_0(x_0, a_0) = \alpha(x_0).$$

Define

$$\Delta_\Pi^t := \prod_{n=0}^{t} q_n^\Pi(x_n, a_n | x_{n-1}, a_{n-1}) e^{\sum_{n=0}^{t-1} r_n(x_n, a_n)}.$$

Note that Δ_Π^t depends upon the vectors $\mathbf{a}_0^t, \mathbf{x}_0^t$. Using the above definition, we can rewrite

$$y_\Pi(t,x,a) = \sum_{\mathbf{a}_0^{t-1}\mathbf{x}_0^{t-1}} \alpha(x_0)\Delta_\Pi^t.$$

To simplify the notations, we represent the action state pair by $z_t := (x_t, a_t)$, for every t. Considering the right hand side (RHS) of the second constraint (10):

$$\begin{aligned}
&\sum_{z_{t-1}=(x_{t-1},a_{t-1})} e^{r_{t-1}(z_{t-1})} p(x_t|z_{t-1}) y_\Pi(t-1, z_{t-1}) \\
&= \sum_{z_{t-1}} e^{r_{t-1}(z_{t-1})} p(x_t|z_{t-1}) \sum_{\mathbf{z}_0^{t-2}=(\mathbf{a}_0^{t-2}\mathbf{x}_0^{t-2})} \alpha(x_0)\Delta_\Pi^{t-1} \\
&= \sum_{\mathbf{z}_0^{t-1}} e^{r_{t-1}(z_{t-1})} \alpha(x_0)\Delta_\Pi^{t-1} \sum_{a_t} \pi_t(z_t) p(x_t|z_{t-1}) \\
&= \sum_{a_t}\sum_{\mathbf{z}_0^{t-1}} \alpha(x_0)\left(e^{r_{t-1}(z_{t-1})}\Delta_\Pi^{t-1} q_t^\Pi(z_t|z_{t-1})\right) \\
&= \sum_{a_t}\sum_{\mathbf{z}_0^{t-1}} \alpha(x_0)\Delta_\Pi^t = \sum_{a_t} y_\Pi(t, z_t).
\end{aligned}$$

Hence the point y_Π satisfies the second constraint (10).

Part (ii): Consider any $\mathbf{y} \in \mathcal{F}$, the policy $\Pi_\mathbf{y}$ defined as in (12) and then the point $\mathbf{y}_{\Pi_\mathbf{y}}$ defined using policy $\Pi_\mathbf{y}$ as in (11). Aim is to prove that

$$y(t, z_t) = y_{\Pi_\mathbf{y}}(t, z_t) \text{ for all } t \le T-1, x_t \in \mathcal{X}, a_t \in \mathcal{A}.$$

Fix (t, z_t). As in previous proof, define

$$\Delta_\Pi^{k,t} := \prod_{n=k}^{t} q_n^\Pi(z_n|z_{n-1}) e^{\sum_{n=k}^{t-1} r_n(z_n)}$$

now including the starting time k. Since $\mathbf{y} \in \mathcal{F}$, it satisfies (9) and by definition (12) of $\Pi_\mathbf{y}$ we have:

$$\begin{aligned}
y_{\Pi_\mathbf{y}}(t, z_t) &= \sum_{\mathbf{z}_0^{t-1}} \alpha(x_0)\Delta_\Pi^{0,t} \\
&= \sum_{\mathbf{z}_0^{t-1}} \Delta_\Pi^{1,t}\alpha(x_0) e^{r_0(z_0)} \pi_{\mathbf{y},0}(z_0) \\
&= \sum_{\mathbf{z}_0^{t-1}} \Delta_\Pi^{1,t}\alpha(x_0) e^{r_0(z_0)} \frac{y(0, z_0)}{\sum_{a_0'} y(0, x_0, a_0')}; \quad \text{using (12)} \\
&= \sum_{\mathbf{z}_0^{t-1}} \Delta_\Pi^{1,t} e^{r_0(z_0)} y(0, z_0); \qquad \text{using (9).}
\end{aligned}$$

Further expanding $\Delta_\Pi^{1,t} = \Delta_\Pi^{2,t} e^{r_1(z_1)} q_1^\Pi(z_1|z_0)$ and simplifying as before, we reduce one pair of elements (z_0) in the summation:

$$\begin{aligned}
y_{\Pi_\mathbf{y}}(t, z_t) &= \sum_{\mathbf{z}_1^{t-1}} \Delta_\Pi^{2,t} e^{r_1(z_1)} \sum_{z_0} e^{r_0(z_0)} q_1^\Pi(z_1|z_0) y(0, z_0) \\
&= \sum_{\mathbf{z}_1^{t-1}} \Delta_\Pi^{2,t} e^{r_1(z_1)} \pi_{\mathbf{y},1}(z_1) \sum_{z_0} e^{r_0(z_0)} p(x_1|z_0) y(0, z_0) \\
&= \sum_{\mathbf{z}_1^{t-1}} \Delta_\Pi^{2,t} e^{r_1(z_1)} \pi_{\mathbf{y},1}(z_1) \sum_{a_1'} y(1, x_1, a_1'); \qquad \text{using (10)} \\
&= \sum_{\mathbf{z}_1^{t-1}} \Delta_\Pi^{2,t} e^{r_1(z_1)} \frac{y(1, z_1)}{\sum_{a_1'} y(1, x_1, a_1')} \sum_{a_1'} y(1, x_1, a_1'); \ \text{using (12)} \\
&= \sum_{\mathbf{z}_1^{t-1}} \Delta_\Pi^{2,t} e^{r_1(z_1)} y(1, z_1).
\end{aligned}$$

Proceeding in a similar way, we reduce one more pair of elements $z_1 = (x_1, a_1)$ summation, that is:

$$\begin{aligned}
y_{\Pi_\mathbf{y}}(t, z_t) &= \sum_{\mathbf{z}_2^{t-1}} \Delta_\Pi^{3,t} e^{r_2(z_2)} \sum_{z_1} e^{r_1(z_1)} q_2^\Pi(z_2|z_1) y(1, z_1) \\
&= \sum_{\mathbf{z}_2^{t-1}} \Delta_\Pi^{3,t} e^{r_2(z_2)} \pi_{\mathbf{y},2}(z_2) \sum_{z_1} e^{r_1(z_1)} p(x_2|z_1) y(1, z_1) \\
&= \sum_{\mathbf{z}_2^{t-1}} \Delta_\Pi^{3,t} e^{r_2(z_2)} \pi_{\mathbf{y},2}(z_2) \sum_{a_2'} y(2, x_2, a_2'); \\
&\qquad\qquad \text{using (10)} \\
&= \sum_{\mathbf{z}_2^{t-1}} \Delta_\Pi^{3,t} e^{r_2(z_2)} \frac{y(2, z_2)}{\sum_{a_2'} y(2, x_2, a_2')} \sum_{a_2'} y(2, x_2, a_2') \\
&= \sum_{\mathbf{z}_2^{t-1}} \Delta_\Pi^{3,t} e^{r_2(z_2)} y(2, z_2).
\end{aligned}$$

Repeating exactly the same steps, we eliminate all the terms till and including (z_{t-2}), to obtain the following (note that $\Delta_\Pi^{t,t} = q_t^\Pi(z_t|z_{t-1})$):

$$\begin{aligned}
&y_{\Pi_\mathbf{y}}(t, z_t) \\
&\quad = \sum_{z_{t-1}} q_t^\Pi(z_t|z_{t-1}) e^{r_{t-1}(z_{t-1})} y(t-1, z_{t-1}) \\
&\quad = \pi_t(z_t) \sum_{z_{t-1}} e^{r_{t-1}(z_{t-1})} p(x_t|z_{t-1}) y(t-1, z_{t-1}) \\
&\quad = \frac{y(t, z_t)}{\sum_{a_t'} y(t, x_t, a_t')} \sum_{a_t'} y(t, x_t, a_t'); \quad \text{using (10)} \\
&\quad = y(t, z_t).
\end{aligned}$$

This is true for all $t \le T-1$. ■

Proof of Lemma 1: By Theorem (3), $y(t, z_t) = y_{\Pi_X}(t, z_t)$ for any $t < T$. Further, using the definition of $y_{\Pi_X}(t, z_t)$ from (10), one can rewrite left hand side

(LHS) of (13)

$$
\begin{aligned}
\sum_{z_t} y(t, z_t) f(z_t) & \\
&= \sum_{z_t} y_{\Pi_X}(t, z_t) f(z_t) \\
&= \sum_{z_t} f(z_t) \sum_{\mathbf{z_0^{t-1}}} \alpha(x_0) \prod_{n=0}^{t} q_n^{\Pi}(z_n | z_{n-1}) e^{\sum_{n=0}^{t-1} r_n(z_n)} \\
&= \sum_{\mathbf{z_0^t}} \Big(f(z_t) e^{\sum_{n=0}^{t-1} r_n(z_n)} \Big) \, \alpha(x_0) \prod_{n=0}^{t} q_n^{\Pi}(z_n | z_{n-1}) \\
&= E^{\Pi_{\mathbf{y}}} \Big[e^{\sum_{n=0}^{t-1} r_n(X_n, A_n)} f(X_t, A_t) \Big] \text{ for any } t < T.
\end{aligned}
$$

One can get the second result (14) in exactly similar lines. ■

Proof of Theorem (4): We begin with proof of (a). Let $g(\mathbf{y})$ represent the dual objective (see (9)), for any $\mathbf{y} \in \mathcal{F}$, i.e.,

$$
g(\mathbf{y}) := \sum_{a_{T-1}} \sum_{x_{T-1}} e^{r_{T-1}(z_{T-1})}
$$

$$
\left[\sum_{x_T \in \mathcal{X}} p(x_T | z_{T-1}) e^{r_T(x_T)} \right] y(T-1, z_{T-1}).
$$

Let $\mathbf{y}^*$ be an optimal solution of the dual LP, and let $\Pi_{\mathbf{y}^*}$ be the corresponding policy given by (12). By optimality and because of equation (15), for any $\Pi \in \mathcal{D}$:

$$
\begin{aligned}
E^{\alpha, \Pi} & \left[e^{\sum_{n=t}^{T-1} r_n(X_n, A_n) + r_T(X_T)} \right] \\
&= \quad g(\mathbf{y}_{\Pi}) \\
&\geq \quad g(\mathbf{y}^*) \\
&= \quad E^{\alpha, \Pi_{\mathbf{y}^*}} \left[e^{\sum_{n=t}^{T-1} r_n(X_n, A_n) + r_T(X_T)} \right],
\end{aligned}
$$

establishing the required optimality.

Part (b) can be proved in a similar way. ■

REFERENCES

[1] Atul Kumar, Veeraruna Kavitha and N. Hemachandra, “Finite horizon risk sensitive MDP and linear programming,” 2015, Technical report available at *h*ttp://www.ieor.iitb.ac.in/files/faculty/kavitha/RiskMDPLP.pdf.

[2] Altman, Eitan. Constrained Markov decision processes. Vol. 7. CRC Press, 1999.

[3] Bertsekas, Dimitri P., and Dimitri P. Bertsekas. Dynamic programming and optimal control. Vol. 1. No. 2. Belmont, MA: Athena Scientific, 1995.

[4] Feinberg, Eugene A., and Adam Shwartz, eds. Handbook of Markov decision processes: methods and applications. Boston, MA: Kluwer Academic Publishers, 2002.

[5] Stefano P Coraluppi and Steven I Marcus. Risk-sensitive queueing. In *Proceedings of the Annual Allerton Conference on Communication Control and Computing*, volume 35, pages 943–952. Citeseer, 1997.

[6] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[7] Howard, Ronald A., and James E. Matheson. ”Risk-sensitive Markov decision processes.” Management Science 18.7 (1972): 356-369.

[8] Coraluppi, Stefano P., and Steven I. Marcus. ”Risk-sensitive and minimax control of discrete-time, finite-state Markov decision processes.” Automatica 35.2 (1999): 301-309.

[9] Osogami, Takayuki. ”Robustness and risk-sensitivity in Markov decision processes.” Advances in Neural Information Processing Systems. 2012.

[10] Borkar, Vivek S., and Sean P. Meyn. ”Risk-sensitive optimal control for Markov decision processes with monotone cost.” Mathematics of Operations Research 27.1 (2002): 192-209.

[11] Atul Kumar, Veeraruna Kavitha and N. Hemachandra, ”Power Constrained DTNs: Risk MDP-LP Approach”. International Workshop on D2D Communications held in conjunction with WiOpt 2015, Mumbai, India.