LPs for Finite Horizon (Constrained) Risk MDPs

Atul Kumar, Veeraruna Kavitha and N. Hemachandra
Industrial Engineering and Operations Research, IIT Bombay, Mumbai, India
Email: atulkr.in@gmail.com, vkavitha@iitb.ac.in and nh@iitb.ac.in

***Abstract*—Finite horizon linear or standard Markov Decision processes (MDPs) consider optimization of expected value of sum of the stage-wise running costs over all the relevant time slots. Two important solution approaches for solving MDPs are Dynamic Programming (DP) and Linear Programming (LP) based methods. In the context of standard MDPs, the connection between the two approaches is well understood and is well established under sufficient general conditions. LP based approach, in addition, facilitates solving the constrained MDPs. Risk sensitive MDPs, introduced to control the fluctuations/variations around the expected value, are relatively less studied objects. Here one considers optimization of expected value of the exponential of the sum cost. This is an example of multiplicative MDPs. 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. The results obtained are applied to two applications: the Delay Tolerant Networks (DTNs) and server selection problem in $Ber/M/K/K$ queues. For DTNs our approach provided solution for power constrained problem. We optimized the expected losses in $Ber/M/K/K$ queue under a constraint on the fast server utilization. Some interesting structural properties of the resulting risk optimal policies are discussed.**

I. Introduction

Markov Decision process (MDP) is a mathematical framework that solves the problem of sequential decision making under uncertainty, to obtain an optimal policy ([6], [3], [1], [4] etc). A policy is a sequence of actions/rules, one for each time slot, depending upon the current state of the system. MDP has a running cost (r_t) associated with every decision epoch/slot, which depends upon the state and the action taken at that time slot. In case of finite horizon problems, it also considers a terminal cost that only depends upon the state at termination.

In general, MDP optimizes the expected value of a function f that defines a way of aggregating the running costs related to all the time slots under consideration, i.e., optimizes $E[f(r_1, r_2 \cdots)]$. The linear MDPs consider expected value of either sum of all the running costs $E[\sum_t r_t]$, or sum of discounted values of all the running costs $E[\sum_t \beta^t r_t]$ with discount factor $\beta < 1$ or time average of the running costs $\lim_{T\to\infty} E[\sum_{t\leq T} r_t]/T$. They are respectively called total cost, discounted cost or average cost problems. Further the MDP problem can either span over a finite time horizon or over the infinite time horizon. The focus of this work is on finite horizon problems and when the aggregation function equals the exponential of the sum of the running costs, $E[e^{-\gamma \sum_{t\leq T} r_t}]$. Thus we are working with a special case of multiplicative and finite horizon MDPs.

In many scenarios it is not sufficient to optimize the expected cost, rather the decision maker likes to reduce the risk on majority of sample path trajectories. Worst case analysis deals with an extreme case in this direction, while risk sensitive framework offers varying degrees of importance to sample path trajectories and the expected value, as controlled by a parameter γ. Depending upon γ, called the risk parameter, it provides importance to higher moments of the sum cost. Note $E[e^{\gamma \sum_{t\leq T} r_t}] = E[\sum_k \gamma^k (\sum_{t\leq T} r_t)^k / k!]$. In all, while the linear MDPs control the first moment (expected value) of the sum cost, the risk sensitive MDPs also control the variability/fluctuations around the expected value, by considering the higher moments. The linear MDPs are also viewed as risk neutral MDPs.

The linear MDP is a well studied topic and many solution approaches are known. Dynamic programming (DP), Linear program (LP), Value iteration are some of them ([1], [3], [4], [6]). DP obtains the value function, the optimal expected total cost to go till termination from any time and any state, using backward induction. Alternatively, value function can be obtained using a solution of an appropriate LP. The dual LP directly provides the optimal policy (e.g., [1], [6]).

The literature on risk sensitive MDP is relatively limited, however is still vast and varied. We give a sample of some of the strands. The pioneering work is done by Howard and Matheson [7], where the generalization of MDP cost as risk cost has been considered. 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 (e.g., [2], [5]). In general, the optimal policies in infinite horizon setting tend to be non stationary, (e.g., [8]). Some papers identify sufficient conditions for stationary optimal policies and also develop algorithms to compute either exact or approximate optimal policies (e.g., [9]).

As mentioned earlier, LP based approach is another alternative to solve linear MDPs. The connections between DPs and LPs are well established (e.g., [6]). To the best of our knowledge such connections are not known in the context of risk sensitive MDPs. In [15], authors provide connection between the dynamic programming equations and a concave programming to solve an infinite horizon risk sensitive MDP problem.

In this paper, we establish the connection between DP equations and two appropriate LPs for finite horizon case: (i) The primal LP provides the value function, the optimal expected cost to go; (ii) The dual LP provides the optimal policy. For linear MDPs, the cost accumulates in a linear fashion across the time slots. Because of this, it is possible

Parts of this work is presented in CDC 2015 ([16]) and in [17]. This journal version combines the two papers and further includes all the proofs.

to construct the linear objective function of the relevant LP using the running costs of all the time slots. However the cost accumulates in a multiplicative manner for risk sensitive MDPs and hence the above approach can not be adapted directly for constructing an LP for risk MDP. We circumvent this problem by incorporating the multiplicative cost term into the mapping, that converts any given Markov policy to a feasible point of the LP.

The LPs automatically provide us a path for solving constrained risk MDPs. However we had to handle the multiplicative cost term in feasible point-policy mapping. In the context of linear MDPs it is straight forward to add additional constraints to the LP, when one requires a solution to a constrained MDP problem. When we attempted the same procedure to obtain an equivalent constraint for LP, the multiplicative cost (incorporated in the transformation) appears in the constraint equation (details are in section VI). We propose to augment the state space with an extra component, which is representative of this multiplicative cost. With the augmented state space we could handle the constraints and hence provide a third LP to solve the finite horizon constrained risk sensitive MDP. Our LP connection thus provides computational methods to solve constrained as well as unconstrained risk sensitive MDPs.

Availability of resources can be captured as suitable constraints and hence solutions to constrained MDPs are important. We apply our results to two different applications with two different purposes. In the first application we consider delay tolerant networks (e.g., [2], [14]), where a message has to be transferred from a source to a far away destination with the help of occasional contacts between the freely moving nodes (that are willing to become the relays) and the source/destination. The contact process turns out to be Poisson ([14]), because of which the delivery-failure (message not delivered within the given deadline) probability takes the form of a finite horizon risk sensitive cost. In [2] authors solve such a problem using risk MDP approach, referred to as soft problem, where they optimize a joint cost that includes a term proportional to the expected power spent. These nodes usually operate using batteries, and hence would have a hard constraint on the available power. Our LP based solution for constrained risk MDP problems, makes it possible to optimize the delivery-failure probability given the constraint on the expected power utilized. We observed very good improvement in delivery-failure probability with hard optimal policies over that with soft optimal policies, when both the policies utilize the same expected power.

For the second application, we consider a lossy (finite buffer) queueing system with two server modes and with a constraint on the utilization of fast server mode. Our purpose here is to compare the risk neutral policies with risk sensitive policies for the constrained MDP problem, that optimizes the expected number of customers lost. As anticipated, the risk neutral policies are (time) threshold type while the risk sensitive policies are no more threshold type. They are not even monotone. Further the risk policy utilizes the fast mode with higher probability when the number waiting is large (i.e., at high risk points, where the new arrivals can be lost), while the risk neutral policies allocate in the opposite manner, leading to smaller expected number of losses.

The remaining part of the paper is organized as follows. A brief description about risk sensitive MDP framework and main results are presented in Section II. In Section III and IV we discuss the two applications. In Sections V and VI we obtain the proof of Theorem 1 and 2 respectively. Appendix contains some parts of the proof.

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 (12), given below. Also $\mathbf{x}_n^t$ 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 and avoided when not required. For example, x represents a realization of random variable X_t for any t. If it is 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 MDP framework and Main Results

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:

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

A finite horizon MDP problem has terminal cost r_T which depends only upon the state $x \in \mathcal{X}$. The state, action spaces $\mathcal{X}$, $\mathcal{A}$ do not depend[1] on the time slot t, i.e., we consider the same set for all the time slots. 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' ($x, x' \in \mathcal{X}$), when action $a \in \mathcal{A}$ is chosen.

A policy $\Pi = (\pi_0, \pi_1 \cdots \pi_{T-1})$ is a sequence of state dependent (i.e., Markov) and possibly randomized actions, given for time slots between 0 and $T-1$. Let $\{X_t\}_{t \le T}$, $\{A_t\}_{t \le T-1}$ respectively represent the trajectories of the state and the action processes. Given a policy Π, the state-action pair evolve randomly, with transitions as below (for $0 < t < T$):

$$\begin{aligned} q_t^\Pi(x', a'|x, a) &= P(X_t = x', A_t = a' | X_{t-1} = x, A_{t-1} = a) \\ &= \pi_t(x', a') p(x'|x, a) \text{ where} \\ p(x'|x,a) &= P(X_t = x' | X_{t-1} = x, \, A_{t-1} = a) \text{ and} \\ \pi_t(x', a') &= P(A_t = a' | X_t = x'). \end{aligned} \tag{1}$$

For ease of notations, define dummy variables (x_{-1}, a_{-1}) and let $q_0^\Pi(x_0, a_0 | x_{-1}, a_{-1}) := P(A_0 = a_0 | X_0 = x) = \pi_0(x_0, a_0)$. The above evolution further depends upon the initial condition, i.e., the starting state X_0. Let $E^{x,\Pi}$ represent the expectation operator with initial condition $X_0 \equiv x$ (or almost surely) and when the policy Π is used. Let $E^{\alpha,\Pi}$ represent the same expectation operator when the initial condition is distributed according to α, written as $X_0 \sim \alpha$. Here $\alpha(x) = P(X_0 = x)$. A standard stochastic decision problem considers the objective function,

$$\tilde{J}_0(\alpha, \Pi) = E^{\alpha,\Pi}\left[\sum_{t=0}^{T-1} r_t(X_t, A_t) + r_T(X_T) \right]. \tag{2}$$

[1] These assumptions can easily be generalized, but are not considered to keep the notations/explanations simple. Further, $t = 0$ is the initial time slot.

The objective (2) is called risk neutral, because it considers only the expected total cost. The above cost does not consider the variations around the mean. But there are many scenarios in which importance is given to various random realizations. In such cases importance has to be given to higher moments of the cost. One of the possible ways to achieve this goal is by considering the following transformation of the risk neutral objective function (2) (e.g., [5], [7]):

$$\begin{aligned}\tilde{J}_0(\alpha,\Pi) &= \gamma^{-1}\log\left(J_0(\alpha,\Pi)\right) \text{ where}\\ J_0(\alpha,\Pi) &= E^{\alpha,\Pi}\left[e^{\gamma(\sum_{t=0}^{T-1} r_t(X_t,A_t)+r_T(X_T))}\right]. \qquad (3)\end{aligned}$$

This type of generalization of MDP objective was first introduced in ([7]) and is called risk sensitive objective. For smaller value of γ the above objective takes the form:

$$\begin{aligned}\tilde{J}_0(\alpha,\Pi) \simeq E^{\alpha,\Pi}&\left[\sum_{t=0}^{T-1} r_t(X_t,A_t)+r_T(X_T)\right]\\ &+\frac{\gamma}{2}Var^{\alpha,\Pi}\left[\sum_{t=0}^{T-1} r_t(X_t,A_t)+r_T(X_T)\right]. \qquad (4)\end{aligned}$$

As $\gamma \to 0$, the impact of the variance term (similarly higher moments) becomes negligible and cost (4) converges towards the risk neutral cost (2). For γ sufficiently different from 0, the variance part (subsequently the higher moments) significantly influences the optimal policy, which penalizes the scenarios with large variability. The parameter γ is called the risk parameter which allows us to consider the risk averse ($-\gamma < 0$) and risk seeking ($-\gamma > 0$) cost criteria (e.g., [5]).

We are interested in optimizing the finite horizon risk sensitive objective (3). The equation (3) represents the cost to go from time slot 0 to T under the policy Π, with $X_0 \sim \alpha$. When $X_0 \equiv x$, we let $J_0(x,\Pi)$, $\tilde{J}_0(x,\Pi)$ represent the corresponding costs to go. The value function, a function[2] of x, is defined as the optimal value of the above risk sensitive objective given the initial condition $X_0 \equiv x$:

$$v^*(x) := \min_{\Pi\in\mathcal{D}} \tilde{J}_0(x,\Pi) \text{ for any } x\in\mathcal{X}, \qquad (5)$$

where $\mathcal{D}$ represents the space of Markov policies, Π.

A. Dynamic Programming

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

$$\begin{aligned}&v_T(x) = r_T(x), \text{ and for any } 0 \le t \le T-1,\\ &v_t(x) = \min_{a\in\mathcal{A}}\left\{r_t(x,a)+\frac{1}{\gamma}\log\left[\sum_{x'\in\mathcal{X}} p(x'|x,a)e^{\gamma v_{t+1}(x')}\right]\right\},\end{aligned}$$

and their solution $\{v_t^*(x)\}_{t,x}$ provides the value function (5), i.e., $v^*(x) = v_0^*(x)$ for all x. 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 $u_0^*(x) = e^{\gamma v_0^*(x)}$ is minimum value of the risk cost J_0 given in (3):

$$u_0^*(x) = \min_{\Pi} J_0(x,\Pi).$$

The DP equations can now be rewritten as:

$$\begin{aligned}u_T(x) &= e^{\gamma r_T(x)} \text{ for any } x\in\mathcal{X} \text{ and} \qquad (6)\\ u_t(x) &= \min_a\left\{e^{\gamma r_t(x,a)}\sum_{x'\in\mathcal{X}} p(x'|x,a)u_{t+1}(x')\right\}\\ &\quad \text{for any } 0\le t\le T-1, \text{ and } x\in\mathcal{X}. \qquad (7)\end{aligned}$$

The solution of the above set of equations provides the translated value function and let this be denoted by the following vector

$$\underline{\mathbf{u}}^* = \{u_t^*(x);\ 0 \le t < T, x\in\mathcal{X}\}. \qquad (8)$$

The optimizers in the minimization step will provide the optimal policy ([6], [5] etc). For ease of notation, we absorb γ into the cost functions $\{r_t\}$.

B. Constrained Risk MDP

Let us consider a stochastic decision making problem where one considers not only the system gain/loss, but also the cost spent for resources. Say there is a constraint on the resources utilized. If this constraint is represented by functions $\{f_t\}$, then we are interested in the following variant of the original risk MDP problem (5):

$$\min_{\Pi} \tilde{J}_0(\alpha,\Pi) \qquad (9)$$

$$\text{Subject to: } \sum_t E^{\alpha,\Pi}\left[f_t(X_t,A_t)\right] \le B, \qquad (10)$$

where $\{f_t\}$ represents a set of integrable functions, B represents the constraint and $\tilde{J}_0(\alpha,\Pi)$ is the running cost (3).

C. Main Results

The first two main results of this paper are the connection with the following two LPs and risk senstivie MDPs:

1) LP for risk MDP: Consider the following two LPs:

Primal Linear Program:

$$\max_{\left\{\{u_t(x)\}_{x\in\mathcal{X},t\le T-1}\right\}} \sum_{x\in\mathcal{X}} \alpha(x)u_0(x) \qquad (11)$$

subject to: $u_{T-1}(x) \le b_{x,a}$ for all x, a,

$$u_t(x) - e^{r_t(x,a)}\sum_{x'\in\mathcal{X}} p(x'|x,a)u_{t+1}(x') \le 0$$

for all a, x and $t \le T-2$

with $b_{x,a} := e^{r_{T-1}(x,a)}\sum_{x'\in\mathcal{X}} p(x'|x,a)e^{r_T(x')}$.

Dual Linear Program:

$$\min_{\mathbf{y}=\{y(t,x,a);t\le T-1,x\in\mathcal{X},a\in\mathcal{A}\}} \sum_a\sum_{x\in\mathcal{X}} b_{x,a}\, y(T-1,x,a) \qquad (12)$$

subject to:

$$\sum_a y(0,x',a) = \alpha(x') \text{ for all } x'\in\mathcal{X}, \qquad (13)$$

$$\sum_a y(t,x',a) = \sum_a\sum_x e^{r_{t-1}(x,a)}p(x'|x,a)y(t-1,x,a)$$

$$\text{for all } 1\le t\le T-1 \text{ and } x'\in\mathcal{X} \qquad (14)$$

[2] In general value function can be defined as a function of (x,t) (optimal cost with initialization $X_t = x$), but we fix $t = 0$ to simplify the notations.

Hemachandra : These are the limit LPs when $\gamma \to \infty$ We basically neglect the terms with γ. Some times all 1 terms are cancelled and then we would both sides divide by γ if possible. For example if $e^r_T e^r_{T-1} \approx 1 + r_T + r_{T-1}$ because $r_T r_{T-1}$ are negligible because they contain γ^2 terms and further term containing 1 cancels out with RHS/LHS.

Primal Linear Program:

$$\max_{\{\{v_t(x)\}_{x \in \mathcal{X}, t \le T-1}\}} \sum_{x \in \mathcal{X}} \alpha(x) v_0(x) \tag{15}$$

subject to: $v_{T-1}(x) \le b_{x,a}$ for all x, a,

$$v_t(x) - r_t(x,a) - \sum_{x' \in \mathcal{X}} p(x'|x,a) v_{t+1}(x') \le 0$$

for all a, x and $t \le T - 2$

with $b_{x,a} := r_{T-1}(x,a) + \sum_{x' \in \mathcal{X}} p(x'|x,a) r_T(x')$.

Dual Linear Program:

$$\min_{\mathbf{y} = \{y(t,x,a); t \le T-1, x \in \mathcal{X}, a \in \mathcal{A}\}} \sum_a \sum_{x \in \mathcal{X}} b_{x,a}\, y(T-1,x,a) \tag{16}$$

subject to:

$$\sum_a y(0,x',a) = \alpha(x') \text{ for all } x' \in \mathcal{X}, \tag{17}$$

$$\sum_a y(t,x',a) = \sum_a \sum_x p(x'|x,a) y(t-1,x,a)$$

$$\text{for all } 1 \le t \le T-1 \text{ and } x' \in \mathcal{X} \tag{18}$$

Theorem 1: (a) Any optimal solution of primal LP (11) equals the value function $\underline{\mathbf{u}}^*$ given by (8) of the risk MDP .
(b) For any feasible vector $\mathbf{y}$ define:

$$\pi_{\mathbf{y},t}(x,a) := \frac{y(t,x,a)}{\sum_{a'} y(t,x,a')} \text{ for all } x \in \mathcal{X}, \text{ and } a \in \mathcal{A}, \tag{19}$$

If $\mathbf{y}^*$ is an optimal solution of the dual LP (12), then $\Pi_{\mathbf{y}^*}$ is an optimal policy for risk MDP (5).
(c) For any $\Pi \in \mathcal{D}$, define vector $\mathbf{y}_\Pi$ as:

$$y_\Pi(0,x_0,a_0) = \alpha(x_0)\pi_0(x_0,a_0) \text{ for all } x_0 \in \mathcal{X}, a_0 \in \mathcal{A},$$

$$y_\Pi(t,x_t,a_t)$$

$$= \sum_{\mathbf{a}_0^{t-1}, \mathbf{x}_0^{t-1}} \alpha(x_0) e^{\sum_{n=0}^{t-1} r_n(x_n,a_n)}\, \Pi_{n=0}^t q_n^\Pi(x_n, a_n | x_{n-1}, a_{n-1})$$

$$\text{for all } x_t \in \mathcal{X}, a_t \in \mathcal{A}, \text{ and } 1 \le t < T. \tag{20}$$

If Π^* is an optimal policy for risk MDP, then $\mathbf{y}_{\Pi^*}$ defined by (20) is an optimal solution of the dual LP (12). ■

2) LP for Constrained Risk MDP: In the case of linear MDPs, one can solve a constrained MDP problem by adding an appropriate additional constraint to the corresponding LP (see for example [6, Example 6.9.1, pp. 229]). However in case of risk MDPs, it is not straight forward to add this extra constraint. We introduce an additional state component ψ_t (details are in the next subsection and in section VI), which is instrumental in translating the constraint of MDP problem to an equivalent LP constraint, and obtain the following LP:

Dual LP for constrained risk MDP:

$$\min \sum_a \sum_x e^{r_{T-1}(x,a)} \left[\sum_{x' \in \mathcal{X}} p(x'|x,a) e^{r_T(x')} \right] y(T-1,x,a) \tag{21}$$

subject to:

$$y(t,x,a) = \sum_{\psi_t} y(t,x,\psi_t,a) \text{ for all } t$$

$$\sum_a y(0,x,\psi_0,a) = \alpha(x) 1_{\{\psi_0 = 1\}} \text{ for all } x, \psi_0$$

$$\sum_a y(t,x',\psi'_t,a)$$

$$= \sum_{a,x,\psi_{t-1}} e^{r_{t-1}(x,a)} \tilde{p}_t(x',\psi'_t | x, \psi_{t-1}, a) y(t-1,x,\psi_{t-1},a)$$

for all $1 \le t \le T-1$ and x', ψ'_t with

$$\tilde{p}_t(x',\psi'_t | x,\psi_{t-1},a) := 1_{\{\psi'_t = \psi_{t-1} e^{-r_{t-1}(x,a)}\}} p(x'|x,a)$$

$$\text{and } \sum_t \sum_{x,\psi_t,a} y(t,x,\psi_t,a)\, \psi_t\, f_t(x,a) \le B. \tag{22}$$

Theorem 2: (a) If $\mathbf{y}^*$ is an optimal solution of the LP (21), then $\Pi_{\mathbf{y}^*}$ defined by (19) is an optimal policy for constrained risk MDP (9).
(b) If Π^* is an optimal policy for constrained risk MDP (9), then $\mathbf{y}_{\Pi^*}$ defined by (20) is an optimal solution of the LP (21).
■

D. *Comparison with Linear MDPs*

The objective function of the LP that provides solution for finite horizon linear MDP is given by (see e.g., []):

$$\sum_{t=1}^{T-1} \sum_{a_t} \sum_{x_t} r_t(x_t,a_t)\, y(t,x_t,a_t)$$

$$+ \sum_a \sum_{x,x'} r_T(x')\, y(T-1,x,a) p(x'|x,a).$$

As seen from above, the running costs are incorporated in the objective function itself. While the objective function of risk MDP-LP (21) does not include running costs of all the time slots. With risk MDPs, the cost does not accumulate in linear fashion over the subsequent time slots and this poses a difficulty in obtaining the linear objective function. We circumvented this problem by incorporating the multiplicative term, formed using the running costs, in the mapping that converts any dual policy π to a feasible point as below (see equation (20) for risk MDPs and [] for linear MDPs):

$$y_\pi(t,x,a)$$

$$= \begin{cases} E^{\alpha,\pi}\left[X_t = x, A_t = a\right], & \text{for linear MDP} \\ E^{\alpha,\pi}\left[e^{\sum_{n=0}^{t-1} r_n(X_n,A_n)}; X_t = x, A_t = a\right], & \text{for risk MDP.} \end{cases}$$

Dual Linear Program for Linear MDP (Atul Check this):

$$\min_{\mathbf{y} = \{y(t,x,a); t \le T-1, x \in \mathcal{X}, a \in \mathcal{A}\}} \sum_{t=1}^{T-2} \sum_{a_t} \sum_{x_t \in \mathcal{X}} r_t(x_t,a_t)\, y(t,x_t,a_t)$$

$$+ \sum_a \sum_{x,x' \in \mathcal{X}} (r_{T-1}(x,a) + r_T(x'))\, y(T-1,x,a) p(x'|x,a)$$

subject to:

$$\begin{aligned}\sum_a y(0,x',a) &= \alpha(x') \text{ for all } x' \in \mathcal{X},\\ \sum_a y(t,x',a) &= \sum_a \sum_x p(x'|x,a) y(t-1,x,a)\\ &\qquad \text{for all } 1 \le t \le T-1 \text{ and } x' \in \mathcal{X}\end{aligned}$$

But because of this when one tries to handle constraints, the multiplicative risk cost term till the time slot under consideration ($e^{\sum_{n=0}^{t-1} r_n(X_n,A_n)}$) appears in the constraint equations (see equation (34) of Lemma 1). The augmented state component ψ_t of dual LP (21) precisely equals the inverse of this multiplicative cost and by multiplying the constraint equations with ψ_t as in (22) we inverted the effect of multiplicative cost term.

We consider two applications of these results, before proceeding with the proof of the two theorems. The first application uses LP based approach to solve a constrained risk sensitive cost that arises naturally in the context of Delay Tolerant Networks (DTNs). The second application considers Bernoulli queueing system with two types (fast and slow) of servers/serving modes. The problem is to chose optimal server speed policy, under a constraint on the fast server utilization. In this application, we investigate the effect of risk sensitive cost on optimal policy.

III. Power constrained Delay Tolerant Networks

We consider a large area with N active and freely moving nodes as in [2], where connection between any two nodes is not guaranteed all the times. The aim is to transfer a message from a static source to a static destination within the given deadline T, using the occasional contacts between the various moving elements. Source transfers the message to any node that comes in contact with it and the message is delivered to the destination if any one of the nodes with message comes in contact with it. One relay node cannot transfer the message to another and this is called the two-hop protocol (e.g., [14]). These networks are called Delay Tolerant Networks (DTNs).

Whenever a node arrives in the range of transmission of the source/destination, a contact occurs. In large areas with small transmission range, the contacts are rare. In such scenarios, the contact process can be modelled by a Poisson process ([14]), for a variety of mobility models like random walk, random waypoint etc. We further assume that the contact time, for any contact, is sufficient to transfer the entire message.

The source transfers the message to the contacted node. We refer these nodes as infected nodes. We consider a time-slotted system and let X_t represent the number of nodes infected at the beginning of time slot t. If λ is the rate of the source-node Poisson contact process, then the probability that any given node does not come in contact with the source is given by $e^{-\lambda}$ (with unit-duration time slots). Using this one can easily compute the transition probabilities (see [2] for more details):

$$\begin{aligned} p(x'|x,\lambda) &:= P(X_{t+1} = x'|X_t = x, \lambda) \qquad (23)\\ &= \begin{cases} \binom{N-x}{x'-x}(1-e^{-\lambda})^{x'-x} e^{-\lambda(N-x')} & \text{if } x \le x' \le N \\ 0 & \text{else} \end{cases}\end{aligned}$$

Basically the number infected increases by $(x'-x)$ if any $(x'-x)$ among the non-infected $(N-x)$ nodes contact the source and the above is the probability of precisely this event.

The contact rates depend upon the transmission range which in turn is proportional to the power used for transmission ([2]). And the probability of successful delivery depends upon the contact rates. However, the source mostly derives power from a battery or other such power constrained devices. Thus the source has to accomplish its goal utilizing the available power *leading to a hard power constrained problem.* If the source transmits using higher power, the contact range increases, which further increases the contact opportunities. However the power is consumed within a shorter time. On the other hand if it transmits with lower power, it can remain active for a longer period, but with smaller contact range. Thus there is an inherent trade-off between remaining active for longer duration and remaining active with larger contact range. One would like to use optimal power level in each time slot, and such a policy is provided by our risk sensitive MDP results as below.

A. Resource allocation policy

A policy represents the decision of power levels transmitted in each time slot and the aim is to obtain a power policy which maximizes the probability of successful delivery or equivalently minimizes the probability of delivery failure, for any given power constraint. In [2], it is shown that the contact rates are proportional to the power used and *hence we consider an equivalent policy in terms of (source) contact rates.* The system has M different choices of transmit powers that can be used in any time slot and let $\mathcal{A} = \{\lambda_0, ..., \lambda_M\}$ represent the corresponding set of source contact rates. Let A_t represent the contact rate chosen in time slot t and let $\Pi = \{\pi_0, \cdots, \pi_{T-1}\}$ represent a randomized policy. For each $0 \le t \le T-1$, π_t is a probability distribution over $\mathcal{A}$:

$$\pi_t(x,\lambda) = P(A_t = \lambda | X_t = x) \text{ for any } \lambda \in \mathcal{A} \text{ and } 0 \le x \le N.$$

B. Probability of failure under a policy

If there is a contact between the destination and an infected node within deadline T, then the message delivery is accomplished. Otherwise, delivery fails. Let ν represent the rate of destination-node contact Poisson process. The probability of failure $P_f(\Pi)$ for a given policy Π is derived in [2] and we briefly summarize the same over here. A failure event occurs, when none of the X_t infected nodes contact the destination in time slot t and if this is true for all the time slots. Probability of failure is calculated by conditioning on Markov chain trajectory $\{X_t\}_{t \le T}$ and is given by (see [2] for details):

$$P_f(\Pi) = E^{\alpha,\Pi}\left[e^{-\nu \sum_t X_t}\right]. \qquad (24)$$

In the above $E^{\alpha,\Pi}$ represents the expectation under policy Π (Markov transitions (23) are governed by policy Π) and when the initial condition X_0 is distributed according to α, $X_0 \sim \alpha$.

C. Total power spent of a policy

The contact rate λ is proportional to $p^{-\beta}$, where p is the transmitted power and β is a path loss factor (a constant) depends upon propagation characteristics of the area in which the nodes are operating (Appendix of [2]). In other words if one chooses rate λ, the power transmitted is proportional to λ^{β}. Without loss of generality, let the constant of proportionality be one. Thus the total (random) power spent over the T slots is given by:

$$\mathcal{P}(\Pi) = \sum_{t=0}^{T-1} A_t^{\beta}. \tag{25}$$

D. Power control problem

The problem is to minimize the probability of failure P_f, given a hard constraint B on the average total power, $E^{\alpha,\Pi}[\mathcal{P}]$, spent by the source:

$$\begin{aligned} &\min_{\Pi} E^{\alpha,\Pi}\left[e^{-\nu \sum_{t=0}^{T} X_t}\right] \\ \text{subject to:} \quad & E^{\alpha,\Pi}\left[\sum_{t=0}^{T-1} A_t^{\beta}\right] \le B. \end{aligned} \tag{26}$$

We refer this as hard constraint (HC) problem. Clearly the above has exactly the same form as that of a finite horizon constrained risk MDP problem specified by (9) (see also (3)) and one can solve it using Theorem 2. In [2] the authors optimized a joint cost, that depends both upon the probability of failure P_f and a term proportional to the total power spent, $e^{h\mathcal{P}}$:

$$\min_{\Pi} E^{\alpha,\Pi}\left[e^{-\nu \sum_{t=0}^{T} X_t + h \sum_{t=0}^{T-1} A_t^{\beta}}\right]. \tag{27}$$

In the above, h defines the weight factor given for total power term in the joint cost. We refer this as a soft constraint (SC) problem and compare the performance of SC optimal policy with that of the HC optimal policy. A direct solution (i.e., HC solution) would obviously perform better, when one considers a hard constraint on the power used. We compare the two solutions in section III-F (provided below) to determine the percentage of improvement obtained, when the two utilize the same power. In section III-F, we also compare the structural properties of the two policies.

E. LP based approach

The soft constraint (SC) problem (27) has the form of a risk sensitive MDP problem. The joint cost in (27), except for the logarithmic function, is similar to the risk sensitive cost $J(\alpha, \Pi)$ of (3), with running and terminal costs given by

$$r_t^{SC}(x, \lambda) = -\nu x + h\lambda^{\beta} \text{ and } r_T^{SC}(x) = -\nu x. \tag{28}$$

By monotonicity, the optimization of a cost is equivalent to optimizing the logarithm of the same cost. Thus Theorem 1 is applicable and one can also solve the SC problem by solving the dual LP given by (12) after substituting the running and terminal costs given by (28). The running and terminal costs corresponding to HC problem (26) are given by:

$$r_t^{HC}(x, \lambda) = -\nu x \text{ and } r_T^{HC}(x) = -\nu x, \tag{29}$$

while the constraint function

$$f_t^{HC}(\lambda) = \lambda^{\beta}. \tag{30}$$

Its optimal policy can be obtained by solving Dual LP given by (21) of Theorem 2.

F. Numerical analysis

We obtain the solution of SC and HC problems, by numerically solving the relevant LPs (12) and (21), with appropriate running and terminal costs. The constraint for HC problem is given by (30). For numerical computations we use Matlab and AMPL. We did most of the coding in Matlab except for Linear programming part. We used AMPL to model the LP and Gurobi solver to solve. The solution $\mathbf{y}^*$ of the LP provides the optimal policy $\Pi_{\mathbf{y}^*}$ as given by equation (19) of Theorem 1.

1) Verification of SC solution: SC problem (27) is analyzed in [2] and [2, Lemma 2] provides some structural properties of the optimal policy. The lemma establishes the existence of a switch off threshold n_{off} on the number infected. The optimal policy switches off (zero contact rate) the transmission, once the number infected reaches the threshold. It also showed that the contact rate chosen below the threshold is always nonzero. However this statement needs a small correction[3]. For every $x < n_{off}$, there exists a threshold T_x^* (depending upon x) such that

$$\lambda_t^*(x) \ge \lambda_1 \text{ for all } t < T_x^* \text{ and } \lambda_t^*(x) = 0 \text{ for all } t \ge T_x^*.$$

Thus the difference is that, beyond n_{off} it is always OFF as in [2]. However below n_{off}, the actual switch off time threshold depends upon the number infected, x.

We verify that our LP based solution satisfies the above structural properties. This also partially validates the LP based solutions that we derived.

We consider an example with $N = 15$, $\mathcal{X} = \{0, 1, \ldots 15\}$, $T = 20$, $h = 20$, $\nu = 0.1$, $\beta = 2.1$ and $\mathcal{A} = \{0, 0.1, 0.2, 0.3\}$. For this example, the $n_{off} = 13.411344$ as given by [2, Lemma 2]. The simulation results are following the structure given by the [2, Lemma 2], as seen from the Table I. For example for all $x \ge n_{off}$, $T_x^* = 0$ and for others it is nonzero. We have conducted few more examples and verified the same.

2) Comparison of SC-HC policies, Method 1: We compute the HC and SC optimal policies and compare them in various aspects. To begin with, we compare their performance when both utilize the same total expected power. We consider the following procedure for this comparison:

(i) We first solve the SC problem for a fixed value of h.

[3] In [2, page 9] in the proof of Lemma 2, the line after the sentence starting with "When $n < n_{off}$, $Q_n \lambda_1^{\beta} < \lambda_1$ ..." need not be true always. There can be scenarios in which $f_{T-1}(0) < f_{T-1}(\lambda_1)$. However the lines after that are correct. Hence for any $n < n_{off}$, if there exists a $t+1$ such that $\lambda_{t+1}^*(n) \ge \lambda_1$ then for all $\tau \le t$ $\lambda_{\tau}^*(n) \ge \lambda_1$. Thus we have the above modification with $T_n^* = t$, the first t for which $\lambda_{t+1}^*(n) \ge \lambda_1$.

States (s)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Threshold time (T_s^*)	19	19	19	19	19	19	18	18	18	17	17	15	13	4	0	0

TABLE I
VERIFICATION OF SC POLICY

(ii) We compute the total power spent by the SC optimal policy Π^*_{SC}, call it $\mathcal{P}^*_{SC}(h)$.

(iii) We then obtain the solution of HC problem (26) with bound B set to $\mathcal{P}^*_{SC}(h)$.

We use Lemma 1 of Section V (which provides the proof of Theorem 1) and the additional state component $\{\Psi_t\}$ of Section VI to compute the total power $\mathcal{P}^*_{SC}(h)$ of step (ii). More details are available in the relevant sections and the power used by SC optimal policy Π^*_{SC} can be computed using equation (37) of Section VI as below:

$$\mathcal{P}^*_{SC}(h) = \sum_t E[A_t^\beta] = \sum_t \sum_{x_t,a_t,\psi_t} y_{\Pi^*_{SC}}(t, x_t, \psi_t a_t)\ \ \psi_t\ \ a_t^\beta.$$

Note here that this procedure is only used for computing the power utilized under the already computed optimal policy Π^*_{SC}, and not for the purpose of constrained optimization. With this procedure we noticed that both the policies consume the same power, i.e., $\mathcal{P}^*_{SC} = \mathcal{P}^*_{HC}$. But there is a good improvement in the performance under HC policy (see Tables II, III). Table II shows the improvement in P_f for HC model over SC model, when both the models operate with equally likely initial conditions, i.e., $\alpha(x) = 1/(|N+1|)$ for $x \in \mathcal{X}$. In Table III we have shown the improvement in P_f for HC model over SC model, when the starting state is given (here $x = 0$), i.e., $\alpha(x) = 1_{\{x=0\}}$ for $x \in \mathcal{X}$. In the limited examples that we conducted, we saw an improvement as high as 26%. In all these examples we set $M = 1$, resulting in a ON-OFF control.

Thus, when both the optimal policies utilize the same average total power, the HC solution performs better. This is obvious because it directly solves the constrained problem. However the more interesting observation is that the improvement can be very significant.

3) Comparison of SC-HC policies, Method 2: We would now like to compare the two policies using a very different perspective. Say we are given any arbitrary total average power constraint B. The requirement is an optimal policy that operates within this power constraint and which minimizes the failure probability P_f, precisely the HC problem. But if one approaches this via SC problem, then one needs to solve the SC problem for various values of weight factors h to obtain various $\{P_f^{SC}(h)\}_h$ and the corresponding total average powers $\{\mathcal{P}^*_{SC}(h)\}_h$. Consider only those h for which total average power is less than given threshold, i.e., $\mathcal{P}^*_{SC}(h) \leq B$. Among these chose the best failure probability $P_f^{SC}(h^*)$ as the solution. That is, one needs to continue the search among SC policies, until they hit upon that value of h for which the total average power is the maximum possible one, which is still below the given limit B.

The SC solutions are supposed to be pure policies i.e., π_t =1 or 0, for all t, which means that the system is either switched 'on' or 'off' respectively. We observe this to be the case in simulations. With pure policies, the various choices of total average power would be discrete. One can have various SC solutions by considering various values of weight factors h. However the set of all possible total average powers obtained even after exhausting the entire range of h, would be finite. On the other hand HC solution is a randomized policy and achieves the bound B with equality, as long as it is feasible.

Thus the improvement seen by directly using our LP based HC solution would be much more significant than that demonstrated in Tables II and III. This effect is shown in Figure 1. This figure plots best P_f performance versus power constraint B, under both HC and SC policies. The curve with dotted marks, represents the best performance facilitated by SC solution, as a function of the power constraint B, obtained by trying all possible values of h. As seen from the figure, the P_f^{SC} performance remains constant over a range of power constraints B, confirming our earlier discussions. This is mainly because the SC policies are pure. The other curve in Figure 1 represents the P_f^{HC} performance under HC policy as a function of power constraint B. The entries of the Tables II and III correspond to the SC and HC pair of points, where SC points are precisely the corner points of the SC curve that are near the HC curve. These entries already showed an improvement (up to 26%), and we have a much larger gains in the performance at the other points (see the horizontal portions of the SC curve in Figure 1).

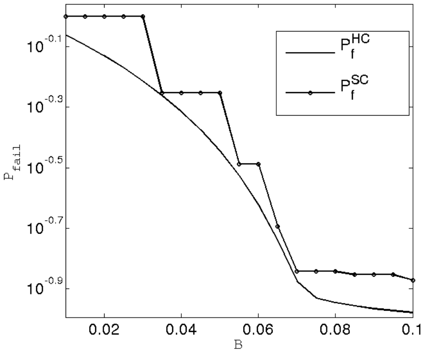

Fig. 1. P_f performance as a function of power bound B

4) Structural properties: We noticed from various examples of the simulations that the SC policies are all pure policies. While the HC policies are randomized. Further, for any given time slot t the policy suggests complete switch ON for all states less than a threshold x_{th}, a randomized switch ON-OFF at the threshold state $x = x_{th}$ and a complete switch OFF for all states, $x > x_{th}$. This threshold depends upon the time slot and of course, the power constraint B.

IV. QUEUEING WITH LOSSES

We consider a queueing system with two possible service modes. 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

T, N	h	ν	β	λ_1	$\mathcal{P}$	P_f^{SC}	P_f^{HC}	% Imp
5, 3	10.2	0.70	2.1	0.20	0.03	0.04	0.03	26.72
5, 3	10.2	0.60	2.1	0.20	0.34	0.04	0.03	21.85
6, 3	8	0.50	2.1	0.20	0.05	0.03	0.02	16.71
6, 4	8	0.30	2.1	0.20	0.05	0.03	0.02	20.22

TABLE II
IMPROVEMENT IN P_f: HC VERSUS SC, WITH EQUALLY LIKELY INITIAL CONDITIONS.

T, N	h	ν	β	λ_1	$\mathcal{P}$	P_f^{SC}	P_f^{HC}	% Imp
5, 3	10	0.50	2.1	0.20	0.08	0.19	0.17	11.94
5, 3	10.2	0.70	2.1	0.20	0.08	0.14	0.11	21.59
6, 3	10	0.50	2.1	0.20	0.10	0.11	0.10	16.74
6, 3	8	0.50	2.1	0.20	0.11	0.10	0.09	15.33

TABLE III
IMPROVEMENT IN P_f: HC VERSUS SC STARTING WITH ZERO INFECTED NODES.

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 expected 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

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

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

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

We have the following simplification with proof in Appendix:

Theorem 3: The required risk sensitive cost has a simpler form as below:

$$\begin{aligned} J(x, \Pi) &= E^{x,\Pi}\left[e^{\beta \sum_{t=0}^{T} 1_{\{X_t=N\}}}\right] \text{ with} \\ \beta &= \ln\left(\delta e^{\gamma} + (1-\delta)\right). \end{aligned} \qquad (31)$$

■

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

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

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.

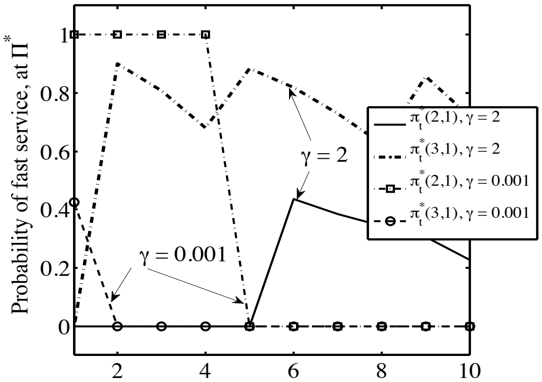

Fig. 2. Risk neutral and risk sensitive optimal policies.

A. Numerical analysis

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

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

by solving the corresponding LP (21), where β depends upon the risk parameter γ as given by Theorem 3. As in previous application, we used Matlab and 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 (19).

In Figure 2, we consider a system with 3 servers. We consider an example with $T = 10$, $N = 3$, $\mu_1 = 0.3$, $\mu_0 = 0.1$ and $B = 2.5$. 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. For both the 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 the time slots. But there is a difference for the remaining two states $x = 2$ or 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 (time) 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 $\gamma = 0.001$.

With $\gamma = 2$, the risk optimal policies are no more threshold. In fact they are not even monotone as seen from the Figure 2. 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 more importance to risk cost, the policy is more cautious at the points of high risk, i.e., when $x = 3$. We estimate the expected 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 expected 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 expected number lost.

We considered some more numerical examples and found similar characterization of the optimal policies (i.e. for smaller value of γ we have threshold optimal policy and for larger value of γ neither the optimal policy is threshold type nor is it monotone).

V. Proof of Theorem 1

Approach

We will study the DP equations (6)-(7) and discuss their relation with the two LPs (11) and (12), to obtain the proof of Theorem 1. This connection is well understand in the context of Linear MDPs (e.g., [6], [1]). We follow a similar approach as in ([6]), but the details are sufficiently different because of the multiplicative nature of risk cost. We obtain this proof using the following four steps:

Step 1: The DP equations (7) are rewritten as an appropriate non-linear transformation (equation (32) given below), such that the fixed point of the later provides the optimal risk MDP cost, the translated value function $\underline{\mathbf{u}}^*$ given by equation (8). The non-linear transformation is converted to an appropriate linear operator with additional constraints. This gives us the primal LP (11) and the proof of part (a) of Theorem 1.

The dual of primal LP (11) equals the second LP, (12). We obtain the remaining two parts of the Theorem 1 by showing that the solution of the dual LP directly provides the MDP optimal policy. This is achieved in the following three steps.

Step 2: We provide a transformation, which transforms every feasible vector of dual LP to an MDP policy and vice versa.

Step 3: We show the equivalence between expected risk cost and dual objective, using Step 2.

Step 4: In the final step we will show the relation between optimal policy and the dual solution.

Step 1: Primal LP (11) and translated value fucntion (8)

Define the following vector to represent all the components of the required value function:

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

Consider the following operator (see terminal cost (6))

$$\begin{aligned} \mathcal{L}\underline{\mathbf{u}} &= [L_0\underline{\mathbf{u}}, L_1\underline{\mathbf{u}}, \cdots, L_{T-1}\underline{\mathbf{u}}] \text{ where,} \\ L_t\underline{\mathbf{u}} &:= \inf_{\pi_t} \sum_a C_{t,a}^{\pi_t} P_a \mathbf{u}_{t+1} \text{ for } t < T \text{ with} \\ u_T(x) &:= e^{r_T(x)} \text{ for all } x \in \mathcal{X}, \end{aligned} \tag{32}$$

and where the matrices $C_{t,a}^{\pi_t}, P_a$ are defined as below:

$$C_{t,a}^{\pi_t} = \begin{bmatrix} \pi_t(1,a)e^{r_t(1,a)} & 0 & \cdots & 0 \\ 0 & \pi_t(2,a)e^{r_t(2,a)} & \cdots & 0 \\ . & . & \cdots & 0 \\ . & . & \cdots & 0 \\ . & . & \cdots & 0 \\ 0 & . & \cdots & \pi_t(N,a)e^{r_t(N,a)} \end{bmatrix}$$

$$P_a = \begin{bmatrix} p(1|1,a) & p(2|1,a) & \cdots & p(N|1,a) \\ p(1|2,a) & p(2|2,a) & \cdots & p(N|2,a) \\ . & . & \cdots & . \\ . & . & \cdots & . \\ . & . & \cdots & . \\ p(1|N,a) & p(2|N,a) & \cdots & p(N|N,a) \end{bmatrix}.$$

In the above, infimum is defined as component wise. The above operator consolidates the right hand side (RHS) of the DP equations (7) and satisfies the following properties (proof in Appendix):

Theorem 4: (i) Any vector $\underline{\mathbf{u}}$ with $\mathcal{L}\underline{\mathbf{u}} \ge \underline{\mathbf{u}}$ (the component wise inequality), satisfies: $\underline{\mathbf{u}}^* \ge \underline{\mathbf{u}}$, where $\underline{\mathbf{u}}^*$ is given by (8).

(ii) Any vector $\underline{\mathbf{u}}$ with $\mathcal{L}\underline{\mathbf{u}} \le \underline{\mathbf{u}}$, satisfies: $\underline{\mathbf{u}}^* \le \underline{\mathbf{u}}$. ■

Thus if a vector $\underline{\mathbf{u}}$ satisfies, $\mathcal{L}\underline{\mathbf{u}} \ge \underline{\mathbf{u}}$, i.e., if

$$\mathbf{u}_t \le \sum_a C_{t,a}^{\pi_t} P_a \mathbf{u}_{t+1} \text{ for all } t \text{ and } \pi_t \in \Pi, \tag{33}$$

then it is (component-wise) lower than the value function of risk MDP, $\underline{\mathbf{u}}^*$. It is trivial to check that $\underline{\mathbf{u}}^*$ also satisfies (33), and with equality. Thus it is the greatest lower bound (component-wise) among all vectors that satisfy (33). It is easy to verify that the constraints posed by (33) are same as the constraints of primal LP (11). Following similar procedure as in ([6]), we chose a non-negative vector $\alpha(x), x \in \mathcal{X}$ that satisfies $\sum_{x \in \mathcal{X}} \alpha(x) = 1$ and obtain the objective function of LP (11). Thus the solution of the primal LP (11) provides $\underline{\mathbf{u}}^*$ of equation (8), establishing the part (a). The vector α can be interpreted as the distribution of initial state, X_0.

Step 2: Dual Feasible Region and Set of MDP Policies

We say that a vector $\mathbf{y}$ is feasible if it satisfies the dual constraints (13)-(14) and let $\mathcal{F}$ represent this feasible region. The one to one correspondence between the two spaces, $\mathcal{F}$ and $\mathcal{D}$ is established by the following (proof in Appendix):

Theorem 5: (i) For any $\Pi \in \mathcal{D}$ vector $\mathbf{y}_{\mathbf{\Pi}}$, defined by (20) in the hypothesis of Theorem 1, satisfies constraints (13)-(14), and so $\mathbf{y}_{\mathbf{\Pi}} \in \mathcal{F}$.

(ii) For any vector $\mathbf{y} \in \mathcal{F}$, the policy $\Pi_{\mathbf{y}}$ defined by (19), satisfies:

$$\mathbf{y}_{\Pi_{\mathbf{y}}} = \mathbf{y},$$

where $\mathbf{y}_{\Pi_{\mathbf{y}}}$ is defined by (20) for the policy $\Pi_{\mathbf{y}}$. ■

Step 3: Expected Risk Cost and Dual Objective Function

We now study the connection between the risk sensitive cost for any policy Π and the dual objective function at the translated feasible point $\mathbf{y}_\Pi$. Towards this, we require the expression for the expected value of any given function f, in terms of the dual variable $\mathbf{y} \in \mathcal{F}$ and that is provided below with the proof in Appendix.

Lemma 1: (i) Let $X_0 \sim \alpha$. For any $\mathbf{y} \in \mathcal{F}$ with corresponding policy $\Pi_\mathbf{y}$, integrable function f and $t < T$:

$$\sum_{x,a} y(t,x,a) f(x,a) \;=\; E^{\alpha,\Pi_\mathbf{y}}\left[\Psi_t^{-1} f(X_t,A_t)\right] \text{ with}$$

$$\Psi_t \;:=\; e^{-\sum_{n=0}^{t-1} r_n(X_n,A_n)}. \tag{34}$$

(ii) For any integrable function f of the last two states x_{T-1}, x_T and the final action a_{T-1}, we have:

$$\sum_{x,a,x'} y(T-1,x,a) p(x'|x,a) f(x,a,x') \tag{35}$$

$$= E^{\alpha,\Pi_\mathbf{y}}\left[\Psi_{T-1}^{-1}\, f(X_{T-1},A_{T-1},X_T)\right]. \quad \blacksquare$$

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

Lemma 2: For any policy $\Pi \in \mathcal{D}$ with corresponding feasible vector $\mathbf{y}_\Pi$ and integrable function f, we have:

$$E^{\alpha,\Pi}\left[\Psi_{T-1}^{-1} f(X_{T-1},A_{T-1},X_T)\right] \tag{36}$$

$$= \sum_{x,a,x'} y_\Pi(T-1,x,a) p(x'|x,a) f(x,a,x'). \quad \blacksquare$$

Step 4: Optimal Policies and the Dual Solutions

Let $g(\cdot)$ represent the dual objective (12). By equation (35) of Lemma 1 with $f(x,a,x') = e^{r_{T-1}(x,a)} e^{r_T(x')}$, we have:

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

Let $\mathbf{y}^*$ be an optimal solution of the dual LP, and let $\Pi_{\mathbf{y}^*}$ be the corresponding policy given by (19). For any $\Pi \in \mathcal{D}$, using equation (36) and by optimality of $\mathbf{y}^*$:

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

establishing the optimality required in part (b) of Theorem 1. Part $(\mathbf{c})$ can be proved using similar logic. $\blacksquare$

VI. Proof of Theorem 2 - Constrained risk MDP

We follow similar steps as in the proof of Theorem 1. The Step 2 which establish the equivalence of the two spaces and Step 3 that maps the objective function to risk cost, are exactly the same. The final step, Step 4, needs to be proved while considering the extra constraint (10). One needs to add this extra constraint to the LP. Towards this, the constraint need to be converted to an appropriate linear constraint.

Equation (34) of Lemma 1 could have been useful in obtaining the expectation defining the constraint, but for the extra factor Ψ_t^{-1} of its RHS. We add Ψ_t as an 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:

$$\sum_{x,\psi_t,a} y(t,x,\psi_t,a)\, \psi_t\, f(x,a) = E^{\alpha,\Pi_\mathbf{y}}\left[f(X_t,A_t)\right]. \tag{37}$$

Thus, the LHS of the above equation is added as an extra constraint (22) to the dual LP (12) to obtain the LP (21) for the constrained risk MDP. Now following similar steps as in Step 4 of proof of Theorem 1, one can complete the proof. $\blacksquare$

Remarks: 1) 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.

2) Equation (37) of section VI can also be used to compute any importance performance measure under the optimal policy. For example, we used this equation to compute the total expected power consumed under optimal policy in Section III, while working with DTN based application. We also used it to compute the average number of customers lost, over all time periods, in Queuing application.

VII. Conclusions and Future Directions

We consider a finite horizon risk MDP problem and establish the connections between the DP and LP approaches. We show that the solution of unconstrained risk MDP problem can be obtained via the solution of any one of the two LPs, a primal and a dual. The primal solution provides value function while the dual solution directly provides 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 found to study Delay Tolerant Networks (DTNs) where there is constraint on power use. We obtained optimal policies for this problem, via the solution of an appropriate LP. Previously, in [2], a joint cost comprising of probability of delivery failure and a term proportional to total power spent is considered. This work is nearest to the problem of [2]. We compared the probability of failure performance of the DTNs under the policy so obtained, with that of the optimal policy obtained in [2]. We observed huge improvement, in scenario with hard power constraint.

We also 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.

Currently finite horizon risk sensitive MDP and LP relation has been studied but this problem can be extended for infinite horizon MDPs also. Further we would like to consider a case where some components of the state need not be finite. Under certain assumptions, we have ideas to extend the LP based approach to this case and this can be explored in future.

References

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

[2] Altman, Eitan, Veeraruna Kavitha, Francesco De Pellegrini, Vijay Kamble, and Vivek Borkar. "Risk sensitive optimal control framework applied to delay tolerant networks." In INFOCOM, 2011 Proceedings IEEE, pp. 3146-3154. IEEE, 2011.

[3] Bertsekas, Dimitri P. "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] Coraluppi, Stefano P., and Steven I. Marcus. "Risk-sensitive queueing." Proceedings of the Annual Allerton Conference on Communication Control and Computing. Vol. 35. University of Illinois, 1997.

[6] Puterman, Martin L. "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] 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.

[10] Altman, Eitan, Tamer Baar, and Francesco De Pellegrini. "Optimal monotone forwarding policies in delay tolerant mobile ad-hoc networks." Performance Evaluation 67.4 (2010): 299-317.

[11] Nagengast, Arne J., Daniel A. Braun, and Daniel M. Wolpert. "Risk-sensitivity and the mean-variance trade-off: decision making in sensorimotor control." Proceedings of the Royal Society of London B: Biological Sciences (2011): rspb20102518.

[12] Littman, Michael L., Thomas L. Dean, and Leslie Pack Kaelbling. "On the complexity of solving Markov decision problems." Proceedings of the Eleventh conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1995.

[13] Papadimitriou, Christos H., and John N. Tsitsiklis. "The complexity of Markov decision processes." Mathematics of operations research 12.3 (1987): 441-450.

[14] Groenevelt, Robin, Philippe Nain, and Ger Koole. "Message delay in manet." ACM SIGMETRICS Performance Evaluation Review. Vol. 33. No. 1. ACM, 2005.

[15] Anantharam, Venkatachalam, and Vivek Shripad Borkar. "A variational formula for risk-sensitive reward." arXiv preprint arXiv:1501.00676 (2015).

[16] Kumar, Atul, Veeraruna Kavitha, and N. Hemachandra. "Finite horizon risk sensitive MDP and linear programming." 2015 54th IEEE Conference on Decision and Control (CDC). IEEE, 2015.

[17] Kumar, Atul, Veeraruna Kavitha, and N. Hemachandra. "Power constrained DTNs: Risk MDP-LP approach." Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2015 13th International Symposium on. IEEE, 2015.

VIII. Appendix

Proof of Theorem 3: 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 $e^\beta = g_\delta$. ■

Proof of Theorem 4: We begin with the proof of Part (i). 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 operator $\mathcal{L}$,

$$\begin{aligned} \mathbf{u}_0 \quad &\le \inf_{\pi_0} \left\{ \sum_{a_0} C_{0,a_0}^{\pi_0} P_{a_0} \mathbf{u_1} \right\} &(38)\\ &\le \sum_{a_0} C_{0,a_0}^{\pi_0'} P_{a_0} \mathbf{u_1} \\ &\le \sum_{a_0} C_{0,a_0}^{\pi_0'} P_{a_0} \sum_{a_1} C_{1,a_1}^{\pi_1'} P_{a_1} \mathbf{u_2} \\ &\vdots \\ &\le \sum_{\mathbf{a}_0^{T-1}} C_{0,a_0}^{\pi_0'} P_{a_0} C_{1,a_1}^{\pi_1'} P_{a_1} \cdots C_{T-1,a_{T-1}}^{\pi_{T-1}'} P_{a_{T-1}} \mathbf{u_T} \\ &= \mathbf{J}_0(\Pi') \text{ with } \mathbf{J}_0(\Pi') := \begin{bmatrix} J_0(1,\Pi') \\ J_0(2,\Pi') \\ \vdots \\ J_0(N,\Pi') \end{bmatrix}. &(39) \end{aligned}$$

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

$$\mathbf{u}_t \le \mathbf{u}_t^* \text{ and hence } \underline{\mathbf{u}} \le \underline{\mathbf{u}}^*.$$

Part (ii): Let us consider any vector $\underline{\mathbf{u}}$ which satisfies $\underline{\mathbf{u}} \ge \mathcal{L}\underline{\mathbf{u}}$. By definition of $\mathcal{L}$ the vector satisfies for all $t < T$:

$$\mathbf{u}_t \ge \inf_{\pi_t} \left\{ \sum_{a_t} C_{t,a_t}^{\pi_t} P_{a_t} \mathbf{u_{t+1}} \right\}. \tag{40}$$

Consider any $\epsilon_0 > 0$, by definition of infimum (with $t = 0$) there exists a π_0 such that:

$$\mathbf{u}_0 \quad \ge \sum_{a_0} C_{0,a_0}^{\pi_0} P_{a_0} \mathbf{u_1} - \epsilon_0 \mathbf{e}, \text{ where } \mathbf{e} := (1, 1, \cdots, 1).$$

For any given $\epsilon_1 > 0$, now using (40) two times ($t = 0$ and 1), one can chose ϵ_1', ϵ_0' and π_1', π_0' such that:

$$\mathbf{u}_1 \quad \geq \sum_{a_1} C_{1,a_1}^{\pi_1'} P_{a_1} \mathbf{u_2} - \epsilon_1' \mathbf{e} \text{ and hence that}$$

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

where by boundedness of the matrices (finite states and actions) involved one can ensure:

$$\sum_{a_0} C_{0,a_0}^{\pi_0'} P_{a_0} \epsilon_1' + \epsilon_0' \mathbf{e} \leq |C_{0,a_0}| \, |P_{a_0}| \, \epsilon_1' \mathbf{e} + \epsilon_0' \mathbf{e} \leq \epsilon_1 \mathbf{e}.$$

In the above $|.|$ represents the norm and note that the matrix norm, $\left| C_{0,a_0}^{\pi_0'} \right|$, can be bounded uniformly of π_0'.

Recursively and using similar logic, for any $t < T$ and for any given $\epsilon_t > 0$, one can chose $\{\epsilon_n'\}_{n \leq t}$ and $\{\pi_n'\}_{n \leq t}$ (if required redefine previous $\{\pi_n'\}$, $\{\epsilon_n'\}$) such that:

$$\mathbf{u}_0 \quad \geq \quad \sum_{\mathbf{a}_0^t} C_{0,a_0}^{\pi_0'} P_{a_0} C_{1,a_1}^{\pi_1'} P_{a_1} \cdots C_{i,a_t}^{\pi_t'} P_{a_t} \mathbf{u_{t+1}} - \epsilon_t \mathbf{e},$$

where again the boundedness ensures:

$$\sum_{\mathbf{a}_0^{t-1}} C_{0,a_0}^{\pi_0'} P_{a_0} C_{1,a_1}^{\pi_1'} P_{a_1} \cdots C_{t-1,a_{t-1}}^{\pi_{t-1}'} P_{a_{t-1}} \epsilon_t' - \epsilon_0' \mathbf{e} \leq \epsilon_t \mathbf{e}.$$

In particular for any $\epsilon = \epsilon_T$, we obtain a policy $\Pi' = [\pi_0', \pi_1', \cdots, \pi_{T-1}']$ such that:

$$\mathbf{u_0} \quad \geq \sum_{\mathbf{a}_0^{T-1}} C_{0,a_0}^{\pi_0'} P_{a_0} C_{1,a_1}^{\pi_1'} P_{a_1} \cdots C_{T-1,a_{T-1}}^{\pi_{T-1}'} P_{a_{T-1}} \mathbf{u_T} - \epsilon \mathbf{e}.$$

Thus for any $\epsilon > 0$ there exists a policy Π' such that

$$\begin{aligned} \mathbf{u_0} \quad &\geq \quad \mathbf{J}_0(\Pi') - \epsilon \mathbf{e} \text{ and thus,} \\ \mathbf{u}_0 \quad &\geq \quad \mathbf{J}_0(\Pi') - \epsilon \mathbf{e} \geq \inf_{\Pi} \mathbf{J}_0(\Pi) - \epsilon \mathbf{e} = \mathbf{u}_0^* - \epsilon \mathbf{e}. \end{aligned}$$

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

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

Following exactly similar logic one can show that

$$\mathbf{u}_t \geq \mathbf{u}_t^* \text{ for all } 1 \leq t < T. \qquad \blacksquare$$

Proof of Theorem 5: We start with the proof of Part (i). It is easy to see that the defined point $\mathbf{y}_\Pi$ satisfies the first constraint (13), 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 notation, we represent the action state pair by $z_t := (x_t, a_t)$, for every t. Considering RHS of the second constraint (14):

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

Hence the vector $\mathbf{y}_\Pi$ satisfies the second constraint (14).

Part (ii): Consider any $\mathbf{y} \in \mathcal{F}$, the policy $\Pi_\mathbf{y}$ defined as in (19) and then the point $\mathbf{y}_{\Pi_\mathbf{y}}$ defined using policy $\Pi_\mathbf{y}$ as in (20). Aim is to prove that

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

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 (13) and by definition (19) of $\Pi_\mathbf{y}$ we have:

$$\begin{aligned} y_{\Pi_\mathbf{y}}(t, z_t) &= \sum_{\mathbf{z}_0^{t-1}} \alpha(x_0) \Delta_\Pi^{0,t} \\ &= \sum_{\mathbf{z}_0^{t-1}} \Delta_\Pi^{1,t} \alpha(x_0) e^{r_0(z_0)} \pi_{\mathbf{y},0}(z_0) \\ &= \sum_{\mathbf{z}_0^{t-1}} \Delta_\Pi^{1,t} \alpha(x_0) e^{r_0(z_0)} \frac{y(0, z_0)}{\sum_{a_0'} y(0, x_0, a_0')}; \quad \text{using (19)} \\ &= \sum_{\mathbf{z}_0^{t-1}} \Delta_\Pi^{1,t} e^{r_0(z_0)} y(0, z_0); \ \text{using (13).} \end{aligned}$$

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:

$$\begin{aligned} &y_{\Pi_\mathbf{y}}(t, z_t) \\ &= \sum_{\mathbf{z}_1^{t-1}} \Delta_\Pi^{2,t} e^{r_1(z_1)} \sum_{z_0} e^{r_0(z_0)} q_1^\Pi(z_1 | z_0) y(0, z_0) \\ &= \sum_{\mathbf{z}_1^{t-1}} \Delta_\Pi^{2,t} e^{r_1(z_1)} \pi_{\mathbf{y},1}(z_1) \sum_{z_0} e^{r_0(z_0)} p(x_1 | z_0) y(0, z_0) \\ &= \sum_{\mathbf{z}_1^{t-1}} \Delta_\Pi^{2,t} e^{r_1(z_1)} \pi_{\mathbf{y},1}(z_1) \sum_{a_1'} y(1, x_1, a_1'); \quad \text{using (14)} \\ &= \sum_{\mathbf{z}_1^{t-1}} \Delta_\Pi^{2,t} e^{r_1(z_1)} \frac{y(1, z_1)}{\sum_{a_1'} y(1, x_1, a_1')} \sum_{a_1'} y(1, x_1, a_1'); \ \text{using (19)} \\ &= \sum_{\mathbf{z}_1^{t-1}} \Delta_\Pi^{2,t} e^{r_1(z_1)} y(1, z_1). \end{aligned}$$

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

$$\begin{aligned}
y_{\Pi_{\mathbf{y}}}(t, z_t) = & \sum_{\mathbf{z}_2^{t-1}} \Delta_{\Pi}^{3,t} e^{r_2(z_2)} \sum_{z_1} e^{r_1(z_1)} q_2^{\Pi}(z_2|z_1) y(1, z_1) \\
= & \sum_{\mathbf{z}_2^{t-1}} \Delta_{\Pi}^{3,t} e^{r_2(z_2)} \pi_{\mathbf{y},2}(z_2) \sum_{z_1} e^{r_1(z_1)} p(x_2|z_1) y(1, z_1) \\
= & \sum_{\mathbf{z}_2^{t-1}} \Delta_{\Pi}^{3,t} e^{r_2(z_2)} \pi_{\mathbf{y},2}(z_2) \sum_{a_2'} y(2, x_2, a_2'); \text{ using (14)} \\
= & \sum_{\mathbf{z}_2^{t-1}} \Delta_{\Pi}^{3,t} e^{r_2(z_2)} \frac{y(2, z_2)}{\sum_{a_2'} y(2, x_2, a_2')} \sum_{a_2'} y(2, x_2, a_2') \\
= & \sum_{\mathbf{z}_2^{t-1}} \Delta_{\Pi}^{3,t} e^{r_2(z_2)} y(2, z_2).
\end{aligned}$$

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

$$\begin{aligned}
y_{\Pi_{\mathbf{y}}}(t, z_t) &= \sum_{z_{t-1}} q_t^{\Pi}(z_t|z_{t-1}) e^{r_{t-1}(z_{t-1})} y(t-1, z_{t-1}) \\
&= \pi_t(z_t) \sum_{z_{t-1}} e^{r_{t-1}(z_{t-1})} p(x_t|z_{t-1}) y(t-1, z_{t-1}) \\
&= \frac{y(t, z_t)}{\sum_{a_t'} y(t, x_t, a_t')} \sum_{a_t'} y(t, x_t, a_t'); \text{ using (14)} \\
&= y(t, z_t).
\end{aligned}$$

This is true for all $t \leq T-1$, z_t. ■

Proof of Lemma 1: By Theorem 5, $y(t, z_t) = y_{\Pi_X}(t, z_t)$ for any $t < T$. Further, using the definition of $y_{\Pi_X}(t, z_t)$ from (14), one can rewrite left hand side (LHS) of (34)

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

One can get the second result (35) in exactly similar lines. ■