Power Constrained DTNs: Risk MDP-LP Approach

Atul Kumar
atulkr.in@gmail.com
IEOR, IIT Bombay, India

Veeraruna Kavitha
vkavitha@iitb.ac.in,
IEOR, IIT Bombay, India

N Hemachandra
nh@iitb.ac.in,
IEOR, IIT Bombay, India.

***Abstract*—Delay Tolerant Networks (DTNs) have gained importance in the recent past, as cost-effective alternative, in scenarios where delays can be accommodated. They work well in discretely connected network, where there is no direct connectivity between some/all components of the system. But the mobility of nodes creates occasional contact opportunities. The randomly moving nodes cooperate to help a fixed source, in delivering message to a far away destination within the given time threshold. The objective is to optimize the delivery success probability, which turns out to be a risk sensitive cost. The success probability depends upon the contact rates, which in turn depend upon the power used by the nodes to remain visible. The more the power used by a node, the larger is the radius for which it is visible. However these nodes are power constrained. This leads to a constrained finite horizon, Risk sensitive Markov Decision Process (MDP). In this paper we propose a linear program (LP) based approach to solve the corresponding dynamic programming equations. This approach enables us in handling the constraints. We showed using numerical simulations that, given a hard power constraint, the solution of the constrained MDP performs significantly superior in comparison with a solution obtained by optimizing a joint cost.**

DTNs, Power allocation, Risk sensitive cost, Linear programs, Dynamic programming.

I. Introduction

We consider a large area with N active and freely moving mobiles as in [1], wherein connectivity between different devices is available only occasionally. The aim is to transfer a message from a static source to a fixed destination within the prescribed deadline, using the occasional contacts between the various moving elements. A source transfers the message to any mobile that comes in contact with it and the message is delivered to the destination if any one of the mobiles with message come in contact with it. One relay mobile cannot transfer the message to another and this is called the two-hop protocol (e.g., [2]). And these networks are called Delay Tolerant Networks (DTNs). Alternatively DTNs can operate using full epidemics, i.e, the relays can transfer the message to any other relay (see for example [5]). The message transmission performance with full epidemics is much better; however, in terms of the power consumed it is inferior. Further there can be flooding of messages across the network.

DTNs are operated in various configurations and using different protocols and there is vast literature analyzing these networks (e.g., [1], [2], [5], etc, and references therein). One of the common techniques to analyze these networks is using mean-field dynamics (e.g., [5]). This approach is valid in scenarios with large population. There are also papers that consider the random system dynamics which is more accurate with limited population (e.g., [1]). We also consider the random dynamics.

An element interested in making a contact transmits beacons (short pulses) regularly and a contact is established with a mobile if the later receives one such beacon. The range of visibility is proportional to the power transmitted. Thus the more the power used, the better are contact opportunities and the better is the probability of successful message transmission. However, the mobiles are power constrained and the main aim of this paper is to maximize the probability of successful message delivery, under the given power constraints.

The contact process is modeled by a Poisson process ([2]), hence the probability of success or equivalently probability of delivery failure includes terms composed of powers of exponent, resulting in a Risk sensitive Markov Decision Process (MDP) cost. Previously in DTN related literature, such costs are handled by exchanging the expected value and the exponent using Jensen's inequality and the solution is obtained by optimizing a bound on the objective function (e.g., [3] etc.,)

Recently in [1] authors solved the direct problem, using risk MDP approach. However, they solved the problem using soft constraints: a joint cost composed of successful delivery probability and the power transmitted is proposed, and is optimized. While in this paper we consider the control problem with hard constraints on the power.

In [4] we showed that the solution of a risk MDP problem can be obtained by solving a corresponding Linear Program (LP). We obtained the solution to the power constrained DTNs by solving the LPs provided in the technical report [4].

When one is interested in a hard power control problem, the solution obtained using a joint cost of [1] is obviously inferior to our direct solution of the constrained problem. More interestingly we noticed a huge improvement in the performance, because a randomized policy optimizes the hard problem while the soft problem (joint cost problem) is optimized by a pure policy. Thus our newly proposed LP based solutions are very useful in the context of hard power constraints.

II. System model and Problem definition

A static source has to transfer a message to a static destination within the given deadline T, and they are sufficiently far away to have any direct communication. The area surrounding the two has N cooperative and moving nodes (mobiles) that assist the source to deliver the message. The source can transmit only to those mobiles that arrive in its range of

978-3-9018-8274-6/15/ ©2015 IFIP

transmission. Similarly a destination can receive information from only those mobiles, that arrive within the range of transmission. And the range depends upon the power used for transmission. We say a contact occurred whenever a mobile comes in the communication range of the source/destination. In large areas with small transmission range, the contacts are rare. In such scenarios, the contact process can be modeled by a Poisson process [2], for a variety of mobility models like random walk, random waypoint, etc. We assume that the contact time is sufficient to transfer the message.

The source transfers the message to the contacted mobile. We refer these mobiles as infected mobiles. If there is a contact between the destination and an infected mobile within deadline T, then the message delivery is accomplished. Otherwise, delivery is failed.

The probability of successful delivery depends upon the power used by the source. The source derives power from a battery and hence is power constrained. In fact, the source is provided with a fixed amount of power and it has to accomplish its goal utilizing the available power *leading to a hard power constraint.* The source spends power for two purposes:

1) for transmitting beacons, to show its presence;
2) to transfer message to the contacted mobiles.

The power spent per transmitting a message could be significantly larger than the power spent for beaconing. However beaconing needs to be done at regular instances while the contacts are rare (in a time frame of few minutes one can at maximum make one contact), making the second component negligible.

If the source transmits beacons with 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 beacons with lower power, it can remain active for a longer period, but with smaller contact range. Thus there is an inherent tradeoff between remaining active for longer duration and remaining active with larger contact range. Mobiles contacted during the earlier stages have better chances of delivery. Hence it might be advantageous to consider varying powers for transmitting beacons, across the entire delivery duration.

A. Resource allocation policy

We consider a time slotted system, and beacons are transmitted with constant power in one time slot. Without loss of generality we assume unit time slots. A policy represents the decisions of power levels transmitted in each time slot and the aim of our paper is to obtain a power policy which maximizes the probability of successful delivery or equivalently minimizes the probability of delivery failure.

The rate of source-mobile contact Poisson process is represented by λ while that of the destination-mobile is given by ν. In [1], 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 $\Lambda = \{\lambda_0, ..., \lambda_M\}$ represent the corresponding set of all possible source contact rates. Let Y_t represent the contact rate chosen in time slot t. Vector $\Pi = \{\pi_1, \cdots, \pi_{T-1}\}$ represents a randomized policy, where π_t for each t, is a probability distribution over Λ:

$$\pi_t(\lambda) = Prob(Y_t = \lambda) \text{ for any } \lambda \in \Lambda.$$

B. Probability of failure given a policy

The probability of failure $P_f(\Pi)$ for a given policy Π is derived in [1] and we briefly summarize the same here. Let X_t be the number of mobiles infected at the beginning of time slot t. The sequence X_t is a controlled Markov chain, controlled by policy Π. The transition probability matrix of this controlled Markov chain is given by ([1]):

$$p(s_1 + s_2 | s_1, \lambda) = \begin{cases} P^{\lambda}_{s_2}(N - s_1), & \text{if } s_1 + s_2 \leq N \\ 0, \text{ else} \end{cases}$$

$$\text{with} \quad P^{\lambda}_{s_2}(r) := \binom{r}{s_2}(1 - e^{-\lambda})^{s_2} e^{-\lambda(r - s_2)}.$$

Basically the number infected increases by s_2 if any s_2 among the non-infected $(N - s_1)$ mobiles contact the source and the above is the probability of precisely this event.

A failure event occurs, when none of the X_t infected mobiles 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 $\{X_t\}_{t \leq T}$ and is given by (see [1] for details):

$$P_f(\Pi) = E^{\alpha, \Pi}\left[e^{-\nu \sum_t X_t}\right]. \tag{1}$$

In the above $E^{\alpha,\Pi}$ represents the expectation under policy Π and when the initial condition X_0 is distributed according to α, written as $X_0 \sim \alpha$. Here $Prob(X_0 = s) = \alpha(s)$.

C. Total power spent given a policy

The contact rate λ is proportional to $p^{-\beta}$, where p is the transmitted power and β is a constant depending upon propagation characteristics of the area in which the mobiles are operating (Appendix of [1]). 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 time slots equals:

$$\mathcal{P}(\Pi) = \sum_{t=0}^{T-1} Y_t^{\beta}. \tag{2}$$

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:

$$\min_{\Pi} E^{\alpha,\Pi}\left[e^{-\nu \sum_{t=0}^{T} X_t}\right] \text{ such that } E^{\alpha,\Pi}\left[\sum_{t=0}^{T-1} Y_t^{\beta}\right] \leq B. \tag{3}$$

Alternatively, in [1] the authors consider a joint cost depending both upon the probability of failure P_f and $e^{h\mathcal{P}}$, a term proportional to the power spent:

$$\min_{\Pi} E^{\alpha,\Pi}\left[e^{-\nu\sum_{t=0}^{T}X_t+h\sum_{t=0}^{T-1}Y_t^{\beta}}\right]. \quad (4)$$

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, because this does not guarantee to operate within a given hard bound on the power spent.

We showed in the technical report ([4]) that the solution of a constrained risk MDP problem can be derived using the solution of an appropriate Linear Program (19) given in the Appendix. Thus we can directly obtain the solution to the hard constraint (HC) problem. The details are given below in section III.

A direct solution would obviously perform better, we compare the two solutions in section IV to determine the percentage of improvement obtained.

III. LP BASED APPROACH

The soft constraint (SC) problem (4) can be cast as a risk sensitive MDP problem of Appendix. The joint cost in (4), except for the logarithmic function, is similar to the risk sensitive cost $J(\alpha,\Pi)$ given as equation (9) of Appendix, with running and terminal costs given by

$$r_t^{SC}(s,\lambda) = -\nu s + h\lambda^{\beta} \text{ and } r_T^{SC}(s) = -\nu s. \quad (5)$$

By monotonicity, the optimization of a cost is equivalent to optimizing the logarithm of the same cost.

The corresponding dynamic programming (DP) equations of the SC problem are given by (11) after substituting the running and terminal costs appropriately. One can obtain the analysis of the optimal policy by solving the dual LP given by (13) of Appendix.

The hard constrained (HC) problem (3) is similar to the constrained risk MDP problem (17) of Appendix. The corresponding running and terminal costs are

$$r_t^{HC}(s,\lambda) = -\nu s \text{ and } r_T^{HC}(s) = -\nu s, \quad (6)$$

while the constraint function

$$f_t^{HC}(s,\lambda) = \lambda^{\beta} \text{ for all } t. \quad (7)$$

Its optimal policy can be obtained by solving Dual LP given by (19).

IV. NUMERICAL ANALYSIS

In [1, Lemma 2], authors obtained structural properties of the policy, that optimizes the SC problem. Lemma 2 establishes the existence of a switch off threshold s_{off} on the number infected. The optimal policy switches off (zero contact rate) the beacon transmission, once the number infected reaches the threshold. It also showed that the contact rate chosen below the threshold is always non-zero. However this statement needs a small correction[1]. For every $s < s_{off}$, there exists a threshold T_s^* (depending upon s) such that

$$\lambda_t^*(s) \geq \lambda_1 \text{ for all } t < T_s^* \text{ and } \lambda_t^*(s) = 0 \text{ for all } t \geq T_s^*.$$

Thus the difference is that, beyond s_{off} it is always OFF as in [1]. However below switch off population threshold, the actual switch off threshold depends upon the number infected s.

A. Verification and comparison of SC, HC problems

We begin with the verification of our solution to SC problem. We obtain the required solution by solving the LP (13) with the running and terminal costs as given by (5). For simulations, we used Matlab and AMPL. We did most of the coding in Matlab except for LP part. We used AMPL to model the LP and Gurobi solver to solve the LP. The solution $\mathbf{x}^*$ of the LP provides the optimal policy Π_{x^*} as given by equation (15) of Appendix.

We then verify that the solution satisfies the Lemma 2 of [1], after the correction. We consider an example with $N = 15$, $S = \{0, 1, \ldots 15\}$, $T = 20$, $h = 20$, $\nu = 0.1$, $\beta = 2.1$ and $\lambda = \{0, 0.1, \ldots, 0.3\}$. For this example, the $s_{off} = 13.411344$ as given by [1, Lemma 2]. The simulation results are following the structure given by the [1, Lemma 2], as seen from the Table I. For example for all $s \geq s_{off}$, $T_s^* = 0$ and for others it is non-zero. We have conducted few more examples and verified the same.

In a similar way we obtain the solution for HC problem, now solving LP (19) with running and terminal costs given by (6). We also consider the constraint given by (7). We consider the following procedure for verification. We first solve SC problem for some value of weight factor h. We compute the total power spent by the SC problem using again the additional state component Ψ of the Appendix. That is, we solve SC problem also using LP (19) with $f_t^{SC} \equiv 0$ for all t. We compute the total average power $\mathcal{P}_{SC}^*$ spent by the system under SC optimal policy Π_{SC}^*, once again using the equation (18) with $f_t(a_t) = a_t^{\beta}$ and $\mathbf{x} = \mathbf{x}_{\Pi_{SC}^*}$. We then obtain the solution of HC problem with bound B set to $\mathcal{P}_{SC}^*$. 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 same power, i.e., $\mathcal{P}_{SC}^* = \mathcal{P}_{HC}^*$. But there is a good improvement in the performance with HC policy (see Tables II-III). In the limited examples that we conducted, we observed an improvement as high as 26% . In all these examples we set $M = 1$, resulting in a ON-OFF control.

Thus, when the two problems obtain optimal policies with the same power constraint, the HC solution performs superior,

[1] In [1, page 9] in the proof of Lemma2, 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 $\lambda_{t+1}^*(n) \geq \lambda_1$ then for all $\tau \leq t$ $\lambda_\tau^*(n) \geq \lambda_1$. Thus we have the above modification with $T_n^* = t$, the first t for which $\lambda_{t+1}^*(n) \geq \lambda_1$.

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 USING [1, LEMMA 2]

T, N	h	ν	β	λ_1	$\mathcal{P}$	P_f^S	P_f^H	% Improvement
5, 3	10.2	0.70	2.1	0.20	0.034481	0.036870	0.028178	26.72488008854999
5, 3	10.2	0.60	2.1	0.20	0.34481	0.043562	0.034982	21.847626807903842
6, 3	8	0.50	2.1	0.20	0.048897	0.028480	0.024088	16.709785420788315
6, 4	8	0.30	2.1	0.2	0.048073	0.031175	0.025450	20.220750551876382

TABLE II
IMPROVEMENT IN P_f: HC VERSUS SC, WITH EQUALLY LIKELY INITIAL CONDITIONS, $\alpha(s) = 1/(|S|)$ FOR $s \in S$.

T, N	h	ν	β	λ_1	$\mathcal{P}$	P_f^S	P_f^H	% Improvement
5, 3	10	0.50	2.1	0.20	0.081042	0.190186	0.168753	11.942419185432626
5, 3	10.2	0.70	2.1	0.20	0.081042	0.140198	0.112873	21.594730332594403
6, 3	10	0.50	2.1	0.20	0.099265	0.113674	0.096116	16.738643405310077
6, 3	8	0.50	2.1	0.20	0.113148	0.104767	0.089853	15.32627684718939

TABLE III
IMPROVEMENT IN P_f: HC VERSUS SC STARTING WITH ZERO INFECTED MOBILES, $\alpha(s) = 1_{\{s=0\}}$ FOR $s \in S$.

obviously because it directly solves the constrained problem. However the more interesting observation is that the improvement gained can be significant.

We would now like to look at the comparison problem with a different perspective. Say we are given any arbitrary total average power constraint B. 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 known to be pure policies: $\pi_t = 1$ or 0, for all t. We also 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 optimal policies of SC problem are always pure. The other curve in Figure 1 represents the P_f 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 much larger gains in the performance at the other points (see the horizontal portions of the SC curve in Figure1).

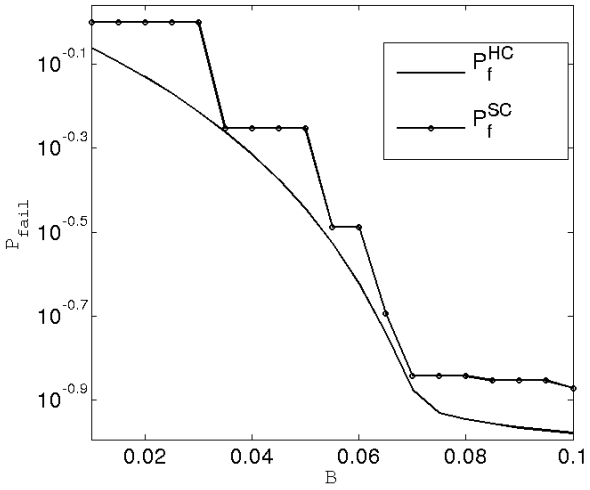

Fig. 1. P_f performance as a function of power bound B

B. 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 s_t, a randomized switch ON-OFF at the threshold state $s =$

s_t and a complete switch OFF for all states, $s > s_t$. This threshold depends upon the time slot and of course, the power constraint B.

ACKNOWLEDGEMENTS:

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

V. CONCLUSIONS

We considered power constrained Delay Tolerant Networks. We obtained optimal policies for this problem, via the solution of an appropriate LP, after modeling it as a constrained risk sensitive MDP. The equivalence of the two is provided in the technical report. Previously a joint cost comprising of probability of delivery failure and a term proportional to total power spent is considered. While in this paper we directly solve the constrained optimization problem. We compared the probability of failure performance of the DTNs under the policy so obtained, with that of the optimal policy obtained by considering an unconstrained problem with the joint cost. We observed huge improvement. This improvement is because of two factors. When the requirement is to operate optimally within a given power constraint our solution provides the optimal solution while the solution of the unconstrained problem with joint cost would be sub-optimal. Secondly and more importantly, the optimization of the unconstrained risk sensitive cost results in pure policies which provide only finite choices of total average power spent. While our proposed constrained risk MDP solution results in optimal policies that are randomized and hence provide a solution, which satisfies with equality the constraint defining the power constraint. This is true as long as the power constraint is achievable. Thus our solution performs significantly superior and is very useful in the scenarios that demand for strict power constraints.

APPENDIX: FINITE HORIZON RISK SENSITIVE MDP AND LINEAR PROGRAMMING

Markov Decision Process (MDP) provides a tool for solving sequential decision making problems in stochastic situations. A typical MDP consists of a set S of all possible states, a set A of all possible actions and an immediate reward function $r_t : S \times A \to \mathcal{R}$ for each time slot t. The terminal cost r_T depends only upon $s \in S$. The set S and A can depend upon the time slot t, however we consider the same set for all the time indices. It is further characterized by a transition function $p : S \times A \to S$, which defines the action dependent state transitions. Here $p(s'|s,a)$ gives the probability of the state transition from s to s', when action a is chosen.

We consider a finite horizon problem and let $\{X_t\}_{t\leq T}$, $\{Y_t\}_{t\leq T-1}$ respectively represent the trajectories of the state and the action. In the last time slot T, there is no requirement for further action and we only have a terminal cost. 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 and initial condition $X_t = \tilde{s}$, the state and action pair evolve randomly over the time slot $t < n < T$, with transitions as governed by the following laws:

$$\begin{aligned} q^{\Pi^t}(s',a'|s,a) &= P(X_n = s', Y_n = a'|X_{n-1} = s, Y_{n-1} = a) \\ &= \pi_n(s',a')p(s'|s,a) \text{ where} \\ p(s'|s,a) &= P(X_n = s'|X_{n-1} = s,\, Y_{n-1} = a) \text{ and} \\ \pi_n(s',a') &= P(Y_n = a'|S_n = a'). \end{aligned} \tag{8}$$

Let E^{s,Π^t} represent the expectation operator with initial condition $X_t = s$ 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 $\alpha(s) = P(X_t = s)$. We are interested in optimizing the following risk sensitive objective:

$$J_t(\alpha, \Pi^t) = \gamma^{-1} \log \left(E^{\alpha,\Pi^t} \left[e^{\gamma \sum_{n=t}^{T-1} r_n(X_n,Y_n) + r_T(X_T)} \right] \right). \tag{9}$$

The above 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 (s,t), is defined as the optimal value of the above risk sensitive objective given the initial condition $X_t = s$:

$$V_t(s) := \min_{\Pi^t} J_t(s, \Pi^t) \text{ for any } s \in S. \tag{10}$$

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 $J_0(s, \Pi^0)$, or equivalently a policy that achieves the value function, i.e., a Π^* such that

$$V_0(s) = J_0(s, \Pi^*) \text{ for all } s \in S.$$

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 ([6]):

$$\begin{aligned} V_T(s) &= r_T(s), \text{ and for any } 0 \leq t \leq T-1, \text{ and } s \in S \\ V_t(s) &= \min_{a\in A} \left\{ r_t(s,a) + \gamma^{-1} \log \left[\sum_{s'\in S} p(s'|s,a) e^{\gamma V_{t+1}(s')} \right] \right\}. \end{aligned}$$

We consider the following translation of the value function:

$$u_t(s) = e^{\gamma V_t(s)} \text{ for all } 0 \leq t \leq T-1, \text{ and } s \in S.$$

The DP equations can now be rewritten as:

$$\begin{aligned} u_t(s) &= e^{\gamma r_T(s)} \text{ and for any } 0 \leq t \leq T-1, \text{ and } s \in S \\ u_t(s) &= \min_a \left\{ e^{\gamma r_t(s,a)} \left[\sum_{s'\in S} p(s'|s,a) u_{t+1}(s') \right] \right\}. \end{aligned} \tag{11}$$

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., [7], [8] in the context of infinite horizon problems). However *this conversion may 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 technical report ([4]), we extend the LP based idea to a finite horizon risk MDPs. In this appendix we briefly summarize the corresponding results, while the details and the proofs are available in ([4]).

We have shown in ([4]) that the solution of the unconstrained risk MDP problem (10) can be obtained via the solution of any one of the two LPs, a primal and a dual. In all the discussions below, we absorb γ into the running costs $r_t(.)$. The primal LP is given by:

$$\max_{\{u_t(s)\}_{s\in S, t\le T-1}\}} \sum_{s\in S} \alpha(s) u_0(s) \tag{12}$$

$$\begin{aligned} \text{subject to:} \qquad u_{T-1}(s) &\le b_{s,a} \quad \text{for all } s,a, \\ u_t(s) - e^{r_t(s,a)} \sum_{s'\in S} p(s'|s,a) u_{t+1}(s') &\le 0 \\ &\text{for all } a, s \text{ and } t \le T-2 \end{aligned}$$

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

In the above $\{\alpha(s); s \in S\}$ is any positive set of weights satisfying $\sum_{s\in S} \alpha(s) = 1$. These can be interpreted as the probability distribution on initial condition. *For example to solve (10) with $t = 0$, the problem with initial condition $X_0 = s$, one needs to set $\alpha(s) = 1$ and $\alpha(s') = 0$ for any $s' \neq s$.*

The solution of the primal gives the translated value functions $\{u_t(s)\}$ while the optimal policy is directly obtained using the Dual LP:

$$\min \sum_a \sum_{s\in S} e^{r_{T-1}(s,a)} \left[\sum_{s'\in S} p(s'|s,a) e^{r_T(s')} \right] x(T-1,s,a) \tag{13}$$

subject to:

$$\begin{aligned} \sum_a x(0,s',a) &= \alpha(s') \text{ for all } s' \in S \\ \sum_a x(t,s',a) &= \sum_a \sum_{s\in S} \left[e^{r_{t-1}(s,a)} p(s'|s,a) x(t-1,s,a) \right] \\ &\qquad \text{for all } 1 \le t \le T-1 \text{ and } s' \in S. \end{aligned}$$

Here again α *represents the probability distribution on the initial condition.*

We have the following results (details in [4]). We discard the notation superscript 0 for risk policies in the following. The bold letters represent the vectors, e.g., $\mathbf{x} = \{x(t,s,a)\}_{t,s,a}$ represents a feasibility vector of Dual LP (13). While $\mathbf{s}_k^n$ represents the vector $\mathbf{s}_k^n = [s_k, \cdots, s_n]$.

Theorem *1:* The following results connecting the Dual LP (13) and the translated risk MDP (11) are true.

1) Feasible region and the set of risk Policies: There is a one to one correspondence between the two as below:
i) For any policy Π of risk MPD, there exists a vector $\mathbf{x_\Pi}$ which satisfies all the constraints of Dual LP (13). The feasible vector is given by the equation (see (8)):

$$\begin{aligned} x_\Pi(0,s_0,a_0) &= \alpha(s_0)\pi_0(s_0,a_0) \text{ for all } s_0 \in S, a_0 \in A, \\ x_\Pi(t,s_t,a_t) &= \sum_{\mathbf{a}_0^{t-1}, \mathbf{s}_0^{t-1}} \alpha(s_0) e^{\sum_{n=0}^{t-1} r_n(s_n,a_n)} \Pi_{n=0}^t q^\Pi(s_n,a_n|s_{n-1},a_{n-1}) \\ &\qquad \text{for all } s_t \in S, a_t \in A, \text{ and } 1 \le t < T. \end{aligned} \tag{14}$$

ii) Given a vector $\mathbf{x}$ in the feasibility region of Dual LP, define a policy $\Pi_{\mathbf{x}}$ using the following rule:

$$\pi_{\mathbf{x},t}(s,a) := \frac{x(t,s,a)}{\sum_{a'} x(t,s,a')} \text{ for all } s \in S, \text{ and } a \in A. \tag{15}$$

The vector $\mathbf{x}_{\Pi_\mathbf{x}}$ defined by equation (14) of point (i) is again in the feasibility region and equals $\mathbf{x}$.

2) Optimal policies and solutions: (a) If $\mathbf{x}^*$ is an optimal solution of the Dual LP, then $\Pi_{\mathbf{x}^*}$ defined by (15) is an optimal policy for risk MDP.

3) Expectation at optimal Policy: For any feasible point $\mathbf{x}$ of Dual LP and for any integrable function f,

$$\begin{aligned} &\sum_{s_t,a_t} x(t,s_t,a_t) f(s_t,a_t) \\ &\qquad = E^{\Pi_\mathbf{x}} \left[e^{\sum_{n=0}^{t-1} r_n(X_n,Y_n)} f(X_t,Y_t) \right]. \end{aligned} \tag{16}$$

■

Constrained risk MDP

We now consider a constrained MDP problem (details are in [4]), with an additional constraint as given below:

$$\min_\Pi J_0(\alpha,\Pi) \tag{17}$$

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

for some set of integrable function $\{f_t\}$, initial distribution α and bound B. The equation (16) of Theorem 1 could have been useful in obtaining the expectation defining the constraint, but for the extra factor Ψ_t^{-1} with $\Psi_t := e^{-\sum_{n=0}^{t-1} r_n(X_n,Y_n)}$, as seen from the right hand side of the equation (16). We propose to add Ψ_t as additional state component to the original Markov process $\{X_t\}$ to tackle this problem. We consider a two component state evolution $\{(X_t,\Psi_t)\}$ and the corresponding probability transition matrix depends explicitly upon time index as below:

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

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

$$\sum_{s_t,\psi_t,a_t} x(t,s_t,\psi_t,a_t)\, \psi_t\, f(s_t,a_t) = E^{\Pi_\mathbf{x}}\left[f(X_t,Y_t)\right]. \tag{18}$$

Thus one can obtain optimal policy of constrained risk MDP (17) by considering an additional state component and by adding an extra constraint to the Dual LP (13) as below:

$$\min \sum_a \sum_s e^{r_{T-1}(s,a)} \left[\sum_{s' \in S} p(s'|s,a) e^{r_T(s')} \right] x(T-1,s,a) \quad (19)$$

subject to:

$$\begin{aligned} x(t,s,a) &= \sum_{\psi_t} x(t,s,\psi_t,a) \\ \sum_a x(0,s,\psi_0,a) &= \alpha(s) 1_{\{\psi_0=1\}} \text{ for all } s, \psi_0 \\ \sum_a x(t,s',\psi_t',a) &= \\ & \sum_{a,s,\psi_{t-1}} e^{r_{t-1}(s,a)} \tilde{p}(s',\psi_t'|s,\psi_{t-1},a) x(t-1,s,\psi_{t-1},a) \end{aligned}$$

for all $1 \leq t \leq T-1$ and s', ψ_t' and

$$\sum_t \sum_{s,\psi_t,a} x(t,s,\psi_t,a) \psi_t f_t(s,a) \leq B.$$

We would like to mention here that ψ_0 is always initialized to one, i.e., $\psi_0 = 1$, ψ_1 can take at maximum $|S| \times |A|$ values while ψ_t for any t can take at maximum $|S|^t \times |A|^t$ possible values. There will also be considerable deletions if the concerned mapping

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

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.

REFERENCES

[1] E. Altman, V. Kavitha, F. De Pellegrini, V. Kamble, and V. Borkar, "Risk sensitive optimal control framework applied to delay tolerant networks," in *INFOCOM, 2011 Proceedings IEEE*. IEEE, 2011, pp. 3146–3154.

[2] R. Groenevelt, P. Nain, and G. Koole, "Message delay in manet," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 33, no. 1. ACM, 2005, pp. 412–413.

[3] E. Altman, T. Başar, and F. De Pellegrini, "Optimal monotone forwarding policies in delay tolerant mobile ad-hoc networks," *Performance Evaluation*, vol. 67, no. 4, pp. 299–317, 2010.

[4] Atul Kumar, Veeraruna Kavitha and N. Hemachandra, "Finite horizon risk sensitive MDP and linear programming," *Manuscript under preparation. Technical report available at http://www.ieor.iitb.ac.in/files/faculty/kavitha/RiskMDPLP.pdf*, 2015.

[5] Khouzani, M. H. R., Eshghi, S., Sarkar, S., Shroff, N. B., and Venkatesh, S. S. (2012, June). "Optimal energy-aware epidemic routing in DTNs," In Proceedings of the thirteenth ACM international symposium on Mobile Ad Hoc Networking and Computing (pp. 175-182). ACM.

[6] S. P. Coraluppi and S. I. Marcus, "Risk-sensitive queueing," in *Proceedings of the Annual Allerton Conference on Communication Control and Computing*, vol. 35. Citeseer, 1997, pp. 943–952.

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

[8] Eitan Altman, *Constrained Markov decision processes*, volume 7, 1999, CRC Press.