IEEE Control Systems Letters paper presented at
2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

# Optimal Markov Policies for Finite-Horizon Constrained MDPs With Combined Additive And Multiplicative Utilities

Uday M. Kumar, Veeraruna Kavitha, Sanjay P. Bhat and Nandyala Hemachandra

***Abstract*— This paper considers the problem of optimizing a finite-horizon constrained Markov decision process (CMDP) where the objective and constraints are sums of additive and multiplicative utilities. To solve this, we construct another CMDP with only additive utilities whose optimal value over a restricted set of policies is equal to that of the original CMDP. Further, we provide a finite-dimensional bilinear program (BLP) whose value equals the CMDP value and whose solution provides the optimal policy. We also suggest an algorithm to solve the proposed BLP.**

## I. Introduction

In this paper, we consider constrained Markov decision processes (CMDPs) whose objective as well as constraints involve combination of multiplicative and additive utilities. We are interested in optimizing the expected value of $\sum_{t=0}^{T} r_t(X_t, A_t) + \alpha \prod_{t=0}^{T} f_t(X_t, A_t)$, a weighted combination of the two. Dynamic programming equations are not known directly for such MDPs. One can solve finite horizon variants by augmenting an additional state $\prod_{t=0}^{\tau} f_t(X_t, A_t)$ that tracks the multiplicative cost upto time $\tau$; however the state space grows exponentially with the time horizon and the paper aims to fill this gap by working towards an implementable solution.

Towards the first motivation for such MDPs, first observe that risk-sensitive costs (a special case of multiplicative costs) provide one way of robust control by optimizing a weighted sum of moments (e.g., robust control of a queuing system as in [1]). There can be examples in which some objectives require robustness, while for others it is sufficient to optimize the first moment. Secondly, there are applications where one or more objectives directly have multiplicative form. For example in [2], [3] while optimizing the failure probability in a delay tolerant networks with a two-hop protocol results in a risk-sensitive MDP, additionally considering resource constraints results in a combined MDP. In [3], the combined MDP is converted to a classical MDP with two additive costs, by optimizing an upper bound on the failure probability using Jensen's inequality.

Unconstrained MDPs with only multiplicative cost components can be solved using dynamic programming [1]. Reference [1] provides an LP-based solution for a finite-horizon MDP having only multiplicative (specifically, risk-sensitive) objective with additive constraint. The solution technique of [1] is based on augmenting the state space with a variable that keeps track of the running multiplicative cost as already mentioned. In principle, using this approach, one can solve the combined-cost MDPs, however the state space and thus the size of the LP grows exponentially with the horizon. One cannot implement a numerical solution even for moderate-sized MDPs, [4] discusses the difficulty even for short horizons like 10-15. A recent algorithm in [5] solves constrained risk-sensitive MDPs, however, they do not consider combined MDPs. The unique contribution of our paper is to find an optimal Markov policy for a finite-horizon CMDP with combined multiplicative and additive costs and also to suggest an implementable algorithm.

We provide a Bi-linear programming (BLP) based solution technique that involves augmenting the state with binary variables, one for every multiplicative component appearing in either the objective or the constraint. We construct a CMDP defined on the augmented state space such that the additive and multiplicative cost components of the original CMDP are absorbed into the additive stage-wise costs and the controlled transition functions, respectively, of the augmented CMDP. We prove that both the CMDPs have the same optimal value under a slight restriction on the augmented policy space. Moreover, an optimal policy for the augmented CMDP can be constructed using the solution of the BLP. We further suggest an iterative algorithm (inspired by [5]) to solve the BLP. Each iteration of the algorithm solves a finite-dimensional linear program. We conclude this section with a relevant application.

*Epidemics and Delay Tolerant Networks*

Consider an area with $N$ individuals where an epidemic is spreading fast. The government has to devise a lock-down policy to efficiently combat the disease, keeping in view lock-down costs in the form of economical losses. The time-frame is divided into $T$-time slots and the policy prescribes the level of lock-down to be imposed for each time slot based on the system state at the beginning of the time slot.

Any infected person can infect a normal/susceptible person when they come in contact with each other (see [6] for similar details). The successive contacts between any two persons are modelled by a Poisson process (as in [2], [6]). The rate of this contact process is determined by the imposed level of lock-down. Let $\Lambda_t$ represent the contact rate chosen in slot $t$ and let $g(\Lambda_t)$ be the economic cost due to the corresponding level of lockdown. Any infected person recovers in a slot with probability $r$; a person infected in a time slot can infect others only from the next slot. Let $X_t$ represent the number of infected individuals at the beginning of time slot $t$. Then the number infected in the next slot $X_{t+1} = \mathcal{B}(N - X_t, q_t) - \mathcal{B}(X_t, r)$, where $\mathcal{B}(\cdot, \cdot)$, is a binomial random variable, $\mathcal{B}(X_t, r)$ represents the number of recoveries and $q_t = 1 - \exp(-\Lambda_t X_t)$ is the probability that a susceptible gets infected. We are interested in the

probability $P_S(\pi)$ that a given typical individual survives without infection in the given time-frame for a given lock-down policy $\pi$. By conditioning on state trajectory $\{X_t\}$, details as in [2], $P_S(\pi) = \mathbb{E}^\pi\left[e^{-\sum_{t=1}^T \Lambda_t X_t}\right]$. Thus in all we would optimize a combined cost that also considers the economic losses due to lock-down ($\alpha > 0$ – trade-off factor):

$$\min_\pi \mathbb{E}^\pi\left[\sum_t g(\Lambda_t) - \alpha e^{-\sum_{t=1}^T \Lambda_t X_t}\right].$$

A similar problem (excluding recoveries) arises in delay tolerant networks ( [2]), when one considers a combined cost related to delivery failure probability and power-budget. One can further have a bound on the expected number of copies.

The aim of this paper is to consider a general class of problems that are similar in nature to the above example.

## II. Model Formulation and Problem Statement

We consider a finite-horizon Markov decision process (MDP) $M := (\mathcal{X}, \mathcal{A}, Q, \{0, 1, \ldots, T\})$, where $\mathcal{X}$ is the finite state space, $\mathcal{A}$ is the finite action space, $T > 0$ is the terminal time, $Q$ is the transition function (or transition law) and $\{0, 1, \cdots, T\}$ is the set of discrete decision epochs. Here $Q(x'|x, a)$ is the probability that the system transitions to state $x'$ when action $a$ is taken at state $x$.

A Markov randomized (MR) decision rule is a map $d : \mathcal{X} \to \mathcal{P}(\mathcal{A})$, where $\mathcal{P}(\mathcal{A})$ is the set of probability distributions on the action space $\mathcal{A}$. We denote by $d(a|x)$ the probability of choosing action $a$ in the state $x$ under the MR decision rule $d$. An MR policy is a sequence of MR decision rules indexed by the decision epochs. We denote the set of MR policies by $\Pi_{\text{MR}}$. For every initial state $s$, a policy $\pi := \{d_t\}_{t=0}^{T-1} \in \Pi_{\text{MR}}$ induces a probability measure $P_s^\pi$ on the space of state-action trajectories ( [7, Ch2.]). Let $\mathbb{E}_s^\pi[\cdot]$ denote the corresponding expectation. To provide implementable algorithms, we restrict ourselves only to the space of MR policies and avoid history-based policies.

This paper considers an MDP involving additive as well as multiplicative stage-wise components with one objective function and $K$ number of constraints. These quantities are defined using $(K+1)$ number of additive and multiplicative components, $r_{t,i} : \mathcal{X}\times\mathcal{A}\times\mathcal{X} \to \mathbb{R}$ and $f_{t,i} : \mathcal{X}\times\mathcal{A}\times\mathcal{X} \to \mathbb{R}$, respectively, where $0 \le i \le K$ and $t \in \{0, 1, \ldots, T-1\}$. We consider the following optimization problem: given $s \in \mathcal{X}$ and $b_i \in \mathbb{R}$ for each $0 \le i \le K$, find

$$\inf_{\pi\in\Pi_{\text{MR}}} w_0^\pi(s), \text{ subject to, } w_i^\pi(s) \le b_i, \ \forall\ 1 \le i \le K, \quad (P)$$

where, for any policy $\pi \in \Pi_{\text{MR}}$, for each $0 \le i \le K$ and $\alpha_i \in \mathbb{R}$, we define

$$w_i^\pi(s) \triangleq \mathbb{E}_s^\pi\left[\sum_{t=0}^{T-1} r_{t,i}(X_t, A_t, X_{t+1}) + \alpha_i \prod_{t=0}^{T-1} f_{t,i}(X_t, A_t, X_{t+1})\right].$$

We assume that, for each $t$ and $i$ the multiplicative cost component $f_{t,i}$ has the same sign everywhere on $\mathcal{X} \times \mathcal{A} \times \mathcal{X}$. By appropriately scaling and then absorbing the scaling factor and common sign into the coefficient $\alpha_i$, we assume without loss of generality that $0 \le f_{t,i} \le 1$ for all $t$ and $i$.

**Special Cases:**

The problem ($P$) is fairly general, and covers the following special cases: a) with $\alpha_i = 0$ for all $i$, we have the well known classical MDP with constraints ( [8]) (e.g., we have discounted-cost MDPs when $r_{t,i} \equiv \beta^t r_i$ and $\alpha_i = 0$ in the problem); b) setting $r_{t,i} \equiv 0$ and $f_{t,i}(x, a, x') = \exp(\beta^t c_i(x, a))$ reduces ($P$) to the well known *risk-sensitive MDP* ( [1]); and c) by setting $f_{t,1}(x, a, x') = \mathbb{1}_{\mathcal{E}}(x, a)$ and $r_{t,i} \equiv 0$, ($P$) reduces to the problem of optimizing the cost under a constraint on the probability of entering a set $\mathcal{E}$ of undesirable "error" states ( [9]).

## III. Equivalent CMDP and Main results

In this section, we provide the two main results of this work. The first result gives the equivalence between the original CMDP ($P$) and a newly constructed CMDP involving only additive costs. The second result provides an equivalence to a finite dimensional optimization problem with linear objective function and with linear and bilinear constraints. The proofs of the results are in the appendix.

We augment the state space with one binary variable per multiplicative cost; each variable starts with 1 and can get absorbed to 0. The core idea is to capture the multiplicative cost via the expectation of terminal value of the binary variable. In the next subsection, we give precise mathematical details of this construction.

**Equivalent CMDP:** We now construct a new CMDP involving only additive cost and constraint components. To simplify the exposition, consider $\alpha_i \neq 0$ for each $i$ in ($P$). Define the augmented state space by $\bar{\mathcal{X}} := \mathcal{X} \times \mathcal{Z}$, where $\mathcal{Z} = \{0, 1\}^{K+1}$. For each $(x, z), (x', z') \in \bar{\mathcal{X}}$, $a \in \mathcal{A}$ and $t \in \{0, \ldots, T-1\}$ the transition function of the new CMDP at epoch $t$, i.e., from $(X_{t-1}, Z_{t-1})$ to $(X_t, Z_t)$ is

$$\bar{Q}_t\Big((x', z')\Big|(x, z), a\Big) := \begin{cases} 0, \text{ if } z_i = 1 - z_i' = 0 \text{ for at least one } i, \\ Q(x'|x, a)\prod_{i=0}^{K} \rho_{t-1,i}(x, x', a, z, z'), \text{ else,} \end{cases}$$

where $\rho_{t,i}(x, x', a, z, z') := (f_{t,i}(x, a, x'))^{z_i z_i'} \times (1 - f_{t,i}(x, a, x'))^{z_i(1-z_i')}$,

for each $t \in \{0, \ldots, T-1\}$ and $0 \le i \le K$.

It is easy to check that the transition law $\bar{Q}_t(\cdot|(x, z), a)$ is indeed a probability distribution on the set $\bar{\mathcal{X}}$ for all $(x, z) \in \bar{X}$ and $a \in \mathcal{A}$. This allows us to define a Markov control model, $\bar{M} := (\bar{\mathcal{X}}, \mathcal{A}, \bar{Q}_t, \{0, 1, \ldots, T\})$ to represent the state evolution of the process $\{\bar{X}_t\}_{t=0}^T$ where $\bar{X}_t := (X_t, \{Z_{t,i}\}_{i=0}^K)$. Here, for every $i$, $Z_{t,i} \in \{0, 1\}$. The central idea of this construction is two fold: (a) the transitions of the original Markov chain $\{X_t\}_{t=0}^T$ are not affected by the process $\{Z_t\}_{t=0}^T$ and (b) as we shall soon see, the expected value of the binary variable $Z_{T,i}$ equals the expected value of the multiplicative cost $\prod_t f_{t,i}(X_t, A_t, X_{t+1})$.

Next, we define $\bar{\Pi}$ to be the set of MR policies w.r.t $\bar{M}$ which are indifferent to values of the augmented state. More precisely, $\bar{\Pi}$ is the set of policies $\bar{\pi} = \{\bar{d}_t\}_{t=0}^{T-1}$ such that $\bar{d}_t(a|(x, z)) = \bar{d}_t(a|(x, \mathbf{1}))$ for all $t \in \{0, \ldots, T-1\}$ where $\mathbf{1} := (1, \ldots, 1) \in \mathbb{R}^{K+1}$. It is easy to observe that $\bar{\Pi}$ is in a one-to-one correspondence with $\Pi_{\text{MR}}$, the space of MR

Let $\mathbf{W}_1 := \{w_t(x,\mathbf{1},a);\ \text{for all } x,a,t<T\}$ and $\mathbf{W}_2 := \{w_t(x,z,a);\ \text{for all } x,a,t<T, z\neq\mathbf{1}\}, \mathbf{W}=(\mathbf{W}_1,\mathbf{W}_2)$.

$$\textbf{BLP } (\mathbf{W}_1,\mathbf{W}_2)\textbf{:}\ \min_{\{\mathbf{W}=(\mathbf{W}_1,\mathbf{W}_2)\}} \sum_{t=0}^{T-1}\sum_{(x,z)\in\bar{\mathcal{X}}}\sum_{a\in\mathcal{A}} \bar{r}_{t,0}((x,z),a)w_t((x,z),a) + \sum_{(x,z)\in\bar{\mathcal{X}}} \bar{r}_{T,0}(x,z)\Phi(x,z;\mathbf{W})$$

$$\text{s.t.} \sum_{a\in\mathcal{A}} w_0((x,z),a) = \mathbb{1}_{\{(s,\mathbf{1})\}}(x,z),\ \Phi(x,z;\mathbf{W}) = \sum_{x',z',a} \bar{Q}_T\Big((x,z)\Big|(x',z'),a\Big) w_{T-1}((x',z'),a) \text{ for all } (x,z)\in\bar{\mathcal{X}},$$

$$\sum_a w_t((x,z),a) - \sum_{x',z',a} \bar{Q}_t\Big((x,z)\Big|(x',z'),a\Big) w_{t-1}((x',z'),a) = 0, (x,z)\in\bar{\mathcal{X}},\ \text{and } 1\le t\le T-1 \tag{1}$$

$$\sum_{t=0}^{T-1}\sum_{(x,z)\in\bar{\mathcal{X}}}\sum_{a\in\mathcal{A}} \bar{r}_{t,i}((x,z),a)w_t((x,z),a) + \sum_{(x,z)\in\bar{\mathcal{X}}} \bar{r}_{T,i}(x,z)\Phi(x,z;\mathbf{W}) \le b_i,\quad \text{for } i=1,\dots,K,$$

$$w_t(x,z,a)\sum_{a'} w_t(x,\mathbf{1},a') = w_t(x,\mathbf{1},a)\sum_{a'} w_t(x,z,a'),\quad w_t((x,z),a)\ge 0,\ \forall (x,z)\in\bar{\mathcal{X}},\ a\in\mathcal{A},\ t\le T-1.$$

policies w.r.t the original model $M$. We set the initial state in $\bar{M}$ to be $(s,\mathbf{1})$, where $s$ is initial state in $M$.

Further, we define the stage-wise costs in $\bar{M}$ by

$$\begin{aligned}\bar{r}_{t,i}((x,z),a) &:= \sum_{x'\in\mathcal{X}} r_{t,i}(x,a,x')Q(x'|x,a),\\ \bar{r}_{T,i}(x,z) &:= \alpha_i z_i \ \text{ for all } i, t<T.\end{aligned} \tag{2}$$

Finally, we define the following optimization problem the newly constructed augmented MDP,

$$\begin{aligned}&\min_{\eta\in\bar{\Pi}} v^{\eta}_{T,0}(s,\mathbf{1})\\ &\text{subject to, } v^{\eta}_{T,i}(s,\mathbf{1})\le b_i,\ \forall\ 1\le i\le K,\ \text{where,}\end{aligned} \tag{$\bar{P}$}$$

$$v^{\eta}_{T,i}(x,z) := \mathbb{E}^{\eta}_{(x,z)}\Big[\sum_{t=0}^{T-1}\bar{r}_{t,i}\Big((X_t,Z_t),A_t\Big) + \bar{r}_{T,i}(X_T,Z_T)\Big].$$

Observe, that the objective and constraints of the problem ($\bar{P}$) are both linear/additive only. We define the projection map $\Gamma:\bar{\Pi}\mapsto\Pi_{\text{MR}}$ as follows; for any $\eta=\{\bar{d}_t\}_{t=0}^{T-1}\in\bar{\Pi}$, define $\Gamma(\eta)=\{d_t\}_{t=0}^{T-1}$ where $d_t(a|x)=\bar{d}_t(a|x,\mathbf{1})$ for all $t$, $a$ and $x$.

### A. Main Results

We now give the first main result of this paper: the equivalence of ($P$) and ($\bar{P}$).

*Theorem 1:* Let $p^*$, $\bar{p}^*$ be the optimal values of ($P$) and ($\bar{P}$), respectively. Then $p^*=\bar{p}^*$. Further, $\eta^*$ is an optimizer of ($\bar{P}$) if and only if $\Gamma(\eta^*)$ is an optimizer of ($P$). ■

Theorem 1 provides an alternative CMDP ($\bar{P}$) where the objective as well as constraints are additive (or linear) as in a standard MDP, but the policy space is restricted to the policies which are indifferent to the augmented state component. There many solution techniques to solve standard MDP ( [8]), however the restricted policy space requires special attention.

The second main result of this paper is the Bi-linear programming (BLP), given by (1) provided at the top of the page, which solves the new problem ($\bar{P}$). The unknown variables of BLP (1) are indexed by $t$, $x$, $z$ and $a$, whose linear constraints represent the initial state conditions, state transitions and constraints of the MDP. The bilinear constraint in (1) ensures that the solution is indifferent to augmented state $z$. The decision function of the optimal policy using these variables are obtained as given below in theorem 2.

*Theorem 2:* **[Solution using BLP]** The value of BLP (1) is the value of the problems ($P$) and ($\bar{P}$). Let $\{w^*_t((x,z),a)\}_{t,x,z,a}$ for $0\le t\le T-1$, $x\in\mathcal{X}$, $z\in\mathcal{Z}$ and $a\in\mathcal{A}$ be the solution of the BLP. Then, the optimal policies $\bar{\pi}^* := \{\bar{d}^*_t\}_t$ for the augmented problem ($\bar{P}$) and $\pi^* := \{d^*_t\}_t$ for the original ($P$) are given by

$$\bar{d}^*_t(a|(x,z)) = d^*_t(a|x) = \frac{w^*_t((x,\mathbf{1}),a)}{\sum_{a'\in\mathcal{A}} w^*_t((x,\mathbf{1}),a')},$$
$$\text{for } x\in\mathcal{X},\ a\in\mathcal{A}, z\in\mathcal{Z}, t<T. \tag{3}$$

## IV. Algorithm

By Theorem 2, constrained global optimization problem of BLP (1) is equivalent to the combined-CMDP problem given in ($\bar{P}$). We now suggest an algorithm to solve the BLP which can rewritten as

$$\inf_{\mathbf{W}} B(\mathbf{W}) \text{ s.t. } C_1(\mathbf{W})\le\mathbf{b} \text{ and } C_2(\mathbf{W})=0, \tag{$BLP$}$$

for appropriate functions $(B,C_1,C_2)$ and $\mathbf{b}$, for example:

$$\begin{aligned}B(\mathbf{W}) &= \sum_{t=0}^{T-1}\sum_{(x,z)\in\bar{\mathcal{X}}}\sum_{a\in\mathcal{A}} \bar{r}_{t,0}((x,z),a)w_t((x,z),a)\\ &\quad + \sum_{(x,z)\in\bar{\mathcal{X}}} \bar{r}_{T,0}(x,z)w_T(x,z).\end{aligned} \tag{4}$$

This BLP has $\mathbf{W}=(\mathbf{W}_1,\mathbf{W}_2)$ as variables, and when one of them is fixed then it is clearly an LP in the other variable. Let $\mathcal{L}_1(\mathbf{W}_2)$ and $\mathcal{L}_2(\mathbf{W}_1)$ represent the solution set of the respective LPs when $\mathbf{W}_2$ and $\mathbf{W}_1$ are fixed. One can define a relevant fixed point equation using the above two LP solutions, to be precise we are interested in the fixed points $\mathbf{W}\in\mathcal{W}(\mathbf{W})$, where the solution set $\mathcal{W}$ is defined as,

$$\mathcal{W}(\mathbf{W}) := \{\tilde{\mathbf{W}} : \tilde{\mathbf{W}}_1\in\mathcal{L}_1(\mathbf{W}_2), \tilde{\mathbf{W}}_2\in\mathcal{L}_2(\mathbf{W}_1)\}. \tag{5}$$

Define $\mathbb{W} = \{\mathbf{W} : \mathbf{W} \in \mathcal{W}(\mathbf{W})\}$, the set of all fixed points. It is not difficult to verify that the solution of the following global fixed point problem ($GF$) provides the solution to BLP (1):

$$\inf_{\mathbf{W}\in\mathbb{W}} B(\mathbf{W}). \qquad (GF)$$

Thus any fixed point iterative algorithm is useful in solving the BLP, however one needs to ensure that it is converging towards the best fixed point (as define above). This is similar to alternate convex search (ACS) algorithm [10]. The global optimization problem defined in ($BLP$) can be useful in this context. This solution approach provided below, exactly parallels that in the recently provided algorithm [5]; the authors in [5] consider constrained risk-sensitive MDP problem and also have a fixed point equation and global optimization problem as in ($BLP$) and ($GF$). We believe the justification of the algorithm can parallel that provided in [5], however skip those details due to lack of space.

This global optimization problem can for example be solved using random restarts [11]. It has two types of update steps (at any iterative step $k$): i) random search step – a random new point is chosen from the feasible region with probability $p_k$, and, ii) local improvement step as in (6) is chosen otherwise. The probability $p_k$ diminishes with $k$.

The aim in the local improvement step is to converge to a fixed point in $\mathbb{W}$ as given by the following:

$$\mathbf{W}_k = \mathbf{W}_{k-1} + \epsilon_k \Big( \Psi(\mathbf{W}_{k-1}) - \mathbf{W}_{k-1} \Big), \epsilon_k = 1/k, \qquad (6)$$

where $\Psi(\mathbf{W}_k)$ (a single solution) is chosen from the set $\mathcal{W}(\mathbf{W}_k)$ according a fixed rule determined by the solver used for LP. The complete procedure is in Algorithm 1.

**Algorithm 1** Global combined-CMDP algorithm

Initialize $\mathbf{W}_0$ randomly, set $B^* = -\infty$, $\hat{\mathbf{W}}^* = \mathbf{W}_0$; choose a constant $w$; $\mathcal{U}$ set of all possible policies.
**For** $k = 1, 2, \ldots$

$$\mathbf{W}_k \leftarrow \begin{cases} \text{random policy chosen from } \mathcal{U} & \text{w.p. } p_k = \frac{w}{k} \\ \text{Local improvement } (\mathbf{W}_{k-1}) \text{ of (6)} & \text{w.p. } 1 - p_k \end{cases}$$

Calculate $C_1(\mathbf{W}_k)$ and $C_2(\mathbf{W}_k)$
**if** $C_2(\mathbf{W}_k) = 0$ and $C_1(\mathbf{W}_k) \leq \mathbf{b}$ **then**
  Calculate $B(\mathbf{W}_k)$ using (4)
  **if** $B(\mathbf{W}_k) \leq B^*$ **then**
    $\hat{\mathbf{W}}^* \leftarrow \mathbf{W}_k$
    $B^* \leftarrow B(\mathbf{W}_k)$
  **end if**
**else**
  Choose random policy from $\mathcal{U}$ (random restart again)
**end if**

*Complexity comparison*

For finite-horizon problem, one may argue that the problem can be converted to standard linear cost MDP problem, by augmenting the accumulated multiplicative cost as an additional state, one for each multiplicative component, e.g., $Y_{\tau,i} := \prod_{t=0}^{\tau-1} f_{t,i}(X_t, A_t, X_{t+1})$, and then the accumulated multiplicative component becomes terminal cost. The complexity of problem obtained after such an augmentation is high, as the state space grows geometrically with $t$. Similar approach was used in [1]. Our solution, the problem ($\bar{P}$) also augments the state space – but as the augmented variables are binary the state space does not increase geometrically.

The number of unknowns (or decision variables) in LP of [1] grows exponentially in $T$ while that in our BLP grows only linearly. Table I compares the two problems for risk-sensitive MDP with one additive constraint. In [4], we observed that one cannot implement LP of [1] for even small time horizons due to curse of dimensionality, while the algorithm in [5] was implemented comfortably for $T$ even as big as 1000. Algorithm discussed in section IV closely resembles that in [5] and solves the BLP. Thus we anticipate its complexity to be on par with that in [5]; therefore we have an implementable algorithm for not just constrained risk-sensitive MDPs, but also for MDPs with combined costs.

TABLE I: Comparison with $m = |\mathcal{X}|$ and $n = |\mathcal{A}|$

| Reference | No. of Decision Variables | No. of Constraints |
|---|---|---|
| LP as in [1] | $\frac{mn((mn)^T-1)}{(mn-1)}$ | $m + \frac{m((mn)^T-1)}{(mn-1)} + 1$ |
| BLP | $2Tmn$ | $2m(T + Tn - n) + 1$ |

## V. Proof of Theorems

The mapping $\eta \mapsto \Gamma(\eta)$ given in the hypothesis between the domains of the problems ($P$) and ($\bar{P}$) is bijective. Therefore, to prove the first result of theorem 1, it is enough to prove that the corresponding objective and constraint costs are also equal, when respectively started in states $s$ and $(s, \mathbf{1})$, which the below theorem asserts .

*Theorem 3:* For the policies $\eta \in \bar{\Pi}$ and its corresponding policy $\Gamma(\eta) \in \Pi_{\text{MR}}$, the following holds:

$$v^{\eta}_{T,i}(s, \mathbf{1}) = w_i^{\Gamma(\eta)}(s) \text{ for all } i. \qquad \blacksquare \qquad (7)$$

Towards Theorem 2, observe that the problem ($\bar{P}$) is a standard CMDP, but for the domain of optimization. LP based techniques are well known to solve standard CMDPs ( [7], [8]). But $\bar{\Pi}$ includes only a special class of policies which are indifferent to the augmented state $z$. This restriction is captured via the bilinear constraint of the BLP (1), where the variables are indifferent to the value of $z$. The rest of the BLP without this particular condition is same as in [12] extended to the constraints. The variable $w_t(x, z, a)$ is representative of the probability that the system is in state $(x, z)$ at time $t$ and decision $a$ is made. The rest of the proof of Theorem 2 is given in the Appendix. $\blacksquare$

## Conclusions

Many applications require sequential decision models involving a combination of multiplicative and additive cost components which can be formulated as CMDP. One can convert them to standard MDP models, that was recently done (for risk-sensitive MDPs) by augmenting the state space by taking the total costs to cover the multiplicative costs; but

the complexity grows exponentially in time horizon making such problems unsolvable even with moderate time horizons. This paper fills this gap and makes an important contribution addressing not only risk-sensitive MDPs, but also MDPs involving a combination of multiplicative and additive cost components in objective and/or constraint. We address this by augmenting state space with binary variables such that the number of unknowns in the resulting optimization problem grow linearly in time horizon. An implementable algorithm is provided to solve such CMDPs.

## REFERENCES


[1] A. Kumar, V. Kavitha, and N. Hemachandra, "Finite horizon risk sensitive MDP and linear programming," in *Proc. 54th IEEE Conference on Decision and Control CDC*, Osaka., Japan, 2015, pp. 7826-7831.

[2] E. Altman, V. Kavitha, F. D. Pellegrini, K. Vijay, and V. Borkar, "Risk sensitive optimal control framework applied to delay tolerant networks", in *Proc. InfoCom*, Shanghai, China, 2011, pp. 3146-3154.

[3] E. Altman, T. Basar, and V. Kavitha. "Adversarial control in a delay tolerant network", in *Proc., Decision and Game Theory for Security: First International Conference, GameSec*, Berlin, Germany, 2010, pp. 87-106.

[4] M. U. Kumar, S. P. Bhat, V. Kavitha and N. Hemachandra, "Approximate solutions to constrained risk-sensitive Markov decision processes," *EJOR. European Journal of Operational Research*, vol. 310, Issue. 1, pp. 249-267, October 2023.

[5] V. Singh and V. Kavitha, "Fixed-Point equations solving risk-sensitive MDP with constraint," in *Proc. American Control Conference*, San Diego, California, USA, 2023, pp. 3409-3414.

[6] V. Singh, K. Agarwal, Shubham and V. Kavitha, "Evolutionary vaccination games with premature vaccines to combat ongoing deadly pandemic", in *Proc. Performance Evaluation Methodologies and Tools. VALUETOOLS*, Guangzhou, China, 2021, pp. 185-206.

[7] M.L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New Jersey, USA, John Wiley & Sons, 2005.

[8] E. Altman, *Constrained Markov Decision Processes*. Chapman & Hall C, 1999.

[9] P. Geibel and F. Wysotzki, "Risk-sensitive reinforcement learning applied to control under constraints," *Journal of Artificial Intelligence,* vol. 24, issue. 1, pp. 81-108, July 2005.

[10] J. Gorski, F. Pfeuffer and K. Klamroth. "Biconvex sets and optimization with biconvex functions: a survey and extensions," *Mathematical Methods of Operations Research*, vol. 66, pp. 373–407, 2007.

[11] A. Pepelyshev, A. Zhigljavsky and A. Žilinskas, "Performance of global random search algorithms for large dimensions," *Journal of Global Optimization,* vol. 71, pp 57–71, 2018.

[12] A. Bhattacharya and J. P. Kharoufeh, "Linear programming formulation for non-stationary, finite-horizon Markov decision process models," *Operations Research Letters,* vol. 45, pp. 570-574, 2017.

[13] M.U. Kumar, S.P. Bhat, V. Kavitha and N. Hemachandra, "Finite-Horizon Constrained MDPs With Both Additive And Multiplicative Utilities", 2023. https://doi.org/10.48550/arXiv.2303.07834


## VI. APPENDIX

We provide below two basic analysis-based lemmas without proof (proofs in [13]). Proof of Lemma 1 is via mathematical induction, while that of Lemma 2 is straight forward.

*Lemma 1:* Let $N \in \mathbb{Z}^+$ and $q \in \mathbb{R}^N$ be given. Then,

$$\sum_{\delta \in \{0,1\}^N} \prod_{i=1}^{N} q_i^{\delta_i} (1-q_i)^{(1-\delta_i)} = 1. \qquad \blacksquare \qquad (8)$$

*Lemma 2:* Let $f : A \to \mathbb{R}$ and $g : B \to \mathbb{R}$ be two mappings. Assume, i) for each $a \in A$ there exists $b_a \in B$ such that $f(a) = g(b_a)$, and, ii) for each $b \in B$ there exists $a_b \in A$ such that $f(a_b) = g(b)$. Then $\min_{a \in A} f(a) = \min_{b \in B} g(b)$ and

$$\text{if } a^* \in \arg\min_{a \in A} f(a) \text{ then } b_{a^*} \in \arg\min_{b \in B} g(b), \text{ and similarly,}$$
$$\text{if } b^* \in \arg\min_{b \in B} g(b), \text{ then } a_{b^*} \in \arg\min_{a \in A} f(a). \qquad \blacksquare$$

*Proposition 1:* Fix $j \in \{0, 1, \ldots, K\}$. Denote $\mathcal{Z}_j := \{z = (z_i)_{i=0}^K \in \{0,1\}^{K+1} : z_j = 1\}$. Let $z \in \mathcal{Z}_j$. For any $t, x, x'$ and $a$, we have the following identity

$$\sum_{z' \in \mathcal{Z}_j} \bar{Q}_t\Big((x', z') \Big| (x, z), a\Big) = f_{t-1,j}(x, a, x') Q(x'|x, a), \quad (9)$$

$$\sum_{z' \in \{0,1\}^{K+1}} \bar{Q}_t\Big((x', z') \Big| (x, z), a\Big) = Q(x'|x, a). \quad (10)$$

**Proof:** Let $I := \{i \in \{0, 1, \ldots, K\} : z_i = 1\}$. Clearly, $I \neq \emptyset$ because $j \in I$. Observe that, in the summation appearing in left hand side of (9), the summands where $z_i = 0 = 1 - z'_i$, for some $i \in \{0, 1, \ldots, K\}$, is 0 and so they don't contribute to the total. Therefore, we can restrict the summation to only those $z' = (z'_i)_{i=0}^K \in \mathcal{Z}_j$ which satisfy the property that $z'_i = 0$ for $i \notin I$. Denote $\mathcal{Z}'_j := \{z' = (z'_i)_{i=0}^K \in \mathcal{Z}_j : z'_i = 0, \text{ for all } i \notin I\}$. Thus, left hand side of (9) simplifies as below, where we suppress the parameters $x, a, x'$ used in immediate cost functions $f_{t-1,i}$:

$$\sum_{z' \in \mathcal{Z}'_j} \bar{Q}_t\Big((x', z') \Big| (x, z), a\Big)$$

$$= Q(x'|x, a) \sum_{z' \in \mathcal{Z}'_j} \left( \prod_{i=0}^{K} (f_{t-1,i})^{z_i z'_i} (1 - f_{t-1,i})^{z_i (1 - z'_i)} \right)$$

$$= f_{t-1,j} Q(x'|x, a) \sum_{z' \in \mathcal{Z}'_j} \left( \prod_{i \in I \setminus \{j\}} (f_{t-1,i})^{z_i z'_i} (1 - f_{t-1,i})^{z_i (1 - z'_i)} \right)$$

$$= f_{t-1,j} Q(x'|x, a) \sum_{z' \in \mathcal{Z}'_j} \left( \prod_{i \in I \setminus \{j\}} (f_{t-1,i})^{z'_i} (1 - f_{t-1,i})^{(1 - z'_i)} \right)$$

$$= f_{t-1,j} Q(x'|x, a).$$

We used Lemma 1 in the last equality. Thus proving (9). The proof of (10) is easy to verify, therefore skipped. For more detailed explanations regarding the proof of this proposition, see [13]. $\blacksquare$

*Proposition 2:* Let $\eta \in \bar{\Pi}$. For $0 \leq t \leq T-1$, we have for all $x, x', a$,

$$\sum_{z \in \mathcal{Z}} P^{\eta}_{(s,1)}((X_t = x, Z_t = z), A_t = a) = P_s^{\Gamma(\eta)}(X_t = x, A_t = a).$$

**Proof:** Let $\eta = \{\bar{d}_t\}_t$ and $\Gamma(\eta) = \{d_t\}_t$. We prove the proposition using mathematical induction. At $t = 0$, it is easy to prove that left and right hand side of the above equation equals $d_0(a|s)\mathbb{1}_{\{x=s\}}$. Suppose at time $t-1$, the proposition holds true. The identity at $t$ holds by following the below steps:

$$\sum_{z\in\mathcal{Z}} P^{\eta}_{(s,\mathbf{1})}(X_t=x,Z_t=z,A_t=a)$$
$$=\sum_{z\in\mathcal{Z}}\bar{d}_t(a|(x,z))P^{\eta}_{(s,\mathbf{1})}\Big(X_t=x,Z_t=z\Big)$$
$$=\sum_{z\in\mathcal{Z}} d_t(a|x)P^{\eta}_{(s,\mathbf{1})}\Big((X_t=x,Z_t=z)\Big)$$
$$=d_t(a|x)\sum_{x',z',a'}\sum_{z\in\mathcal{Z}}\bar{Q}_t\Big((x,z)|(x',z'),a'\Big)$$
$$P^{\eta}_{(s,\mathbf{1})}\Big(X_{t-1}=x',Z_{t-1}=z',A_{t-1}=a'\Big)$$
$$=d_t(a|x)\sum_{x',z',a'}Q(x|x',a')P^{\eta}_{(s,\mathbf{1})}\Big(X_{t-1}=x',Z_{t-1}=z',A_{t-1}=a'\Big)$$
$$=d_t(a|x)\sum_{x',a'}Q(x|x',a')P^{\Gamma(\eta)}_{s}(X_{t-1}=x',A_{t-1}=a')$$
$$=P^{\Gamma(\eta)}_{s}(X_t=x,A_t=a).$$

Induction hypothesis and (10) gives penultimate two equalities.■

**Proof of Theorem 3:** The left hand side of (7) equals,

$$v^{\eta}_{T,i}(s,\mathbf{1})=\sum_{t=0}^{T-1}\mathbb{E}^{\eta}_{(s,\mathbf{1})}[\bar{r}_{t,i}((X_t,Z_t),A_t)]+\mathbb{E}^{\eta}_{(s,\mathbf{1})}\Big[\bar{r}_{T,i}(X_T,Z_T)\Big]. \tag{11}$$

Using Proposition 2, the first term in (11) simplifies to

$$\sum_{t=0}^{T-1}\mathbb{E}^{\eta}_{(s,\mathbf{1})}[\bar{r}_{t,i}((X_t,Z_t),A_t)]$$
$$=\sum_{t=0}^{T-1}\sum_{x,z,a}\bar{r}_{t,i}((x,z),a)P^{\eta}_{(s,\mathbf{1})}\Big((X_t=x,Z_t=z),A_t=a\Big)$$
$$=\sum_{\substack{t,x,\\x',z,a}} r_{t,i}(x,a,x')Q(x'|x,a)P^{\eta}_{(s,\mathbf{1})}\Big(X_t=x,Z_t=z,A_t=a\Big)$$
$$=\sum_{t,x,x',a} r_{t,i}(x,a,x')Q(x'|x,a)P^{\Gamma(\eta)}_{s}(X_t=x,A_t=a)$$
$$=\sum_{t=0}^{T-1}\mathbb{E}^{\Gamma(\eta)}_{s}[r_{t,i}(X_t,A_t,X_{t+1})]=\mathbb{E}^{\Gamma(\eta)}_{s}\left[\sum_{t=0}^{T-1} r_{t,i}(X_t,A_t,X_{t+1})\right].$$

Define for each $i\in\{0,1,\ldots,K\}$, $\mathcal{Z}_i:=\{z\in\mathcal{Z}:z_i=1\}\subset\{0,1\}^{K+1}$. We denote the decisions of the policy $\eta$ by $\{\bar{d}_t\}_t$ and that of policy $\Gamma(\eta)$ by $\{d_t\}_t$. To simplify the terminal cost in (11), we sum the sample path probabilities:

$$\mathbb{E}^{\eta}_{(s,\mathbf{1})}\Big[\bar{r}_{T,i}(X_T,Z_T)\Big]=\alpha_i\mathbb{E}^{\eta}_{(s,\mathbf{1})}[Z_{T,i}]=\alpha_i P^{\eta}_{(s,\mathbf{1})}[Z_{T,i}=1]$$
$$=\alpha_i\sum_{\substack{x_t\in\mathcal{X},\,t=0\\ z_t\in\mathcal{Z}_i,\\ a_t\in\mathcal{A}}}^{T-1}\prod\bar{d}_t(a_t|(x_t,z_t))\bar{Q}_{t+1}((x_{t+1},z_{t+1})|(x_t,z_t),a_t)$$
$$=\alpha_i\sum_{\substack{x_t\in\mathcal{X},\,t=0\\ z_t\in\mathcal{Z}_i,\\ a_t\in\mathcal{A}}}^{T-1}\prod d_t(a_t|x_t)\bar{Q}_{t+1}((x_{t+1},z_{t+1})|(x_t,z_t),a_t)$$
$$=\alpha_i\sum_{\substack{x_t\in\mathcal{X},\\ a_t\in\mathcal{A}}} d_t(a_t|x_t)Q(x_{t+1}|x_t,a_t)f_{t,i}(x_t,a_t,x_{t+1})$$
$$=\alpha_i\mathbb{E}^{\Gamma[\eta]}_{s}\left[\prod_{t=0}^{T-1} f_{t,i}(X_t,A_t,X_{t+1})\right].$$

The penultimate equality above is due to successive application of (9) backwards in time for $t=T,T-1,\ldots,1$. Replacing back each of the expectation operation terms in (11) proves the identity (7). ■

**Proof of Theorem 2:** Let $\bar{\Pi}_{\mathrm{MR}}$ denote space of all MR policies in the MDP $\bar{M}$. Recall $\bar{\Pi}$ is a set of policies that are indifferent to augmented component $z$. Note that $\bar{\Pi}\subset\bar{\Pi}_{\mathrm{MR}}$. Denote the feasible region of the BLP by $\mathcal{Q}$. Observe that objective function and all the constraints except the bilinear constraints are linear and therefore the problem BLP without the bilinear constraints is indeed a LP that solves the CMDP ($\bar{P}$) with the domain $\bar{\Pi}_{\mathrm{MR}}$, instead of $\bar{\Pi}$. Denote the feasible region of this LP by $\mathcal{L}$. Clearly $\mathcal{Q}\subseteq\mathcal{L}$.

We first claim that the feasible region of the problem ($\bar{P}$) is bijective to $\mathcal{Q}$ by defining the mappings $\pi\mapsto w_\pi$ and $w\mapsto\pi_w$ respectively between these two sets.

It is easy to see that, given a feasible vector $w=\{\mathbf{W}_1,\mathbf{W}_2\}\in\mathcal{Q}$, constructing the policy $\pi_w=\{d_t\}_t$ by rationalising over $a$ and applying the bilinear constraints immediately as below

$$\begin{aligned} d_t(a|(x,z)) &:= \frac{w_t(x,z,a)}{\sum_{a'}w_t(x,z,a')}=\frac{w_t(x,\mathbf{1},a)}{\sum_{a'}w_t(x,\mathbf{1},a')}\\ &= d_t(a|(x,\mathbf{1})),\end{aligned}$$

makes $\pi_w\in\bar{\Pi}$. Also we know that, for a given feasible policy of the (augmented) problem ($\bar{P}$), say, $\pi=\{d_t\}_t\in\bar{\Pi}\subset\bar{\Pi}_{\mathrm{MR}}$, there exists a feasible vector $w_\pi=\{\mathbf{W}_1,\mathbf{W}_2\}\in\mathcal{L}$ with $w_t(x,z,a):=P^{\pi}_{s}(X_t=x,Z_t=z,A_t=a)$ ( [7]). Now, to prove $w_\pi\in\mathcal{Q}$, it is enough to prove that $w_\pi$ satisfies the bilinear constraints.

From the literature on linear MDPs ( [1], [7], [12]), we know that, the mappings $w\mapsto\pi_w$ and $\pi\mapsto w_\pi$ are such that $w_{\pi_w}=w$ and $\pi_{w_\pi}=\pi$.

Choose any arbitrary $z\in\mathcal{Z}$, a $K+1$ dimensional vector of 0s and 1s. Then, the $t$-th decision w.r.t the policy $\pi_{w_\pi}$ for the two states $(x,z_1),(x,\mathbf{1})\in\bar{\mathcal{X}}$ is given by

$$d_t(a|(x,z))=\frac{(w_\pi)_t(x,z,a)}{\sum_{a'}(w_\pi)_t(x,z,a')},\ d_t(a|(x,\mathbf{1}))=\frac{(w_\pi)_t(x,\mathbf{1},a)}{\sum_{a'}(w_\pi)_t(x,\mathbf{1},a')}. \tag{12}$$

Because $\pi_{w_\pi}=\pi\in\bar{\Pi}$, the decisions are indifferent to the vector $z$ and $\mathbf{1}$, that is, $d_t(a|(x,z))=d_t(a|(x,\mathbf{1}))$ implying, $(w_\pi)_t(x,z,a)\sum_{a'}(w_\pi)_t(x,\mathbf{1},a')=(w_\pi)_t(x,\mathbf{1},a)\sum_{a'}(w_\pi)_t(x,z,a')$. Thus satisfying the bilinear constraint.

We state below two claims whose proof is given in [13].

**Claim 1:** For a given feasible policy $\pi$ of the augmented problem ($\bar{P}$), the objective function of ($\bar{P}$), $v^{\pi}_{T,0}(s,\mathbf{1})$ at $\pi$ is equal the objective function of the BLP when evaluated at its feasible point $w_\pi$.

**Claim 2:** For a given feasible vector $w=(\mathbf{W}_1,\mathbf{W}_2)$ of the BLP, the objective function BLP$(\mathbf{W}_1,\mathbf{W}_2)$ at $w$ is equal to the objective function of augmented problem ($\bar{P}$) $v^{\pi_w}_{T,0}(s,\mathbf{1})$ at $\pi_w$.

The above two claims complete the proof of the theorem, by applying the Lemma 2. ■