Risk-constrained Markov Decision Processes

Vivek Borkar & Rahul Jain

Abstract—We propose a new constrained Markov decision process framework with risk-type constraints. The risk metric we use is Conditional Value-at-Risk (CVaR), which is gaining popularity in finance. It is a conditional expectation but the conditioning is defined in terms of the level of the tail probability. We propose an iterative offline algorithm to find the risk-constrained optimal control policy. A stochastic approximation-inspired ‘learning’ variant is also sketched.

Index Terms—Constrained Markov decision processes; Risk measures; Stochastic Approximations.

I. INTRODUCTION

The theory of single-stage stochastic programming is fairly well-developed [6]. Equally well developed is the theory of (unconstrained) stochastic dynamic programming for sequential optimization [4]. Marrying the two theories to develop a theory of sequential stochastic optimization with general constraints has proved to be challenging [31]. Extending stochastic programming methods to multi-stage problems has remained challenging [32] while attempts to extend dynamic programming approaches to optimization with constraints have been successful only when constraints have special form [7], [2].

In the theory of constrained Markov decision processes developed so far [2], both the objective function and constraints have the same form, usually an expectation of a sum, discounted or averaged. This allows for introduction of an occupation measure. The sequential optimization problem can then be formulated as a convex program. When the objective and constraint functions have different forms, this technique does not work. Moreover, the occupation measure technique cannot be generalized to probabilistic and general conditional expectation (of the form we will shortly see).

Algorithms for stochastic programs with probabilistic/chance or conditional expectation constraints have been developed [25]. However, as with other stochastic programs, it is very difficult to extend these to a multi-stage sequential optimization setting. Our interest is in solving a sequential stochastic optimization problem with conditional expectation constraints. Our motivation comes from financial risk management. Thus, the constraint is on a risk measure (i.e., Conditional Value at Risk) which has a conditional expectation form, but is a bit unusual in that the conditioning event is determined by a constraint on a probability (i.e., Value at Risk).

Value-at-Risk (VaR) is a popular risk metric in finance. For a real-valued random variable $Y$ on some probability space $(\Omega, \mathcal{F}, P)$, the VaR at level $\alpha$, $\beta_\alpha = \beta_\alpha(Y)$ is defined to be $\arg \sup_{\beta} P(Y > \beta) \leq \alpha$. Unfortunately, it has many shortcomings including the fact that it is not subadditive, i.e., the VaR of a portfolio may be greater than the sum of VaRs of the portfolio constituents. Thus, another measure called Conditional VaR (CVaR) was introduced by Artzner, et al. [3], which is defined as $\mathbb{E}[Y | Y > \beta_\alpha]$, and depends on the VaR at level $\alpha$. It has been shown that CVaR is a coherent risk measure, i.e., it is convex, monotone, positively homogeneous and translation equivariant, and is amenable to standard methods of stochastic programming for its optimization [26]. However, often the risk must be measured and optimized not just of the current portfolio, but over time in a sequential manner. Motivated by this, we introduce a constrained Markov decision model where the constraint is on a CVaR-type conditional expectation.

Consider a Markov decision process (MDP) defined on state space $X$, a control space $U$, with a reward function $r(x,u)$ and a cost function $c(x,u)$ where $x \in X$ and $u \in U$, a transition function $P_u(dx|x)$ and a finite horizon $T$. Let $Y(T+1) = \sum_{t=0}^{T} c(X_t, u_t)$. Then the CVaR at level $\alpha$ is given by $\Xi_\alpha(Y(T+1)) = \mathbb{E}[Y(T+1)|Y(T+1) > \beta_\alpha(Y(T+1))]$. We would like to maximize the expected reward over a finite time horizon subject to an upper bound on the conditional expectation, CVaR: $\max_u \{ \mathbb{E}[\sum_{t=0}^{T} r(X_t, u_t)] : \Xi_\alpha(Y(T+1)) \leq C_\alpha \}$.

This is useful when decision-making may involve multiple objectives, one say a reward, and another a cost. The goal is to maximize the expected total reward over a certain time horizon while making sure the conditional expectation of the total cost given the total cost exceeds some given level remains bounded. An example of this is the re-insurance business where the re-insurance companies want to collect premiums (the rewards) by providing re-insurance coverage while ensuring that in the case of rare (that have probability less than $\alpha$) but catastrophic events (e.g., natural calamities such as devastating hurricanes or floods), the expected payouts (the costs) remain bounded.

Standard methods for solving Constrained MDPs such as the occupation measure technique and the Lagrangian method [2] cannot deal with conditional expectation constraints of CVaR type. In this paper, we give an offline multiple time-scale iterative algorithm to solve this problem. We prove its convergence under certain assumptions. We then propose an online stochastic approximations-based learning...
algorithm.

**Literature overview of CMDPs and Stochastic Programming for Risk Analysis**

Constrained Markov Decision Process (MDP) models were first introduced by Derman and Klein [14] in the 1960s. It had been noticed that such models were not generally amenable to solution by dynamic programming. Thus, linear programming based solutions using an occupation measure approach (which had already been developed for dynamic programming formulations [13]) were proposed. This was extended by Kallenberg and others [18] to discounted cost, total cost, and average cost criterion for MDPs with a unichain structure. Borkar [7], Altman and others [16], [2] further generalized this approach to average cost with general multi-chain ergodic structure. A second method based on a Lagrangian approach was developed in [5] for MDPs with a single constraint, and extended to multiple constraints in [1]. A third method based on a linear program mixing stationary deterministic policies was developed in [1], [27], [15]. CMDP models with different discount factors for different constraints tend to be much more difficult, and usually optimal stationary policies do not exist though it has been shown that the optimal policies are eventually stationary [16]. Sample path formulations of MDPs were introduced in [28] and shown to satisfy Bellman’s principle of optimality [17]. Alternative solution approaches based on stochastic approximations have also been proposed in [22]. However, almost all of this literature [2] is focused on expectation approximations.

In many applications, other kinds of constraints also appear, such as probabilistic constraints, and more generally stochastic dominance constraints [12]. For example, it is said that ‘Wall Street wants a lot of potential profit on the upside, and not much risk of losses on the downside’; i.e., it wants to maximize profits subject to bounds on the risk of losses exceeding a certain amount. One popular measure of risk is the Value-at-Risk (VaR) metric, which is the smallest exceeding a certain amount. One popular measure of risk, and not much risk of losses on the downside’, i.e., it wants ‘Wall Street wants a lot of potential profit on the upside, and not much risk of losses on the downside’, i.e., it wants ‘Wall Street wants a lot of potential profit on the upside, and not much risk of losses on the downside’, i.e., it wants ‘Wall Street wants a lot of potential profit on the upside, and not much risk of losses on the downside’, i.e., it wants ‘Wall Street wants a lot of potential profit on the upside, and not much risk of losses on the downside’, i.e., it wants ‘Wall Street wants a lot of potential profit on the upside, and not much risk of losses on the downside’, i.e., it wants ‘Wall Street wants a lot of potential profit on the upside, and not much risk of losses on the downside’, i.e., it wants ‘Wall Street wants a lot of potential profit on the upside, and not much risk of losses on the downside’, i.e., it wants ‘Wall Street wants a lot of potential profit on the upside, and not much risk of losses on the downside’, i.e., it wants ‘Wall Street wants a lot of potential profit on the upside, and not much risk of losses on the downside’, i.e., it wants ‘Wall Street wants a lot of potential profit on the upside, and not much risk of losses on the downside’, i.e., it wants ‘Wall Street wants a lot of potential profit on the upside, and not much risk of losses on the downside’, i.e., it wants ‘Wall Street wants a lot of potential profit on the upside, and not much risk of losses on the downside’, i.e., it wants ‘Wall Street wants a lot of potential profit on the upside, and not much risk of losses on the downside’, i.e., it wants ‘Wall Street wants a lot of potential profit on the upside, and not much risk of losses on the downside’, i.e., it wants ‘Wall Street wants a lot of potential profit on the upside, and not much risk of losses on the downside’, i.e., it wants ‘Wall Street wants a lot of potential profit on the upside, and not much risk of losses on the downside’, i.e., it wants ‘Wall Street wants a lot of potential profit on the upside, and not much risk of losses on the downside’, i.e., it wants ‘Wall Street wants a lot of potential profit on the upside, and not much risk of losses on the downside’, i.e., it wants ‘Wall Street wants a lot of potential profit on the upside, and not much risk of losses on the downside’, i.e., it wants ‘Wall Street wants a lot of potential profit on the upside, and not much risk of losses on the downside’, i.e., it wants

The paper is organized as follows. In II, we introduce the problem formulation. Section III then presents preliminary results while section IV presents an offline iterative quasi-gradient stochastic-approximation inspired algorithm. In section V, we present an online learning algorithm. These results are all for a finite horizon. Section VI discusses further work.

**II. Problem Formulation**

Consider a compact metric state space $X$, a compact metric control space $U$, a continuous reward function $r(x,u)$, a continuous cost function $c(x,u)$ where $x \in X$ and $u \in U$, a controlled transition function $P_u(dx'|x,u)$ continuous in $(x,u)$, and a finite horizon $T$. Time is discrete and starts at 0. We will denote a policy by $u = u^T = (u_1, \ldots, u_T)$, where $u_t$ is the control applied at time $t$ according to this policy. We will denote $P_u$ as the probability measure on the finite horizon process $X^T = (X_0, \ldots, X_T)$ under control policy $u$. Only noisy observations of the cost are available. Thus, given a zero-mean i.i.d. noise process $\{\xi_t\}$ with strictly positive density $\phi$, we define the cumulative cost process $\{Y_t\}$ as

$$Y_0 = 0, Y_{t+1} = Y_t + c(X_t, u_t) + \xi_{t+1}. \quad (1)$$

We define the Value-at-Risk (VaR) function $\beta_\alpha = \beta_\alpha(Y)$ for a random variable $Y$ as

$$\beta_\alpha(Y) = \arg \sup_\beta \mathbb{P}(Y > \beta) \leq \alpha,$$

with $\alpha \in (0,1)$ and is typically close to 0 such as 0.1, 0.05, or 0.01.

The Conditional-Value-at-Risk (CVaR) function $\Xi_\alpha$ is defined as

$$\Xi_\alpha(Y) = \mathbb{E}[Y | Y > \beta_\alpha(Y)].$$

Then, our objective is to maximize the expected total reward over the finite horizon subject to the CVaR of the terminal cost being bounded by some constant $C_\alpha$.

$$\text{rMDP}_\alpha : \max_u \quad \mathbb{E}[\sum_{t=0}^{T} r(X_t, u_t)] \quad \text{s.t.} \quad \Xi_\alpha(Y_{T+1}) \leq C_\alpha. \quad (2)$$

As discussed earlier, such a formulation is useful when the decision-maker wants to maximize expected total reward (first objective) over a finite horizon, while making sure that the conditional expectation of the total cost (the second objective) given that it lies in the $\alpha$-probability tail does not exceed a deterministic bound $C_\alpha$.

**Proposition 1:** If $\{\xi_t\}$ is an i.i.d. noise process with strictly positive density $\phi$, then a solution $u^*$ of $\text{rMDP}_\alpha$ exists.

**Proof:** Existence of an optimal solution of the above follows by standard ‘compactness–continuity’ arguments as follows. $\mathbb{E}[\sum_{t=0}^{T} r(X_t, u_t)]$ is clearly a continuous functional of the law $L$ of $(X_t, Y_t, u_t), t \geq 0$. Since $Y_{T+1}$ has a strictly
positive density by our hypotheses on \( \{ \xi_t \} \), its distribution function \( F(\cdot) \) is continuous and strictly increasing. Thus \( \beta_\alpha = F^{-1}(1-\alpha) \). Now

\[
\Xi_\alpha(Y_{T+1}) = E[|Y_{T+1} > \beta_\alpha(Y_{T+1})|
\]
\[
= E[|Y_{T+1} | > F^{-1}(1-\alpha)] / P(Y_{T+1} > F^{-1}(1-\alpha))
\]
\[
= \frac{1}{\alpha} E[|Y_{T+1} | > F^{-1}(1-\alpha)]
\]

In view of the foregoing, this is a continuous functional of \( \mathcal{L} \). Since \( x_0 \) is fixed, \( (x,u) \rightarrow P(dx'|x,u) \) is continuous, and \( \mathcal{X} \) and \( \mathcal{U} \) compact, it follows that the set of attainable \( \mathcal{L} \) is compact. Thus the constraint set is also compact and hence the objective attains its maximum on it. 

III. Preliminaries

Consider a controlled Markov chain \( Z_t = (X_t, Y_t, v_t), t \geq 0 \), with control process \( u_t, t \geq 0 \), where \( v_0 = u_{t-1} \). The combined evolution of this three-component controlled Markov chain is determined by the transition kernel \( p(dx'|x_t, u_t) \), the evolution equation (1) for \( Y_t \), and the equation \( v_t = u_{t-1} \). We will assume \( z_0 = (x_0, 0, v_0) \) to be deterministic. The combined transition kernel will be denoted by \( \bar{p}(dz'|z, u) \).

We shall assume the cost function to be separable, i.e.,
\[
c(x,u) = c_1(x) + c_2(u).
\]
We will set \( u_{t-1} = u_T = \bar{u}, \bar{u} \) being an additional element we add to \( \mathcal{U} \) with \( c_2(\bar{u}) = 0 \). This does not alter the definition of \( Y_{T+1} \), the terminal cost. For a given \( \beta \), define the state-value-at-risk function as
\[
V_t(z) := P(Y_{T+1} > \beta | Z_t = z).
\]
Then, we can write a backward recursion equation for \( V_t(z) \) as
\[
V_t(z) = \int \bar{p}(dz'|z, v) V_{t+1}(z').
\]
Denote \( \bar{c}(z_t) = c(x_t, v_t) = c_1(x_t) + c_2(v_t) \). Then, similarly, we can express CVaR in terms of the state-value function.

**Lemma 1:** If \( V_0(z_0) > 0 \), then we have that
\[
\Xi_\alpha(Y_{T+1}) = \frac{1}{V_0(z_0)} E \left[ \sum_{t=0}^{T} \bar{c}(Z_t) V_t(Z_t) \right].
\]

**Proof:** By definition, we have
\[
E[Y_{T+1} | Y_{T+1} > \beta] = \sum_{t=0}^{T} \sum_{z_t} P(\bar{c}(z_t) P(Z_t = z_t | Y_{T+1} > \beta) c(x_t, v_t),
\]

where we have used the separability of the cost function. Then,
\[
\Xi_\alpha(Y_{T+1}) = E[Y_{T+1} | Y_{T+1} > \beta_\alpha]
\]
\[
= \sum_{t=0}^{T} \sum_{z_t} \bar{c}(z_t) P(Z_t = z_t | Y_{T+1} > \beta_\alpha) \]
\[
= \sum_{t=0}^{T} \sum_{z_t} \bar{c}(z_t) P(Y_{T+1} > \beta_\alpha | Z_t = z_t)
\]
\[
\times P(Z_t = z_t) / P(Y_{T+1} > \beta_\alpha | Z_t = z_t)
\]
\[
= \frac{1}{\alpha} \sum_{t=0}^{T} \bar{c}(Z_t) V_t(Z_t).
\]

Note that \( V_0(z_0) = P(Y_{T+1} > \beta | Z_t = z_0) = \alpha \). Thus, the constraint in \( \text{rMDP}_\alpha \) becomes
\[
\frac{1}{\alpha} E \left[ \sum_{t=0}^{T} \bar{c}(z_t) V_t(z_t) \right] \leq C_\alpha.
\]

IV. An Offline Iterative Algorithm for the RMDP Problem

We now present a multiple time-scale iterative algorithm to solve the RMDP problem: Let \( \{ \gamma_n \}, \{ \eta_n \} \) be strictly positive stepsizes satisfying the conditions \( \gamma_n, \eta_n \downarrow 0, \sum_n \gamma_n = \sum_n \eta_n = \infty \). We specify non-obvious initial conditions. Other variables and functions are assumed to be initiated at the beginning of the algorithm.

**Remarks.** (1) The innermost loop over \( t \) computes \( V_0^n(x_0) = P(Y_{T+1} > \beta), J_0^n(x_0, 0) = \min_u \left( E[\sum_{t=0}^{T} r(X_t, u_t)] + \lambda(C_\alpha - E[Y_{T+1} | Y_{T+1} > \beta]) \right) \) over all non-anticipative controls, for fixed \( \beta = \beta_n^m, \lambda = \lambda_m \). It also computes \( Q^n(x_0, 0) = E[Y_{T+1} | Y_{T+1} > \beta] \) for fixed \( \lambda = \lambda_m \) though it should be possible to do this less often by moving it outside the \( t \) and \( n \) loops.

(2) The middle loop over \( n \) adjusts \( \beta \) till \( V_0^n(z_0) = P(Y_{T+1} > \beta) = \alpha \).

(3) The outer loop over \( m \) adjusts \( \lambda \) till \( E[Y_{T+1} | Y_{T+1} > \beta] \leq \alpha C_\alpha \).

(4) The \( \beta \) and \( \lambda \) iterations can also be done concurrently if we use \( \eta_n = o(\gamma_n) \). Convergence analysis below then still holds using a two two-scale argument [10].

**Convergence Analysis:** All iterations in the innermost loop involve finitely many steps, so do not need any convergence analysis. In practice, the iterations for \( \{ \beta_n, n \geq 0 \}; \{ \lambda_m, m \geq 0 \} \) will also be stopped after finitely many steps according to some stopping rule, but a convergence analysis is required nevertheless to justify such a procedure.

Unfortunately, it does not seem possible to establish convergence of \( \{ \beta_n \} \) in general. (Note that we have suppressed the superscript for notational ease, taking advantage of the
For $m = 1, 2, \cdots$, till convergence:
For $n = 1, 2, \cdots$, till convergence:
For $t = T, \cdots, 0$:
1) $J_{t,1}^m(x, y) = \max_u \left( r(x, u) + \int p(dx'_{t+1}|x, u) J_{t+1}^m(x', y + c(x, u) + s) \phi(s) ds \right)$ with $J_{T+1}^m(x, y) = \lambda_m (C_a - y I (y > \beta_m^m)/\alpha$)
2) $u_{t,1}^m(z) \in \arg \max_u \left( r(x, u) + \int p(dx'|x, u) J_{t+1}^m(x', y + c(x, u) + s) \phi(s) ds \right)$
3) $V_{t,1}^m(z) = \int \hat{p}(dz'|z, u_{t,1}^m(z)) V_{t+1}^m(z')$ with $V_{T+1}^m(z) = I(y > \beta_m^m)$
4) $Q_{t,1}^m(z) = \hat{c}(z) V_{t,1}^m(z)/\alpha + \int \hat{p}(dz'|z, u_{t,1}^m(z)) Q_{t+1}^m(z')$, with $Q_{T+1}^m(z) = y I(y > \beta_m^m)/\alpha$  
\[ \beta_m^m = \beta_m - \gamma_n (C_a - Q_{0}^m(z_0)) \]
End $t$
\[ \lambda_{m+1} = (\lambda_m - \eta_m (C_a - Q_{0}^m(z_0)))^+ \]
End $m$

Algorithm 1 iRMDP

For $m = 1, 2, \cdots$, till convergence:
For $n = 1, 2, \cdots$, till convergence:
For $t = T, \cdots, 0$:
1) $J_{t}^m(x, y) = \max_u \left( r(x, u) + \int p(dx'_{t+1}|x, u) J_{t+1}^m(x', y + c(x, u) + s) \phi(s) ds \right)$ with $J_{T+1}^m(x, y) = \lambda_m (C_a - y I (y > \beta_m^m)/\alpha$)
2) $u_{t}^m(z) \in \arg \max_u \left( r(x, u) + \int p(dx'|x, u) J_{t+1}^m(x', y + c(x, u) + s) \phi(s) ds \right)$
3) $V_{t}^m(z) = \int \hat{p}(dz'|z, u_{t}^m(z)) V_{t+1}^m(z')$ with $V_{T+1}^m(z) = I(y > \beta_m^m)$
4) $Q_{t}^m(z) = \hat{c}(z) V_{t}^m(z)/\alpha + \int \hat{p}(dz'|z, u_{t}^m(z)) Q_{t+1}^m(z')$, with $Q_{T+1}^m(z) = y I(y > \beta_m^m)/\alpha$  
\[ \beta_m^m = \beta_m - \gamma_n (C_a - Q_{0}^m(z_0)) \]
End $t$
\[ \lambda_{m+1} = (\lambda_m - \eta_m (C_a - Q_{0}^m(z_0)))^+ \]
End $m$

fact that it represents a slower time scale and its effect therefore is 'quasi-static' (see [10], section 6.1.) We shall consider a special case, viz., when the function $\beta \rightarrow h(\beta) := P(Y_{T+1} > \beta)$ is Lipschitz continuous. The difficulty is that in addition to the explicit dependence of this probability on $\beta$, there is also the hidden dependence via the underlying control policy, which is harder to decipher. Nevertheless, we assume this condition.

Lemma 2: Suppose the function $h(\beta) = P(Y_{T+1} > \beta)$ is Lipschitz-continuous, then for fixed $m$, as $n \rightarrow \infty$, $V_{0}^n(z_0) \rightarrow \alpha$ for all $z_0$.

Proof: Observe that with Lipschitz-continuity, the fact $\gamma_n \downarrow 0$ implies that the iterates will have the same asymptotic behavior as that of the o.d.e.,
\[ \dot{\beta}(t) = \alpha - h(\beta(t)). \]
For $\beta \approx 0$, $h(\beta) > \alpha$ and for $\beta >> 0$, $h(\beta) < \alpha$. Thus the trajectories of (6) remain bounded and for a scalar o.d.e., this implies convergence to an equilibrium, i.e., a point where $h(\beta) = \alpha$. In view of the aforementioned properties of $h$ and its continuity, at least one such point exists.

In fact, all such points that correspond to downward crossings of the level $\alpha$ by $h$ will be stable equilibria and the rest unstable, and the smallest and the largest equilibrium will be necessarily stable. In general, however, continuity of $h$ cannot be guaranteed. Hence, the iterates track not the o.d.e. (6), but the differential inclusion
\[ \dot{\beta}(t) \in \alpha - H(\beta(t)), \]
where
\[ H(\beta) := c(d(\cap_{\epsilon > 0} \{ P(Y_{T+1} > \beta') : ||\beta' - \beta|| < \epsilon \}). \]
(See, e.g., [10], Chapter 4.) This, in fact is one of the solution concepts for o.d.e.s with discontinuous right hand side, due to Krasovskii [19]. In some cases, this can yield useful information.

Convergence analysis for $\{\lambda_m\}$ is easier.

Lemma 3: Let $\{\eta_m\}$ be strictly positive step-sizes such that $\eta_m \downarrow 0$ and $\sum_m \eta_m = \infty$. Then, as $m \rightarrow \infty$, we obtain $\lambda_m \rightarrow \lambda^*$ and $Q_0^m \rightarrow Q_0^*$ such that
\[ \lambda^* \cdot (C_a - Q_0^*(z_0)) = 0 \quad \text{and} \quad C_a - Q_0^*(z_0) \geq 0, \quad \forall z_0. \]

Proof: Let
\[ G(\lambda) := \max \left( \mathbb{E} \left[ \sum_{t=0}^T r(X_t, u_t) \right] + \lambda (C_a - \mathbb{E}[Y_{T+1}]) \right) \]
for $G(\lambda)$ so that the maximum is over all admissible $\mathcal{L}$. Suppose the maximum is attained by $\mathcal{L}^* := \lambda^*$ and $Q^*_0$. Then, as $m \rightarrow \infty$, the $\mathcal{L}$-iteration is simply an instance of the classical subgradient descent that is known to converge to a global minimum of $g$, in this case the desired Lagrange multiplier.

V. AN ONLINE LEARNING ALGORITHM FOR FINITE RMDPs

We now consider state space $\mathcal{X}$ and control space $\mathcal{U}$ to be finite which will make $\mathcal{Y}$ to be finite as well. We present an online learning algorithm for this setting that finds the optimal control given a sequence of samples. A sample $k$ here is $(X_t^k, Y_t^k, u_t^k, T_k)$ where $X_t^k = (X_0^k, \cdots, X_t^k)$, i.e., the entire state trajectory over the $T$ horizon, $Y_t^k$ is the entire cost trajectory and $u_t^k$ the control sequence that generates it. Besides being an online algorithm, the algorithm below differs from the iterative algorithm presented earlier, namely as it operates at multiple time-scales and thus, along with the Milgrom-Segal envelope theorem [23], $C_a - \mathbb{E}[Y_{T+1}]$ is a valid subgradient thereof. Thus, the $\lambda$-iteration is simply an instance of the classical subgradient descent that is known to converge to a global minimum of $g$, in this case the desired Lagrange multiplier.
Algorithm 2 oRMDP

For $k=1,2,\ldots$

For $t = T, T-1, \ldots, 0$:

1) $J_t^k(x, y, u) = J_{t+1}^{k-1}(x, y, u) + a_k I\{X_{t+1}^k = x, Y_{t+1}^k = y, u_{t+1}^k = u\} \left(r(x, u) + \max_{u'} J_{t+1}^k(X_{t+1}^k, Y_{t+1}^k, u') - J_t^k(x, y, u)\right)$, with $J_{T+1}^k(x, y) = I(Y_{T+1}^k = y) \lambda_k (C_\alpha - yI(y > \beta_k))/\alpha$

2) $u_t^k := v_{t+1}^k = \arg\max J_t^k(X_t^k, Y_t^k, \cdot)$

3) $V_t^k(z) = V_{t+1}^{k-1}(z) + a_k I\{Z_{t+1}^k = z\} (V_{t+1}^k(Z_{t+1}^k) - V_t^k(z))$

4) $Q_t^k(z) = Q_{t+1}^{k-1}(z) + a_k I\{Z_{t+1}^k = z\} (V_{t+1}^k(z)\gamma(z) + Q_{t+1}^k(Z_{t+1}^k) - Q_t^k(z))$ with $Q_{T+1}^k(z) = yV_{T+1}^k(z)/\alpha$

End $t$

$\beta_k = \beta_k - 1 - \gamma_k (\alpha - V_0(z))$

$\lambda_k = (\lambda_k - 1) (C_\alpha - Q_0(z)) + 1$

End $k$

'exploratory' randomization of the action choice will yield a near-optimal rather than optimal control.

VI. DISCUSSION AND FURTHER WORK

In this paper, we have introduced a new class of constrained Markov decision processes. The constraints are of conditional expectation of the terminal value of a total cost functional. The motivation comes from finance, in particular insurance, wherein an insurance company wants to maximize its revenue from premiums subject to a constraint on the conditional expectation of the claims the insurer might have to pay. Since our interest is in risk from catastrophic events such as floods, hurricanes, market crashes which have small probability but result in large claims when they happen. Thus, conditioning in the expectation is on the tail probability (i.e., on events that have the largest claims but together have less than say 5% total probability mass). The problem as formulated is of tremendous interest in finance and risk management.

It is, however, not amenable to solutions techniques available either in stochastic programming, nor theory of constrained Markov decision processes. We, thus, give an iterative/stochastic approximation-based algorithm to solve it. We are currently in the process of acquiring relevant insurance data, and would test the algorithm on such data in further work. In future work, we will also extend the methodology to the infinite-horizon case, both discounted as well as average-case. Our proof of convergence of the offline algorithm also needs an additional assumption. In future work, we will also try to derive a proof that does not need such an assumption. The assumption on separability of the cost function is also crucial. In the future, we shall also consider the general cost structure.

We hope our most immediate contribution is to re-stimulate research in the community on the further development of the theory of constrained MDPs whereby we can handle more general constraints than the current framework can.

ACKNOWLEDGEMENTS

The problem was introduced to the second author by Pu Huang and Dharmashankar Subramanian of the Risk Group in the Mathematical Sciences division of the IBM Watson Research Center. The second author is very grateful to them, and to Alan J. King and Jayant Kalagnanam, also of IBM Watson Research, and to Erim Kardes of USC for many helpful discussions.

REFERENCES


