Método Evolutivo em Horizonte de Planejamento Reduzido e Amostrado para Solução de Processos Markovianos de Decisão

Luiz G. N. Nunes[1, 2], Rita C. M. Rodrigues[2], Solon V. Carvalho[2]

[1]Rede Sarah de Hospitais de Reabilitação
SMHS-501, Conjunto A, Brasília, DF, 70335-901

[2]Instituto Nacional de Pesquisas Espaciais
Av. dos Astronautas, 1.758, Jd. Granja, São José dos Campos, SP, 12227-010

guilherme@sarah.br, rita@lac.inpe.br; solon@lac.inpe.br

Resumo. *Os Processos Markovianos de Decisão (PMDs) têm provado sua utilidade como modelo para problemas de decisões seqüenciais com característica estocástica e propriedades markovianas. Os métodos tradicionais de solução dos PMDs (algoritmos de iteração de políticas e de valores), quando os espaços de estados e/ou de ações são muito grandes, são computacionalmente onerosos ou mesmo impraticáveis. Novas abordagens baseadas em métodos de buscas especializadas, técnicas de amostragem e horizontes de planejamento reduzidos têm sido desenvolvidas recentemente. Neste estudo pretende-se apresentar os conceitos de um novo método para solucionar PMDs com espaço de estados e de ações simultaneamente grandes. Denomina-se o método proposto por evolutivo em horizonte de planejamento reduzido e amostrado (EHRA).*

Abstract. *Markovian Decision Processes (MDPs) have proved their usefulness as models for sequential decision problems with stochastic characteristics and Markovian properties. Traditional methods, as policy and value iteration algorithms, are computationally difficult, or even impracticable, when states and/or action spaces are large. New approaches, based on specialized searches, sample techniques, and receding horizon control, have been developed recently. In this study we present the concepts of a new method to be developed in order to find solutions for MDPs with states and action spaces simultaneously large. We have named the proposed method evolutionary with receding sampled horizon (ERSH).*

1. Introdução

Metodologias para solucionar problemas que envolvem tomadas de decisão sob incerteza são de interesse de diversas áreas de estudo, especialmente a área de pesquisa operacional, e encontram ampla abrangência de aplicações. Dentre as técnicas exploradas nesse contexto, destacam-se os bem conhecidos *processos markovianos de decisão* (PMDs). Os PMDs têm provado sua utilidade como modelo para problemas de decisões seqüenciais com característica estocástica, que apresentam a propriedade markoviana, qual seja, os estados e decisões futuros são independentes dos estados e decisões passados, dado o conhecimento do estado presente do sistema considerado.

As metodologias empregadas para solucionar os PMDs são estudadas há mais de 50 anos, existindo algoritmos bem conhecidos capazes de determinar políticas de decisão ótimas. Dentre os métodos tradicionais para solução de PMDs destacam-se o algoritmo de iteração de políticas (AIP) e o algoritmo de iteração de valores (AIV) (Puterman (1994)). Uma das limitações desses métodos é o tamanho dos espaços de estados e de ações (Littman et al. (1995)). Quando os espaços de estados e/ou de ações são muito grandes, como freqüentemente ocorre em situações práticas, os métodos tradicionais são computacionalmente onerosos ou impraticáveis. Diversas abordagens têm sido

estudadas nas últimas décadas para contornar as dificuldades envolvidas no tratamento desses modelos explosivos, incluindo métodos de decomposição de espaços de estados e de ações e métodos que empregam técnicas de Inteligência Artificial (Bertsekas e Castañon (1989), Boutilier et al. (1995), Dean et al. (1997), Givan et al. (2003), Parr (1998), Singh e Cohn (1998)). No entanto, mesmo essas técnicas tornam-se ineficientes ou inviáveis quando nenhuma representação favorável da estrutura do sistema modelado (espaços de estados e de ações) é conhecida.

Novas linhas de pesquisa promissoras em relação à viabilização de soluções para PMDs de grandes dimensões, que dependem menos da estrutura do sistema modelado, foram desenvolvidas a partir da segunda metade da década de 90. Os métodos dessa nova linha, ao contrário dos métodos tradicionais, têm por objetivo evitar a necessidade de se avaliar exaustivamente os espaços de estados e de ações a cada iteração do processo de solução dos PMDs e, ao mesmo tempo, garantir uma solução próxima à solução ótima. Esta aproximação é realizada através de métodos de buscas especializadas para espaços de ações grandes e de técnicas de amostragem que reduzem o horizonte de planejamento e permitem o tratamento de espaços de estados grandes.

Neste estudo pretende-se apresentar os conceitos para o desenvolvimento de um novo método de solução de PMDs que combina os métodos amostrados em horizonte reduzido e os métodos evolutivos de iteração de políticas. Esta nova metodologia se destina especificamente à solução de *PMDs com espaço de estados e de ações simultaneamente grandes*. Antes que definir uma política ótima com ações predeterminadas para cada estado em cada instante de decisão através de um processo de otimização "off-line", o objetivo do método é, através de uma busca evolutiva no espaço de ações, determinar uma boa *ação de elite* a cada instante de decisão, dado o estado observado do sistema. A cada iteração, a determinação de uma ação de elite pertencente a uma *população de ações* será realizada através de um método amostrado em um horizonte de planejamento reduzido.

O restante deste texto está organizado da seguinte forma. Na seção 2 apresentam-se os fundamentos da teoria dos PMDs, dos métodos amostrados em horizonte reduzido e dos métodos evolutivos de iteração de políticas. Na seção 3 enunciam-se os conceitos para um método combinado, evolutivo em horizonte de planejamento reduzido e amostrado. Nas seções 4 e 5 apresentam-se, respectivamente, uma breve discussão sobre o método proposto e as considerações finais sobre este trabalho.

2. Fundamentos

2.1. Processos Markovianos de Decisão

Um PMD pode ser definido de forma geral por uma *n-upla* (X, A, P, R), onde: X é um conjunto de estados; A é o conjunto de ações aplicáveis em função dos estados; P representa as distribuições de probabilidades de transições entre estados em função de um estado observado e de uma ação adotada em um instante de decisão; e R determina a recompensa associada ao estado observado e à ação adotada. Em outras palavras, em um PMD, dado um instante de decisão qualquer t, observa-se o estado do sistema $x_t \in X$ e adota-se uma ação $a_t \in A(x_t)$, gera-se assim uma recompensa $R(x_t, a_t)$ e estabelece-se a probabilidade de o sistema passar a um estado x_{t+1}, representada por $P(x_{t+1} \mid (x_t, a_t))$, $\forall\ x_{t+1} \in X$.

Pode-se definir uma política de decisão π como sendo uma série de funções $\pi = \{\pi_0, \pi_1, ...\}$, onde $\pi_t : X \to A$ tal que $\pi_t(x_t) = a_t \in A(x_t)$. Se π_t é invariante em relação à t, então π é dita uma política de decisão estacionária. Seja Π o conjunto de todas as políticas de decisão possíveis, o problema que se deseja resolver ao formular um sistema como um PMD é, dado um estado inicial x_0 em um instante de decisão $t = 0$, encontrar uma política de decisão $\pi^* \in \Pi$ que otimize uma função objetiva pertinente ao problema. Os critérios mais utilizados para otimização dos PMDs são:

maximização de uma função de *recompensa total descontada* (com horizonte de planejamento finito ou infinito) e maximização de uma função de *recompensa média* (com horizonte de planejamento infinito).

Considerando-se a maximização de uma função de recompensa total descontada em um horizonte de planejamento infinito, o que se deseja é encontrar uma política ótima $\pi^* \in \Pi$ que maximize a recompensa total descontada esperada sobre um horizonte de planejamento infinito, calculada através da expressão:

$$V_\infty^\pi(x) = E\left\{ \sum_{t=0}^{\infty} \gamma^t R(x_t, \pi_t(x_t)) \right\},$$

onde $\gamma \in (0,1)$ é o fator de desconto e x_0 é o estado inicial observado.

Da teoria geral dos PMDs (Puterman (1994)) é bem conhecido que para o caso da otimização utilizando-se a função de recompensa total descontada em um horizonte de planejamento infinito, se existir pelo menos uma política ótima, existe pelo menos uma política ótima que é estacionária.

Define-se o valor ótimo da função recompensa $V_\infty^* : X \to R$, por:

$$V_\infty^*(x) = \sup_{\pi \in \Pi} \left(V_\infty^\pi(x) \right), \ \forall x \in X,$$

que satisfaz as seguintes equações, conhecidas como *equações de otimalidade de Bellman*:

$$V_\infty^*(x) = \max_{a \in A(x)} \left\{ R(x,a) + \gamma E_y \left(V_\infty^*(y) \right) \right\}, \ \forall x \text{ e } \forall y \in X.$$

Defini-se $Q_\infty(x,a) = R(x,a) + \gamma E_y \left(V_\infty^*(y) \right), \ \forall y \in X$, como a função de utilidade máxima de uma ação $a \in A$ em um estado $x \in X$.

A política definida pela expressão $\pi^*(x) = \arg\max_{a \in A(x)} Q_\infty(x,a)$ é uma política ótima que atinge o valor máximo de retorno esperado $V_\infty^*(x), \ \forall x \in X$.

Pode-se definir também uma função $Q_\infty^\pi(x,a)$ de utilidade da ação a para o estado x, sob uma política $\pi \in \Pi$, substituindo-se na expressão anterior o valor da recompensa máxima esperada alcançável no horizonte de planejamento infinito, $V_\infty^*(y)$, pelo valor de $V_\infty^\pi(y)$.

Pode-se definir ainda uma função $Q_H(x,a)$ de utilidade máxima da ação a para o estado x, a um horizonte de planejamento finito H, tal que

$$Q_H(x,a) = R(x,a) + \gamma E_y \left(V_{H-1}^*(y) \right), \ \forall y \in X.$$

2.2. Controle “on-line” em horizonte de planejamento reduzido

O controle ótimo de um PMD formulado originalmente a horizonte de planejamento infinito pode ser aproximado por um resultado simplificado a horizonte de planejamento reduzido finito (Hernández-Lerma e Lasserre (1990)). Para tal, a cada instante de decisão seleciona-se um horizonte pequeno, mas típico do PMD original, e, otimizando-se a função de recompensa descontada no horizonte reduzido finito, obtém-se uma *ação ótima corrente*. Assim, controla-se o sistema “on-line”.

Define-se formalmente uma política de decisão a horizonte de planejamento reduzido como $\pi_{H,\gamma}^{re} = \{ \pi_t \}, \ t = 0,1,2...$, tal que para todo t:

$$\pi_t(x) = \arg\max_{a \in A(x)} Q_H(x,a), \ \forall x \in X.$$

Bes e Lasserre (1986) demonstraram que sempre existe um horizonte reduzido finito mínimo H, tal que, para $\gamma \in (0,1)$, $\pi^{re}_{H,\gamma}(x) = \pi^*(x)$, $\forall x \in X$, onde π^* é a política ótima, que atinge $\sup_{\pi \in \Pi}\left(V^{\pi}_{\infty}(x_0)\right)$.

2.3. Métodos amostrados em horizonte de planejamento reduzido

Os métodos amostrados em horizonte de planejamento reduzido, H, empregam aproximações do valor da função de utilidade máxima de uma ação a em um estado x, $Q_H(x,a)$, para encontrar políticas de decisão "on-line" e $\varepsilon - ótimas$. A cada instante de decisão, escolhe-se a ação que apresentar o maior valor aproximado $\hat{Q}_H(x,a)$. Os métodos amostrados em horizonte de planejamento reduzido são aplicados especificamente a *PMDs compostos por espaço de estados grande e espaço de ações pequeno*. Isto porque a cada iteração desses métodos é necessário avaliar todas as ações disponíveis, o que os inviabiliza ou os torna desvantajosos quando o espaço de ações é grande.

Nos métodos amostrados em horizonte de planejamento reduzido substitui-se o cálculo oneroso, sobre todo o espaço de estados, da recompensa esperada $E_y\left(V^*_{H-1}(y)\right)$, presente na expressão $Q_H(x,a) = R(x,a) + \gamma E_y\left(V^*_{H-1}(y)\right)$, pelo cálculo simplificado de um valor estimado, obtido a partir de uma amostragem do espaço de estados.

O **método KMN**, desenvolvido por Kearns et al. (2002), obtém $\hat{Q}_H(x,a)$ através da exploração de uma árvore de busca com profundidade limitada (horizonte reduzido) formada por estados aleatoriamente amostrados do espaço de estados original a cada nível da árvore. Explorando a árvore com uma adequada profundidade "H" e largura "C" (tamanho da amostra para cada ação em cada nível da árvore) desenvolve-se uma metodologia de tomada de decisão que determina uma política não estacionária $\varepsilon - ótima$. O erro ϵ em relação à aproximação do valor ótimo da função recompensa fica limitado em função dos valores de H e C. A aproximação do valor da função recompensa converge para o valor ótimo com probabilidade um, à medida que se ampliam a largura e a profundidade da árvore. Neste método a política de decisão é definida por:

$$\hat{\pi}_H(x) = \arg\max_{a \in A(x)} \hat{Q}_H(x,a), \ \forall x \in X,$$

onde $\hat{Q}_H(x,a) = R(x,a) + \gamma \, 1/C \sum_{y \in S_{x,a}} \hat{V}_{H-1}(y)$, sendo $S_{x,a}$ o conjunto de C estados amostrados a partir da distribuição de probabilidade de transição para o próximo estado correspondente à adoção da ação a no estado x, e $\hat{V}_{H-1}(y) = \max_{u \in A(y)}\left\{\hat{Q}_{H-1}(y,u)\right\}$, sendo que $\hat{Q}_H(x,a)$ pode ser obtido recursivamente.

O **método "rollout"**, desenvolvido por Bertsekas e Castañon (1999), parte de uma política de base predeterminada $\pi \in \Pi$ e, através de avaliações de caminhos amostrados do funcionamento do sistema em um horizonte de planejamento reduzido, determina "on-line" uma ação corrente a ser adotada. Mais especificamente, no método "rollout" supõe-se a cada instante de decisão que, a partir do próximo instante de decisão, o sistema será controlado por uma *política de base* π. Esta política de base deve ser determinada previamente através da utilização de uma *boa* heurística. Em um instante de decisão, dado o estado observado $x \in X$, escolhe-se uma ação:

$$\hat{\pi}_H(x) = \arg\max_{a \in A(x)} \hat{Q}_H(x,a),$$

onde $\hat{Q}_H(x,a) = R(x,a) + \gamma \, 1/C \sum_{i=1}^{C} V_i^{\pi,H-1}$, sendo $V_i^{\pi,H-1}$ o valor de retorno do controle da política de base π para o caminho i, e C o número de caminhos amostrados para cada ação avaliada.

O **método "parallel rollout"**, proposto por Chang (2001), oferece uma variação para o método "rollout", empregando não uma única política de base π, mas um conjunto de políticas de base Λ. Garante-se que o desempenho alcançado em cada estado não é pior que o melhor desempenho determinado pela aplicação de qualquer uma das políticas de decisão pertencentes ao conjunto de políticas de base. Em um instante de decisão, escolhe-se uma ação:

$$\hat{\pi}_H(x) = \arg\max_{a \in A(x)} \hat{Q}_H(x,a),$$

onde $\hat{Q}_H(x,a) = R(x,a) + \gamma \frac{1}{C} \sum_{i=1}^{C} \max_{\pi \in \Lambda} V_i^{\pi,H-1}$, sendo $V_i^{\pi,H-1}$ o valor de retorno do controle da política de base $\pi \in \Lambda$ para o caminho i, e C o número de caminhos amostrados para cada ação avaliada.

O **método "hindsight"**, desenvolvido por Chong et al. (2000), apresenta uma abordagem alternativa aos métodos "rollout" e "paralell rollout". Este método também emprega caminhos amostrados de estados subseqüentes do sistema com profundidade H, mas não é necessário determinar políticas de base para controlar o sistema após o primeiro instante de decisão. O método "hindsight" determina a cada caminho amostrado o valor estimado da função recompensa para a ação avaliada através da soma entre o valor esperado até o próximo instante de decisão, dado que a ação em questão é adotada no estado presente, e o valor máximo da função recompensa nos $H-1$ estados subseqüentes. Para tal, admite-se a hipótese de que os estados que compõem o caminho amostrado são conhecidos de antemão pelo controlador do sistema e, assim, as ações adotadas por este são ótimas ao longo do horizonte de planejamento reduzido. Sendo conhecidos os caminhos amostrados, a determinação de uma solução ótima para os $H-1$ estados subseqüentes torna-se um problema determinístico. Obtendo-se a média da avaliação dos caminhos amostrados, determina-se o valor "ótimo hindsight" da função recompensa para cada ação avaliada. Empregando-se uma profundidade H e um número de caminhos amostrados adequados, o valor "ótimo hindsight" fornece um limitante superior para o melhor valor da função recompensa da ação avaliada. Em um instante de decisão, escolhe-se uma ação:

$$\hat{\pi}_H(x) = \arg\max_{a \in A(x)} \hat{Q}_H(x,a),$$

onde $\hat{Q}_H(x,a) = R(x,a) + \gamma \frac{1}{C} \sum_{i=1}^{C} \left(\max_{\{a_1,\ldots,a_{H-1}\} | a \in A} \sum_{t=1}^{H-1} \gamma^{t-1} R(x_t, a_t) \right)$,

sendo $\max_{\{a_1,\ldots,a_{H-1}\} | a \in A} \sum_{t=1}^{H-1} \gamma^{t-1} R(x_t, a_t)$ o maior valor de retorno dentre todas as seqüências de ações possíveis para o caminho amostrado i, e C o número de caminhos amostrados para cada ação avaliada.

2.4. Métodos evolutivos de iteração de políticas

Os métodos evolutivos de iteração de políticas empregam técnicas de algoritmos genéticos para empreender buscas no espaço de ações capazes de convergir para a política ótima de controle do sistema modelado. Estes métodos se aplicam especificamente a PMDs *com espaço de estados pequeno e espaço de ações grande*. O espaço de estados pequeno é requerido em função da necessidade de se obter de forma exata o valor de retorno das políticas de decisão consideradas ao longo do processo iterativo destes métodos. Portanto, espaços de estados grandes dificultam ou inviabilizam os métodos evolutivos de iteração de políticas. Por outro lado, com os métodos evolutivos evita-se a necessidade, existente para os métodos tradicionais de otimização, de se calcular a recompensa esperada sobre todas as possíveis políticas compostas a partir do espaço de ações. A cada iteração, concentra-se a busca em populações com um número pequeno de políticas.

O **método "evolutionary policy iteration"** (EPI), desenvolvido por Chang et al. (2005), parte de uma população de políticas $\Lambda(0)$, onde $\Lambda(k)$ representa a $k-ésima$ população de políticas

pertencentes a $\prod$. Define-se um tamanho para as populações $n = |\Lambda(k)| \geq 2$, o qual será mantido constante. Denota-se a probabilidade do tipo de mutação por P_m, a probabilidade de mutação global por P_g e a probabilidade de mutação local por P_l. Define-se a distribuição de probabilidade μ para escolha de uma ação, tal que $\sum_{a \in A} \mu(a) = 1$ e $\mu(a) > 0$, $\forall a \in A$. A probabilidade do tipo de mutação P_m determina se a mutação aplicada a uma política $\pi \in \Lambda(k)$, $k = 0,1,...$, será global ou local, as probabilidades de mutação, global P_g e local P_l, determinam se a ação para um estado $\pi(x)$ será ou não alterada (mutada), e a distribuição de probabilidade para a escolha de uma ação μ é empregada para alterar $\pi(x)$. Define-se também o valor médio de uma política π, dada uma distribuição de probabilidade δ para o estado inicial do sistema, por $G_\delta^\pi = \sum_{x \in X} V_\infty^\pi(x)\delta(x)$, observando que uma política ótima π^* satisfaz a relação $G_\delta^{\pi^*} \geq G_\delta^\pi$, $\forall \pi \in \prod$. G_δ^π é empregado no método EPI para se estabelecer o critério de parada para o processo iterativo.

O processo iterativo do método EPI é composto pela repetição sucessiva de dois passos: (1) a escolha de uma política de elite e (2) a geração de uma nova população. Este processo é interrompido de acordo com uma regra de parada bastante simples: se o valor médio G_δ^π da política de elite se repetir em K populações consecutivas, encerra-se o processo.

Para a escolha de uma política de elite π_k^*, dada uma população de políticas $\Lambda(k)$, é empregado o método *"policy switching"*, o qual escolhe as ações para composição da política de elite da seguinte forma

$$\pi_k^*(x) = \underset{\pi(x) | \pi \in \Lambda(k)}{\arg\max} \left\{ V_\infty^\pi(x) \right\}, \ \forall x \in X .$$

π_k^* apresenta um desempenho sempre melhor ou igual a qualquer política de $\Lambda(k)$, ou seja, $V_\infty^{\pi_k^*}(x) \geq \max_{\pi \in \Lambda(k)} \left\{ V_\infty^\pi(x) \right\}$, $\forall x \in X$, e consequentemente $G_\delta^{\pi_k^*} \geq \max_{\pi \in \Lambda(k)} G_\delta^\pi$.

A política de elite π_k^* pode não estar contida em $\Lambda(k)$, e pode também não ser a melhor política em $\Lambda(k+1)$, onde estará incluída em conjunto com outras políticas alteradas (mutadas) a partir de $\Lambda(k)$. Desta forma a nova população $\Lambda(k+1)$ contém uma política que tem desempenho igual ou melhor que o de todas as políticas da população prévia $\Lambda(k)$. Garante-se assim a monotocidade da convergência de EIP para uma solução ótima, ou seja, para qualquer δ e $k \geq 0$, $G_\delta^{\pi_k^*} \geq G_\delta^{\pi_{k-1}^*}$.

Para se obter uma nova população de políticas, primeiramente geram-se $n-1$ subconjuntos de $\Lambda(k)$, S_i, $i = 1,...,n-1$, como segue: seleciona-se, com igual probabilidade, um número $m \in \{2,...,n-1\}$; seleciona-se, então, com igual probabilidade, m políticas em $\Lambda(k)$, repetindo-se esse passo na geração de cada subconjunto S_i; e pela aplicação do método "policy switching", geram-se $n-1$ políticas definidas por

$$\pi_i^*(x) = \underset{\pi(x) | \pi \in S_i}{\arg\max} \left\{ V_\infty^\pi(x) \right\}, \ \forall x \in X , \ i = 1,...,n-1.$$

As $n-1$ políticas geradas desta forma são então mutadas. Determina-se se a mutação de uma política π_i será global ou local, com probabilidade P_m. Se a mutação for global (local), para cada estado $x \in X$, $\pi_i(x)$ será alterada com probabilidade P_g (P_l). Se uma alteração for determinada, uma nova ação $a \in A$ será selecionada de acordo com a distribuição de probabilidade μ. As ações que não sofrem mutações permanecem inalteradas.

Na $k-ésima$ geração, a nova população de políticas $\Lambda(k+1)$ será o conjunto formado pela política de elite π_k^* e pelas $n-1$ políticas mutadas. Este método de obtenção de gerações de populações de políticas permite ao EPI tanto se aproximar de uma política ótima quanto escapar de mínimos locais.

O método EPI converge para a política ótima independentemente da população inicial escolhida. Assim, pode-se escolher uma população inicial $\Lambda(0)$ qualquer. Por exemplo, $\Lambda(0)$ pode ser composta por políticas que adotam sempre a mesma ação em todos os estados, sendo que cada política da população adota uma ação distinta.

O **método "evolutionary random policy search"** (ERPS), desenvolvido por Hu et al. (2005), é similar ao método EPI, porém, emprega metodologia distinta para a geração da política de elite de uma população de políticas e para a alteração (mutação) das políticas na geração de uma nova população de políticas. O método ERPS é composto pelos passos descritos a seguir. (1) Definem-se os parâmetros iniciais: distribuição de probabilidade para escolha de uma ação μ; tamanho das populações de políticas $n=|\Lambda(k)|\geq 2$; probabilidade de exploração p_0; amplitude de busca de ações r_i para cada estado $x_i \in X$, $i=1,...,|X|$, onde $|X|$ é o tamanho do espaço de estados. (2) Estabelece-se uma política inicial qualquer $\Lambda(0)$, determina-se o número K de iterações "sem melhora" para o critério de parada e inicia-se o processo iterativo. (3) Determina-se $V_\infty^\pi(x)$, $\forall \pi \in \Lambda(k)$ e $\forall x \in X$, e escolhe-se uma política de elite através de um método denominado pelos autores por "policy improvement with cost swapping" (PICS). (4) Checa-se o critério de parada e: ou se encerra o processo, ou se gera uma nova população de políticas retornando ao passo anterior.

Em relação à escolha da política de elite, o método de escolha PICS pode ser decomposto em duas etapas: (1) primeiramente determina-se para cada $\pi \in \Lambda(k)$ a recompensa descontada esperada $V_\infty^\pi(x)$ sobre o horizonte de planejamento infinito dado que o estado inicial do sistema é x, $\forall x \in X$; (2) em seguida computa-se a política de elite da seguinte forma:

$$\pi_k^*(x) = \underset{\pi(x)\,|\,\pi\in\Lambda(k)}{\arg\max} \left\{ R(x,\pi(x)) + \gamma \sum_{y\in X} \left(p(y\,|\,(x,\pi(x))) \left[\max_{\pi\in\Lambda(k)} \left(V_\infty^\pi(y) \right) \right] \right) \right\},\ \forall x \in X\ .$$

O desempenho de uma política de elite, pela forma como é gerada, é sempre igual ou melhor que o desempenho da política de elite das iterações prévias. Além desta qualidade, o ERPS mantém o melhoramento contínuo das políticas de elite ao longo das iterações através da aplicação de um método que gera novas populações de políticas alternando mutações sobre a política de elite corrente, ora explorando a vizinhança da política de elite (amostra viesada), ora explorando todo o espaço de ações (amostra não viesada).

O balanço entre estes dois tipos de busca, local viesada e global não viesada, é regulado pela probabilidade de exploração p_0. A cada estado $x \in X$ e para cada nova política $\pi_{k+1}^i \in \Lambda(k+1)$, $i=2,...,n$, com probabilidade p_0, $\pi_{k+1}^i(x)$ é escolhida em uma vizinhança próxima a $\pi_k^*(x)$ através de uma heurística denominada "nearest-neighbor", e com probabilidade $1-p_0$, $\pi_{k+1}^i(x)$ é selecionada dentre uma das ações de A, de acordo com a distribuição de probabilidade para escolha de uma ação, μ.

O método ERPS converge para a política ótima independentemente da população inicial escolhida. Em função dos melhoramentos introduzidos pela utilização do PICS para determinação de uma política de elite e pelo emprego de busca local através da heurística "nearest-neighbor", o método ERPS converge mais rapidamente para uma política ótima que o método EPI.

3. Método evolutivo em horizonte de planejamento reduzido e amostrado

Neste trabalho, propõe-se um método para solucionar PMDs com espaços de estados e de ações simultaneamente grandes, o qual se denomina por **método evolutivo em horizonte de planejamento reduzido e amostrado** (EHRA). Mais especificamente, propomos a combinação entre os métodos "parallel rollout" e ERPS. Combinando os conceitos desses dois métodos é possível determinar, antes que uma política ótima com ações previamente determinadas para todos os estados do sistema, uma ação a cada instante de decisão dado o estado observado do sistema. Escolhendo-se parâmetros adequados (horizonte de planejamento reduzido, quantidade de caminhos amostrados, tamanho das populações de ações, probabilidades para as mutações da ação de elite, critério de parada) pode-se garantir que as ações obtidas através do ERHA são melhores ou, pelo menos, tão boas quanto as ações determinadas pelas políticas de base empregadas no método.

O método evolutivo é aplicado para gerar a cada iteração uma nova população de ações a partir da ação de elite da população de ações imediatamente anterior. O método amostrado em horizonte reduzido é aplicado para determinar uma ação de elite a cada população de ações avaliada. Descreve-se abaixo a seqüência de passos para a combinação proposta.

3.1. Passos de inicialização

Especificações de entrada:

- estado observado, $x \in X$;
- fator de desconto, γ ;
- horizonte de planejamento reduzido, H ;
- número de simulações de funcionamento do sistema, C ;
- θ heurísticas geradoras do conjunto de políticas de base, $\Lambda = \{\pi_1, ..., \pi_\theta\}$;
- número m de $n-uplas$ para compor o *conjunto de ações de melhor desempenho* $\Omega = \{(a_1, d_1, f_1), ..., (a_m, d_m, f_m)\}$ sendo que, para $i = 1, ..., m$, a_i representa uma ação pertencente a $A(x)$, d_i registra o número de vezes que a ação a_i foi escolhida como ação de elite e f_i registra o valor médio de utilidade da ação a_i ;
- Ω inicial, sendo a_i uma ação qualquer pertencentes a $A(x)$, $d_i = 0$ e $f_i = 0$, para $i = 1, ..., m$;
- número máximo K de repetições da escolha de uma ação como ação de elite, para verificação do critério de parada;
- número n de ações para compor as populações de ações $\Gamma_k(x) = \{a_1^k, ..., a_n^k\}$, a serem avaliadas em cada iteração k, $k = 1, 2, ...$; $(\theta + m) < n < |A(x)|$;
- função de recompensa esperada $R(x, a)$ em um período entre dois instantes de decisão consecutivos, dado o estado e a ação adotada no instante inicial;
- função $f(x, a)$ para simular o próximo estado, dado um estado e uma ação adotada, baseada no conjunto de distribuições de probabilidades de transições entre estados P, definido para o PMD;
- população de ações inicial $\Gamma_0(x) = \{a_1^0, ..., a_n^0\}$, incluindo nas posições de $n-(\theta-1)$ até n as ações estabelecidas pelas θ políticas de base, dado o estado observado x, nas posições de $n-(\theta+m-1)$ até $n-\theta$ as ações correspondentes às $n-uplas$ de Ω e nas posições de 1 até $n-(\theta+m)$ ações quaisquer pertencentes à $A(x)$;
- distribuição de probabilidades P_x para a escolha de ações em $A(x)$, dado o estado observado x, utilizada na busca global;
- probabilidade q_o para realização de uma busca local na vizinhança da ação de elite, ou $1-q_o$ para realização de uma busca global entre as ações de $A(x)$;
- amplitude r para o intervalo de busca local na vizinhança da ação de elite, através da heurística "nearest neighbor";

Vetores auxiliares:

- vetor $x[H]$ de estados, para armazenar temporariamente um caminho de funcionamento do sistema de tamanho H ;
- vetor $EstQ[n]$, para armazenar a estimativa do valor de utilidade de cada uma das n ações pertencentes à população de ações;
- vetor $EstV[\theta]$, para armazenar a estimativa do valor de retorno descontado de um caminho de funcionamento amostrado para cada uma das θ políticas de base pertencente à Λ ;

3.2. Passos do processo iterativo

"Seja x o estado observado do sistema:"

fazer o contador $k = 0$

fazer o contador $l = 1$

(0) repetir enquanto $d_l < K$ (até que a regra de parada seja atingida, onde d_l é o número de vezes que a ação a_l foi escolhida como ação de elite, sendo a_l e d_l componentes da $l-ésima \;\; n-upla$ de Ω)

"Determinação de uma ação de elite para a população de ações $\Gamma_k(x)$:"

(1) para o contador $j = 1$ até n fazer

$EstQ[j] = 0$

(2) repetir C vezes

(3) para o contador $i = 1$ até θ fazer

$EstV[i] = 0$

$x[0] = x$

(4) para o contador $t = 1$ até H fazer

apenas quando $t = 1$, fazer $\pi_i(x[0]) = a_j^k$

$EstV[i] = EstV[i] + \gamma^{t-1} R(x[t-1], \pi_i(x[t-1]))$

$x[t] = f(x[t-1], \pi_i(t-1))$

(4) finalizar

(3) finalizar

$EstQ[j] = EstQ[j] + \max_{i=1,...,\theta} (EstV[i])$

(2) finalizar

$EstQ[j] = EstQ[j] / C$

(1) finalizar

fazer $e = \arg\max_{j=1,...,n} EstQ[j]$

"O índice da ação de elite da população k é dado por e ."

"Atualização de $\Omega = \{(a_1, d_1, f_1), ..., (a_m, d_m, f_m)\}$:"

fazer o contador $l = 1$

(1) repetir enquanto $l < m+1$ e $a_l \neq a_e^k$

$l = l + 1$

(1) finalizar

(1) se $l < m+1$, então

$f_l = \{(f_l \times d_l) + EstQ[e]\} / d_l + 1$

$d_l = d_l + 1$

(1) finalizar

(1) se $l = m + 1$, então

localizar no conjunto Ω o índice s correspondente ao menor f_s, valor médio de utilidade da ação a_s pertencente à $s-ésima\ \ n-upla$ de Ω, e então fazer

$a_s = a_e^k$

$f_s = EstQ[e]$

$d_s = 1$

$l = s$

(1) finalizar

(1) se $d_l < K$, o critério de parada não foi atingido, então

"Determinação de uma nova população de ações, $k+1$:"

(2) para o contador $i = 1$ até $n - (\theta + m)$ fazer

gerar um número aleatório u, seguindo a distribuição uniforme $U \sim [0,1]$

(3) se $u \leq q_o$, então

escolher a ação a_i^{k+1} na vizinhança de a_e^k, através da heurística *"nearest-neighbor"*

(3) finalizar

(3) se $u > q$, então

escolher a ação a_i^{k+1} empregando a distribuição de probabilidades P_x para escolha de ações em $A(x)$

(3) finalizar

(2) finalizar

(2) para o contador $i = (n+1) - (\theta + m)$ até $n - \theta$ fazer

$a_i^{k+1} = a_{(n+1)-(i+\theta)}$

(2) finalizar

"A nova população de ações é $\Gamma_{k+1}(x) = \{a_1^{k+1}, ..., a_n^{k+1}\}$, composta por $n - (\theta + m)$ ações escolhidas randomicamente de $A(x)$, por m ações que compõem as $n-uplas$ de Ω e por θ ações correspondentes às decisões do conjunto de políticas de base Λ."

fazer o contador $k = k + 1$

(1) finalizar

(0) finalizar

"Após k iterações, a ação escolhida pelo método EHRA para o estado observado x é a_e^k, com um valor de utilidade estimado, em relação às políticas de base e ao horizonte de planejamento reduzido H, igual a f_l."

4. Discussão

Como demonstraram Kearns e seus colegas (2002), o método KMN é capaz de atingir uma solução ótima, desde que se empreguem um horizonte de planejamento, H, e um número de estados amostrados, C, suficientemente grandes. A complexidade computacional para obtenção de uma ação, dado um estado $x \in X$, é $O(C|A|)^H$, onde $|A|$ é a cardinalidade do espaço de ações.

Os métodos "rollout" e "parallel rollout", como demonstraram, respectivamente, Bertsekas e Castañon (1999) e Chang (2001), geram soluções sub-ótimas, que fornecem limitantes inferiores para o retorno esperado ótimo. Esses métodos garantem soluções melhores ou tão boas quanto as soluções das políticas de base empregadas em seus processos iterativos. A complexidade

computacional para a obtenção de uma ação em ambos os métodos, dado um estado $x \in X$, é $O(|A||\Lambda|CH)$, onde $|\Lambda|$ é a cardinalidade do conjunto de políticas de base e C é o número de caminhos amostrados para cada ação avaliada.

O método "hindsight" não emprega políticas de base. Como demonstraram Chong e seus colegas (2000), pela assunção de visão antecipada e conseqüente inserção de problemas determinísticos no processo de resolução, este método gera soluções que fornecem limitantes superiores para o retorno esperado ótimo. Os problemas determinísticos associados ao método tornam-se computacionalmente onerosos à medida que são abordados modelos com espaços de ações maiores e que o horizonte de planejamento é ampliado. A complexidade computacional para se obter uma ação através do método "hindsight", dado um estado $x \in X$, é $O(C|A|^H)$.

Chang et al. (2005) e Hu et al. (2005) demonstraram, respectivamente, que os métodos evolutivos EIP e ERPS convergem em probabilidade para uma solução ótima. A complexidade computacional para obtenção de uma política de elite em uma iteração do método EPI é polinomial de ordem três sobre o espaço de estados, definida por $O(nm|X|^3)$, onde n é o tamanho da população de políticas avaliada a cada iteração, m é o número de políticas selecionadas da população de políticas avaliada para o processo de geração de novas políticas (mutação). A complexidade computacional de uma iteração do método ERPS é $O(n|X|^3)$.

As complexidades computacionais dos quatro métodos amostrados em horizonte de planejamento reduzido são independentes do tamanho do espaço de estados e as complexidades computacionais dos dois métodos evolutivos são independentes do tamanho do espaço de ações. O método proposto, evolutivo e amostrado em horizonte de planejamento reduzido, tem complexidade computacional independente dos espaços de ações e de estados. A complexidade do método ERHA a cada iteração do processo evolutivo para obtenção de uma ação de elite é da ordem de $O(n(|\Lambda|CH+1))$, sendo n o tamanho da população de ações a ser avaliada.

Em testes realizados com a implementação de um modelo de admissões de pacientes em hospital eletivo formulado como um PMD, as soluções do método ERHA tiveram desempenho sempre superior ao das políticas de base empregadas. Diversas instâncias do modelo foram testadas. Para uma proporção entre 92% e 98% dos estados as ações geradas pelo ERHA concordaram com as ações da política ótima gerada pelo algoritmo de iteração de valores. A melhor política de base concordou com a política ótima em proporções entre 85% e 88% dos estados. O ERHA foi capaz de gerar ações para instâncias do modelo nas quais os algoritmos tradicionais não puderam gerar soluções por falta de memória computacional.

5. Conclusão

A modelagem de sistemas como PMDs, principalmente quando são considerados modelos próximos à realidade, gera dificuldades em relação ao tratamento computacional de grandes dimensionalidades. Neste estudo um novo método para solução de PMDs com espaços de estados e de ações simultaneamente grandes foi apresentado.

O método EHRA combina um horizonte de planejamento reduzido e amostrado a uma técnica de busca evolutiva, sendo capaz de gerar uma política de decisão (1) *não estacionária*, visto que, por se tratar de um processo iterativo amostrado, é possível a ocorrência de ações distintas para os mesmos estados em instantes de decisão subseqüentes, (2) *"on-line"*, pois a cada instante de decisão, dado o estado observado, o método deve ser executado completamente, e (3) *sub-ótima*, que melhora as políticas de decisão de base empregadas no processo iterativo do método.

Referências

Bertsekas, D. P. e Castañon, D. A. (1989). Adaptive aggregation methods for infinite horizon dynamic programming. In *IEEE Transactions on Automatic Control*, 34(6), 589-598.

Bertsekas, D. P. e Castañon, D. A. (1999). Rollout algorithms for stochastic scheduling problems. In *Journal of Heuristics*, 5, 89-108.

Bes, C. e Lasserre, J. B. (1986). An on-line procedure in discounted infinite-horizon stochastic optimal control. In *Journal of Optimization Theory and Applications*, 50, 61-67.

Boutilier, C., Dearden, R. e Goldszmidt, M. (1995). Exploiting Structure in Policy Construction. In *Proc. of the Fourteenth International Joint Conference on Artificial Intelligence*, 14, 1104-1113.

Chang, H. S. (2001), On-line sampling-based control for network queueing problems, Purdue University, West Lafayette, Indiana.

Chang, H. S., Lee, H., Fu, M. C. e Marcus, S. I. (2005). Evolutionary policy iteration for solving Markov decision processes. In *IEEE Transactions on Automatic Control*, 50(11), 1804-1808.

Chong, E. K. P., Givan, R. e Chang, H. S. (2000). A framework for simulation-based network control via hindsight optimization. In *Proc. of the 39th IEEE Conference on Decision and Control*, 1433-1438.

Dean, T., Givan, R. e Leach, S. (1997). Model reduction techniques for computing approximately optimal solutions for Markov decision processes. In *Proc. of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, 124-131.

Givan, R., Dean, T. e Greig, M. (2003). Equivalence notions and model minimization in Markov decision processes. In *Artificial Intelligence***,** 147(1-2), 163-223.

Hernández-Lerma, O. e Lasserre, J. B. (1990). Error bounds for rolling horizon politicies in discrete-time Markov control processes. In *IEEE Transactions on Automatic Control*, 35(10), 1118-1124.

Hu, J., Fu, M. C., Ramezani, V. R. e Marcus, S. I. (2005) "An evolutionary random policy search algorithm for solving Markov decision processes", Institute for Systems Research, Technical Report, University of Maryland, http://techreports.isr.umd.edu/reports/2005/TR_2005-3.pdf.

Kearns, M., Mansour, Y. e Ng, A. Y. (2002). A sparse sampling algorithm for near-optimal planning in large Markov decision processes. In *Machine Learning*, 49, 193-208.

Littman, M. L., DEAN, T. e Kaelbling, L. P. (1995). On the complexity of solving Markov decision problems. In *Proc. of the Eleventh International Conference on Uncertainty in Artificial Intelligence*, 394-402.

Parr, R. E. (1998). Flexible decomposition algorithms for weakly coupled Markov decision problems. In *Proc. of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, 422-430.

Puterman, M. L. (2005), Markov decision processes: discrete stochastic dynamic programming, John Wiley & Sons, New York.

Singh, S. P. e Cohn, D. (1998). How to dynamically merge Markov decision processes. In *Advances in Neural Information Processing Systems*, 10, 1057-1063.