Analysis of the Risk Sensitive Value Iteration Algorithm

Igor Oliveira Borges[1], **Karina Valdivia Delgado**[1] **e Valdinei Freire**[1]

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (USP)
São Paulo, SP – Brasil

{igor.borges,kvd,valdinei.freire}@usp.br

Abstract. *This paper shows an empirical study of Value Iteration Risk Sensitive algorithm proposed by Mihatsch and Neuneier (2002). This approach makes use of a risk factor that allows dealing with different types of risk attitude (prone, neutral or averse) by using a discount factor. We show experiments with the domain of Crossing the River in two different scenarios and we analyze the influence of discount factor and risk factor under two aspects: optimal policy and processing time to convergence. We observed that: (i) the processing cost in extreme risk policies is high with both risk-averse and risk-prone attitude; (ii) a high discount increases time to convergence and reinforces the chosen risk attitude; and (iii) policies with intermediate risk factor values have a low computational cost and show a certain sensitivity to risk based on the discount factor.*

Keywords. Risk Sensitive Markov Decision Process, Stochastic Planning, Risk Sensitive Policy.

Resumo. *Este artigo apresenta um estudo empírico do algoritmo de Iteração de Valor Sensível a Risco proposto por Mihatsch e Neuneier (2002). Essa abordagem usa um fator de risco que permite lidar com diferentes tipos de risco (propenso, neutro ou averso) e usa fator de desconto. Foram realizados experimentos no domínio de Travessia do Rio em dois cenários de recompensas distintos e feita uma análise da influência do fator de desconto e do fator de risco na política estacionária obtida e no tempo de processamento necessário para convergência nestes cenários. Observou-se que: (i) o custo de processamento de políticas extremas a risco, tanto de aversão quanto de propensão é elevado; (ii) um desconto elevado aumenta o tempo de convergência do algoritmo e reforça a sensibilidade ao risco adotada; e (iii) políticas com valores para o fator de risco intermediários possuem custo computacional baixo e já possuem certa sensibilidade ao risco dependendo do fator de desconto utilizado.*

Palavras-Chave. Processo de Decisão Markoviano Sensível a Risco, Planejamento Estocástico, Política Sensível a Risco.

1. Introdução

Um modelo comumente utilizado em planejamento probabilístico é o Processo de Decisão Markoviano (em inglês *Markovian Decision Process* – MDP), neste um agente deve encontrar uma política que minimiza o custo esperado [Puterman 1994].

Um ponto a ser considerado quando da tomada de decisão em planejamento probabilístico é como levar em conta os riscos associados a estocasticidade no resultado de

decisões. Um agente que minimiza o custo esperado pode ser considerado um agente neutro ao risco, enquanto um agente sensível a risco deve escolher entre duas atitudes: aversa ou propensa ao risco [Shen et al. 2014].

O desenvolvimento de algoritmos sensíveis ao risco, i.e. que consideram a sensibilidade ao risco na tomada de decisão, é um tema pouco explorado na literatura [García and Fernández 2015]. Existem diferentes abordagens para quantificar o risco como: utilidade exponencial esperada [Howard and Matheson 1972, Jaquette 1976, Denardo and Rothblum 1979, Rothblum 1984, Patek 2001], ponderação entre esperança e variância [Sobel 1982, Filar et al. 1989], estimação de desempenho em um intervalo de confiança [Filar et al. 1995, Yu et al. 1998, Hou et al. 2014, Hou et al. 2016].

Nos trabalhos baseados em utilidade exponencial esperada, deve-se especificar um fator de risco, sendo que os valores factíveis para esse fator de risco dependem do problema de decisão em questão [Patek 2001], uma alternativa é considerar fator de desconto, mas nesse caso a política ótima torna-se não estacionária [Chung and Sobel 1987]. [Mihatsch and Neuneier 2002] propõem uma equação de ponto fixo baseado em uma função de transformação escalar por partes e fator de desconto, que permite uma política ótima estacionária como solução, assim como uma escolha arbitrária para o fator de risco.

Embora [Mihatsch and Neuneier 2002] foquem em propor algoritmos de aprendizado por reforço, operadores de ponto-fixo e com propriedade de contração são propostos, o que permite especificar algoritmos de planejamento probabilístico para esse modelo. No entanto, os autores não realizaram nenhuma avaliação empírica desses operadores. Por outro lado, [Freire 2016] explora o papel que o fator de desconto produz em diferentes modelos sensíveis ao risco. Em especial é mostrado que o fator de desconto apresenta uma característica de propensão ao risco quando minimização de custo é considerada.

O objetivo deste trabalho é analisar de forma empírica o impacto da escolha do fator de risco e do fator de desconto sob dois aspectos: (*i*) quão aversa ao risco é a política obtida sob tais parâmetros, e (*ii*) como o tempo de convergência do planejamento é afetado por tais parâmetros.

2. MDPs neutros ao risco

Processos de decisão markovianos (MDPs) permitem modelar problemas da área de planejamento probabilístico e de aprendizado por reforço. Em MDPs as transições entre estados são definidas probabilisticamente [Puterman 1994] e o processo é chamado de markoviano pois o efeito de uma ação em um dado estado depende somente da ação escolhida naquele estado, não levando em conta o histórico de tomadas de decisão sequenciais já realizadas [Bellman 1957]. Formalmente um MDP é uma tupla: $\langle S, A, T, R \rangle$, no qual:

- S é o conjunto de estados pertencentes ao processo.
- A é o conjunto de ações que podem ser executadas durante as épocas de decisão.
- $T : S \times A \times S \rightarrow [0, 1]$ é uma função que define a probabilidade de transição dos estados no sistema, sendo que $T(s'|s, a)$ representa a probabilidade de chegar no estado $s' \in S$, dado que o agente está no estado $s \in S$ e foi escolhida a ação $a \in A$.
- $R : S \times A \rightarrow R$ é uma função recompensa que define a recompensa recebida no estado $s \in S$ ao tomar uma ação $a \in A$.

A quantidade de épocas de decisão é chamada de horizonte e pode ser finito (definido por um número fixo), infinito (repetido seguidamente sem parada) ou ainda indeterminado (repetido seguidamente com possibilidade de parada, por exemplo, quando o agente atinge um estado meta).

A solução de um MDP infinito ou indeterminado é uma política estacionária π, isto é, uma função que mapeia estados em ações; e o valor $V^{\pi}(s)$ de uma política em um estado $s \in S$ é determinado por:

$$V^{\pi}(s) = \mathrm{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \middle| \pi, s_0 = s\right],$$

em que $\gamma \in [0, 1)$ é o fator de desconto.

Seja Π o conjunto de políticas estacionárias, a função valor ótima $V^*(s) = \max_{\pi \in \Pi} V^{\pi}(s)$ é a solução da equação de Bellman:

$$V^*(s) = \max_{a \in A(s)} \sum_{s' \in S} T(s'|s, a)[R(s, a) + \gamma V^*(s')] \forall s \in S.$$

A política ótima pode ser obtida com base na função valor ótima por:

$$\pi^*(s) = \arg\max_{a \in A(s)} \sum_{s' \in S} T(s'|s, a)[R(s, a) + \gamma V^*(s')]$$

.

Algoritmo de Iteração de Valor O algoritmo de Iteração de Valor (Value Iteration – VI) é um algoritmo de programação dinâmica. Em cada iteração i é calculado o valor $V^i(s)$ baseado no valor $V^{i-1}(s)$ para cada estado $s \in S$ do MDP, isto é:

$$V^i(s) \leftarrow \max_{a \in A} \sum_{s' \in S} T(s'|s, a)[R(s, a) + \gamma V^{i-1}(s')]$$

Um possível critério de parada é considerar o resíduo $\max_{s \in S} |V^i(s) - V^{i-1}(s)|$ e iterar enquanto o resíduo for maior que um erro mínimo desejado ϵ.

3. MDPs Sensíveis ao Risco

MDP Sensível a Risco (em inglês *Risk Sensitive* MDP – RSMDP) é uma extensão de MDP que inclui alguma forma de modelar risco. Neste trabalho estamos interessados na abordagem de [Mihatsch and Neuneier 2002] que está focada principalmente em Aprendizado por Reforço e no lugar de transformar a recompensa acumulada, como é feito na abordagem de utilidade exponencial, a diferença temporal é transformada. Em [Mihatsch and Neuneier 2002] são propostas duas versões com risco dos algoritmos Q-Learning e Temporal Difference. Além disso, os autores demonstram que os algoritmos propostos convergem e para isso são definidos diferentes operadores. A seguir é descrita a abordagem proposta em [Mihatsch and Neuneier 2002].

Formalmente um RSMDP é definido por uma tupla $\langle \text{MDP}, k, \gamma \rangle$, sendo $-1 < k < 1$ o fator de risco e $\gamma \in [0, 1)$ o fator de desconto. Seja a função de transformação $\mathcal{X}^k$ que depende do sinal da entrada x e do fator de risco, definida por:

$$\mathcal{X}^k(x) = \begin{cases} (1-k)x & \text{se } x > 0 \\ (1+k)x & \text{caso contrário} \end{cases} \tag{1}$$

Dada uma política estacionária π, a função valor $V_k^\pi(s)$ correspondente pode ser obtida resolvendo o seguinte sistema de equações para todo $s \in S$:

$$0 = \sum_{s' \in S} T(s'|s, \pi(s)) \chi^k \Big(R(s, \pi(s)) + \gamma V_k^\pi(s') - V_k^\pi(s) \Big) \tag{2}$$

Seja $x = R(s, \pi(s)) + \gamma V_k^\pi(s') - V_k^\pi(s)$ a diferença temporal, se k for positivo, então diferenças temporais negativas tem peso de ponderação maior que as positivas. Em outras palavras se: $R(s, \pi(s)) + \gamma V_k^\pi(s') - V_k^\pi(s) < 0$, transições para estados sucessores em que a recompensa imediata passou a ser menor que a média recebem um peso extra maior. Enquanto, as transições para estados que prometem um retorno maior do que a média recebem uma ponderação menor. Ou seja, se $k > 0$, a função objetivo $V_k^\pi(s)$ é aversa ao risco e é propensa ao risco, se $k < 0$, em $k = 0$ é neutra a risco e existe a equivalência com o critério de MDPs clássicos. No limite de extrema aversão ao risco, quando $k \to 1$, a função objetivo resolve um problema equivalente a otimização no pior dos casos. No limite de extrema propensão a risco, quando $k \to -1$, o agente é muito otimista assumindo que para todos os possíveis próximos estados, aquele que acontece é sempre o melhor.

Teorema 1 *(Solução única e casos limites) [Mihatsch and Neuneier 2002] Para cada $k \in (-1, 1)$ existe uma solução única $V_k^\pi(s)$ obtida pela resolução da Equação 2. Assim a função valor sensível a risco está bem definida. Para k=0, $k \to 1$ e $k \to -1$, temos:*

$$V_0^\pi(s) = E\left(\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, \pi(s_t)) \middle| s_0 = s \right), \tag{3}$$

$$\lim_{k \to 1} V_k^\pi(s) = \inf \left(\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, \pi(s_t)) \middle| s_0 = s \right), \tag{4}$$

$$\lim_{k \to -1} V_k^\pi(s) = \sup \left(\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, \pi(s_t)) \middle| s_0 = s \right). \tag{5}$$

O teorema mostra que a abordagem permite interpolar entre os critérios de melhor caso, neutro e pior caso. Uma política estacionária π^* é ótima, se $V_k^{\pi^*}(s) \geq V_k^\pi(s)$ $\forall \pi \in \Pi, s \in S$.

3.1. Operador Sensível a Risco

Os Lemas 1 e 2 apresentados nesta seção foram adaptados de [Mihatsch and Neuneier 2002].

Lema 1 *Seja $k \in (-1, 1)$. Para cada par de números reais a, b, existe uma $\xi_{(a,b,k)} \in [1 - |k|, 1 + |k|]$ tal que $\chi^k(a) - \chi^k(b) = \xi_{(a,b,k)}(a - b)$.*

Em [Mihatsch and Neuneier 2002] foram definidos diferentes operadores, entre eles, o operador $\mathcal{N}_{\alpha,k}$ sobre o espaço de funções $Q(s,a)$, tal que:

$$\mathcal{N}_{\alpha,k}[Q](s,a) := Q(s,a) + \alpha \sum_{s' \in S} T(s'|s,a)\chi^k \Big(R(s,a) + \gamma \max_{u \in A(s)} Q(s',u) - Q(s,a) \Big), \tag{6}$$

em que α denota um tamanho de passo positivo. Esse operador é um mapeamento de contração relacionado à norma máxima $|Q| := \max_{s \in S, a \in A(s)} |Q(s,a)|$ desde que α seja pequeno o suficiente, como especificado no próximo lema.

Lema 2 *Seja $k \in (-1,1)$, $0 \leq \gamma < 1$ e $0 < \alpha \leq (1+|k|)^{-1}$ para todas as funções Q_1 e Q_2, é verdade que:*

$$|\mathcal{N}_{\alpha,k}[Q_1] - \mathcal{N}_{\alpha,k}[Q_2]| \leq \rho |Q_1 - Q_2|,$$

em que $\rho = (1 - \alpha(1-|k|)(1-\gamma)) \in (0,1)$. Assim o operador $\mathcal{N}_{\alpha,k}$ é um mapeamento de contração.

Note que α está definido em $0 < \alpha \leq (1+|k|)^{-1}$, resultado obtido com este estudo. Para garantir que α possa ser igual a $(1+|k|)^{-1}$, basta definir que o desconto γ seja necessariamente inferior a 1.

4. Algoritmo de Iteração de Valor Sensível a Risco

A partir da Equação 6 e do Lema 2, é possível definir a função de atualização de Q a seguir, a qual é utilizada no algoritmo de Iteração de Valor Sensível a Risco:

$$Q^i(s,a) \leftarrow Q^{i-1}(s,a) + \alpha \sum_{s' \in \mathcal{S}} \mathcal{T}(s'|s,a) \mathcal{X}^k \Big(\mathcal{R}(s,a) + \gamma \max_a Q^{i-1}(s',a) - Q^{i-1}(s,a) \Big) \tag{7}$$

Nessa equação, a função escalar $\mathcal{X}^k$ é aplicada também diretamente nas diferenças temporais dos valores. O fator α é usado para garantir convergência e segundo o Lema 2, os valores possíveis são $0 < \alpha \leq (1+|k|)^{-1}$. De maneira intuitiva, o fator α assegura que os valores parciais não cresçam muito, em especial quando o fator de desconto γ estiver próximo de 1 e o fator de risco k for negativo.

Dado $Q^i(s,a)$, é possível obter a função valor na iteração i:

$$V^i(s) = \max_a \{Q^i(s,a)\}$$

e também é possível obter uma política gulosa:

$$\pi_{\chi^k}(s) = \arg\max_a \{Q^i(s,a)\}.$$

O critério de parada usado no algoritmo de Iteração de Valor Sensível a Risco é baseado no seguinte residual relativo:

$$\text{residual}(s) = \left| \frac{V^i(s) - V^{i-1}(s)}{V^{i-1}(s)} \right|.$$

Dado um erro mínimo desejado ϵ, se $\max_{s \in \mathcal{S}}\{\text{residual}(s)\} \leq \epsilon$, o algoritmo para. Note que a utilização do residual relativo é melhor para o algoritmo de Iteração de Valor Sensível a Risco do que o residual absoluto uma vez que o algoritmo pode ter valores muito diferentes para $V^i(s)$ e $V^{i-1}(s)$.

O algoritmo de Iteração de Valor Sensível a Risco (Algoritmo 1) recebe como entrada um RSMDP, o erro mínimo desejado ϵ e o fator α; e devolve como saída a função valor V ótima sensível a risco. Nas linhas 1 a 4 são inicializados o valor V^0 com a maior recompensa para cada estado e o valor Q_0 com 0 para todo par estado-ação. Em cada iteração i o algoritmo atualiza os valores usando a Equação 7 (Linha 9). Na linha 11, V^i é calculada com base em Q^i e na linha 12 é calculado o residual relativo entre as iterações i e $i-1$ de cada estado. Por fim, na linha 16 o algoritmo devolve a função valor V.

Algoritmo 1 Algoritmo de Iteração de Valor Sensível a Risco

Entrada: Um RSMDP ($\mathcal{S}$, $\mathcal{A}$, $\mathcal{T}$, $\mathcal{R}$, k, γ), ϵ, α
Saída : Função valor V para o RSMDP de entrada

1 **para cada** $s \in \mathcal{S}$ **faça**
2 $\quad V^0(s) \leftarrow \max_{a \in \mathcal{A}}\{\mathcal{R}(s,a)\}$
$\quad$ **para cada** $a \in \mathcal{A}$ **faça**
3 $\quad\quad Q^0(s,a) = 0$
4 $\quad$ **fim**
5 **fim**
6 $i \leftarrow 1$
enquanto $\max_{s \in \mathcal{S}}\{\text{residual}(s)\} > \epsilon$ **faça**
7 $\quad$ **para cada** $s \in \mathcal{S}$ **faça**
8 $\quad\quad$ **para cada** $a \in \mathcal{A}$ **faça**
9 $\quad\quad\quad Q^i(s,a) = Q^{i-1}(s,a) + \alpha \sum_{s' \in \mathcal{S}} \mathcal{T}(s'|s,a)\mathcal{X}^k[\mathcal{R}(s,a) + \gamma \max_{a \in \mathcal{A}} Q^{i-1}(s',a) - Q^{i-1}(s,a)]$
10 $\quad\quad$ **fim**
11 $\quad\quad V^i(s) = \max_{a \in \mathcal{A}}\{Q^i(s,a)\}$
12 $\quad\quad \text{residual}(s) = \left|\frac{V^i(s) - V^{i-1}(s)}{V^{i-1}(s)}\right|$
13 $\quad$ **fim**
14 $\quad i \leftarrow i + 1$
15 **fim**
16 **retorna** V

5. Experimentos

Nesta seção analisamos o algoritmo de Iteração de Valor Sensível a Risco no domínio de Travessia do Rio. O código-fonte[1] que inclui a modelagem do domínio do rio, o algoritmo de iteração de valor sensível a risco e a interface gráfica foi desenvolvido em Octave/Matlab. Os experimentos foram realizados no Matlab R2015a, na versão acadêmica de 64 bits no Windows 10, em um processador Intel Core i7-3537U CPU @3.10Ghz, 8 GB de memória RAM @1600Mhz, 256 GB de armazenamento SSD. A interface gráfica foi desabilitada para o experimento.

[1]Disponível em: https://github.com/RSVI

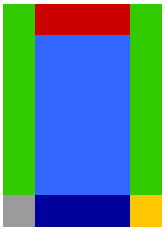
5.1. Domínio de Travessia no Rio

O problema da Travessia do Rio [Freire and Delgado 2017] é representado como um *grid* de tamanho $N_x \times N_y$. Um exemplo pode ser visualizado na Tabela 1. Nesta tabela está em azul claro, o rio; em azul escuro, a cachoeira; em vermelho, a ponte; em verde, o solo; em cinza, o agente propriamente dito no estado inicial; e por fim a meta é destacada em amarelo.

O domínio consiste em apenas um agente no extremo do canto inferior esquerdo da matriz, o qual tem como objetivo chegar ao canto inferior direito do *grid*, sendo que apenas 4 movimentos são possíveis: ir para o norte (N), sul (S), leste (E), e oeste (W). Assim o agente pode chegar no objetivo de duas formas: (i) nadando a partir de qualquer ponto do rio; ou (ii) subindo o *grid* até a ponte que está na posição extrema superior.

Tabela 1. Instância do Domínio de Travessia no rio

Grid 5x7



Estado
Solo
Ponte
Rio
Cachoeira
Inicial
Meta

5.2. Configurações do Experimento

Para o experimento adotou-se a condição de parada $\epsilon = 0.00001$; fator de desconto $\gamma \in \{0.6; 0.7; 0.8; 0.9; 0.99\}$; e fator de risco $k \in \{-0.99; -0.8; -0.5; 0; 0.5; 0.8; 0.99\}$.

A probabilidade de transição fora do rio é de 99% da ação ter o efeito desejado e 1% de ficar parado. No rio, o agente tem 20% de chance de ser arrastado pela correnteza e 80% de sucesso na execução da ação escolhida. A probabilidade de ir para o estado inicial dado que o agente está na cachoeira é 1, isto é, sempre que o agente cai na cachoeira ele retorna para o estado inicial.

Os experimentos foram realizados em dois cenários de recompensa distintos, o primeiro com recompensa 0 em cada estado e $+1$ no estado meta, o qual chamaremos de cenário de recompensas acumuladas positivas $(+)$; e o segundo com recompensa negativa -1 em cada estado e 0 na meta, o qual chamaremos de cenário de recompensas acumuladas negativas $(-)$. Além disso, foram realizados experimentos com $\alpha = 0.5$ e com α definido baseado no valor de k, como mostrado na Tabela 2.

Tabela 2. Relação de valor entre fator de risco k e fator α adotado no experimento.

k	α
-0.99 ou 0.99	0.50
-0.8 ou 0.8	0.56
-0.5 ou 0.5	0.67
0	1

5.3. Políticas Obtidas

Foram realizados experimentos com diferentes tamanhos de *grid*, entre eles 3×7, 5×7, 7×7 e 10×7. Nesta seção são avaliados os resultados obtidos para o *grid* 10×7 pois ele apresenta uma variedade maior de políticas. A Tabela 3 mostra as políticas obtidas para esse *grid*, variando k e γ no algoritmo de Iteração de Valor sensível a risco nos cenário de recompensa $(+)$ e recompensa $(-)$, parte superior e inferior da tabela, respectivamente. Tanto ao adotar $\alpha = 0.5$ quanto com o α relativo a k, foram obtidas as mesmas políticas relativas ao cenário de recompensa adotado, isto ocorre pois para ambos os valores de α utilizados, é garantida a convergência pelo Lema 2.

Em ambos cenários se observa um comportamento de risco esperado para os parâmetros de propensão, neutralidade e aversão ao risco, sendo que para $k = -0.8$, $k = -0.5$, $k = 0$, $k = 0.5$ e $k = 0.8$ as políticas encontradas variando γ são as mesmas. A diferença entre ambos cenários acontece nos extremos, com $k = -0.99$ e $k = 0.99$.

Para $k = -0.99$, no cenário $(-)$ há algumas políticas com atitudes mais propensas ao risco que não aparecem no cenário de recompensa $(+)$, em especial saltar da ponte a fim de chegar mais rápido na meta (veja no cenário de recompensa acumulada negativa $k = -0.99$ e $\gamma \geq 0.7$). Para $k = 0.99$ aparecem duas políticas diferentes no cenário $(-)$ e no cenário positivo três políticas diferentes. Nessas políticas o agente tenta atravessar pela ponte e caso caia no rio tenta voltar a borda mais perto ou se está perto da ponte tenta subir para ter uma travessia mais segura evitando cair na cachoeira.

Em valores de $k \geq -0.5$, observasse políticas mais conservadoras que tentam atravessar pela ponte e evitam entrar no rio para assegurar uma transição mais segura. Ao tomar uma política mais propensa ($k = -0.99$ e $k = -0.8$), o agente tende a se arrisca mais atravessando pelo rio e saltando da ponte em direção a meta.

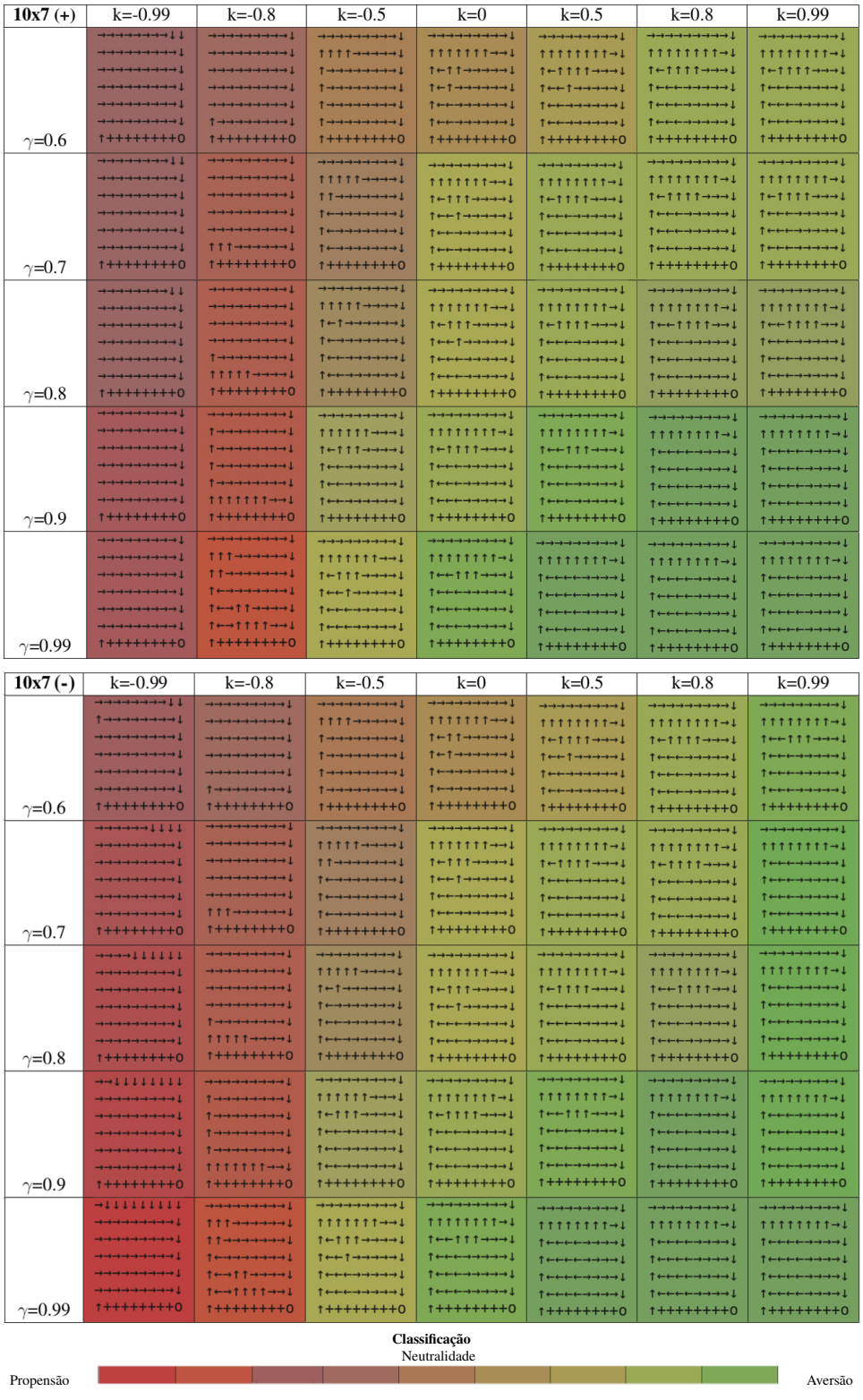
Nota-se ainda que nos extremos de valor de risco $k = -0.99$ para propensão e $k = 0.99$ para aversão obtém-se políticas mais reforçadas para o respectivo tipo de atitude esperada, em especial quando o γ é elevado ($\gamma \geq 0.9$). Note que o fator de desconto tem o papel de atenuar as atitudes se o desconto for pequeno ou realçar as atitudes se o desconto for grande. O parâmetro de risco por sua vez infere na política obtida um comportamento de propensão, neutralidade ou aversão conforme o valor do parâmetro escolhido.

5.4. Tempo de Processamento

As Figuras 1 e 2 apresentam o tempo de processamento para o cenário de recompensas $(+)$ e $(-)$ com duas configurações de α distintas ($\alpha = 0.5$ e α relativo a k) no *grid* 10×7. Como comentado anteriormente as políticas obtidas para α fixo e α relativo para cada cenário de recompensa adotado foram iguais entre si. Porém, o tempo de processamento necessário para a convergência foi diferente, sendo mais eficiente escolher um fator α relativo ao fator de risco escolhido, especialmente para $k = 0$.

No *grid* 10×7, $\gamma = 0.99$ e α relativo, o tempo necessário para convergência para $k = -0.8$ e $k = 0.8$ é aproximadamente $4s$ e $12s$ no cenário $(+)$ e de $9s$ e $10s$ no cenário $(-)$, respectivamente. Enquanto no intervalo de risco $-0.5 \leq k \leq 0.5$ notou-se uma execução rápida, independente do fator de desconto adotado. Para valores extremos de k ($k = -0.99$ e $k = 0.99$) têm-se uma maior demanda de tempo para processamento quando comparado com outros valores de k.

Tabela 3. Políticas obtidas pelo algoritmo de Iteração de Valor Sensível a Risco em um grid 10×7 conforme cenário de recompensa adotado.



O fator de desconto por sua vez tem o papel de garantir que a convergência ocorra e tem influência no tempo necessário para convergência, porém de forma mais tênue se

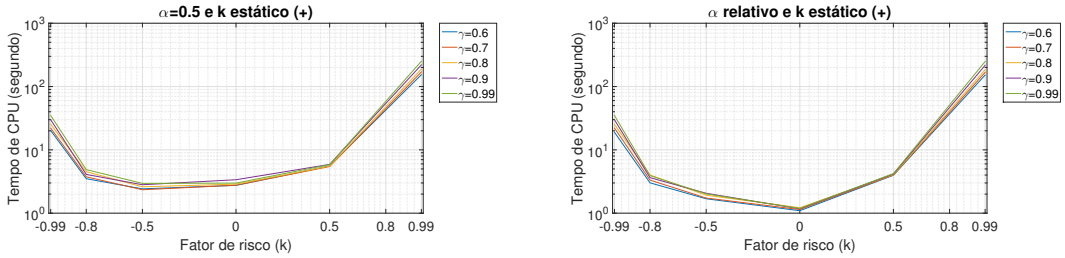


Figura 1. Tempo de convergência em um rio de tamanho 10×7 no cenário de recompensas (+).

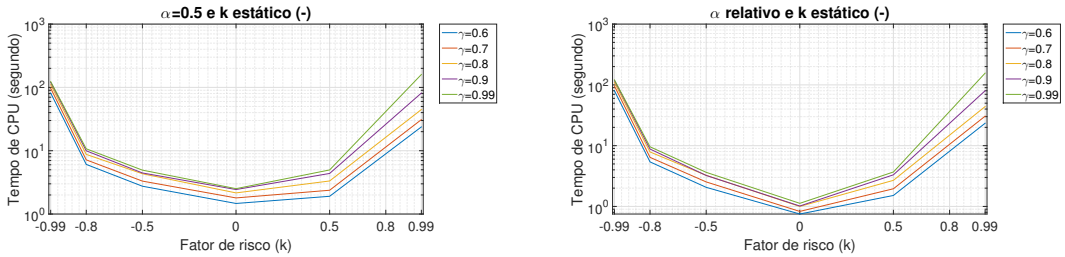


Figura 2. Tempo de convergência em um rio de tamanho 10×7 no cenário de recompensas (-).

comparado ao impacto do fator de risco no tempo. Quanto maior o desconto, maior o tempo necessário para convergência.

Na Figura 3 é mostrado o tempo para diferentes tamanhos de *grid* considerando $\gamma = 0.99$, α relativo nos cenários de recompensas $(+)$ e $(-)$ para $k = -0.99$ (lado esquerdo da figura) e $k = 0.99$ (lado direito da figura). A figura mostra que o tempo de processamento depende do tamanho do problema, quanto maior a quantidade de estados, maior o tempo necessário para convergência. Note que para $k = -0.99$, a adoção de um cenário $(+)$ pode reduzir o tempo de convergência do algoritmo quando comparado com o cenário $(-)$, o contrário ocorre para $k = 0.99$.

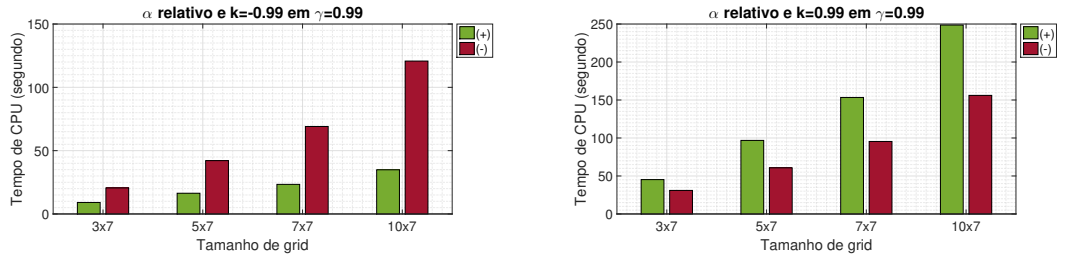


Figura 3. Tempo de processamento por tamanho de grid e cenário de recompensa para $k = -0.99$ e $k = 0.99$.

5.5. Convergência com $\alpha = 1$

A convergência do algoritmo ocorre tanto no cenário $(+)$ quanto $(-)$ com $\alpha = 0.5$ e α relativo a k. Também foram feitos testes com $\alpha = 1$ e para problemas com um número pequeno de estados o algoritmo converge muito rápido. Contudo vale ressaltar que em problemas com um número de estados elevado no cenário $(+)$ e um fator de desconto alto

(γ próximo de 1), o algoritmo não consegue convergir, sendo que os valores crescem até o infinito.

6. Considerações Finais

Neste trabalho, foi analisado de forma empírica o impacto da escolha do fator de risco e o fator de desconto na política e no tempo de convergência do algoritmo de Iteração de Valor de [Mihatsch and Neuneier 2002]. O algoritmo de Iteração de Valor Sensível a Risco consegue encontrar políticas sensíveis ao risco conforme o fator de risco escolhido, tanto para aversão ($k > 0$) quanto propensão ($k < 0$) ao risco. Além disso, a característica de neutralidade é mantida quando $k = 0$ sendo equivalente ao algoritmo de Iteração de Valor clássico. Nesta abordagem é necessário a escolha de um valor de risco k arbitrário coerente ao domínio de aplicação, para garantir que ocorra a ponderação esperada e permita a construção de uma política π com a sensibilidade ao risco desejada.

Foram realizados experimentos no domínio de Travessia do Rio com diferentes tamanhos de *grid* em dois cenários de recompensas distintos: recompensas positivas acumuladas e recompensas negativas acumuladas. O fator de desconto tem o papel de atenuar as atitudes se o desconto for pequeno ou realçar as atitudes se o desconto for grande. No limite de k ($k = -0.99$ para propensão extrema ao risco e $k = 0.99$ para aversão extrema ao risco), embora o algoritmo encontre políticas com as respectivas atitudes de risco esperadas, o algoritmo demora mais tempo para convergir. Para $k = -0.99$, a adoção de um cenário $(+)$ pode reduzir o tempo de convergência do algoritmo quando comparado com o cenário $(-)$, o contrário ocorre para $k = 0.99$.

Os experimentos mostram que o tempo de processamento está diretamente relacionado ao fator de risco escolhido, o fator de desconto γ, o fator α e o tamanho da instância. Enquanto o fator de desconto tem uma influência fraca no tempo de convergência, o uso de um α dependente de k traz bastante ganho no desempenho do algoritmo reduzindo o número de iterações necessárias para o algoritmo convergir.

Em trabalhos futuros pretende-se analisar o uso de α variando durante as iterações do algoritmo, começando com $\alpha = 1$; bem como alterar o fator de risco durante o algoritmo, por exemplo começar com $k = 0$ e incrementar (decrementar) o fator de risco durante as iterações até chegar no valor k desejado.

Agradecimentos

Os autores agradecem à Capes pela concessão da bolsa de mestrado para as atividades de pesquisa e à FAPESP pelo apoio financeiro (processo #2015/01587-0).

Referências

[Bellman 1957] Bellman, R. (1957). A Markovian decision process. *Indiana Univ. Math. J.*, 6:679–684.

[Chung and Sobel 1987] Chung, K.-J. and Sobel, M. J. (1987). Discounted mdp's: distribution functions and exponential utility maximization. *SIAM J. Control Optim.*, 25:49–62.

[Denardo and Rothblum 1979] Denardo, E. V. and Rothblum, U. G. (1979). Optimal stopping, exponential utility, and linear programming. *Mathematical Programming*, 16(1):228–244.

[Filar et al. 1989] Filar, J. A., Kallenberg, L. C. M., and Lee, H.-M. (1989). Variance-penalized Markov decision processes. *Mathematics of Operations Research*, 14(1):147–161.

[Filar et al. 1995] Filar, J. A., Krass, D., Ross, K. W., and Ross, K. W. (1995). Percentile performance criteria for limiting average Markov decision processes. *IEEE Transactions on Automatic Control*, 40(1):2–10.

[Freire 2016] Freire, V. (2016). The role of discount factor in risk sensitive markov decision processes. In *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 480–485.

[Freire and Delgado 2017] Freire, V. and Delgado, K. V. (2017). GUBS: a utility-based semantic for Goal-Directed Markov Decision Processes. In *Sixteenth International Conference on Autonomous Agents & Multiagent Systems*, pages 741–749.

[García and Fernández 2015] García, J. and Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *J. Mach. Learn. Res.*, 16(1):1437–1480.

[Hou et al. 2014] Hou, P., Yeoh, W., and Varakantham, P. (2014). Revisiting risk-sensitive MDPs: New algorithms and results. In *Proceedings of the Twenty-Fourth International Conference on Automated Planning and Scheduling, ICAPS 2014, Portsmouth, New Hampshire, USA, June 21-26, 2014*.

[Hou et al. 2016] Hou, P., Yeoh, W., and Varakantham, P. (2016). Solving risk-sensitive POMDPs with and without cost observations. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 3138–3144.

[Howard and Matheson 1972] Howard, R. A. and Matheson, J. E. (1972). Risk-sensitive Markov decision processes. *Management science*, 18(7):356–369.

[Jaquette 1976] Jaquette, S. C. (1976). A utility criterion for Markov decision processes. *Management Science*, 23(1):43–49.

[Mihatsch and Neuneier 2002] Mihatsch, O. and Neuneier, R. (2002). Risk-sensitive reinforcement learning. *Machine Learning*, 49(2):267–290.

[Patek 2001] Patek, S. D. (2001). On terminating markov decision processes with a risk-averse objective function. *Automatica*, 37(9):1379–1386.

[Puterman 1994] Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition.

[Rothblum 1984] Rothblum, U. G. (1984). Multiplicative Markov decision chains. *Mathematics of Operations Research*, 9(1):6–24.

[Shen et al. 2014] Shen, Y., Tobia, M. J., Sommer, T., and Obermayer, K. (2014). Risk-sensitive reinforcement learning. *Neural computation*, 26(7):1298–1328.

[Sobel 1982] Sobel, M. J. (1982). The variance of discounted Markov decision processes. *Journal of Applied Probability*, 19(4):794–802.

[Yu et al. 1998] Yu, S. X., Lin, Y., and Yan, P. (1998). Optimization models for the first arrival target distribution function in discrete time. *Journal of Mathematical Analysis and Applications*, 225(1):193 – 223.