Compromissos arbitrários entre custo e probabilidade à meta e algoritmo de busca heurística em planejamento probabilístico sob o critério GUBS

Gabriel Nunes Crispino[1], **Valdinei Freire**[1], **Karina Valdivia Delgado**[1]

[1]Universidade de São Paulo – São Paulo, Brasil

Abstract. *Stochastic Shortest Path Markov Decision Processes (SSP-MDPs) are used to model probabilistic sequential decision problems where the objective is to minimize the expected accumulated cost to goal. However, in the presence of dead-ends, this criterion can become ill-defined. GUBS, a criterion based on Expected Utility Theory that makes trade-offs between costs and probability-to-goal, was proposed to address these problems. The dissertation contains a comparison between GUBS and other related criteria for solving SSP-MDPs in the presence of dead-end states, and introduces theoretical definitions and results that show that GUBS, unlike these other criteria, allows for a trade-off between accumulated costs and probability-to-goal, and also guarantees arbitrary trade-offs that can be tuned from its parameters without previous knowledge of the problem. Also, this dissertation introduces eGUBS-AO*, an optimal heuristic search algorithm for solving the eGUBS criterion. As subproducts of this masters dissertation, α-MCMP, a new criterion, and UCT-GUBS, an algorithm for solving SSP-MDPs under the GUBS criterion, were also proposed.*

Resumo. *Processos de Decisão Markovianos de Caminho Estocástico mais Curto (SSP-MDPs) são utilizados para modelar problemas de decisões sequenciais probabilísticas em que o objetivo é minimizar o custo acumulado esperado para a meta. No entanto, na presença de estados a partir dos quais não se pode alcançar a meta (dead ends), esse critério pode deixar de ser bem-definido. GUBS, um critério baseado em Teoria da Utilidade Esperada que realiza compromissos entre custos e probabilidade à meta, foi proposto para resolver esses problemas. A dissertação de mestrado tem como contribuições uma comparação entre o GUBS e outros critérios para resolver SSP-MDPs na presença de dead ends, e introduz definições e resultados teóricos que mostram que o GUBS, ao contrário desses outros critérios, possibilita a realização de compromissos entre custos acumulados sobre perdas em probabilidade à meta, e também garante compromissos arbitrários que podem ser configurados a partir dos seus parâmetros sem conhecimento prévio do problema a ser resolvido. Além disso, o presente trabalho introduz o eGUBS-AO*, um algoritmo ótimo de busca heurística para resolver o critério eGUBS. Como subprodutos da dissertação de mestrado, também foram propostos o α-MCMP, um novo critério, e o UCT-GUBS, um algoritmo de Monte Carlo Tree Search para resolver SSPs sob o critério GUBS.*

1. Introdução

Processos de Decisão Markovianos de Caminho Estocástico mais Curto (*Stochastic Shortest Path Markov Decision Processes* – SSP-MDPs) [Bertsekas 1995] são utilizados para

modelar problemas em que um agente interage com um ambiente por meio de ações com resultados estocásticos, com o objetivo de minimizar o custo esperado para a meta. Em SSP-MDPs, podem existir estados tal que a partir deles a probabilidade de alcançar a meta (probabilidade à meta) é menor que 1. Esses estados são chamados de *dead ends* e, quando são inevitáveis no problema, o critério convencional que minimiza o custo esperado para a meta é mal-definido. Por essa razão, novos modelos ou adaptações desse critério precisam ser definidos. No caso do critério MAXPROB [Kolobov et al. 2011], políticas ótimas maximizam a probabilidade à meta, sem avaliar custos. Outros critérios lexicográficos minimizam o custo acumulado esperado à meta considerando apenas caminhos que maximizam a probabilidade à meta, como em iSSPUDEs [Kolobov et al. 2012], S^3Ps [Teichteil-Königsbuch 2012] e no critério MCMP [Trevizan et al. 2017]. Também é possível utilizar um fator de desconto $\gamma \in (0, 1)$ [Teichteil-Königsbuch et al. 2011] ou uma penalidade finita D para desistir do processo [Kolobov et al. 2012] para que o critério de custo acumulado esperado seja bem-definido na presença de *dead ends* inevitáveis.

Para ilustrar esses problemas, considere o exemplo de uma pessoa que precisa pegar um voo em um determinado horário. Isso pode ser considerado um problema de raciocínio sob incerteza, já que certas ações que a pessoa pode tomar com esse objetivo são estocásticas. Por exemplo, ao esperar em um ponto de ônibus para pegar um ônibus para o aeroporto, não se pode definir deterministicamente o tempo que o próximo ônibus demorará para chegar, ou até mesmo o tempo que levará para a pessoa chegar no aeroporto considerando qualquer que seja o meio de transporte utilizado para esse fim. Critérios que consideram políticas que maximizam a probabilidade à meta (como MAXPROB [Kolobov et al. 2011] e critérios lexicográficos como iSSPUDE [Kolobov et al. 2012], S^3P [Teichteil-Königsbuch 2012] e MCMP [Trevizan et al. 2017]) não resultariam em decisões realísticas para esse problema, já que haveria uma grande chance de tais decisões fazerem com que o agente saia para o aeroporto muito cedo (ou até imediatamente) ao tentar maximizar a probabilidade de chegar lá, ao invés de sair com um intervalo de tempo mais realístico. Outro problema dessa abordagem lexicográfica é que ela não permite que compromissos sejam feitos entre políticas que levam a custos grandes e outras que não maximizam a probabilidade à meta, mas que levam a custos consideravelmente menores. Ao longo desse trabalho esse tipo de compromisso será referido como *compromisso infinito-infinitesimal*. No caso do exemplo dado, um possível caso em que esse tipo de compromisso não seria realizado seria se o agente preferisse uma política que tem como ação pegar um táxi para o aeroporto pagando um valor arbitrariamente grande por isso, comparada com outra que leva o agente a pegar um ônibus com um custo arbitrariamente menor, desde que a probabilidade à meta da primeira opção seja maior que a segunda. A primeira opção seria a política ótima mesmo que essa diferença entre as probabilidades seja infinitesimal, e o tempo de chegada de ambas seja similar.

Os outros critérios mencionados (uso de um fator de desconto ou de uma penalidade finita) podem realizar compromissos ao ter seus parâmetros ajustados para refletir em decisões mais realistas se comparadas apenas à escolha de políticas que maximizam a probabilidade à meta. No entanto, para alguns problemas esses modelos podem dar preferência igual ou superior a políticas que não alcançam metas se comparadas a políticas que alcançam, o que é uma propriedade indesejada.

Baseado nos problemas desses critérios presentes na literatura, o GUBS (*Go-*

als with Utility-Based Semantics) [Freire and Delgado 2017], um critério que combina priorização de metas sobre históricos com Teoria da Utilidade Esperada, foi proposto. O critério eGUBS, um caso particular do GUBS em que uma função de utilidade exponencial é utilizada com um fator de risco negativo, foi também proposto em [Freire et al. 2019]. No mesmo trabalho, o eGUBS-VI, um algoritmo de iteração de valor, foi introduzido para resolver SSP-MDPs de maneira ótima sob o critério eGUBS.

1.1. Objetivos

A dissertação de mestrado tem como principais objetivos identificar e introduzir propriedades teóricas ao critério GUBS e outros critérios da literatura para resolver SSP-MDPs com *dead ends* inevitáveis e propor algoritmos que resolvam SSP-MDPs sob o GUBS.

1.2. Contribuições

Na dissertação de mestrado são introduzidos dois conceitos principais para realizar uma comparação teórica entre o critério GUBS e outros critérios: primeiro, uma definição do que decisões em SSP-MDPs que permitem compromissos infinito-infinitesimais são; e segundo, uma propriedade da preferência entre pares de políticas em um critério. Essa propriedade, nomeada de propriedade de priorização de probabilidade à meta α-forte, pode ser garantida por um critério se a razão entre valores de probabilidade à meta de todos pares de políticas estão limitados por um valor α a partir da preferência entre elas para todo SSP-MDP, em que $0 \leq \alpha \leq 1$. Entre outros resultados, o presente trabalho demonstra que o GUBS é o único critério entre todos os analisados que não apenas permite compromissos infinito-infinitesimais, como também garante a propriedade de priorização de probabilidade à meta α-forte para um α arbitrário, tal que $0 \leq \alpha \leq 1$.

Além de novas contribuições relacionadas às propriedades teóricas do GUBS e dos critérios relacionados, alguns resultados no formato de definições, corolários e teoremas são modificados com adições referentes aos conceitos de priorização de probabilidade à meta α-forte e da realização de compromissos infinito-infinitesimais. Ademais, o trabalho propõe o algoritmo eGUBS-AO*, e o seu desempenho comparado com o eGUBS-VI em diferentes domínios de SSP-MDPs com dead ends é analisado. Por questões de espaço, os resultados e discussões desse conjunto de experimentos não estão disponíveis no presente resumo. Uma análise empírica para comparar o critério GUBS com outros critérios presentes na literatura com o mesmo propósito é também realizada. Também por razões de espaço, as provas de todos teoremas, corolários, e proposições contidas na dissertação não foram incluídas no presente trabalho.

Como resultado do presente trabalho de mestrado, também foi proposto um novo critério chamado α-MCMP, uma extensão do critério MCMP [Trevizan et al. 2017] que também garante que a priorização de probabilidade à meta α-forte seja mantida para $0 \leq \alpha \leq 1$. Além disso, foi também proposto um algoritmo de *Monte Carlo Tree Search* para resolver SSP-MDPs sob o critério GUBS, o UCT-GUBS. Ambas essas contribuições foram introduzidas respectivamente em [Crispino et al. 2023] e [Crispino et al. 2020]. No presente resumo, por questões de espaço, essas contribuições não são descritas.

2. Caminho Estocástico mais Curto

Um MDP de Caminho Estocástico mais Curto (SSP-MDP) [Bertsekas 1995] é uma tupla $\mathcal{M} = \langle \mathcal{S}, s_0, \mathcal{A}, P, c, \mathcal{G} \rangle$, em que: $\mathcal{S}$ é o conjunto de estados; $s_0 \in \mathcal{S}$ é o estado inicial; $\mathcal{A}$

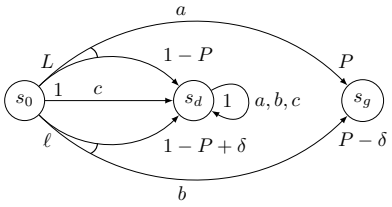


Figura 1. Exemplo de um SSP-MDP

é o conjunto de ações que podem ser executadas em cada estado; $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ é a função de transição, tal que $P(s, a, s') = \Pr(s_{t+1} = s' | s_t = s, a_t = a)$ indica a probabilidade do agente estar em um estado s' no próximo passo, dado que a ação a é executada no estado s no passo atual; $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_{>0}$ é a função de custo, que define um custo $c(s, a)$ ao tomar uma ação a em um estado s; e $\mathcal{G} \subset \mathcal{S}$ é o conjunto de estados meta absorvedores.

Em um SSP-MDP, a interação do agente com o ambiente, considerando cada passo de tempo $t \in \{0, 1, ..., T\}$, pode ser resumida em um histórico $h = \{\langle s_0, a_0, c_0\rangle, ..., \langle s_{T-1}, a_{T-1}, c_{T-1}\rangle, s_T\}$. A solução de um SSP-MDP é uma política π que é estacionária e markoviana. Tradicionalmente, SSP-MDPs convencionais consideram o critério de otimalidade de custo acumulado esperado. Nesse critério, uma política π é avaliada pela função valor $V^\pi(s) = \lim_{T\to\infty} \mathbb{E}\left[\sum_{t=0}^{T-1} c_t \mid \pi, s_0 = s\right]$. A partir dessa função, pode-se derivar a função valor **ótima**, denotada por V^*, tal que $V^*(s) = 0$ se $s \in \mathcal{G}$, e $V^*(s) = \min_{a\in\mathcal{A}}\{c(s, a) + \sum_{s'\in\mathcal{S}}[P(s, a, s')V^*(s')]\}$, caso contrário [Bellman et al. 1957]. Assim, a política ótima π^* pode ser obtida a partir de V^*. A medida de custo acumulado esperado não é bem-definida para SSP-MDPs quando *dead ends* são inevitáveis, já que nesses casos, a função valor V^π dessas políticas diverge.

3. Políticas que Priorizam Metas

Enquanto critérios que maximizam probabilidade à meta tem a vantagem de serem bem-definidos para SSP-MDPs com *dead ends* inevitáveis e de não possuírem parâmetros, é razoável questionar o quão racionais são decisões tomadas sob eles. Por outro lado, critérios que permitem compromissos arbitrários entre custo e probabilidade à meta o fazem a partir de uma escolha apropriada de valores para seus parâmetros, o que pode originar o questionamento do quão difícil pode ser escolher esses valores apropriadamente.

Por exemplo, considere o SSP-MDP na Figura 1, em que $1 \ll \ell \ll L$ são custos arbitrários, $P \in (0, 1]$ é uma probabilidade arbitrária, $\delta \in (0, P)$ é um escalar arbitrário, s_g é um estado meta, e s_d um *dead end*. No estado inicial s_0, o agente tem três ações para tomar: a, b e c. A ação a é a que tem a probabilidade à meta máxima P, mas que leva o agente a pagar um custo grande L. A ação b tem probabilidade à meta $P - \delta$, um valor muito próximo de P, e que faz com que o agente pague ℓ, um custo muito menor que L. Uma questão importante é a de se maximizar probabilidade à meta é uma escolha racional quando $\delta \to 0$ e $L - \ell \to \infty$. A ação c leva o agente diretamente para o *dead end* s_d ao pagar 1. Assim que o agente alcança s_d, ele paga 1 para os próximos passos de tempo. Outra pergunta que pode ser feita é se é racional a escolha de c em s_0, mesmo quando $P \to 0$ e $\ell \to \infty$. A ação a seria a escolha ótima para os critérios MCMP e S^3P, enquanto que para critérios que realizam compromissos como fSSPUDE e o critério de

custo descontado, qualquer uma das ações a, b, ou c poderia ser a ótima, dependendo dos valores escolhidos para seus parâmetros.

3.1. Quanto Pagar por um Aumento Infinitesimal de Probabilidade à Meta?

Apesar de critérios baseados em maximizar probabilidade à meta como MAXPROB, MCMP, e S^3P parecerem não ter grandes desvantagens, eles não permitem compromissos entre grandes aumentos em custos acumulados para pequenas perdas em probabilidade à meta. Isso acontece porque, como esses critérios maximizam a probabilidade à meta, não faz diferença se existe uma ação que tem uma medida de custo muito menor que outra ação com uma probabilidade à meta ligeiramente menor. Logo, critérios que maximizam probabilidade à meta não permitem a realização de compromissos infinito-infinitesimais. A prova formal disso está disponível na dissertação.

3.2. Até Quanto um Aumento de Probabilidade à Meta Compensa na Existência de Custos Infinitos?

Se compromissos infinito-infinitesimais são desejáveis, pode-se perguntar para o quanto de acréscimo em probabilidade à meta se torna razoável pagar um custo arbitrariamente grande (infinito). A próxima definição mensura tal acréscimo de maneira relativa. Para essa definição, considere a seguinte notação para políticas e históricos: $A \succ B$ significa que o tomador de decisões prefere A a B, $A \sim B$ significa que o tomador de decisões tem preferência igual para A e B, e $A \succeq B$ significa que o tomador de decisões tem preferência maior ou igual por A em relação a B. Além disso, considere $P^{\pi}_{G}(s)$ como a função de probabilidade à meta, que descreve a probabilidade de alcançar um estado meta ao seguir a política π a partir de um estado s.

Definição 1 (Priorização de probabilidade à meta α-forte). *Considerando $0 \leq \alpha \leq 1$, um critério de decisão tem priorização de probabilidade à meta α-forte se, para todos SSP-MDPs $\mathcal{M}$ e para todos pares de políticas $(\pi, \pi') \in \Pi \times \Pi$, a condição $\pi \succeq \pi' \implies \frac{P^{\pi}_{G}(s_0)}{P^{\pi'}_{G}(s_0)} \geq \alpha$, é verdadeira, tal que $P^{\pi'}_{G}(s_0) > 0$. Se π^* é a política ótima, então para toda política $\pi' \in \Pi$, $\pi^* \succeq \pi' \implies \frac{P^{\pi^*}_{G}(s_0)}{P^{\pi'}_{G}(s_0)} \geq \alpha$.*

Entre os seus possíveis tipos, existem dois casos especiais de priorização de probabilidade à meta α-forte: α-forte com $\alpha = 1$, chamada de **1-forte**, e α-forte com $\alpha \not> 0$, chamada de **apenas-0-forte**. Considere novamente o SSP-MDP da Figura 1. Se o critério de decisão utilizado tem priorização de probabilidade à meta 1-forte, então a ação a é ótima para qualquer valor de P, L, ℓ, e δ. Se o critério tem priorização de probabilidade à meta apenas-0-forte, c pode ser ótima mesmo que ela não leve o agente a um estado meta. Se o critério de decisão tem priorização de probabilidade à meta α-forte com $0 < \alpha < 1$, b pode ser ótima dependendo dos valores de P, L, ℓ, e δ e/ou dos parâmetros do critério.

Como demonstrado formalmente na dissertação, o critério MAXPROB tem priorização de probabilidade à meta 1-forte e, por outro lado, os critérios de custo descontado e fSSPUDE garantem somente priorização apenas-0-forte. Ou seja, para esses dois últimos critérios, enquanto um α arbitrário pode ser garantido para um SSP-MDP específico, não é possível escolher um $\alpha > 0$ que seja suficiente para qualquer SSP-MDP. Isso significa que uma política π que não alcança a meta ($P^{\pi}_{G}(s_0) = 0$) pode ser escolhida no lugar de qualquer outra política π' que tem alguma chance de alcançar a meta ($P^{\pi'}_{G}(s_0) > 0$).

Tabela 1. Tipos de priorização de probabilidade à meta α-forte e compromissos infinito-infinitesimais.

	Priorização de probabilidade à meta α-forte	Falta de compromissos infinito-infinitesimais
MAXPROB [Kolobov et al. 2011]	1-forte	Sim
Lexicográficos [Kolobov et al. 2012, Teichteil-Königsbuch 2012, Trevizan et al. 2017]	1-forte	Sim
fSSPUDE [Kolobov et al. 2012]	apenas-0-forte	Não
Custo descontado [Teichteil-Königsbuch et al. 2011]	apenas-0-forte	Não
GUBS [Freire and Delgado 2017]	α-forte, $0 \leq \alpha \leq 1$	Não

4. Critério GUBS (*Goals with Utility-Based Semantics*)

O critério GUBS [Freire and Delgado 2017] avalia um histórico baseado no seu custo acumulado e na condição de um estado meta ter sido alcançado ou não nesse histórico. No GUBS, então, são atribuídos pesos para essas duas propriedades do histórico baseados, respectivamente, em uma função de utilidade u sobre custos acumulados C_T, e uma utilidade constante de alcançar metas K_g. Assim, a função de utilidade sobre históricos utilizada pelo critério GUBS é: $U(C_T, \beta_T) = u(C_T) + K_g\beta_T$, tal que β_T é uma variável indicadora que vale 1 caso até o tempo T um estado meta tenha sido alcançado, e 0 caso contrário. Ademais, um agente segue o critério GUBS se ele avalia uma política π sob a função valor $V^{\pi}_{GUBS}(s) = \lim_{T\to\infty} \mathbb{E}[U(C_T, \beta_T)|\pi, s_0 = s]$.

Uma política π^* é ótima sob o critério GUBS se ela maximiza a função V^{π}_{GUBS}, ou seja, $V^{\pi^*}_{GUBS}(s) \geq V^{\pi}_{GUBS}(s), \forall\pi \in \Pi$ e $\forall s \in \mathcal{S}$. Note que uma maximização é utilizada porque no critério GUBS valores de utilidade são considerados ao invés de custos, como é o caso do critério convencional para SSP-MDPs definido na Seção 2.

4.1. Critério GUBS e Priorização de Probabilidade à Meta α-forte

Considerando que a sua função utilidade utilizada como parâmetro seja da forma $u : \mathbb{R} \cup \{\infty\} \to [U_{min}, U_{max}]$, o critério GUBS tem priorização de probabilidade à meta α-forte para $0 \leq \alpha \leq 1$ se U_{max}, U_{min} e a constante K_g forem escolhidos apropriadamente. Isso é demonstrado formalmente na dissertação. A Tabela 1 resume como a priorização de probabilidade à meta α-forte é garantida para cada critério, assim como se cada um deles permite a realização de compromissos infinito-infinitesimais. Note que o GUBS é o único que garante priorização de probabilidade à meta α-forte para $0 \leq \alpha \leq 1$ e permite a realização de compromissos infinito-infinitesimais.

4.2. Critério eGUBS

O *GUBS Exponencial* (*Exponential GUBS* - eGUBS) [Freire et al. 2019] é um caso do GUBS em que u é uma a função de utilidade exponencial. O uso dela é baseado em SSP-MDPs sensíveis ao risco [Patek 2001]. Essa função utilidade u é então definida como $u(C_T) = 0$, se $C_T = \infty$, e $u(C_T) = e^{\lambda C_T}$, caso contrário, para um custo acumulado C_T e fator de risco $\lambda < 0$.

Para resolver o critério GUBS, o custo acumulado precisa ser considerado para computar valores de utilidade. Por essa razão, é necessário definir uma função valor

para o eGUBS que depende não apenas em um estado s, mas em um custo acumulado C. Essa função é definida como função valor estado-custo. Considere que o agente já pagou um custo acumulado C, está no estado s, e então segue a política π a partir desse ponto. A função valor estado-custo é definida por $V^{\pi}_{GUBS}(s, C) = \lim_{T\to\infty} \mathbb{E}[U(C + C_T, \beta_T)|\pi, s_0 = s]$, descrevendo a utilidade esperada obtida pelo agente.

A função valor de uma política estacionária avaliada sob o critério eGUBS pode ser expressada da seguinte maneira, em termos de $V^{\pi}_{\lambda}(s)$ (a função valor da política π para SSP-MDPs sensíveis ao risco): $V^{\pi}_{GUBS}(s, C) = e^{\lambda C} V^{\pi}_{\lambda}(s) + K_g P^{\pi}_{G}(s)$ (a demonstração formal desse resultado está disponível na dissertação).

A principal propriedade do critério eGUBS é a de que um custo máximo pode ser obtido tal que quando o agente alcança esse custo acumulado, a política ótima a partir desse ponto é estacionária. Na dissertação, é demonstrado formalmente que esse custo máximo a partir do qual a política ótima sob o eGUBS é estacionária é dado pela equação $\bar{C}_{max}(s) = \max\{W(s), \max_{s' \in \mathcal{S}_{succ}(s), a \in \mathcal{A}_s(s')}[\bar{C}_{max}(s') - c(s, a)]\}$, em que $\mathcal{S}_{succ}(s) = \{s' \mid \exists a \in \mathcal{A}$ tal que $P(s, a, s') > 0, \forall s' \in \mathcal{S}\}$, $\mathcal{A}_s(s') = \{a \in \mathcal{A} \mid P(s, a, s') > 0\}, \forall (s, s') \in \mathcal{S} \times \mathcal{S}$ é o conjunto de ações que levam o agente de s a s', e $W(s)$ é o custo mínimo a partir do qual a política ótima é estacionária considerando apenas s (sua definição formal está disponível na dissertação).

4.3. Algoritmos Exatos para Resolver o Critério eGUBS

O eGUBS-VI [Freire et al. 2019] é o primeiro algoritmo exato a resolver SSP-MDPs sob o critério eGUBS. De maneira geral, o eGUBS-VI realiza o cálculo do C_{max} (uma versão menos eficiente da função $\bar{C}_{max}$ introduzida na Seção 4.2) a partir do valor da política ótima de um critério lexicográfico sensível ao risco (definido formalmente na dissertação), e realiza o cálculo da política ótima sob o eGUBS de trás para frente para estados aumentados a partir de uma política ótima do critério lexicográfico sensível ao risco.

Na dissertação de mestrado é introduzido o algoritmo eGUBS-AO*, um algoritmo de busca heurística baseado no AO* que encontra políticas ótimas para SSP-MDPs sob o eGUBS. Ele tem três fases principais: (i) computar o valor do critério lexicográfico sensível ao risco $V_{\lambda}(s)$ para cada estado alcançável; (ii) computar $\bar{C}_{max}(s)$ para o mesmo conjunto de estados; e (iii) computar V^{*}_{GUBS} ao realizar busca heurística no SSP-MDP.

Uma diferença entre os algoritmos eGUBS-VI e eGUBS-AO* é que o último computa a função $\bar{C}_{max}(s)$ para cada estado alcançável de s_0, enquanto que o eGUBS-VI computa um único valor escalar C_{max} e o utiliza para cada estado. Considerando isso, o passo (iii) é a principal diferença entre os dois algoritmos. Para executar a busca, o eGUBS-AO* mantém dois grafos: G, o grafo da melhor solução parcial do algoritmo, e G', o grafo explícito, que guarda informações sobre todos estados aumentados visitados durante a busca. Na dissertação, é demonstrado que o eGUBS-AO* retorna uma política ótima sob o critério eGUBS quando utilizada uma heurística admissível h_v.

5. Experimentos e Resultados

Experimentos foram realizados em três domínios: *Navigation* [Sanner and Yoon 2011], *River* [Freire and Delgado 2017] e *Triangle Tireworld* [Little et al. 2007]. Os experimentos tem como objetivo responder às seguintes questões: *(i)* Qual é a influência dos

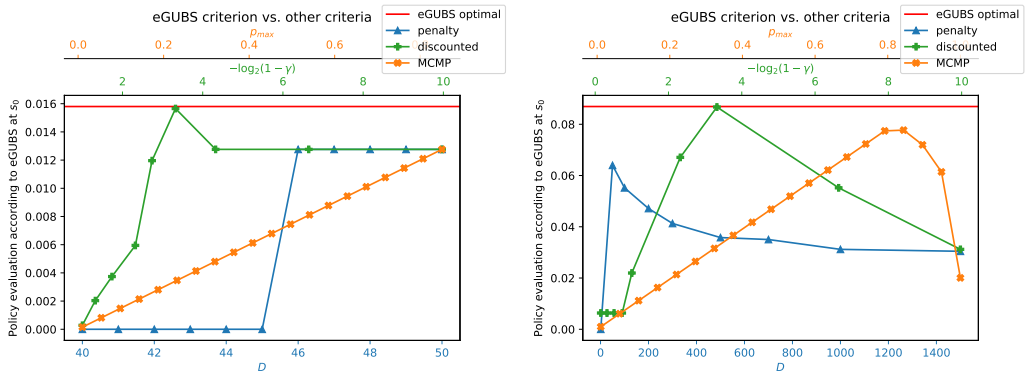
parâmetros λ e K_g na diferença entre C_{max} (utilizado no eGUBS-VI) e $\bar{C}_{max}(s_0)$ (utilizado no eGUBS-AO*)? *(ii)* Qual é a influência de λ e K_g no número de estados armazenados em memória por eGUBS-VI e eGUBS-AO*? *(iii)* Qual é a influência de λ e K_g no número de atualizações realizadas por eGUBS-VI e eGUBS-AO*? *(iv)* Como eGUBS-VI e eGUBS-AO* escalam de acordo com o número de estados do problema em termos de tempo? *(v)* Existem direções sobre qual algoritmo utilizar entre o eGUBS-VI ou o eGUBS-AO*? *(vi)* Como o critério eGUBS se compara a outros critérios? Por questões de espaço, o presente resumo apresenta apenas a discussão sobre os experimentos para comparar o eGUBS com outros critérios, com o objetivo de responder à pergunta (vi).

A seguir, serão mostrados resultados de experimentos feitos em instâncias selecionadas dos domínios para comparar políticas obtidas sob o critério eGUBS com políticas obtidas sob os critérios fSSPUDE, critério de custo descontado e MCMP. Para eles, os seus parâmetros (D, γ, e p_{max}, respectivamente) foram variados do menor valor para o maior no conjunto de valores escolhidos, exceto para os valores de p_{max} para o critério MCMP, que foram variados do maior valor para o menor. Por fim, essas políticas foram avaliadas sob o eGUBS e os valores resultantes disso foram então comparados com o valor da política ótima sob esse critério. Por essa comparação, pode-se avaliar se diferentes critérios podem encontrar políticas ótimas para o eGUBS apenas ao escolher parâmetros apropriados e, se não, o quão próximo do valor ótimo eles chegam.

As figuras 2a e 2b mostram esses resultados para a instância 10 do domínio *Navigation* e instância 5 do domínio *River*, respectivamente. Os valores de K_g e λ para a instância 10 do domínio *Navigation* foram fixados em 10^{-12} e -0.1, e para a instância 5 do domínio *River*, 0.01 e -0.1, respectivamente. Para ambas configurações de domínios, o único critério que foi capaz de encontrar uma política que o valor chegou próximo do ótimo sob o eGUBS foi o critério de custo descontado para o valor de $\gamma = 0.9$ em ambas instâncias. Como não existe um procedimento claro para escolher um certo valor para os parâmetros desses critérios com alguma garantia dos valores das políticas resultantes sob o eGUBS, um conjunto indefinido de valores precisa ser variado para esses parâmetros com esse propósito. Em algumas situações, isso pode ser impraticável se é desejado ter garantias no valor das políticas resultantes sob o critério eGUBS.

Por outro lado, se apenas uma aproximação é suficiente, é possível que poucos valores de cada parâmetro precisem ser testados para cada critério para que se aproxime do valor ótimo do critério eGUBS. Além disso, novamente, para os critérios fSSPUDE e de custo descontado, existem garantias de que no limite o critério MAXPROB é atingido (para o critério MCMP, isso acontece por padrão). Na Figura 2a ambos critérios fSSPUDE e de custo descontado obtém a política que maximiza a probabilidade à meta no último ponto testado ($D = 50$, $\gamma = 0.999$), enquanto que na Figura 2b, o valor exato não é obtido ($D = 1500$, $\gamma = 0.999$).

É também importante ressaltar que um tomador de decisão pode não ter recursos suficientes para resolver diferentes critérios para variados valores de cada um de seus parâmetros, a fim de encontrar uma aproximação razoável ao valor ótimo do critério eGUBS, especialmente sem garantias. Para a instância 10 do domínio *Navigation* o tempo gasto para computar o valor ótimo do critério eGUBS foi suficiente para resolver os critérios de custo descontado, fSSPUDE, e MCMP para respectivamente 5, 8 e 6 valores diferentes de seus parâmetros. Para a instância 5 do domínio *River*, o tempo foi



(a) Instância 10 do domínio *Navigation* **(b) Instância 5 do domínio *River***

Figura 2. Comparação do valor da política ótima para o eGUBS com o valor das políticas ótimas para os critérios fSSPUDE, de custo descontado e MCMP.

suficiente para resolver os critérios base para 6, 2, e 3 valores de parâmetros diferentes.

6. Considerações Finais

A dissertação de mestrado fornece uma análise do critério GUBS, um critério que combina priorização de metas sobre históricos com Teoria da Utilidade Esperada, comparando-o com outros critérios da literatura que resolvem SSP-MDPs com *dead ends*, e também apresenta resultados teóricos referentes às suas propriedades. Entre outros resultados, é demonstrado que o GUBS é o único critério que garante a propriedade de priorização de probabilidade à meta α-forte para $0 \leq \alpha \leq 1$, e que permite compromissos infinito-infinitesimais. O critério eGUBS, uma instância do GUBS que utiliza uma função exponencial com um fator de risco negativo, também é analisado. É possível demonstrar que, a partir de um certo custo acumulado, a política ótima sob o critério eGUBS é ótima sob o critério lexicográfico sensível ao risco.

Na dissertação de mestrado é proposto o eGUBS-AO*, um algoritmo de busca heurística que realiza uma busca do estado inicial guiada por uma função heurística, também tendo como entrada uma política ótima do critério lexicográfico sensível ao risco. A busca pode parar quando valores do custo acumulado de estados aumentados (s, C) alcançam $\bar{C}_{max}(s)$, podendo parar antes dependendo dos valores obtidos durante a busca.

Experimentos foram realizados nos domínios *Navigation*, *River* e *Triangle Tireworld*. Os seus resultados indicam que, para esses domínios, o algoritmo eGUBS-AO* pode em geral ter um melhor desempenho quando a função heurística é boa o suficiente, além de sempre processar um número igual ou menor de estados aumentados. Em outros casos, o eGUBS-VI pode ser uma melhor escolha em função do menor processamento das estruturas de dados auxiliares necessárias para busca no eGUBS-AO*. O segundo conjunto de experimentos, que analisa políticas obtidas que são ótimas sob o eGUBS comparando-as com políticas obtidas sob outros critérios indicam que é possível se aproximar de políticas ótimas sob o eGUBS por meio de outros critérios e algoritmos para resolvê-los. No entanto, não existem garantias teóricas de quanto processamento precisa ser realizado para que isso aconteça em qualquer domínio.

Referências

Bellman, R., Corporation, R., and Collection, K. M. R. (1957). *Dynamic Programming*. Rand Corporation research study. Princeton University Press.

Bertsekas, D. (1995). *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont, Mass.

Crispino, G., Freire, V., and Delgado, K. V. (2020). Algoritmo de Monte Carlo Tree Search para SSPs sob o critério GUBS. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 402–413, Porto Alegre, RS, Brasil. SBC.

Crispino, G. N., Freire, V., and Delgado, K. V. (2023). α-MCMP: Trade-offs between probability and cost in SSPs with the MCMP criterion. In *Brazilian Conference on Intelligent Systems*, pages 112–127. Springer.

Freire, V. and Delgado, K. V. (2017). GUBS: a utility-based semantic for goal-directed Markov decision processes. In *Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems*, pages 741–749.

Freire, V., Delgado, K. V., and Reis, W. A. S. (2019). An exact algorithm to make a trade-off between cost and probability in SSPs. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 29, pages 146–154.

Kolobov, A., Weld, D., et al. (2012). A theory of goal-oriented MDPs with dead ends. In *Uncertainty in artificial intelligence : proceedings of the Twenty-eighth Conference [on uncertainty in artificial intelligence] (2012)*, pages 438–447.

Kolobov, A., Weld, D. S., and Geffner, H. (2011). Heuristic search for generalized stochastic shortest path MDPs. In *Proceedings of the Twenty-First International Conference on International Conference on Automated Planning and Scheduling*, pages 130–137.

Little, I., Thiebaux, S., et al. (2007). Probabilistic planning vs. replanning. In *ICAPS Workshop on IPC: Past, Present and Future*, pages 1–10.

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

Sanner, S. and Yoon, S. (2011). IPPC results presentation. In *International Conference on Automated Planning and Scheduling*. Acessado em março de 2024.

Teichteil-Königsbuch, F. (2012). Stochastic safest and shortest path problems. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pages 1825–1831.

Teichteil-Königsbuch, F., Vidal, V., and Infantes, G. (2011). Extending classical planning heuristics to probabilistic planning with dead-ends. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pages 1017–1022.

Trevizan, F. W., Teichteil-Königsbuch, F., and Thiébaux, S. (2017). Efficient solutions for stochastic shortest path problems with dead ends. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (UAI) (2017)*, pages 1–10.