Algoritmos para Multi-Armed Bandits: Teoria e Aplicação à Precificação Dinâmica

Ismael Sampaio Bastos

Universidade Federal do Rio de Janeiro

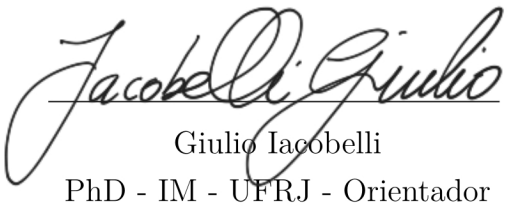
Instituto de Matemática

Ismael Sampaio Bastos

Algoritmos para Multi-Armed Bandits: Teoria e Aplicação à Precificação Dinâmica

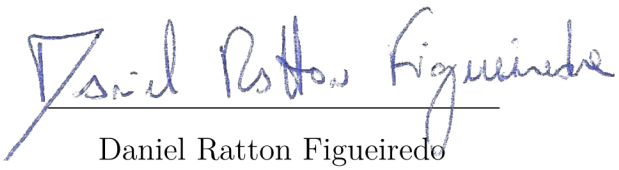
Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Estatística do Instituto de Matemática da Universidade Federal do Rio de Janeiro — UFRJ, como parte dos requisitos necessários à obtenção do título de Mestre em Estatística.

Aprovada em:



Giulio Iacobelli
PhD - IM - UFRJ - Orientador

Guilherme Ost
PhD - IM - UFRJ



Daniel Ratton Figueiredo
PhD - PESC/COPPE - UFRJ

Departamento de Métodos Estatísticos

2025

CIP - Catalogação na Publicação

S327a

Sampaio Bastos, Ismael
Algoritmos para multi-armed bandits: teoria e aplicação à precificação dinâmica / Ismael Sampaio Bastos. -- Rio de Janeiro, 2025.
69 f.

Orientador: Giulio Iacobelli.
Dissertação (mestrado) - Universidade Federal do Rio de Janeiro, Instituto de Matemática, Programa de Pós-Graduação em Estatística, 2025.

1. multi-armed bandits. 2. exploração vs explotação. 3. exploração pura. 4. precificação dinâmica. I. Iacobelli, Giulio, orient. II. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

"Mesmo sendo os piores alunos
na escola deste mundão,
nunca vamos repetir
nenhum inverno nem verão."

— Wisława Szymborska

AGRADECIMENTOS

Agradeço imensamente ao meu orientador, Giulio Iacobelli, que me acompanhou desde a minha chegada ao mestrado até sua finalização, sendo sempre solícito e atento às minhas dúvidas, ainda que muitas vezes fundamentais. Sou profundamente grato por ele ter acreditado em mim e por sua paciência nos momentos de dificuldade. Sempre o levarei como exemplo pelo resto da minha existência. Agradeço também aos alunos das disciplinas que tive o prazer de ministrar ao longo dos últimos dois semestres. Sem dúvida, as discussões em sala e o interesse demonstrado por eles me motivaram a seguir em frente e a não me deixar abater.

RESUMO

Este trabalho versa sobre o problema da tomada de decisões sequenciais, focando especificamente no problema de multi-armed bandit. Em sua concepção clássica, o problema de multi-armed bandits é caracterizado pela existência de um agente que se encontra diante de uma fileira de máquinas caça-níquéis (*bandits*), possuindo um número limitado de vezes que pode puxar a alavanca (*arm*) das máquinas, tendo por objetivo realizar a sequência de ações que maximize a recompensa obtida. O desafio consiste em equilibrar a escolha entre a ação que parece ser a mais lucrativa até aquele momento e a busca por informações sobre outras alternativas ainda não exploradas. Esse dilema é chamado de exploração (*exploration*) versus explotação (*exploitation*). Neste trabalho estudaremos vários algoritmos para auxiliar a tomada de decisões no problema de multi-armed bandits. Veremos também uma aplicação dessa teoria ao problema de precificação dinâmica, i.e., a determinação de preços de venda ótimos para produtos e serviços. Nesse caso, o vendedor ocupa o papel do agente que deseja vender um determinado produto, possuindo um conjunto finito de possíveis preços, sem saber nem a demanda do produto nem o comportamento do consumidor, cabendo ao vendedor adotar uma estratégia que vise encontrar o preço ótimo.

Palavras-chaves: multi-armed bandits, exploração vs explotação, exploração pura, precificação dinâmica.

ABSTRACT

This work addresses the problem of sequential decision-making, focusing specifically on the multi-armed bandit (MAB) framework. In its classical formulation, the MAB problem involves an agent facing a row of slot machines (bandits), with a limited number of pulls (arms) available. The agent's goal is to determine a sequence of actions that maximizes the total reward. The core challenge lies in balancing the trade-off between choosing the action that currently appears to yield the highest reward and exploring lesser-known alternatives (a dilemma known as exploration versus exploitation). In this study, we explore several algorithms designed to support decision-making within the multi-armed bandit setting. We also examine an application of this theory to the problem of dynamic pricing, i.e., determining optimal selling prices for products and services. In this context, the seller takes the role of the agent who aims to sell a product by selecting from a finite set of possible prices, without prior knowledge of demand or consumer behavior. The seller must therefore adopt a strategy that enables the identification of the optimal price over time.

Keyworsds: multi-armed bandits, exploration vs exploitation, pure exploration, dynamic pricing.

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1 INTRODUÇÃO

1.1 Motivação

O problema do bandit de múltiplos braços (do inglês *multi-armed bandits*) consiste em uma situação na qual um agente precisa escolher repetidamente entre diversas opções fixas (chamadas, neste contexto, de ações ou "braços"). Após cada decisão, o agente recebe uma recompensa (gerada aleatoriamente), e a ação escolhida pode também influenciar as decisões futuras.

De modo geral, o objetivo é selecionar a sequência de ações que maximize a recompensa total esperada ao longo do tempo ou, de forma equivalente, minimize o arrependimento. A medida de arrependimento quantifica o quanto o agente deixou de ganhar ao optar por ações subótimas durante o processo decisório. Assim, o problema de maximizar a recompensa total esperada se torna equivalente ao de minimizar o arrependimento. Essa medida desempenha um papel central ao longo do estudo do presente tema, dado que, na maioria das vezes, o foco está em desenvolver algoritmos que ofereçam garantias teóricas sobre o arrependimento, tais como limites inferiores e superiores.

Em linhas gerais, o desafio consiste em equilibrar a escolha entre a ação que aparenta ser a mais lucrativa até o momento e a busca por informações sobre outras alternativas ainda não exploradas, mas potencialmente mais promissoras. Essa problemática ilustra o dilema clássico entre exploração — do inglês *exploration* (*i.e.*, selecionar ações ainda não testadas) e explotação[1] — do inglês *exploitation* (*i.e.*, escolher a ação mais lucrativa até aquele instante).

O termo *multi-armed bandit* advém da analogia com um jogador (agente) diante de uma fileira de máquinas caça-níqueis (em inglês também chamadas de *armed bandits*), devendo decidir em quais máquinas jogar, quantas vezes acionar cada uma e em que ordem, ponderando se deve insistir na atual (explotação) ou experimentar outra (exploração).

Vale destacar que, nesse cenário, cada máquina gera uma recompensa aleatória segundo uma distribuição de probabilidade específica, a qual é desconhecida pelo agente. Ademais, assume-se que o agente dispõe de recursos limitados, geralmente representados por um número máximo de decisões, sendo esse fator denominado horizonte.

Em certos contextos, o dilema entre exploração e explotação dá lugar a um foco exclusivo na exploração, em que o agente precisa apenas investigar ao máximo o ambiente com o objetivo de, ao final, identificar a melhor ação. Nesse caso, não é necessário explorar alguma ação que pareça vantajosa, mas sim examinar todas as alternativas possíveis para descobrir a melhor. Problemas desse tipo são denominados problemas de exploração pura (do inglês *pure exploration*) e apresentam características distintas dos problemas que equilibram exploração e explotação.

Um exemplo simples que ilustra essa distinção é o de um agente que entra em um cassino e se depara com diversas máquinas caça-níqueis, tendo um número limitado de rodadas. O agente pode

[1] Neste contexto, o termo "explotação" é empregado de forma análoga ao seu uso na geologia, referindo-se ao ato de extrair recursos para obter benefício econômico.

escolher puxar o braço de cada máquina para descobrir qual oferece maior recompensa enquanto se concentra nas mais vantajosas, buscando maximizar os ganhos ao final. Alternativamente, pode estar no cassino apenas com o intuito de identificar qual máquina proporciona a melhor recompensa, com o objetivo de informar um terceiro interessado em maximizar os próprios lucros. O primeiro cenário reflete o equilíbrio entre exploração e explotação, enquanto o segundo exemplifica a exploração pura.

O problema de *multi-armed bandits* é relevante por diversos motivos, sendo o principal o fato de oferecer um modelo simples para o desafio da tomada de decisões sob incerteza. Além disso, possui aplicações práticas significativas, como o uso de algoritmos de bandit em sistemas de recomendação, precificação dinâmica e detecção de anomalias, conforme discutido em (BOUNEFFOUF; RISH; AGGARWAL, 2020). Em (SLIVKINS, 2024), também são apresentados exemplos voltados para o posicionamento de anúncios e precificação.

Observa-se que a maioria das aplicações práticas se concentra em problemas que exploram o dilema entre exploração e explotação, sendo menos comuns as aplicações associadas à exploração pura. Neste contexto, destaca-se como aplicação relevante a seleção de hiperparâmetros, especialmente na abordagem de (LI et al., 2018), que utiliza algoritmos de exploração pura para escolher hiperparâmetros durante o treinamento de redes neurais.

A formulação matemática dos problemas de *multi-armed bandits* leva a uma estrutura rica, conectada a diversas áreas da matemática, como probabilidade, teoria da decisão e teoria da informação. Assim, a presente dissertação seguirá o princípio de inter-relacionar esses campos ao longo do texto, apresentando os resultados de forma detalhada e destacando os conceitos fundamentais de cada área.

1.2 Definição do problema

O problema do bandit de múltiplos braços foi inicialmente apresentado por William R. Thompson (THOMPSON, 1933). Neste artigo, o autor abordou a questão ética dos ensaios clínicos, ressaltando as implicações de conduzir experimentos de maneira cega, sem adaptar os tratamentos à medida que surgem novas evidências sobre a eficácia dos medicamentos.

Outros trabalhos relevantes sobre decisões sequenciais incluem os estudos de Robbins (ROBBINS, 1952) e de Bather e Chernoff (BATHER; CHERNOFF, 1967). Existem também diversos livros dedicados ao tema dos bandits, com destaque para o de Bubeck e Cesa-Bianchi (BUBECK; CESA-BIANCHI, 2012), o recente trabalho de Tor Lattimore e Csaba Szepesvári (LATTIMORE; SZEPESVáRI, 2020), bem como os livros de Cesa-Bianchi e Lugosi (CESA-BIANCHI; LUGOSI, 2006) e Slivkins (SLIVKINS, 2024).

O problema dos multi-armed bandits pode ser formalizado considerando o conjunto $\mathcal{A}$, que contém todas as possíveis ações, sendo este um conjunto finito. Um bandit estocástico[2] com conjunto de ações $\mathcal{A}$ é uma coleção de distribuições $\nu = (P_a : a \in \mathcal{A})$, onde P_a é uma medida de probabilidade em $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ para todo $a \in \mathcal{A}$.

A dinâmica do problema consiste em um agente que interage com o ambiente, e a cada rodada

[2] Estocástico aqui se refere ao fato de que as recompensas são variáveis aleatórias.

$t \in \{1, \ldots, n\}$ seleciona uma ação $A_t \in \mathcal{A}$, sendo n o número total de rodadas (também conhecido como horizonte[3]). O ambiente, por sua vez, amostra uma recompensa $X_t \in \mathbb{R}$ segundo a distribuição P_{A_t}.

O objetivo é maximizar a soma total das recompensas obtidas $S_n = \sum_{t=1}^{n} X_t$, quantidade aleatória que depende das ações tomadas e das recompensas recebidas. Ressalta-se que este não é um problema de otimização clássica, visto que o bandit $\nu = (P_a : a \in \mathcal{A})$ é desconhecido, sabe-se apenas que ν pertence a um conjunto $\mathcal{E}$, denominado classe de ambiente.

1.3 Contribuição

O principal desafio relacionado ao problema de *multi-armed bandits* é identificar a sequência de ações que maximiza a recompensa ao longo de diversas interações. Diversos autores propuseram algoritmos que definem regras de decisão com o objetivo de otimizar esse ganho. Cada algoritmo apresenta características próprias de funcionamento e fornece garantias teóricas quanto a limites superiores e inferiores para a medida de arrependimento.

Dessa forma, o presente trabalho oferece as seguintes contribuições:

- Estudo dos principais algoritmos que consideram o dilema entre exploração e explotação;
- Investigação da classe específica dos algoritmos de exploração pura;
- Avaliação numérica desses algoritmos, analisando seu desempenho em diferentes cenários;
- Desenvolvimento e análise de métodos baseados no problema do bandit de múltiplos braços aplicados à precificação dinâmica.

1.4 Organização

Este texto está estruturado da seguinte forma: no Capítulo 2, são apresentados conceitos preliminares essenciais para a compreensão da teoria desenvolvida nos capítulos seguintes. O Capítulo 3 trata da formulação matemática do problema de *multi-armed bandits*, bem como dos principais algoritmos que abordam o dilema exploração/explotação e os respectivos resultados teóricos. No Capítulo 4, são estudados os algoritmos de exploração pura, com ênfase na identificação da melhor ação dentro de um horizonte pré-determinado. No Capítulo 5, realiza-se uma análise experimental visando compreender o comportamento dos algoritmos. No Capítulo 6, discute-se a precificação dinâmica e sua relação com os algoritmos de *bandits*, apresentando uma aplicação prática. Por fim, as conclusões do trabalho estão no Capítulo 7.

[3] Nesta dissertação, considera-se apenas o caso no qual o horizonte é finito.

2 CONCEITOS PRELIMINARES

Ao longo do desenvolvimento da presente dissertação será visto que um dos principais objetivos ao se estudar o problema do bandit de múltiplos braços é apresentar teoremas que forneçam cotas superiores e inferiores para a medida de arrependimento. Nesse sentido, se faz necessário inicialmente apresentar um conjunto de conhecimentos fundamentais para que o desenvolvimento desses teoremas possa ser realizado de forma sólida e bem estruturada.

Definição 2.1 (Variáveis aleatórias sub-gaussianas)**.** *Uma variável aleatória X com média μ é dita σ-sub-gaussiana se para todo $\lambda \in \mathbb{R}$ vale que:*

$$\mathbb{E}[\exp(\lambda(X-\mu))] \leq \exp\left\{\frac{\lambda^2\sigma^2}{2}\right\}.$$

A proposição seguinte fornece algumas propriedades simples de variáveis aleatórias sub-gaussianas.

Proposição 2.1. *a) Seja X uma variável aleatória σ-sub-gaussiana então:*

1. *$Var(X) \leq \sigma^2$;*
2. *cX é $|c|\sigma$-sub-gaussiana para todo $c \in \mathbb{R}$.*

b) Sejam X_1, X_2 duas variáveis aleatórias independentes σ_1-sub-gaussiana e σ_2-sub-gaussiana respectivamente, então $X_1 + X_2$ é $\sqrt{\sigma_1^2+\sigma_2^2}$-sub-gaussiana.

Prova.

a) 1. Seja $\mu = \mathbb{E}[X]$. Temos

$$Var(X) = Var(X-\mu) = \mathbb{E}[(X-\mu)^2] = \mathcal{M}''_{X-\mu}(0) \leq \sigma^2,$$

onde, $\mathcal{M}_{X-\mu}(\lambda) = \mathbb{E}[\exp(\lambda(X-\mu))]$ e as derivadas são com respeito a λ.

2.

$$\mathbb{E}[\exp(\lambda(cX - c\mu))] = \mathbb{E}[\exp(\lambda c(X-\mu))].$$

Definindo $\lambda_1 = c\lambda$ e usando a Definição 2.1

$$\mathbb{E}[\exp(\lambda\cdot c(X-\mu))] = \mathbb{E}[\exp(\lambda_1(X-\mu))] \leq \exp\left\{\frac{\lambda_1\sigma^2}{2}\right\} = \exp\left\{\frac{\lambda^2(c\sigma)^2}{2}\right\},$$

logo cX é $\sqrt{c^2\sigma^2}-$sub-gaussiana e, portanto, $|c|\sigma-$sub-gaussiana.

b)

$$\mathcal{M}_{X_1+X_2}(\lambda) = \mathbb{E}\left[\exp\{\lambda(X_1+X_2)\}\right] = \mathbb{E}\left[\exp\{\lambda X_1\}\right]\cdot\mathbb{E}\left[\exp\{\lambda X_2\}\right].$$

Usando a Definição 2.1:

$$\mathcal{M}_{X_1+X_2}(\lambda) \leq \exp\left\{\frac{\lambda^2\sigma_1^2}{2}\right\}\exp\left\{\frac{\lambda^2\sigma_2^2}{2}\right\} = \exp\left\{\frac{\lambda^2(\sigma_1^2+\sigma_2^2)}{2}\right\}.$$

□

Alguns exemplos de variáveis sub-gaussianas são:

(a) Se $X \sim \mathcal{N}(0, \sigma^2)$ então X é σ-sub-gaussiana.

(b) Se $X \in [a, b]$ quase certamente, então X é $(b-a)/2$-sub-gaussiana.

O item 1 pode ser verificado uma vez que se $X \sim \mathcal{N}(0, \sigma^2)$ então $\mathbb{E}[\exp(\lambda X)] = \exp\left\{\frac{\sigma^2\lambda^2}{2}\right\}$. Já o item 2 pode ser verificado por meio do seguinte lema.

Lema 2.1 (Lema de Hoeffding)**.** *Seja X uma variável aleatória tal que $a \leq X \leq b$ quase certamente. Então, para todo $\lambda \in \mathbb{R}$:*

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leq \exp\left\{\frac{\lambda^2(b-a)^2}{8}\right\}.$$

Prova. Observe que é suficiente provar o resultado para $\lambda > 0$, enquanto o resultado para $\lambda < 0$ segue deste último aplicado a $-X$ e o caso $\lambda = 0$ é trivial. Como a função $x \mapsto e^{\lambda x}$ é convexa, no intervalo $a \leq x \leq b$ vale

$$e^{\lambda x} = e^{\frac{x-a}{b-a}\lambda b + \frac{b-x}{b-a}\lambda a} \leq \frac{x-a}{b-a}e^{\lambda b} + \frac{b-x}{b-a}e^{\lambda a}.$$

Então,

$$\begin{aligned}\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] &\leq \frac{\mathbb{E}[X]-a}{b-a}e^{\lambda(b-\mathbb{E}[X])} + \frac{b-\mathbb{E}[X]}{b-a}e^{\lambda(a-\mathbb{E}[X])} \\ &= \gamma e^{(1-\gamma)u} + (1-\gamma)e^{-\gamma u} = e^{-\gamma u}(1-\gamma+\gamma e^u) = e^{g(u)},\end{aligned}$$

com $u = \lambda(b-a)$, $\gamma = \dfrac{\mathbb{E}[X]-a}{b-a}$ e $g(u) = -\gamma u + \log(\gamma e^u + (1-\gamma))$. Como $g(u) \geq 0$ para todo $u \geq 0$, para concluir a prova basta mostrar que $g(u) \leq \frac{u^2}{8}$ para todo $u \geq 0$. Vamos representar $g(u)$ em uma série de Taylor com resto:

$$g(u) = g(0) + g'(0)u + \frac{u^2}{2}g''(\xi), \text{ por algum } \xi \in [0, u].$$

Temos que

$$g'(u) = -\gamma + \frac{\gamma e^u}{\gamma e^u + (1-\gamma)} \implies g'(0) = 0,$$

$$g''(\xi) = \frac{\gamma(1-\gamma)e^{-\xi}}{(\gamma + (1-\gamma)e^{-\xi})^2} = \rho_\xi(1-\rho_\xi),$$

onde $\rho_\xi = \frac{\gamma}{\gamma+(1-\gamma)e^{-\xi}} \in (0,1)$ para todo $\xi \geq 0$. Então, temos que $g''(u) \leq 1/4$ o que implica que $g(u) \leq \frac{u^2}{8}$. □

Uma outra consequência do Lema 2.1 é que se $X \sim Ber(p)$ então X é $\dfrac{1}{2}$-sub-gaussiana, uma vez que $\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leq \exp\left\{\lambda^2\left(\frac{1}{2}\right)^2/2\right\}$.

2.1 Desigualdades de concentração

Seja $\{X_i\}_{i=1}^n$ uma sequência de variáveis aleatórias independentes e identicamente distribuídas com média $\mu = \mathbb{E}[X_1]$ e variância σ^2. Além disso, denotemos por $\hat{\mu}$ a média empírica, sendo definida da seguinte forma:

$$\hat{\mu} := \frac{1}{n}\sum_{i=1}^{n} X_i.$$

É possível verificar que a média empírica é um estimador não viesado de μ, ou seja, $\mathbb{E}[\hat{\mu}] = \mu$, além disso $Var(\hat{\mu}) = \dfrac{\sigma^2}{n}$, o que indica que a distância entre μ e $\hat{\mu}$ diminui a medida que n aumenta. Apesar disso, em muitos casos é interessante estudar a distribuição do erro, ou seja, a probabilidade da média empírica superestimar ou subestimar a verdadeira média. Para isso surgem as desigualdades de concentração, oferecendo a possibilidade de encontrar cotas inferiores ou superiores para a probabilidade da média empírica se distanciar um valor $\varepsilon > 0$ da verdadeira média.

Uma das formas diretas e mais simples de cotar a probabilidade de $\hat{\mu}$ se distanciar um valor $\varepsilon > 0$ de μ é através da desigualdade de Chebyshev.

Teorema 2.1 (Desigualdade de Chebyshev)**.** *Seja X uma variável aleatória com média μ, então para todo $\varepsilon > 0$:*

$$\mathbb{P}(|X - \mu| \geq \varepsilon) \leq \frac{Var(X)}{\varepsilon^2}.$$

Aplicando o teorema acima para a média empírica é possível notar que:

$$\mathbb{P}(|\hat{\mu} - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

O resultado apresentado, embora simples, é bastante útil, pois não exige a suposição de uma distribuição específica para a variável aleatória envolvida, dependendo apenas da finitude da média e da variância. Por outro lado, há resultados que fornecem limitantes mais precisos, mas que, em contrapartida, demandam pressupostos mais robustos. O principal deles, que será amplamente utilizado nas demonstrações dos teoremas desta dissertação, é o seguinte:

Teorema 2.2. *Seja X uma variável aleatória σ-sub-gaussiana com média μ, então para todo $\varepsilon > 0$:*

$$\mathbb{P}(X - \mu \geq \varepsilon) \leq \exp\left\{-\frac{\varepsilon^2}{2\sigma^2}\right\}.$$

Prova. Sem perda de generalidade, vamos assumir $\mu = 0$. Para todo $\lambda > 0$

$$\begin{aligned}\mathbb{P}(X \geq \varepsilon) = \mathbb{P}(e^X \geq e^\varepsilon) = \mathbb{P}(e^{\lambda X} \geq e^{\lambda\varepsilon}) &\leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda\varepsilon}} \\ &\leq \frac{\exp\left\{\frac{\lambda^2\sigma^2}{2}\right\}}{e^{\lambda\varepsilon}} = \exp\left\{\frac{\lambda^2\sigma^2}{2} - \lambda\varepsilon\right\},\end{aligned}$$

portanto,

$$\mathbb{P}(X \geq \varepsilon) \leq \min_{\lambda>0} \exp\left\{\frac{\lambda^2\sigma^2}{2} - \lambda\varepsilon\right\},$$

que é alcançado quando $\lambda = \dfrac{\varepsilon}{\sigma^2}$, dessa forma:

$$\mathbb{P}(X \geq \varepsilon) \leq \exp\left\{\frac{\varepsilon^2}{2\sigma^2} - \frac{\varepsilon^2}{\sigma^2}\right\} = \exp\left\{-\frac{\varepsilon^2}{2\sigma^2}\right\}.$$

□

De forma análoga, é possível também verificar que $\mathbb{P}(X - \mu \leq \varepsilon) \leq \exp\left\{-\dfrac{\varepsilon^2}{2\sigma^2}\right\}$, o que conduz ao seguinte lema:

Lema 2.2. *Seja X uma variável aleatória σ-sub-gaussiana com média μ, então para todo $\varepsilon > 0$:*

$$\mathbb{P}(|X - \mu| \geq \varepsilon) \leq 2\exp\left\{-\frac{\varepsilon^2}{2\sigma^2}\right\}.$$

Prova. Sem perda de generalidade vamos assumir $\mu = 0$.

$$\mathbb{P}(|X| \geq \varepsilon) = \mathbb{P}(\{X \geq \varepsilon\} \cup \{X \leq -\varepsilon\}) = \mathbb{P}(X \geq \varepsilon) + \mathbb{P}(X \leq -\varepsilon) \leq 2\exp\left\{-\frac{\varepsilon^2}{2\sigma^2}\right\}.$$

□

Assim como foi feito anteriormente para o Teorema 2.1, é possível aplicar o Teorema 2.2 para a média empírica por meio do seguinte corolário:

Corolário 2.1. *Seja $\{X_i\}_{i=1}^n$ uma sequência de variáveis aleatórias σ-sub-gaussianas independentes e identicamente distribuídas com média μ, então para todo $\varepsilon \geq 0$:*

$$\mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) \leq \exp\left\{-\frac{n\varepsilon^2}{2\sigma^2}\right\} \quad e \quad \mathbb{P}(\hat{\mu} \leq \mu - \varepsilon) \leq \exp\left\{-\frac{n\varepsilon^2}{2\sigma^2}\right\}.$$

Prova.

$$\mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) = \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \geq \mu + \varepsilon\right) = \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mu) \geq \varepsilon\right),$$

como X_i é σ-sub-gaussiana, então $\dfrac{1}{n}\sum_{i=1}^n X_i$ é $\dfrac{\sqrt{n\sigma^2}}{n}$-sub-gaussiana, e, portanto, $\dfrac{\sigma}{\sqrt{n}}$-sub-gaussiana. Logo, pela aplicação do Teorema 2.2 segue que:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mu) \geq \varepsilon\right) \leq \exp\left\{-\frac{\varepsilon^2 n}{2\sigma^2}\right\}.$$

□

Lema 2.3 (Desigualdade de Hoeffding)**.** *Seja $\{X\}_{i=1}^n$ uma sequência de variáveis aleatórias independentes tal que $a_i \leq X_i \leq b_i$ quase certamente e seja $S_n = \sum_{i=1}^n X_i$. Então, temos que:*

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) \leq \exp\left\{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}.$$

Prova.

$$\begin{aligned}
\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) &= \mathbb{P}(\exp\{\lambda(S_n - \mathbb{E}[S_n])\} \geq \exp\{\lambda t\}) \\
&\leq \frac{\mathbb{E}[\exp\{\lambda(S_n - \mathbb{E}[S_n])\}]}{\exp\{\lambda t\})} \\
&= \frac{\prod_{i=1}^{n} \mathbb{E}[\exp\{\lambda(X_i - \mathbb{E}[X_i])\}]}{\exp\{\lambda t\}} \\
&\leq \prod_{i=1}^{n} \exp\left\{\frac{\lambda^2(b_i - a_i)^2}{8}\right\} \cdot \exp\{-\lambda t\} \\
&= \exp\left\{\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2 - \lambda t\right\}.
\end{aligned}$$

Agora basta escolher o λ que maximiza $\exp\left\{\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2 - \lambda t\right\}$, que nesse caso é $\frac{4t}{\sum_{i=1}^{n}(b_i - a_i)^2}$, portanto:

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) \leq \exp\left\{-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right\}.$$

□

Antes de introduzir a próxima desigualdade que será útil em problemas futuros, precisamos do conceito de divergência de Kullback-Leibler.

Definição 2.2. *Dado um espaço mensurável $(\Omega, \mathcal{F})$ e duas medidas P e Q definidas nesse espaço, dizemos P é absolutamente contínua em relação a Q ($P \ll Q$) se $P(A) = 0$ para todo conjunto $A \in \mathcal{F}$ no qual $Q(A) = 0$.*

Definição 2.3 (Divergência de Kullback-Leibler)**.** *Seja $(\Omega, \mathcal{F})$ um espaço mensurável e seja P e Q medidas definidas nesse espaço. A entropia relativa de P com respeito a Q é definida por:*

$$\mathcal{D}(P,Q) = \begin{cases} \int \log\left(\frac{dP}{dQ}(\omega)\right) dP(\omega), & \text{se } P \ll Q \\ \infty, & \text{caso contrário} \end{cases} = \begin{cases} \mathbb{E}_P\left[\log\left(\frac{dP}{dQ}(\omega)\right)\right], & \text{se } P \ll Q \\ \infty, & \text{caso contrário} \end{cases}$$

Para o caso em que P e Q são medidas discretas, assumindo $p_i = P(X = i)$ e $q_i = Q(X = i)$ a expressão acima pode ser escrita da seguinte forma:

$$\mathcal{D}(P,Q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right).$$

Se $P \ll \lambda$ e $Q \ll \lambda$ e $P \ll Q$, então definindo $p = \frac{dP}{d\lambda}$ e $q = \frac{dQ}{d\lambda}$, podemos escrever a divergência de Kullback-leibler utilizando as funções densidade:

$$\mathcal{D}(P,Q) = \int p \log\left(\frac{p}{q}\right) d\lambda.$$

Um dos resultados úteis para a presente dissertação é o caso da distribuição Normal:

Exemplo 2.1. *Seja $X \sim \mathcal{N}(\mu_1, \sigma^2)$ definida no espaço de probabilidade $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \nu)$ e $Y \sim \mathcal{N}(\mu_2, \sigma^2)$ definida em $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \rho)$:*

$$\begin{aligned}
\mathcal{D}(\nu, \rho) &= \mathbb{E}_\nu \left[\log \left(\frac{d\nu}{d\rho} \right) \right] = \mathbb{E}_\nu \left[\log \left(\frac{\frac{d\nu}{d\lambda}}{\frac{d\rho}{d\lambda}} \right) \right] = \mathbb{E}_\nu \left[\log \left(\frac{f_X}{f_Y} \right) \right] \\
&= \mathbb{E}_\lambda \left[\log \left(\frac{f_X}{f_Y} \right) \frac{d\nu}{d\lambda} \right] = \mathbb{E}_\lambda \left[\log \left(\frac{f_X}{f_Y} \right) f_X \right] = \int_{-\infty}^{\infty} \log \left(\frac{f_X}{f_Y} \right) f_X dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \log \left(e^{-\frac{(x-\mu_1)^2 + (x-\mu_2)^2}{2\sigma^2}} \right) dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \log \left(e^{-\frac{1}{2\sigma^2}((x-\mu_1+x-\mu_2)(x-\mu_1-x+\mu_2))} \right) dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \log \left(e^{-\frac{1}{2\sigma^2}((2x-(\mu_1+\mu_2))(\mu_2-\mu_1))} \right) dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \frac{-2x(\mu_2-\mu_1)}{2\sigma^2} dx + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \frac{(\mu_2^2-\mu_1^2)}{2\sigma^2} dx \\
&= \frac{-2(\mu_2-\mu_1)}{2\sigma^2} \cdot \mu_1 + 1 \cdot \frac{(\mu_2^2-\mu_1^2)}{2\sigma^2} = \frac{-2\mu_2\mu_1 + 2\mu_1^2 + \mu_2^2 - \mu_1^2}{2\sigma^2} = \frac{(\mu_2-\mu_1)^2}{2\sigma^2}.
\end{aligned}$$

Além da distribuição Normal, outra distribuição que será utilizada de maneira recorrente no presente texto é a Bernoulli:

Exemplo 2.2. *Seja $X \sim Ber(p_1)$ definida no espaço de probabilidade $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \nu)$ e $Y \sim Ber(p_2)$ definida em $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \rho)$:*

$$\mathcal{D}(\nu, \rho) = \log \left(\frac{1-p_1}{1-p_2} \right) (1-p_1) + \log \left(\frac{p_1}{p_2} \right) p_1,$$

com $0 \cdot \log(\cdot) = 0$.

Agora podemos introduzir a última desigualdade usada na presente dissertação, a desigualdade de Bretagnolle-Huber.

Teorema 2.3 (Desigualdade de Bretagnolle-Huber)**.** *Sejam P e Q duas medidas de probabilidade definidas no mesmo espaço mensurável $(\Omega, \mathcal{F})$ e seja $A \in \mathcal{F}$ um evento arbitrário. Então:*

$$P(A) + Q(A^c) \geq \frac{1}{2} \exp\{-\mathcal{D}(P, Q)\}.$$

A prova do resultado acima é omitida no presente texto, podendo ser encontrada ao final do Capítulo 14 em (LATTIMORE; SZEPESVáRI, 2020).

Por fim, um dos conceitos primordiais para a construção da dinâmica de funcionamento do problema dos bandits com múltiplos braços é o conceito de kernel de probabilidade.

Definição 2.4 (Kernel de probabilidade)**.** *Sejam $(X, \mathcal{A})$ e $(Y, \mathcal{B})$ dois espaços mensuráveis. Um kernel de probabilidade K é uma função $K : X \times \mathcal{B} \longrightarrow [0, 1]$ que apresenta as seguintes propriedades:*

1. *$\forall x \in X,\ K(x, \cdot)$ é uma medida de probabilidade em $(Y, \mathcal{B})$.*
2. *$\forall A \in \mathcal{B},\ K(\cdot, A)$ é mensurável com respeito a $\mathcal{A}$.*

Exemplo 2.3. *Sejam $(X, \mathcal{A})$ e $(Y, \mathcal{B})$ dois espaços mensuráveis, onde:*

$$X = \{a, b\} \quad e \quad \mathcal{A} = 2^X = \{\{a\}, \{b\}, \emptyset, \{a, b\}\},$$

$$Y = \{\alpha, \beta\} \quad e \quad \mathcal{B} = 2^Y = \{\{\alpha\}, \{\beta\}, \emptyset, \{\alpha, \beta\}\}.$$

Consideramos a seguinte função $K : X \times \mathcal{B} \longrightarrow [0, 1]$:

$$\begin{aligned} &K(a, \{\alpha\}) = t_1, \quad K(a, \{\beta\}) = t_2, \quad K(a, \emptyset) = t_3, \quad K(a, Y) = t_4 \\ &K(b, \{\alpha\}) = s_1, \quad K(b, \{\beta\}) = s_2, \quad K(b, \emptyset) = s_3, \quad K(b, Y) = s_4. \end{aligned}$$

Para que a condição 1 seja verificada, temos que:

$$\begin{aligned} &t_4 = s_4 = 1 \quad e \quad t_3 = s_3 = 0, \\ &t_1 + t_2 = 1 \quad e \quad s_1 + s_2 = 1. \end{aligned}$$

Para verificar a condição 2 é preciso primeiramente relembrar a definição de mensurabilidade: uma função $f : X \longrightarrow [0, 1]$ é $\mathcal{F}$-mensurável se $\forall B \in \mathcal{B}([0, 1])$, $f^{-1}(B) \in \mathcal{F}$, onde:

$$f^{-1}(B) = \{x \in X : f(x) \in B\},$$

dessa forma, sendo $B \in \mathcal{B}([0, 1])$

$$K(\cdot, \{\alpha\}) = \begin{cases} \{a\} & t_1 \in B, s_1 \notin B \\ \{b\} & t_1 \notin B, s_1 \in B \\ \{a, b\} & t_1 \in B, s_1 \in B \\ \emptyset & t_1 \notin B, s_1 \notin B \end{cases} \quad e \quad K(\cdot, \{\beta\}) = \begin{cases} \{a\} & t_2 \in B, s_2 \notin B \\ \{b\} & t_2 \notin B, s_2 \in B \\ \{a, b\} & t_2 \in B, s_2 \in B \\ \emptyset & t_2 \notin B, s_2 \notin B \end{cases},$$

$$K(\cdot, \emptyset) = \begin{cases} \{a, b\} & t_3 \in B, s_3 \in B \\ \emptyset & t_3 \notin B, s_3 \notin B \end{cases} \quad e \quad K(\cdot, Y) = \begin{cases} \{a, b\} & t_4 \in B, s_4 \in B \\ \emptyset & t_4 \notin B, s_4 \notin B \end{cases}.$$

3 MULTI-ARMED BANDITS

Seja $\mathcal{A}$ o conjunto de todas as ações possíveis (no contexto da presente dissertação, será assumido que $|\mathcal{A}|$ é finito e $\mathcal{A} = \{1, 2, \ldots, |\mathcal{A}|\} \subset \mathbb{N}$). Um bandit estocástico é uma coleção de distribuições $\nu = (P_a : a \in \mathcal{A})$ onde P_a é uma medida de probabilidade em $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ para todo $a \in \mathcal{A}$.

A dinâmica de funcionamento do problema do bandit com múltiplos braços consiste em um agente que interage com o ambiente e em cada rodada $t \in \{1, \ldots, n\}$ seleciona uma ação $A_t \in \mathcal{A}$, sendo n o número total de rodadas (também conhecido como horizonte). O ambiente, em contrapartida, amostra uma recompensa $X_t \in \mathbb{R}$ de acordo com a distribuição P_{A_t}. Sendo $\Omega_n := (\mathcal{A} \times \mathbb{R})^n$ o espaço das sequências de ações e recompensas, as variáveis aleatórias $A_1, X_1, \ldots, A_n, X_n$ são definidas sobre $(\Omega_n, \mathcal{B}(\Omega_n))$ como:

$$A_t(a_1, x_1, \ldots, a_n, x_n) = a_t, \qquad X_t(a_1, x_1, \ldots, a_n, x_n) = x_t,$$

para todo $t \in \{1, \ldots, n\}$. Seja $\mathbb{P}$ a medida de probabilidade sobre $\Omega_n := (\mathcal{A} \times \mathbb{R})^n$ tal que para todo $t \in \{1, \ldots, n\}$:

(a) a distribuição condicional da recompensa X_t dado $A_1, X_1, \ldots, A_{t-1}, X_{t-1}, A_t$ é P_{A_t} quase certamente;

(b) a distribuição condicional da ação A_t dado $A_1, X_1, \ldots, A_{t-1}, X_{t-1}$ é $\pi_t(\cdot|A_1, X_1, \ldots, A_{t-1}, X_{t-1})$ quase certamente, onde $\pi_1, \pi_2, \ldots$ é uma sequência de kernels de probabilidade que caracterizam o agente. A sequencia $\pi = (\pi_t)_{t=1}^n$ é chamada de *política*.

O ponto (a) codifica o fato que a distribuição da recompensa no tempo t depende apenas da ação escolhida no tempo t. O ponto (b) codifica o fato que a escolha da ação no tempo t depende das recompensas e ações passadas e que o agente não pode utilizar observações futuras na decisão do presente. É importante notar que a medida $\mathbb{P}$ depende da política π e do bandit ν.

O objetivo é maximizar a soma total de recompensas recebidas $S_n = \sum_{t=1}^n X_t$, que é uma quantidade aleatória que depende das ações a das recompensas. É importante ressaltar que este não é um problema de otimização uma vez que o bandit $\nu = (P_a : a \in \mathcal{A})$ é desconhecido. Tudo o que o agente sabe é que ν pertence a algum conjunto $\mathcal{E}$, chamado classe de ambiente.

Para avaliar o desempenho de cada braço em um problema de bandit, uma medida fundamental é a média empírica das recompensas, que fornece uma estimativa do retorno esperado com base nas recompensas observadas até o momento. A média empírica $\hat{\mu}_i(t)$ da ação i no tempo t é calculada como a média aritmética das recompensas obtidas ao selecionar essa ação ao longo das tentativas anteriores. Essa medida é essencial para algoritmos de tomada de decisão, pois permite comparar a eficiência das ações e direcionar as escolhas futuras. Formalmente, a média empírica $\hat{\mu}$ de cada braço i no tempo t é definida através da seguinte expressão:

$$\hat{\mu}_i(t) := \frac{1}{T_i(t)} \sum_{s=1}^{t} \mathbb{1}_{\{A_s = i\}} X_s, \tag{3.1}$$

onde $T_i(t) := \sum_{s=1}^{t} \mathbb{1}_{\{A_s=i\}}$ representa o número de vezes que a ação i foi selecionada até a rodada t.

3.1 Bandits estruturados x Bandits não-estruturados

Uma importantes distinção que precisa ser feita é entre bandits estruturados e não-estruturados. A característica fundamental que os diferencia é a definição da *classe do ambiente* $\mathcal{E}$. Para os bandits não estruturados é assumido que $\mathcal{E} = \times_{a\in\mathcal{A}}\mathcal{M}_a$ onde $\mathcal{M}_a$ denota o conjunto de distribuições associadas a ação $a \in \mathcal{A}$. O produto cartesiano codifica o fato que quando o agente escolhe uma ação a, essa escolha não revela nada das distribuições de outras ações. Classes de ambiente típicas para os bandits não estruturados são: Bernoulli - $\mathcal{E}_{\text{Ber}}^{|\mathcal{A}|} := \{(Ber(\mu_a))_{a\in\mathcal{A}} : \mu_a \in [0,1]\}$, Gaussiana - $\mathcal{E}_{\text{Gau}}^{|\mathcal{A}|}(\sigma^2) := \{(\mathcal{N}(\mu_a,\sigma^2))_{a\in\mathcal{A}} : \mu_a \in \mathbb{R}\}$, Subgaussiana - $\mathcal{E}_{\text{SGau}}^{|\mathcal{A}|}(\sigma^2) := \{(P_a)_{a\in\mathcal{A}} : P_a \text{ é } \sigma\text{-subgaussiana}\}$. Os bandits cuja classe de ambiente não respeita a regra supracitada são chamados estruturados. Para o caso dos bandits estruturados, o agente pode tomar uma ação e obter informação sobre a distribuição das demais ações, mesmo sem tê-las tomado. Na presente dissertação serão explorados apenas bandits não estruturados.

3.2 Arrependimento

Um dos conceitos fundamentais no problema dos bandits com múltiplos braços é o conceito de arrependimento (do inglês *regret*). O arrependimento pode ser entendido como a diferença entre o total de recompensas obtido caso o agente tomasse a ação ótima (*i.e.*,a ação com a maior média) em todas as rodadas e a esperança do somatório das recompensas obtidas em cada uma das rodadas. Denotando por $\mu_a(\nu) := \int_{-\infty}^{\infty} x dP_a(x)$ a média da recompensa para a ação $a \in \mathcal{A}$, temos que o arrependimento da política π para o bandit ν é definido como:

$$R_n(\pi,\nu) := n\mu^*(\nu) - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right], \tag{3.2}$$

onde, $\mu^*(\nu) := \max_{a\in\mathcal{A}} \mu_a(\nu)$ é a maior média dentre todos os braços e a esperança é com respeito à $\mathbb{P}$. Vale ressaltar que no presente texto será assumido que $\mu_a(\nu)$ existe e é finita para todo a $\in \mathcal{A}$. Uma forma alternativa de representar o arrependimento é por meio do uso do lema de decomposição do arrependimento apresentado abaixo.

Lema 3.1 (Lema 4.5 em (LATTIMORE; SZEPESVáRI, 2020))**.** *Para qualquer política π e bandit ν com conjunto de ações $\mathcal{A}$ finito ou contável e horizonte $n \in \mathbb{N}$, o arrependimento R_n satisfaz:*

$$R_n(\pi,\nu) = \sum_{a\in\mathcal{A}} \Delta_a \mathbb{E}[T_a(n)], \tag{3.3}$$

onde, $T_a(n) := \sum_{t=1}^{n} \mathbb{1}_{\{A_t=a\}}$ conta o número de vezes que a ação a foi selecionada nas n rodadas e $\Delta_a := \mu^*(\nu) - \mu_a(\nu)$ é chamado de intervalo sub-ótimo.

Prova.

$$
\begin{aligned}
R_n(\pi,\nu) &= n\mu^* - \mathbb{E}[S_n] = n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right] = n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n}\sum_{i=1}^{k} X_t \mathbb{1}_{\{A_t=i\}}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n}\left(-\sum_{i=1}^{k} X_t \mathbb{1}_{\{A_t=i\}} + \mu^*\right)\right] = \mathbb{E}\left[\sum_{t=1}^{n}\sum_{i=1}^{k}(\mu^* - X_t)\mathbb{1}_{\{A_t=i\}}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\sum_{t=1}^{n}\sum_{i=1}^{k}(\mu^* - X_t)\mathbb{1}_{\{A_t=i\}}|A_t\right]\right] = \mathbb{E}\left[\sum_{t=1}^{n}\sum_{i=1}^{k}\mathbb{E}\left[(\mu^* - X_t)\mathbb{1}_{\{A_t=i\}}|A_t\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n}\sum_{i=1}^{k}\mathbb{1}_{\{A_t=i\}}\mathbb{E}\left[(\mu^* - X_t)|A_t\right]\right] = \mathbb{E}\left[\sum_{t=1}^{n}\sum_{i=1}^{k}\mathbb{1}_{\{A_t=i\}}(\mathbb{E}[\mu^*|A_t] - \mathbb{E}[X_t|A_t])\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n}\sum_{i=1}^{k}\mathbb{1}_{\{A_t=i\}}(\mu^* - \mu_{A_t})\right] = \mathbb{E}\left[\sum_{t=1}^{n}\sum_{i=1}^{k}\mathbb{1}_{\{A_t=i\}}(\mu^* - \mu_i)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n}\sum_{i=1}^{k}\mathbb{1}_{\{A_t=i\}}\Delta_i\right] = \sum_{i=1}^{k}\Delta_i\mathbb{E}\left[\sum_{t=1}^{n}\mathbb{1}_{\{A_t=i\}}\right] = \sum_{i=1}^{k}\Delta_i\mathbb{E}\left[T_i(n)\right].
\end{aligned}
$$

□

3.3 Exploração x Explotação

Para resolver o problema dos bandits com múltiplos braços, foram propostos diversos algoritmos, sendo uma característica comum entre eles o objetivo de estabelecer uma política que ao ser adotada minimize o arrependimento. Dentro desse contexto, um aspecto objetivado pelos proponentes dos algoritmos é encontrar cotas superiores e inferiores para o arrependimento, conseguindo dessa forma ter um maior controle sobre o resultado obtido ao adotar o algoritmo.

Na literatura existem vários algoritmos que abordam esse problema, dentre eles é possível citar: algoritmo guloso (do inglês *greedy algorithm*), algoritmo ε-guloso (do inglês *ε-greedy algorithm*),algoritmo *Explore-Then-Commit* - ETC e o algoritmo *Upper Confidence Bound* - UCB. Sendo o último um dos principais algoritmos adotados para solucionar o problema dos bandits com múltiplos braços.

3.3.1 Algoritmo Guloso

O algoritmo guloso é o algoritmo mais simples que foca na relação entre exploração e explotação. Seu funcionamento é apresentado no Algoritmo 1.

Algoritmo 1 Guloso

Entrada: $n, k := |\mathcal{A}|$
  Escolha cada ação em $\mathcal{A}$ exatamente uma vez *(exploração)*
Para cada $t \in \{k+1, \ldots, n\}$ faça:
  Escolha $A_t = \arg\max_{i\in\mathcal{A}} \hat{\mu}_i(t-1)$ *(explotação)*
fim-para

É possível perceber que o Algoritmo 1 consiste apenas em escolher a ação que possui maior média empírica até a rodada t. Nesse sentido, é perceptível que trata-se de um algoritmo que em um

primeiro momento não parece ser confiável, uma vez que, dado que as recompensas são aleatórias, flutuações iniciais dos valores de recompensa em relação a sua média podem incorrer na escolha de ações sub-ótimas ao longo de todo o horizonte.

A fim de ilustrar o aspecto da confiabilidade do algoritmo, suponha um ambiente $\mathcal{E} = \{Ber(0,5), Ber(0,8)\}$, evidentemente a segunda ação possui maior média, entretanto, se a primeira amostra for 1 para a primeira ação e 0 para a segunda, então a ação 1 será escolhida até o final do horizonte, ainda que a segunda ação seja de fato a ótima. Esse cenário ilustra o quão frágil o algoritmo é em relação a possibilidade de escolhas subótimas.

3.3.2 Algoritmo ε-guloso

O algoritmo ε-guloso é um modificação do algoritmo guloso que tenta resolver a fraqueza apresentada pelo último por meio da atribuição de uma probabilidade ε de que o algoritmo selecione uniformemente qualquer uma das ações contidas no conjunto de ações $\mathcal{A}$. Uma descrição detalhada de seu funcionamento é apresentada no Algoritmo 2, sendo perceptível que, dada sua construção, sempre se torna possível que o algoritmo não fique eternamente preso em uma ação sub-ótima.

Algoritmo 2 ε-guloso

Entrada: $n, k := |\mathcal{A}|, \varepsilon$

Escolha cada ação em $\mathcal{A}$ exatamente uma vez

Para cada $t \in \{k+1, \ldots, n\}$ faça:

$$\text{Escolha } A_t = \begin{cases} \arg\max_{i \in \mathcal{A}} \hat{\mu}_i(t-1) & \text{com probabilidade } 1-\varepsilon \\ \text{uniformemente em } \mathcal{A} & \text{com probabilidade } \varepsilon \end{cases}$$

fim-para

Cabe ressaltar que, apesar de ser uma possível solução para o problema apresentado pelo algoritmo guloso, o algoritmo ε-guloso introduz um novo parâmetro ε que precisa ser ajustado de maneira adequada, sendo essa uma tarefa de bastante importância, pois escolhas elevadas de ε levam ao algoritmo escolher ações sub-ótimas de forma desnecessária, por outro lado, escolhas de valores baixos para ε podem levar o algoritmo a ficar preso em ações sub-ótimas por longos períodos.

Resultados envolvendo limitantes superiores para o algoritmo ε-guloso podem ser encontrados em (LATTIMORE; SZEPESVáRI, 2020).

3.3.3 Algoritmo Explore-Then-Commit (ETC)

O algoritmo ETC pode ser visto como uma evolução natural do algoritmo guloso, haja vista que, por meio da observação de seu funcionamento apresentado no Algoritmo 3, nota-se que ainda se baseia na escolha da ação que apresenta maior média empírica, porém, durante a fase de exploração, o algoritmo consiste em selecionar cada uma das ações um número fixo de vezes e em seguida, na fase de explotação, é realizada a escolha da ação que apresentou maior média empírica na fase de exploração para ser executada pelo resto do horizonte.

De maneira específica, considera-se um conjunto $\mathcal{A}$ de ações onde $|\mathcal{A}| = k$, o algoritmo irá explorar cada uma das k ações m vezes antes de selecionar a que apresenta maior média empírica,

onde $1 \leq m \leq \frac{n}{k}$ é um parâmetro do algoritmo.

Algoritmo 3 ETC

Entrada: $m, n, k = |\mathcal{A}|$

Para cada $t \in \{1, \ldots, n\}$ faça:

Escolha a ação A_t tal que $A_t = \begin{cases} (t \mod k) + 1, & \text{se } t \leq mk \quad \textit{(exploração)} \\ \arg\max_i \hat{\mu}_i(mk), & \text{se } t > mk \quad \textit{(explotação)} \end{cases}$

fim-para

Conforme mencionado na introdução do presente texto, um dos interesses principais ao estabelecer um algoritmo é ter controle sobre o arrependimento. Dessa forma, o Teorema 3.1 oferece um limitante superior para o arrependimento dependendo apenas do horizonte n, do número de vezes que cada ação será selecionada no período de exploração (m), do número total de ações e do intervalo sub-ótimo de cada ação (Δ_i), onde $\Delta_i = \mu^* - \mu_i$, sendo μ^* a média da ação ótima e μ_i a média da i-ésima ação.

Teorema 3.1 (Teorema 6.1 em (LATTIMORE; SZEPESVáRI, 2020))**.** *Assumindo o algoritmo ETC com $1 \leq m \leq n/k$ e um ambiente $\mathcal{E}^k_{SGau}(\sigma)$, segue que:*

$$R_n \leq m \sum_{i=1}^{k} \Delta_i + (n - mk) \sum_{i=1}^{k} \Delta_i \exp\left\{-\frac{m\Delta_i^2}{4\sigma^2}\right\}.$$

Prova. A ideia da prova se concentrará em primeiramente encontrar um limitante superior para o número médio de vezes que cada ação é selecionada e, através do Lema 3.1, encontrar um limitante superior para o arrependimento. Com base no Algoritmo 3, dado m e n é possível representar $T_i(n)$ da seguinte forma:

$$\begin{aligned} T_i(n) &= \sum_{s=1}^{n} \mathbb{1}_{\{A_s = i\}} = m + \sum_{s=mk+1}^{n} \mathbb{1}_{\{A_s = i\}} = m + (n - mk)\mathbb{1}_{\{A_{mk+1} = i\}} \\ &= m + (n - mk)\mathbb{1}_{\{\hat{\mu}_i(mk) \geq \max_{i \neq j} \hat{\mu}_j(mk)\}}, \end{aligned}$$

por meio dessa representação é possível obter o número médio de vezes que cada ação será realizada:

$$\begin{aligned} \mathbb{E}[T_i(n)] &= \mathbb{E}[m + (n - mk)\mathbb{1}_{\{\hat{\mu}_i(mk) \geq \max_{i \neq j} \hat{\mu}_j(mk)\}}] = m + (n - mk)\mathbb{E}[\mathbb{1}_{\{\hat{\mu}_i(mk) \geq \max_{i \neq j} \hat{\mu}_j(mk)\}}] \\ &= m + (n - mk)\mathbb{P}\left(\hat{\mu}_i(mk) \geq \max_{i \neq j} \hat{\mu}_j(mk)\right). \end{aligned}$$

Sem perda de generalidade, será assumido que a ação 1 é ótima, dessa forma, intuitivamente é esperado que a média empírica da ação 1 seja maior que a média empírica das demais ações, nesse sentido, o pior cenário possível ocorre quando a média empírica de uma ação $i \neq 1$ é maior que a média empírica da ação 1, nesse caso,

$$\mathbb{P}\left(\hat{\mu}_i(mk) \geq \max_{i \neq j} \hat{\mu}_j(mk)\right) \leq \mathbb{P}(\hat{\mu}_i(mk) \geq \hat{\mu}_1(mk)), \tag{3.4}$$

essa relação é valida porque $\hat{\mu}_1(mk) \leq \max_{i \neq j} \hat{\mu}_j(mk)$, ou seja, se o máximo é o próprio $\hat{\mu}_1(mk)$, então $\hat{\mu}_1(mk) = \max_{i \neq j} \hat{\mu}_j(mk)$ e temos a igualdade, caso não seja, então $\hat{\mu}_1(mk) < \max_{i \neq j} \hat{\mu}_j(mk)$, o que nos leva a concluir que: $\hat{\mu}_1(mk) \leq \hat{\mu}_i(mk) \iff \max_{i \neq j} \hat{\mu}_j(mk) \leq \hat{\mu}_i(mk)$, garantindo dessa forma

que $\mathbb{P}\left(\hat{\mu}_i(mk) \geq \max_{i\neq j} \hat{\mu}_j(mk)\right) \leq \mathbb{P}(\hat{\mu}_i(mk) \geq \hat{\mu}_1(mk)$. Utilizando a Equação (3.1) é possível obter o seguinte resultado:

$$\hat{\mu}_i(mk) = \frac{\sum_{t=1}^{mk} \mathbb{1}_{\{A_t=i\}} X_t}{T_i(mk)}.$$

Pela definição do algoritmo, segue que $T_i(mk) = m$, além disso, considerando um ambiente $\mathcal{E} = \mathcal{E}^k_{SGau}(\sigma)$, segue que X_t é σ-sub-gaussiana para todo t. Como

$$\hat{\mu}_i(mk) = \frac{\sum_{t=1}^{mk} \mathbb{1}_{\{A_t=i\}} X_t}{m} = \sum_{t=1}^{mk} \mathbb{1}_{\{A_t=i\}} \frac{X_t}{m},$$

e $\{X_t\}_{t=1}^n$ são independentes. Segue pela Proposição2.1 que $\dfrac{X_t}{m}$ é $\dfrac{\sigma}{m}$-sub-gaussiana, portanto:

$$\sum_{t=1}^{mk} \mathbb{1}_{\{A_t=i\}} \frac{X_t}{m} \text{ é } \sqrt{\underbrace{\frac{\sigma^2}{m^2} + \frac{\sigma^2}{m^2} + \cdots + \frac{\sigma^2}{m^2}}_{m \text{ vezes}}}\text{-sub-gaussiana} \implies \sum_{t=1}^{mk} \mathbb{1}_{\{A_t=i\}} \frac{X_t}{m} \text{ é } \frac{\sigma}{\sqrt{m}}\text{-sub-gaussiana},$$

garantindo, também por meio da Proposição 2.1 que:

$$\left(\hat{\mu}_i(mk) - \mu_i\right) \text{ é } \frac{\sigma}{\sqrt{m}}\text{-sub-gaussiana} \quad \text{e} \quad \left(\hat{\mu}_1(mk) - \mu_1\right) \text{ é } \frac{\sigma}{\sqrt{m}}\text{-sub-gaussiana},$$

o que por fim garante que:

$$\left(\hat{\mu}_i(mk) - \hat{\mu}_1(mk)\right) \text{ é } \sqrt{\frac{2\sigma^2}{m}}\text{-sub-gaussiana.}$$

Utilizando o Teorema 2.2, segue que:

$$\mathbb{P}(\hat{\mu}_i(mk) \geq \hat{\mu}_1(mk)) = \mathbb{P}\left(\hat{\mu}_i(mk) - \mu_i - (\hat{\mu}_1(mk) - \mu_1) \geq \Delta_i\right) \leq \exp\left\{-\frac{\Delta_i^2}{2 \cdot \dfrac{2\sigma^2}{m}}\right\} = \exp\left\{-\frac{m\Delta_i^2}{4\sigma^2}\right\}.$$

Unindo o resultado acima com o encontrado em (3.4), é possível encontrar o seguinte limitante superior para $\mathbb{E}[T_i(n)]$:

$$\mathbb{E}[T_i(n)] \leq m + (n - mk) \exp\left(-\frac{m\Delta_i^2}{4\sigma^2}\right),$$

consequentemente, também é possível definir um limitante superior para o arrependimento por meio do Lema (3.1):

$$R_n \leq m \sum_{i=1}^{k} \Delta_i + (n - mk) \sum_{i=1}^{k} \Delta_i \exp\left(-\frac{m\Delta_i^2}{4\sigma^2}\right).$$

□

Em linhas gerais, o limitante superior oferecido pelo Teorema 3.1 garante que, para o algoritmo ETC, o arrependimento é sempre menor que algo que cresce linearmente em n, onde, a primeira parcela da soma evidencia o arrependimento decorrente do período de exploração, enquanto a segunda parcela se relaciona ao arrependimento durante o período de explotação.

Uma maneira de visualizar o resultado acima é considerar um caso básico em que $k = 2$. Para esse caso específico, assumindo a ação 1 como ótima, temos o seguinte resultado:

$$R_n \leq m\Delta_2 + (n - 2m) \exp\left(-\frac{m\Delta_2^2}{4\sigma^2}\right)\Delta_2 \leq m\Delta_2 + n \exp\left(-\frac{m\Delta_2^2}{4\sigma^2}\right)\Delta_2.$$

Neste caso, uma escolha possível para m é escolhe-lo de forma a minimizar o valor a direita da desigualdade acima, obtendo assim o m que promove o menor limitante superior. Dessa forma, definindo

$$f(x) = x\Delta_2 + n\exp\left(-\frac{x\Delta_2^2}{4\sigma^2}\right)\Delta_2. \tag{3.5}$$

segue que:

$$\frac{\partial f}{\partial x} = \Delta_2\left(1 - n\frac{\Delta_2^2}{4\sigma^2}\exp\left(-\frac{x\Delta_2^2}{4\sigma^2}\right)\right),$$

portanto,

$$\Delta_2\left(1 - n\frac{\Delta_i^2}{4\sigma^2}\exp\left(-\frac{x\Delta_2^2}{4\sigma^2}\right)\right) = 0 \implies 1 - n\frac{\Delta_2^2}{4\sigma^2}\exp\left(-\frac{x\Delta_2^2}{4\sigma^2}\right) = 0$$

$$\implies \exp\left(-\frac{x\Delta_2^2}{4\sigma^2}\right) = \frac{1}{\frac{n\Delta_2^2}{4\sigma^2}} \implies -\frac{x\Delta_2^2}{4\sigma^2} = \log\left(\frac{1}{\frac{n\Delta_2^2}{4\sigma^2}}\right) \implies -\frac{x\Delta_2^2}{4\sigma^2} = -\log\left(\frac{n\Delta_2^2}{4\sigma^2}\right)$$

$$\implies x = \frac{4\sigma^2}{\Delta_2^2}\log\left(\frac{n\Delta_2^2}{4\sigma^2}\right),$$

Substituindo m por x em 3.5 tem-se que:

$$R_n \leq \frac{4\sigma^2}{\Delta_2}\log\left(\frac{n\Delta_2^2}{4\sigma^2}\right) + n\exp\left(-\log\left(\frac{n\Delta_2^2}{4\sigma^2}\right)\right)\Delta_2.$$

Dessa forma, é possível notar que para o caso em que $k = 2$, a escolha de m apresentada garante um limitante superior que cresce de maneira logarítmica a medida que n aumenta. Sendo esse um resultado melhor que o encontrado inicialmente no Teorema 3.1, evidenciando que a escolha de m é um fator de grande importância.

3.3.4 Algoritmo Upper Confidence Bound (UCB)

O algoritmo UCB, assim como o ETC, também se concentra na relação entre exploração e explotação, mas, diferentemente do último, realiza essas etapas de maneira simultânea, assim como apresentado no Algoritmo 4.

Algoritmo 4 UCB

Entrada: $n, k, \delta \in (0, 1)$

Escolha cada ação A_t exatamente uma vez *(exploração)*

Para cada $t \in \{k+1, \ldots, n\}$ faça:

Calcule $UCB_i(t-1) = \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}}$ *(exploração/explotação)*

Escolha $A_t = \arg\max_i(UCB_i(t-1))$

fim-para

A definição do Algoritmo 4 surge de duas principais ideias:

1. Selecionar a ação com maior média empírica.

2. Evitar que ações sub-ótimas sejam selecionadas perpetuamente.

O primeiro ponto é de fato coerente, uma vez que, pela lei dos grandes números, espera-se que, para amostras grandes, a média empírica convirja para a verdadeira média. No entanto, no presente cenário esse tipo de raciocínio precisa ser tomado com bastante cuidado, uma vez que deve ser aplicado a todas as ações, dessa forma, se uma ação for escolhida poucas vezes, não é possível ter nenhuma garantia. Portanto, a adição do termo $\sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}}$ visa garantir que não ocorra de alguma ação ser amostrada de maneira sub-representativa.

Dessa maneira, se uma ação é selecionada um menor número de vezes em relação as demais, então $T_i(t-1)$ será pequeno, tornando, consequentemente, $\sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}}$ grande, favorecendo a seleção dessa ação perante as demais. Ainda nesse sentido, é possível verificar versões diferentes do algoritmo UCB, como por exemplo a apresentada em (BUBECK; CESA-BIANCHI, 2012) em que aparece o termo $\log(t)$ multiplicado por um fator α no numerador ao invés do $\log(1/\delta)$, contribuindo assim para que conforme o número de iterações do algoritmo aumente, maior seja a chance de escolher ações pouco escolhidas.

Já a parte de explotação é representada pelo termo $\hat{\mu}_i(t-1)$, refletindo no fato de que quanto maior o número de amostras geradas ao longo da execução do algoritmo, menor será a influência do fator $\sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}}$, fazendo com que a média amostral impacte de forma mais preponderante na escolha da ação. É importante ressaltar que, apesar disso, escolhas muito pequenas de δ podem fazer com que o termo $\sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}}$ domine e o algoritmo acabe explorando demais.

O surgimento do fator $\sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}}$ decorre do Teorema 2.2 ao definir $\varepsilon = \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}}$. Além disso, deve ser notado que, nesse caso, $T_i(t-1)$ é uma variável aleatória e não uma constante assim como definido no Teorema 2.2. No entanto, de acordo com (LATTIMORE; SZEPESVáRI, 2020) isso se trata apenas de uma tecnicalidade.

Focando especificamente em um ambiente sub-gaussiano, por meio do Teoremas 2.2 é possível definir:

$$UCB_i(t-1) = \hat{\mu}_i(t-1) + \sqrt{\frac{2\sigma^2\log(1/\delta)}{T_i(t-1)}}.$$

A partir disso e de acordo com a discussão promovida para o algoritmo ETC, no caso do algoritmo UCB também estamos interessados em saber quanto, em média, cada ação será selecionada ao longo da execução do algoritmo e, através disso, ter controle sobre o arrependimento. Assim como visto na seção anterior, com efeito não é possível obter diretamente um número fixo para cada um desses objetos de interesse, sendo necessário obter limitantes.

Teorema 3.2 (Teorema 7.1 em (LATTIMORE; SZEPESVáRI, 2020))**.** *Considere um ambiente* $\mathcal{E}^k_{SGau}(\sigma)$, $\delta = 1/n^2$ *e que a política adotada siga o algoritmo UCB, então:*

$$R_n \leq 3\sum_{i=1}^{k}\Delta_i + \sum_{i:\Delta_i>0}\frac{16\sigma^2\log(n)}{\Delta_i}.$$

Prova. Diferentemente do abordado em (LATTIMORE; SZEPESVáRI, 2020), esta dissertação não se limita ao caso específico de $\sigma = 1$.

Para provar esse teorema, primeiro se faz necessário introduzir algumas notações que serão utilizadas ao longo da prova.

Será definido por $\{X_{ti}\}_{t\in[n],i\in[k]}$ a coleção de variáveis aleatórias independentes, onde cada elemento X_{ti} representa a recompensa obtida ao tomar a ação i no tempo t. Além disso, será utilizada a seguinte definição:

$$\hat{\mu}_{is} := \frac{1}{s}\sum_{u=1}^{s} X_{ui}. \tag{3.6}$$

Será assumido sem perda de generalidade que a ação 1 é a melhor ação.

Definamos o seguinte evento:

$$G_i := \left\{\mu_1 < \min_{t\in[n]} UCB_1(t,\delta)\right\} \cap \left\{\hat{\mu}_{iu_i} + \sqrt{\frac{2\sigma^2\log(1/\delta)}{u_i}} < \mu_1\right\}.$$

Onde u_i é uma constante que será definida posteriormente. Em palavras, G_i é o evento onde μ_1 nunca é subestimado pelo limitante superior da primeira ação, ou seja, o limitante superior em qualquer tempo t está sempre acima de μ_1. Ao mesmo tempo, o limitante superior para a média da ação i depois de u_i observações serem realizadas é inferior a μ_1.

Nesse caso, se faz necessário provar que:

1. Se G_i ocorre, então a ação i será selecionada no máximo u_i vezes, ou seja, $T_i(n) \leq u_i$.
2. O complementar de G_i (G_i^c) ocorre com baixa probabilidade.

Sabemos que sempre é verdade que $T_i(n) \leq n$, então podemos escrever:

$$\mathbb{E}[T_i(n)] = \mathbb{E}[\mathbb{1}_{\{G_i\}}T_i(n)] + \mathbb{E}[T_i(n)\mathbb{1}_{\{G_i^c\}}] \leq u_i + n\mathbb{P}(G_i^c).$$

Nessa desigualdade estamos usando o fato de que quando G_i ocorre, então $T_i(n) \leq u_i$, ou seja:

$$\mathbb{1}_{\{G_i\}} = 1 \Rightarrow T_i(n) \leq u_i,$$

o que pode ser verificado da seguinte forma:

Vamos supor que a proposição acima seja falsa, ou seja:

$$(\mathbb{1}_{\{G_i\}} = 1) \text{ e } (T_i(n) > u_i),$$

por definição:

$$UCB_i(t-1,\delta) = \hat{\mu}_i(t-1) + \sqrt{\frac{2\sigma^2\log(1/\delta)}{T_i(t-1)}}.$$

Suponha que $T_i(n) > u_i$, isso significa dizer que a ação i foi selecionada mais do que u_i vezes ao longo das n rodadas, então deve existir uma rodada $t \in [n]$ onde $T_i(t-1) = u_i$ e $A_t = i$ e:

$$UCB_i(t-1,\delta) = \hat{\mu}_{iu_i} + \sqrt{\frac{2\sigma^2\log(1/\delta)}{u_i}}.$$

Pela definição de G_i segue que:

- $UCB_i(t-1,\delta) < u_1$
- $UCB_i(t-1,\delta) < UCB_1(t-1,\delta)$

Pela definição do algoritmo, $A_t = \arg\max_j UCB_j(t-1,\delta) \neq i$, uma vez que será igual a ação 1 ou outra ação diferente de i, o que é uma contradição e, portanto:

$$\mathbb{1}_{\{G_i\}} = 1 \implies T_i(n) \leq u_i.$$

Para provar o ponto 2, temos que, por definição:

$$G_i^c = \left\{\mu_1 \geq \min_{t\in[n]} UCB_1(t,\delta)\right\} \cup \left\{\hat{\mu}_{iu_i} + \sqrt{\frac{2\sigma^2\log(1/\delta)}{u_i}} \geq \mu_1\right\}.$$

Focando apenas no primeiro conjunto $\{\mu_1 \geq \min_{t\in[n]} UCB_1(t,\delta)\}$, temos que:

$$\begin{aligned}\left\{\mu_1 \geq \min_{t\in[n]} UCB_1(t,\delta)\right\} &\subset \left\{\mu_1 \geq \min_{s\in[n]} \hat{\mu}_{1s} + \sqrt{\frac{2\sigma^2\log(1/\delta)}{s}}\right\} \\ &= \bigcup_{s\in[n]} \left\{\mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2\sigma^2\log(1/\delta)}{s}}\right\},\end{aligned}$$

portanto:

$$\begin{aligned}\mathbb{P}\left(\mu_1 \geq \min_{t\in[n]} UCB_1(t,\delta)\right) &\leq \mathbb{P}\left(\bigcup_{s\in[n]} \left\{\mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2\sigma^2\log(1/\delta)}{s}}\right\}\right) \\ &\leq \sum_{s=1}^{n} \mathbb{P}\left(\mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2\sigma^2\log(1/\delta)}{s}}\right) \leq n\delta.\end{aligned}$$

Focando agora no segundo conjunto e assumindo u_i suficientemente grande tal que:

$$\Delta_i - \sqrt{\frac{2\sigma^2\log(1/\delta))}{u_i}} \geq c\Delta_i,$$

para um $c \in (0,1)$ a ser escolhido posteriormente. Usando o fato que $\mu_1 = \mu_i + \Delta_i$, segue que:

$$\begin{aligned}\mathbb{P}\left(\hat{\mu}_{iu_i} + \sqrt{\frac{2\sigma^2\log(1/\delta)}{u_i}} \geq \mu_1\right) &= \mathbb{P}\left(\hat{\mu}_{iu_i} + \sqrt{\frac{2\sigma^2\log(1/\delta)}{u_i}} \geq \mu_i + \Delta_i\right) \\ &= \mathbb{P}\left(\hat{\mu}_{iu_i} - \mu_i \geq \Delta_i - \sqrt{\frac{2\sigma^2\log(1/\delta)}{u_i}}\right) \\ &\leq \mathbb{P}\left(\hat{\mu}_{iu_i} - \mu_i \geq c\Delta_i\right) \leq \exp\left(-\frac{u_i c^2\Delta_i^2}{2\sigma^2}\right).\end{aligned}$$

Juntando ambos os resultados:

$$\mathbb{P}(G_i^c) \leq n\delta + \exp\left(-\frac{u_i c^2\Delta_i^2}{2\sigma^2}\right),$$

e, portanto:

$$\mathbb{E}[T_i(n)] \leq u_i + n\left(n\delta + \exp\left(-\frac{u_i c^2\Delta_i^2}{2\sigma^2}\right)\right).$$

A escolha de u_i permanece em aberto. Uma escolha natural é escolher o menor inteiro tal que a seguinte relação se faz verdadeira:

$$\Delta_i - \sqrt{\frac{2\sigma^2 \log(1/\delta))}{u_i}} \geq c\Delta_i \implies -\sqrt{\frac{2\sigma^2 \log(1/\delta))}{u_i}} \geq \Delta_i(c-1)$$
$$\implies \sqrt{\frac{2\sigma^2 \log(1/\delta))}{u_i}} \leq \Delta_i(1-c) \implies \frac{2\sigma^2 \log(1/\delta))}{u_i} \leq \Delta_i^2(1-c)^2$$
$$\implies \frac{2\sigma^2 \log(1/\delta))}{\Delta_i^2(1-c)^2} \leq u_i.$$

Logo, a escolha natural de u_i é:

$$u_i = \left\lceil \frac{2\sigma^2 \log(1/\delta))}{\Delta_i^2(1-c)^2} \right\rceil.$$

Escolhendo $\delta = \frac{1}{n^2}$ e escolhendo u_i da forma estabelecida acima, segue que:

$$\mathbb{E}[T_i(n)] \leq u_i + 1 + n\exp\left(-\frac{c^2\log(n^2)}{(1-c)^2}\right) = \left\lceil \frac{2\sigma^2 \log(n^2))}{\Delta_i^2(1-c)^2} \right\rceil + 1 + n(\exp\left(-\log(n^2)\right))^{\frac{c^2}{(1-c)^2}}$$
$$= \left\lceil \frac{2\sigma^2 \log(n^2))}{\Delta_i^2(1-c)^2} \right\rceil + 1 + n^{1-\frac{2c^2}{(1-c)^2}}.$$

Escolhendo $c = \frac{1}{2}$

$$\mathbb{E}[T_i(n)] \leq \left\lceil \frac{2\sigma^2 \log(n^2))}{\sigma^2\Delta_i^2 \cdot \frac{1}{4}} \right\rceil + 1 + \frac{1}{n} = \frac{16\sigma^2\log(n)}{\Delta_i^2} + 1 + \frac{1}{n} \leq 3 + \frac{16\sigma^2\log(n)}{\Delta_i^2}.$$

Utilizando o Lema 3.1:

$$R_n(\pi,\nu) \leq 3\sum_{i=1}^{k}\Delta_i + \sum_{i:\Delta_i>0}\frac{16\sigma^2\log(n)}{\Delta_i}.$$

□

Dessa forma, ao adotar o algoritmo UCB tem-se que o arrependimento é sempre inferior a algo que cresce de forma logarítmica, independentemente da escolha do δ, sendo essa uma vantagem obtida em relação ao algoritmo ETC, que depende da escolha de m para ser possível obter uma cota sub-linear.

3.3.5 Algoritmo UCB Assintoticamente Ótimo

O algoritmo UCB assintoticamente ótimo tem seu funcionamento bastante similar ao UCB comum, entretanto, se diferencia na escolha do parâmetro δ, garantindo, dessa forma, um limitante superior assintoticamente menor do que o encontrado anteriormente. O funcionamento do algoritmo é descrito através do Algoritmo 5.

Algoritmo 5 UCB Assintoticamente Ótimo

Entrada: $n, k := |\mathcal{A}|$

Escolha cada ação em $\mathcal{A}$ exatamente uma vez

Para cada $t \in \{k+1, \ldots, n\}$ faça:

Calcule $UCB_i(t-1) = \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(f(t))}{T_i(t-1)}}$

Escolha $A_t = \arg\max_i(UCB_i(t-1))$

onde $f(t) = 1 + t\log^2(t)$

fim-para

Antes de encontrar um limitante para o arrependimento e verificar que de fato é menor que o encontrado para a versão inicial do algoritmo, se faz necessário verificar o seguinte lema:

Lema 3.2 (Lema 8.2 em (LATTIMORE; SZEPESVáRI, 2020)). *Seja $\{X_i\}_{i=1}^n$ uma sequência de variáveis aleatórias independentes e 1-sub-gaussianas, $\varepsilon > 0$ e $a > 0$. Definamos $E(t) = \left\{\hat{\mu}_t + \sqrt{\frac{2a}{t}} \geq \varepsilon\right\}$, com $\mu_t = \frac{1}{t}\sum_{s=1}^t X_s$ e as seguintes quantidades:*

$$k = \sum_{t=1}^{n} \mathbb{1}_{E(t)},$$
$$k' = u + \sum_{t=\lceil u \rceil}^{n} \mathbb{1}_{E(t)}, \qquad \text{onde } u = 2a\varepsilon^{-2}.$$

Então vale:

$$\mathbb{E}[k] \leq \mathbb{E}[k'] \leq 1 + \frac{2}{\varepsilon^2}(a + \sqrt{\pi a} + 1).$$

Prova.

$$\mathbb{E}[k] \leq \mathbb{E}[k'] = u + \sum_{t=\lceil u \rceil}^{n} \mathbb{P}\left(\hat{\mu}_t + \sqrt{\frac{2a}{t}} \geq \varepsilon\right) \leq u + \sum_{t=\lceil u \rceil}^{n} \exp\left\{-\frac{t}{2}\left(\varepsilon - \sqrt{\frac{2a}{t}}\right)^2\right\}.$$

É possível aproximar e limitar a soma apresentada acima através da técnica de substituição da soma por uma integral:

$$\sum_{t=\lceil u \rceil}^{n} \exp\left\{-\frac{t}{2}\left(\varepsilon - \sqrt{\frac{2a}{t}}\right)^2\right\} \leq 1 + \int_u^{\infty} \exp\left\{-\frac{t}{2}\left(\varepsilon - \sqrt{\frac{2a}{t}}\right)^2\right\} dt,$$

por sua vez, a última integral pode ser limitada por:

$$\begin{aligned}
\int_u^{\infty} \exp\left\{-\frac{t}{2}\left(\varepsilon - \sqrt{\frac{2a}{t}}\right)^2\right\} dt &= \int_u^{\infty} \exp\left\{-\frac{t\varepsilon^2 - 2t\sqrt{\frac{2a}{t}}\varepsilon + 2a}{2}\right\} dt \\
&= \int_u^{\infty} \exp\left\{-\frac{\varepsilon\sqrt{t}(\varepsilon\sqrt{t} - 2\sqrt{2a}) + 2a}{2}\right\} dt \\
&\leq \int_u^{\infty} \exp\left\{-\frac{(\varepsilon\sqrt{t} - \sqrt{2a})^2}{2}\right\} dt,
\end{aligned}$$

Fazendo a mudança de variável $s = \varepsilon\sqrt{t} - \sqrt{2a}$ temos:

$$\begin{aligned}
\int_u^{\infty} \exp\left\{-\frac{(\varepsilon\sqrt{t} - \sqrt{2a})^2}{2}\right\} dt &= \int_{\varepsilon\sqrt{u} - \sqrt{2a}}^{\infty} \exp\left\{-\frac{s^2}{2}\right\} \frac{2\left(\frac{s + \sqrt{2a}}{\varepsilon}\right)}{\varepsilon} ds \\
&= \int_{\varepsilon\sqrt{u} - \sqrt{2a}}^{\infty} \exp\left\{-\frac{s^2}{2}\right\} \frac{2(s + \sqrt{2a})}{\varepsilon^2} ds \\
&= \int_0^{\infty} \left(\exp\left\{-\frac{s^2}{2}\right\} \frac{2s}{\varepsilon^2} + \exp\left\{-\frac{s^2}{2}\right\} \frac{2\sqrt{2a}}{\varepsilon^2}\right) ds \\
&= \int_0^{\infty} \exp\left\{-\frac{s^2}{2}\right\} \frac{2s}{\varepsilon^2} ds + \int_0^{\infty} \exp\left\{-\frac{s^2}{2}\right\} \frac{2\sqrt{2a}}{\varepsilon^2} ds \\
&= \int_0^{\infty} \exp\left\{-\frac{s^2}{2}\right\} \frac{2s}{\varepsilon^2} ds + \frac{2\sqrt{2a}}{\varepsilon^2} \int_0^{\infty} \exp\left\{-\frac{s^2}{2}\right\} ds.
\end{aligned}$$

Fazendo a mudança de variável $k = s^2$ na integral que compõe a primeira parcela da soma:

$$\int_0^\infty \exp\left\{-\frac{s^2}{2}\right\}\frac{2s}{\varepsilon^2}ds = \frac{1}{\varepsilon^2}\int_0^\infty \exp\left\{-\frac{k}{2}\right\}dk = \frac{1}{\varepsilon^2}\left[-2\exp\left\{\frac{-k}{2}\right\}\right]_0^\infty = \frac{2}{\varepsilon^2}.$$

Para a integral que constitui a segunda parcela da soma, o resultado pode ser obtido diretamente pela definição de função densidade de probabilidade da distribuição Normal padrão. Dessa forma, escolhendo $u = 2a\varepsilon^{-2}$ segue que:

$$\mathbb{E}[k] \leq 1 + 2a\varepsilon^{-2} + \frac{2}{\varepsilon^2} + \frac{4\sqrt{a\pi}}{2\varepsilon^2} = 1 + 2\varepsilon^{-2}(a + 1 + \sqrt{a\pi}).$$

□

A partir do Lema 3.2 podemos verificar o Teorema 3.3.

Teorema 3.3 (Teorema 8.1 em (LATTIMORE; SZEPESVáRI, 2020))**.** *Assumindo um ambiente* $\mathcal{E}^k_{SGau}(1)$ *e assumindo o algoritmo UCB Assintoticamente Ótimo, a seguinte desigualdade é statisfeita:*

$$R_n \leq \sum_{i:\Delta_i>0} \inf_{\varepsilon\in(0,\Delta_i)} \Delta_i \left(1 + \frac{5}{\varepsilon^2} + \frac{2\left(\log(f(n) + \sqrt{\pi\log(f(n))} + 1\right)}{(\Delta_i - \varepsilon)^2}\right).$$

Prova. Primeiramente, se valendo da definição de $T_i(n)$ temos que:

$$T_i(n) = \sum_{t=1}^n \mathbb{1}_{\{A_t=i\}} \leq \sum_{t=1}^n \mathbb{1}_{\left\{\hat{\mu}_1(t-1)+\sqrt{\frac{2\log(f(t))}{T_1(t-1)}}\leq\mu_1-\varepsilon\right\}} + \sum_{t=1}^n \mathbb{1}_{\left\{\hat{\mu}_i(t-1)+\sqrt{\frac{2\log(f(t))}{T_i(t-1)}}\geq\mu_1-\varepsilon \text{ e } A_t=i\right\}},$$

onde o resultado acima é verificado uma vez que apenas a função indicadora da segunda parcela da soma leva em consideração o fato de que $A_t = i$. Para realizar a prova, primeiro será verificado um limitante superior para a primeira parcela da soma e em seguida para a segunda parcela.

Para a primeira parcela temos que:

$$\mathbb{E}\left[\sum_{t=1}^n \mathbb{1}_{\left\{\hat{\mu}_1(t-1)+\sqrt{\frac{2\log(f(t))}{T_1(t-1)}}\leq\mu_1-\varepsilon\right\}}\right] = \sum_{t=1}^n \mathbb{P}\left(\hat{\mu}_1(t-1) + \sqrt{\frac{2\log(f(t))}{T_1(t-1)}} \leq \mu_1 - \varepsilon\right),$$

pela Equação (3.1) temos que:

$$\sum_{t=1}^n \mathbb{P}\left(\hat{\mu}_1(t-1) + \sqrt{\frac{2\log(f(t))}{T_1(t-1)}} \leq \mu_1 - \varepsilon\right) = \sum_{t=1}^n \mathbb{P}\left(\frac{\sum_{s=1}^{t-1} X_1\mathbb{1}_{\{A_s=1\}}}{T_1(t-1)} + \sqrt{\frac{2\log(f(t))}{T_1(t-1)}} \leq \mu_1 - \varepsilon\right).$$

Usando o fato de que $T_1(n) \leq n$ temos que:

$$\begin{aligned}\frac{\sum_{s=1}^{t-1} X_1\mathbb{1}_{\{A_s=1\}}}{T_i(t-1)} + \sqrt{\frac{2\log(f(t))}{T_1(t-1)}} &\leq \mu_1 - \varepsilon \\ \implies \exists j \in [1,n] : \frac{1}{j}\sum_{u=1}^{j} X_{u1}\mathbb{1}_{\{A_s=1\}} + \sqrt{\frac{2\log(f(t))}{j}} &\leq \mu_1 - \varepsilon,\end{aligned}$$

dessa forma:

$$\begin{aligned}\sum_{t=1}^n \mathbb{P}\left(\frac{\sum_{s=1}^{t-1} X_1\mathbb{1}_{\{A_s=1\}}}{T_1(t-1)} + \sqrt{\frac{2\log(f(t))}{T_1(t-1)}} \leq \mu_1 - \varepsilon\right) &\leq \sum_{t=1}^n \mathbb{P}\left(\exists j \in [1,n] : \frac{\sum_{u=1}^{j} X_{u1}}{j} + \sqrt{\frac{2\log(f(t))}{j}} \leq \mu_1 - \varepsilon\right) \\ &= \sum_{t=1}^n \mathbb{P}\left(\bigcup_{j=1}^n \left\{\frac{\sum_{u=1}^{j} X_{u1}}{j} + \sqrt{\frac{2\log(f(t))}{j}} \leq \mu_1 - \varepsilon\right\}\right),\end{aligned}$$

Utilizando a cota da união:

$$\sum_{t=1}^{n}\mathbb{P}\left(\bigcup_{j=1}^{n}\left\{\frac{\sum_{u=1}^{j}X_{u1}}{j}+\sqrt{\frac{2\log(f(t))}{j}}\leq\mu_1-\varepsilon\right\}\right)\leq\sum_{t=1}^{n}\sum_{j=1}^{n}\mathbb{P}\left(\frac{\sum_{u=1}^{j}X_{u1}}{j}+\sqrt{\frac{2\log(f(t))}{j}}\leq\mu_1-\varepsilon\right),$$

Utilizando (3.6):

$$\begin{aligned}\sum_{t=1}^{n}\sum_{j=1}^{n}\mathbb{P}\left(\frac{\sum_{u=1}^{j}X_{u1}}{j}+\sqrt{\frac{2\log(f(t))}{j}}\leq\mu_1-\varepsilon\right)&=\sum_{t=1}^{n}\sum_{j=1}^{n}\mathbb{P}\left(\hat{\mu}_{1j}+\sqrt{\frac{2\log(f(t))}{j}}\leq\mu_1-\varepsilon\right)\\&=\sum_{t=1}^{n}\sum_{j=1}^{n}\mathbb{P}\left(\hat{\mu}_{1j}\leq\mu_1-\left(\varepsilon+\sqrt{\frac{2\log(f(t))}{j}}\right)\right)\\&=\sum_{t=1}^{n}\sum_{j=1}^{n}\mathbb{P}\left(-(\hat{\mu}_{1j}-\mu_1)\geq\left(\varepsilon+\sqrt{\frac{2\log(f(t))}{j}}\right)\right).\end{aligned}$$

A partir disso, utilizando a Proposição 2.1 segue que $\hat{\mu}_{1j}-\mu_1$ é $\frac{1}{\sqrt{j}}$ sub-gaussiana e, ainda por meio da mesma Proposição, segue que $-(\hat{\mu}_{1j}-\mu_1)$ é $\frac{1}{\sqrt{j}}$-sub-gaussiana, portanto:

$$\begin{aligned}\sum_{t=1}^{n}\sum_{j=1}^{n}\mathbb{P}\left(-(\hat{\mu}_{1j}-\mu_1)\geq\left(\varepsilon+\sqrt{\frac{2\log(f(t))}{j}}\right)\right)&\leq\sum_{t=1}^{n}\sum_{j=1}^{n}\exp\left\{-\frac{j\left(\varepsilon+\sqrt{\frac{2\log(f(t))}{j}}\right)^2}{2}\right\}\\&=\sum_{t=1}^{n}\sum_{j=1}^{n}\exp\left\{-\frac{j\left(\varepsilon^2+2\varepsilon\sqrt{\frac{2\log(f(t))}{j}}+\frac{2\log(f(t))}{j}\right)}{2}\right\}\\&=\sum_{t=1}^{n}\sum_{j=1}^{n}\exp\left\{-\frac{j\varepsilon^2}{2}\right\}\exp\left\{-\frac{2j\varepsilon\sqrt{\frac{2\log(f(t))}{j}}+\frac{2j\log(f(t))}{j}}{2}\right\}\\&=\sum_{t=1}^{n}\sum_{j=1}^{n}\exp\left\{-\frac{j\varepsilon^2}{2}\right\}\exp\left\{-\varepsilon\sqrt{2j\log(f(t))}\right\}\exp\left\{-\log(f(t))\right\}\\&=\sum_{t=1}^{n}\frac{1}{f(t)}\sum_{j=1}^{n}\exp\left\{-\frac{j\varepsilon^2}{2}\right\}\exp\left\{-\varepsilon\sqrt{2j\log(f(t))}\right\}\leq\sum_{t=1}^{n}\frac{1}{f(t)}\sum_{j=1}^{n}\exp\left\{-\frac{j\varepsilon^2}{2}\right\}.\end{aligned}$$

Além disso,

$$\sum_{s=1}^{n}\exp\left\{-\frac{s\varepsilon^2}{2}\right\}=\frac{\exp\left\{-\frac{\varepsilon^2}{2}\right\}(1-\exp\left\{-\frac{\varepsilon^2}{2}\right\}^n)}{1-\exp\left\{-\frac{\varepsilon^2}{2}\right\}}\leq\frac{\exp\left\{-\frac{\varepsilon^2}{2}\right\}}{1-\exp\left\{-\frac{\varepsilon^2}{2}\right\}}\leq\frac{1}{\frac{\varepsilon^2}{2}}=\frac{2}{\varepsilon^2}.$$

A desigualdade acima é válida uma vez que:

$$\frac{\exp\{-x\}}{1-\exp\{-x\}}\leq\frac{1}{x}\implies x\exp\{-x\}+\exp\{-x\}\leq 1\implies 1+x\leq\exp\{x\}.$$

Ademais,

$$\begin{aligned}\sum_{t=1}^{n}\frac{1}{f(t)}&=\sum_{t=1}^{n}\frac{1}{(1+t\log^2(t))}=1+\sum_{t=2}^{n}\frac{1}{(1+t\log^2(t))}\leq 1+\sum_{t=1}^{n}\frac{1}{t\log^2(t)}\\&\approx 1+\int_2^{\infty}\frac{1}{t\log^2(t)}dt\approx 1+\lim_{b\to\infty}\left[-\frac{1}{\log(x)}\right]_2^b=\frac{1}{\log(2)}\approx 2,44,\end{aligned}$$

portanto,

$$\mathbb{E}\left[\sum_{t=1}^{n}\mathbb{1}_{\left\{\hat{\mu}_1(t-1)+\sqrt{\frac{2\log(f(t))}{T_1(t-1)}}\leq\mu_1-\varepsilon\right\}}\right]\leq\frac{5}{\varepsilon^2}.$$

Para encontrar um limitante superior para a segunda parte,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{n}\mathbb{1}_{\left\{\hat{\mu}_1(t-1)+\sqrt{\frac{2\log(f(t))}{T_1(t-1)}}\geq\mu_1-\varepsilon\text{ e }A_t=i\right\}}\right]\leq\mathbb{E}\left[\sum_{t=1}^{n}\mathbb{1}_{\left\{\hat{\mu}_1(t-1)+\sqrt{\frac{2\log(f(n))}{T_1(t-1)}}\geq\mu_1-\varepsilon\text{ e }A_t=i\right\}}\right]\\
&\leq\mathbb{E}\left[\sum_{s=1}^{n}\mathbb{1}_{\left\{\hat{\mu}_{is}+\sqrt{\frac{2\log(f(n))}{s}}\geq\mu_1-\varepsilon\right\}}\right]=\mathbb{E}\left[\sum_{s=1}^{n}\mathbb{1}_{\left\{\hat{\mu}_{is}-\mu_i+\sqrt{\frac{2\log(f(n))}{s}}\geq\mu_1-\mu_i-\varepsilon\right\}}\right]\\
&=\mathbb{E}\left[\sum_{s=1}^{n}\mathbb{1}_{\left\{\hat{\mu}_{is}-\mu_i+\sqrt{\frac{2\log(f(n))}{s}}\geq\Delta_i-\varepsilon\right\}}\right]\leq 1+\frac{2}{(\Delta_i-\varepsilon)^2}\left(\log(f(n))+\sqrt{\pi\log(f(n))}+1\right),
\end{aligned}$$

onde a última desigualdade segue do Lema 3.2. Por fim, juntando os dois resultados temos que:

$$T_i(n)\leq\frac{5}{\varepsilon^2}+1+\frac{2}{(\Delta_i-\varepsilon)^2}\left(\log(f(n))+\sqrt{\pi\log(f(n))}+1\right).$$

Utilizando (3.1) e tomando $\varepsilon=\log^{-1/4}(n)$:

$$\begin{aligned}
R_n&\leq\sum_{i:\Delta_i>0}\Delta_i\left(\frac{5}{\varepsilon^2}+1+\frac{2}{(\Delta_i-\varepsilon)^2}\left(\log(f(n))+\sqrt{\pi\log(f(n))}+1\right)\right)\\
&\leq\sum_{i:\Delta_i>0}\inf_{\varepsilon\in(0,\Delta_i)}\Delta_i\left(\frac{5}{\varepsilon^2}+1+\frac{2}{(\Delta_i-\varepsilon)^2}\left(\log(f(n))+\sqrt{\pi\log(f(n))}+1\right)\right).
\end{aligned}$$

□

Com efeito, é possível mostrar que o limitante apresentado no Teorema 3.3 é menor, assintoticamente em n, que o oferecido no Teorema 3.2, para evidenciar isso será considerado um ambiente $\mathcal{E}_{SGau}(1)$.

Retomando o limitante superior oferecido no Teorema 3.3 temos que:

$$R_n\leq\sum_{i:\Delta_i>0}\inf_{\varepsilon\in(0,\Delta_i)}\Delta_i\left(1+\frac{5}{\varepsilon^2}+\frac{2\left(\log(f(n))+\sqrt{\pi\log(f(n))}+1\right)}{(\Delta_i-\varepsilon)^2}\right),$$

onde $f(n) = 1 + n\log^2(n)$. Escolhendo $\varepsilon = \alpha\Delta_i$ com $\alpha \in (0,1)$, temos que:

$$
\begin{aligned}
&= \sum_{i:\Delta_i>0} \inf_{\alpha\in(0,1)} \Delta_i \left(1 + \frac{5}{(\alpha\Delta_i)^2} + \frac{2\left(\log(f(n)) + \sqrt{\pi\log(f(n))} + 1\right)}{(\Delta_i - \alpha\Delta_i)^2}\right) \\
&= \sum_{i:\Delta_i>0} \Delta_i + \inf_{\alpha\in(0,1)} \left(\frac{5}{\alpha^2\Delta_i} + \frac{2\left(\log(f(n)) + \sqrt{\pi\log(f(n))} + 1\right)}{\Delta_i(1-\alpha)^2}\right) \\
&= \sum_{i:\Delta_i>0} \Delta_i + \inf_{\alpha\in(0,1)} \frac{\frac{2\log(f(n))+2\sqrt{\pi\log(f(n))}+1}{(1-\alpha)^2} + \frac{5}{\alpha^2}}{\Delta_i} \\
&= \sum_{i:\Delta_i>0} \Delta_i + \sum_{i:\Delta_i>0} \inf_{\alpha\in(0,1)} \frac{\frac{2\log(f(n))+2\sqrt{\pi\log(f(n))}+1}{(1-\alpha)^2} + \frac{5}{\alpha^2}}{\Delta_i} \\
&= \sum_{i:\Delta_i>0} \Delta_i + \sum_{i:\Delta_i>0} \inf_{\alpha\in(0,1)} \frac{\frac{2\log(1+n\log^2(n))}{(1-\alpha)^2} + \frac{2\sqrt{\pi\log(1+n\log^2(n))}}{(1-\alpha)^2} + \frac{1}{(1-\alpha)^2} + \frac{5}{\alpha^2}}{\Delta_i}. \\
&= \sum_{i:\Delta_i>0} \Delta_i + \sum_{i:\Delta_i>0} \inf_{\alpha\in(0,1)} \frac{2\log(n)}{\Delta_i} \frac{\frac{\log(1+n\log^2(n))}{(1-\alpha)^2} + \frac{\sqrt{\pi\log(1+n\log^2(n))}}{(1-\alpha)^2} + \frac{1}{2(1-\alpha)^2} + \frac{5}{2\alpha^2}}{\log(n)},
\end{aligned}
$$

além disso, temos que:

$$
\begin{aligned}
\lim_{n\to\infty} \frac{\frac{\log(1+n\log^2(n))}{(1-\alpha)^2} + \frac{\sqrt{\pi\log(1+n\log^2(n))}}{(1-\alpha)^2} + \frac{1}{2(1-\alpha)^2} + \frac{5}{2\alpha^2}}{\log(n)} &= \lim_{n\to\infty} \frac{1}{(1-\alpha)^2} \frac{\log(n\log^2(n)) + \sqrt{\pi\log(n\log^2(n))}}{\log(n)} \\
&= \frac{1}{(1-\alpha)^2} \lim_{n\to\infty} \frac{\log(n) + 2\log(\log(n)) + \sqrt{\pi(\log(n) + 2\log(\log(n)))}}{\log(n)} \\
&= \frac{1}{(1-\alpha)^2} \lim_{n\to\infty} \frac{\log(n)}{\log(n)} + \frac{2\log(\log(n))}{\log(n)} + \frac{\sqrt{\pi(\log(n) + 2\log(\log(n)))}}{\log(n)} = \frac{1}{(1-\alpha)^2}.
\end{aligned}
$$

Portanto:

$$
R_n \leq \sum_{i:\Delta_i>0} \Delta_i + \sum_{i:\Delta_i>0} \inf_{\alpha\in(0,1)} \frac{2\log(n)}{\Delta_i(1-\alpha)^2}, \text{quando } n \to \infty.
$$

Dessa forma, observa-se que o principal benefício da cota fornecida pelo Teorema 3.3 está na redução da constante multiplicativa, no entanto, esse ganho é obtido de maneira assintótica, o que justifica o nome do algoritmo.

3.3.6 Limitantes inferiores

De acordo com (LATTIMORE; SZEPESVáRI, 2020), existem essencialmente dois tipos de cotas inferiores: aquelas que, para qualquer política, focam em informar uma instância de um problema de bandit na qual o arrependimento é no mínimo L, sendo essas chamadas de cotas inferiores de pior caso. O outro tipo consiste em, dada a política, então seu arrependimento em qualquer instância ν é no mínimo $L(\nu)$. Ainda de acordo com os autores, apesar do segundo tipo ser mais forte, é preciso salientar que apenas funciona com política específicas.

Antes de encontrar os limitantes inferiores, é preciso primeiro definir o arrependimento no pior caso[1], sendo esse definido da seguinte maneira:

$$R_n(\pi, \mathcal{E}) := \sup_{\nu \in \mathcal{E}} R_n(\pi, \nu). \tag{3.7}$$

A partir disso, denotando por Π o conjunto de todas as políticas, é possível definir o arrependimento minimax da seguinte forma:

$$R_n^*(\mathcal{E}) := \inf_{\pi \in \Pi} R_n(\pi, \mathcal{E}). \tag{3.8}$$

Uma política é chamada de minimax para um ambiente $\mathcal{E}$ se $R_n(\pi, \mathcal{E}) = R_n^*(\mathcal{E})$. Um dos resultados principais envolvendo as cotas inferiores de pior caso é apresentada no Teorema 3.4.

Teorema 3.4 (Teorema 15.2 em (LATTIMORE; SZEPESVáRI, 2020))**.** *Seja $k > 1$ e $n \geq k - 1$. Então, para qualquer política π e para um ambiente $\mathcal{E}_{Gau}^k(1)$, existe um vetor de médias $\mu \in [0,1]^k$ e uma instância de bandit ν_μ, na qual o i-ésimo braço segue a distribuição $\mathcal{N}(\mu_i, 1)$, tal que:*

$$R_n(\pi, \nu_\mu) \geq \frac{1}{27}\sqrt{(k-1)n}.$$

Uma vez que o teorema se aplica para qualquer política π e que $\nu_\mu \in \mathcal{E}_{\text{Gau}}^k(1)$, o resultado acima também garante que:

$$R_n^*(\mathcal{E}_{\text{Gau}}^k) \geq \frac{1}{27}\sqrt{(k-1)n}.$$

Antes de provar o teorema acima, é preciso apresentar um resultado auxiliar.

Lema 3.3 (Lema 15.1 em (LATTIMORE; SZEPESVáRI, 2020))**.** *Seja $\nu = (P_1, \ldots, P_k)$ a distribuição da recompensa associada a um bandit com k braços, e seja $\nu' = (P_1, \ldots, P_k)$ a distribuição das recompensas associadas a outro bandit também com k braços. Fixada uma política π e sendo $\mathbb{P}_\nu = \mathbb{P}_{\nu\pi}$ e $\mathbb{P}_{\nu'} = \mathbb{P}_{\nu'\pi}$ medidas de probabilidade no modelo de bandit definido na Seção 3 induzidos pela interconexão entre π e ν (respectivamente, π e ν'). Então:*

$$\mathcal{D}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \sum_{i=1}^{k} \mathbb{E}_\nu[T_i(n)]\mathcal{D}(P_i, P_i').$$

Prova. [Prova do Teorema 3.4] Para provar esse teorema, primeiro fixemos uma política π qualquer e seja $\Delta \in [0, 1/2]$. Por simplificação, vamos assumir um ambiente $\mathcal{E}_{\text{Gau}}^k(1)$ com vetor de médias $\tilde{\mu} = (\Delta, 0, 0, \ldots, 0)$. Selecionemos outro bandit que seja difícil de distinguir do primeiro e que ao mesmo tempo possua vetor de médias próximo.

A ideia por detrás dessa estratégia recai nos conceitos de similaridade e competição. A competição está relacionada ao fato de que uma sequência de ações que é boa para uma instância de bandit, não é boa para outra. Já a ideia de similaridade se relaciona ao fato de que as instâncias devem ser próximas o suficiente para que a política interagindo com qualquer uma delas não seja capaz de identificar o verdadeiro bandit com acurácia estatística razoável.

[1] Esse nome é proveniente do fato que de se tratar de um cenário em que se deseja obter a instância que maximiza o arrependimento por uma política fixada.

Dado isso, o vetor de média do segundo bandit (μ') será definida como sendo igual a do primeiro, exceto na ação que o primeiro explora menos, sendo esse vetor definido da seguinte maneira:

$$\mu'_j = \begin{cases} \tilde{\mu}_j, & sej \neq i; \\ 2\Delta, & \text{caso contrário.} \end{cases}$$

onde

$$i = \arg\min_{j>1} \mathbb{E}[T_j(n)].$$

A partir do Lema 3.1, segue que:

$$\begin{aligned} R_n(\pi,\nu) &= \sum_{a\in\mathcal{A}} \Delta_a \mathbb{E}_\nu[T_a(n)] = \Delta(n - \mathbb{E}_\nu[T_1(n)]) = \Delta(n - \mathbb{E}_\nu[T_1(n)(\mathbb{1}_{\{T_1(n)>n/2\}} + \mathbb{1}_{\{T_1(n)\leq n/2\}})]) \\ &= \Delta(n - (\mathbb{E}_\nu[T_1(n)\mathbb{1}_{\{T_1(n)>n/2\}}] + \mathbb{E}_\nu[T_1(n)\mathbb{1}_{\{T_1(n)\leq n/2\}}])) \\ &\geq \Delta\left(n - n\cdot\mathbb{P}_\nu(T_1(n) > n/2) - \frac{n}{2}\cdot\mathbb{P}_\nu(T_1(n)\leq n/2)\right) \\ &= \Delta n(1 - \mathbb{P}_\nu(T_1(n) > n/2)) - \Delta\frac{n}{2}\cdot\mathbb{P}_\nu(T_1(n)\leq n/2) \\ &= \Delta n(\mathbb{P}_\nu(T_1(n)\leq n/2)) - \Delta\frac{n}{2}\cdot\mathbb{P}_\nu(T_1(n)\leq n/2) = \frac{\Delta n}{2}(\mathbb{P}_\nu(T_1(n)\leq n/2)). \end{aligned}$$

Além disso,

$$\begin{aligned} R_n(\pi,\nu') &= \sum_{a\in\mathcal{A}} \Delta_a \mathbb{E}_{\nu'}[T_a(n)] \geq \Delta_1 \mathbb{E}_{\nu'}[T_1(n)] \geq (2\Delta - \Delta)\mathbb{E}_{\nu'}[T_1(n)\cdot\mathbb{1}_{\{T_1(n)>n/2\}}] \\ &\geq \Delta\cdot\mathbb{E}_{\nu'}\left[\frac{n}{2}\mathbb{1}_{\{T_1(n)>n/2\}}\right] = \frac{\Delta n}{2}\mathbb{P}_{\nu'}(T_1(n)\geq n/2). \end{aligned}$$

logo:

$$R_n(\pi,\nu) + R_n(\pi,\nu') \geq \Delta\frac{n}{2}\left(\mathbb{P}_\nu(T_1(n)\leq n/2) + \mathbb{P}_{\nu'}(T_1(n) > n/2)\right).$$

Aqui podemos pensar em $A = \{T_1(n)\leq n/2\}$ e $A^c = \{T_1(n) > n/2\}$ e aplicar o Teorema 2.3:

$$R_n(\pi,\nu) + R_n(\pi,\nu') \geq \frac{n\Delta}{4}\exp\{-\mathcal{D}(\mathbb{P}_\nu, \mathbb{P}_{\nu'})\}.$$

Utilizando o Lema 3.3:

$$\mathcal{D}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \sum_{i=1}^{k} \mathbb{E}_\nu[T_i(n)]\mathcal{D}(P_i, P'_i),$$

portanto,

$$\mathcal{D}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \mathbb{E}_\nu[T_i(n)]\mathcal{D}(\mathcal{N}(0,1), \mathcal{N}(2\Delta,1)) = \mathbb{E}_\nu[T_i(n)]\frac{(2\Delta)^2}{2}.$$

Usando o fato de que $\mathbb{E}[T_i(n)] \leq \dfrac{n}{k-1}$:

$$\mathbb{E}_\nu[T_i(n)]\frac{(2\Delta)^2}{2} \leq \frac{(2\Delta)^2}{2}\frac{n}{k-1},$$

e retornado a expressão anterior:

$$R_n(\pi, \nu_\mu) + R_n(\pi, \nu') \geq \frac{n\Delta}{4} \exp\left\{-\frac{2\Delta^2 n}{k-1}\right\}.$$

Escolhendo $\Delta = \sqrt{(k-1)/4n} \leq 1/2$:

$$R_n(\pi, \nu) + R_n(\pi, \nu') \geq \frac{n\sqrt{(k-1)/4n)}}{4} \exp\left\{-\frac{2(k-1)/4)}{k-1}\right\} = \frac{\sqrt{n(k-1)}}{8} \exp\{-2/4\}$$

$$\implies 2\max(R_n(\pi, \nu), R_n(\pi, \nu')) \geq \frac{\sqrt{n(k-1)}}{8} \exp\{-1/2\}.$$

□

Dentro da tarefa de encontrar limitantes inferiores para a medida de arrependimento, a noção de consistência de uma política desempenha um papel fundamental. A consistência de uma política π em relação a uma classe de bandits $\mathcal{E}$ é uma propriedade que garante que, à medida que o número de interações n cresce, o arrependimento acumulado $R_n(\pi, \nu)$ cresce a uma taxa assintoticamente mais lenta do que qualquer potência n^p, para $p > 0$.

Definição 3.1. *Uma política π é dita consistente para uma classe de bandits $\mathcal{E}$ se para todo $\nu \in \mathcal{E}$ e $p > 0$ é verificado que:*

$$\lim_{n\to\infty} \frac{R_n(\pi, \nu)}{n^p} = 0.$$

A partir disso, é possível estabelecer uma cota inferior que se encaixa no segundo tipo apresentado no início da presente seção, ou seja, uma cota inferior que, dada uma política π, então o arrependimento em qualquer instância para essa política é no mínimo um valor que depende apenas da instância, sendo essa apresentada no Teorema 3.5.

Teorema 3.5 (Teorema 16.2 em (LATTIMORE; SZEPESVáRI, 2020))**.** *Seja $\mathcal{E} = \mathcal{M}_1 \times \cdots \times \mathcal{M}_k$ e $\pi \in \Pi_{cons}(\mathcal{E})$ uma política consistente, onde $\mathcal{M}_1, \ldots, \mathcal{M}_k$ são conjuntos de distribuições com médias finitas. Então, para todo $\nu = (P_i)_{i=1}^k \in \mathcal{E}$ segue que:*

$$\liminf_{n\to\infty} \frac{R_n(\pi, \nu)}{\log(n)} \geq c^*(\nu, \mathcal{E}), \tag{3.9}$$

onde $c^*(\nu, \mathcal{E}) = \sum_{i:\Delta_i>0} \frac{\Delta_i}{d_{\inf}(P_i, \mu^*, \mathcal{M}_i)}$ com $d_{\inf}(P, \mu^*, \mathcal{M}) := \inf_{P' \in \mathcal{M}}\{\mathcal{D}(P, P') : \mu(P') > \mu^*\}$ e $\mu^* = \max_{i\in[k]} \mu_i(\nu)$.

Prova. Seja $d_i = d_{\inf}(P_i, \mu^*, \mathcal{M}_i)$. Fixemos uma ação subótima i, seja $\varepsilon > 0$ arbitrário, $\nu' = (P'_j)_{j=1}^k \in \mathcal{E}$ um bandit tal que $P'_j = P_j$ para $j \neq i$ e $P'_i \in \mathcal{M}_i$ seja tal que $\mathcal{D}(P_i, P'_i) \leq d_i + \varepsilon$ e $\mu(P'_i) > \mu^*$, que existe pela definição de d_i.

Aqui é possível notar similaridade com a prova do Teorema 3.4, a principal diferença é que na prova anterior foi definido que a primeira ação do primeiro bandit era a ótima e possuía média Δ.

Seja $\mu' \in \mathbb{R}^k$ o vetor de médias das distribuições de ν'. Utilizando o Lema 3.3 temos que:

$$\mathcal{D}(\mathbb{P}_{\nu\pi}, \mathbb{P}_{\nu'\pi}) \leq \mathbb{E}_{\nu\pi}[T_i(n)](d_i + \varepsilon).$$

Utilizando o Teorema 2.3 segue que:

$$\mathbb{P}_{\nu\pi}(A) + \mathbb{P}_{\nu\pi}(A^c) \geq \frac{1}{2}\exp\{-\mathcal{D}(\mathbb{P}_{\nu\pi}, \mathbb{P}_{\nu'\pi}\} \geq \frac{1}{2}\exp\{-\mathbb{E}_{\nu\pi}[T_i(n)](d_i+\varepsilon)\}.$$

Dessa forma, assumindo i^* como a ação ótima (no ambiente ν):

$$\begin{aligned} R_n + R'_n &= \sum_{a\in\mathcal{A}} \Delta_{a,\nu}\mathbb{E}_\nu[T_a] + \sum_{a\in\mathcal{A}} \Delta_{a,\nu'}\mathbb{E}_{\nu'}[T_a] \\ &\geq \left(\min_{a\in\mathcal{A}:a\neq i^*} \Delta_{a,\nu}\right) \sum_{a\in\mathcal{A}:a\neq i*} \mathbb{E}_\nu[T_a] + \Delta_{i^*,\nu'}\mathbb{E}_{\nu'}[T_{i^*}] \\ &= \left(\min_{a\in\mathcal{A}:a\neq i^*} \Delta_{a,\nu}\right) \mathbb{E}_\nu\left[\sum_{a\in\mathcal{A}:a\neq i*} T_a\right] + (\mu'_i - \mu^*)\mathbb{E}_{\nu'}[T_{i^*}]. \end{aligned}$$

Definamos $A = \left\{\sum_{a\in\mathcal{A}:a\neq i^*} T_a > \frac{n}{2}\right\}$, dessa forma A^c é o evento em que a ação i^* é selecionada no máximo $\frac{n}{2}$ vezes. Além disso, através da desigualdade Markov é possível notar que:

$$\sum_{a\in\mathcal{A}:a\neq i*} \mathbb{E}_\nu[T_a] = \mathbb{E}_\nu\left[\sum_{a\in\mathcal{A}:a\neq i*} T_a\right] \geq \frac{n}{2}\mathbb{P}(A).$$

Além disso,

$$\mathbb{E}[T_{i^*}] \geq \mathbb{E}[T_{i^*}\mathbb{1}_{\{T_{i^*}\geq\frac{n}{2}\}}] \geq \frac{n}{2}\mathbb{P}\left(T_{i^*} \geq \frac{n}{2}\right) = \frac{n}{2}\mathbb{P}(A^c).$$

Dessa forma, se valendo do Teorema 2.3,

$$\begin{aligned} R_n + R'_n &\geq \min\left\{\min_{a\neq i^*} \Delta_{a,\nu}, \mu'_i - \mu^*\right\} \cdot \frac{n}{2}(\mathbb{P}_\nu(A) + \mathbb{P}_{\nu'}(A^c)) \\ &\geq \frac{1}{4}\min\left\{\min_{a\neq i^*} \Delta_{a,\nu}, \mu'_i - \mu^*\right\} \exp\{-\mathbb{E}_{\nu\pi}[T_i(n)(d_i+\varepsilon)]\}. \end{aligned}$$

Reorganizando os termos e tomando o limite inferior dos dois lados segue que:

$$\begin{aligned} \liminf_{n\to\infty} \frac{\mathbb{E}_{\nu\pi}[T_i(n)]}{\log(n)} &\geq \frac{1}{d_i+\varepsilon} \liminf_{n\to\infty} \frac{\log\left(\frac{n\min\{\min_{a\neq i^*}\Delta_{a,\nu}, \mu'_i - \mu^*\}}{4(R_n+R'_n)}\right)}{\log(n)} \\ &= \frac{1}{d_i+\varepsilon}\left(1 - \limsup_{n\to\infty} \frac{\log(R_n+R'_n)}{\log(n)}\right) = \frac{1}{d_i+\varepsilon}. \end{aligned}$$

□

4 EXPLORAÇÃO PURA

De acordo com (BUBECK; MUNOS; STOLTZ, 2009), a primeira aplicação do problema de multi-armed bandits foi feita no ramo da realização de ensaios clínicos. Nesse contexto, em um cenário de uma doença grave, apenas pacientes doentes são incluídos no ensaio, dessa forma, uma escolha incorreta do tratamento adotado pode gerar a perda de uma vida. Nesse cenário, fica claro que se torna importante minimizar o arrependimento acumulado, uma vez que a fase de teste e de cura coincidem. Por outro lado, ainda de acordo com os autores, para produtos cosméticos, existe uma fase de teste antes que o produto seja lançado, o que acaba favorecendo o objetivo de minimizar o arrependimentos do produto comercializado ao invés de minimizar o arrependimento acumulado durante a fase de testes, que é irrelevante.

Com isso, surge o campo da exploração pura (do inglês *pure exploration*), que, em contraste com os métodos vistos no capítulo anterior, os quais visam balancear o dilema entre exploração e explotação, se concentra primariamente em explorar o máximo possível o ambiente e em seguida oferecer a ação que gera uma maior recompensa esperada. Esse cenário se encaixa exatamente no exemplo dos produtos cosméticos apresentado no parágrafo anterior e, em linhas gerais, poderia ser traduzido em um cenário em que um jogador se vê em frente a várias máquinas caça-niqueis e precisa descobrir qual delas é a que oferece uma maior recompensa, sem que seja necessário ao mesmo tempo se aproveitar de máquinas que parecem ser promissoras.

Assim como ocorre nos contextos que envolvem exploração e explotação, para o contexto da exploração pura também existem diversos algoritmos que possibilitam encontrar políticas que permitam gerar uma correta identificação da ação ótima, entretanto, não é mais tão adequado avaliar esses algoritmos através da mesma medida de arrependimento definida na Equação (3.2), uma vez que, nesse novo contexto, o agente não se arrepende mais de escolher mais vezes uma ação sub-ótima, mas sim de recomendar uma ação subótima ao final das rodadas.

Nesse sentido, sendo ν um bandit estocástico com k braços e $\pi = (\pi_t)_{t=1}^{n+1}$ uma política, uma forma de mensurar a performance de uma política no contexto de exploração pura é o arrependimento simples, definido como:

$$R_n^{\text{SIMPLES}}(\pi, \nu) := \mathbb{E}\left[\Delta_{A_{n+1}}(\nu)\right], \tag{4.1}$$

nesse caso, a ação escolhida na rodada $n+1$ recebe uma atenção especial, uma vez que se trata da recomendação fornecida pelo algoritmo e, uma vez fornecida, jamais será alterada, ou seja, nesse caso não existe um cenário adaptativo como havia em parte dos algoritmos citados na capítulo anterior.

Um dos algoritmos mais simples para esse contexto é o algoritmo de exploração uniforme - UE (do inglês *Uniform Exploration*), tendo seu funcionamento apresentado no Algoritmo 6.

Algoritmo 6 Algoritmo de Exploração Uniforme

Entrada: n
Para cada $t \in \{1, \dots, n\}$ faça:
    Escolha a ação $A_t = 1 + (t \mod k)$
fim-para
Escolha $A_{n+1} = \arg\max_{i \in [k]} \hat{\mu}_i(n)$

Dado o Algoritmo 6, é possível encontrar um limitante superior para o arrependimento simples através do Teorema 4.1.

Teorema 4.1 (Teorema 33.1 em (LATTIMORE; SZEPESVáRI, 2020)). *Seja $\nu \in \mathcal{E}^k_{Gau}(1)$ arrependimento satisfaz a seguinte desigualdade:*

$$R_n^{SIMPLES}(\pi, \nu) \leq \min_{\Delta \geq 0} \left(\Delta + \sum_{i:\Delta_i(\nu) > \Delta} \Delta_i(\nu) \exp\left\{ -\frac{\lfloor n/k \rfloor \Delta_i(\nu)^2}{4} \right\} \right).$$

Prova. Por simplificação, consideremos $\Delta_i = \Delta_i(\nu)$ e $\mathbb{P} = \mathbb{P}_{\nu\pi}$. Além disso, será considerado, sem perda de generalidade, que a ação ótima é a 1, ou seja, $\Delta_1 = 0$ e que i seja uma ação subótima com $\Delta_i \leq \Delta$, logo:

$$\begin{aligned}
\mathbb{P}(\hat{\mu}_i(n) \geq \hat{\mu}_1(n)) &= \mathbb{P}(\hat{\mu}_i(n) - \hat{\mu}_1(n) \geq 0) \\
&= \mathbb{P}(\hat{\mu}_i(n) - \mathbb{E}[\hat{\mu}_i(n)] - \hat{\mu}_1(n) + \mathbb{E}[\hat{\mu}_1(n)] \geq -\mathbb{E}[\hat{\mu}_i(n)] + \mathbb{E}[\hat{\mu}_1(n)]) \\
&= \mathbb{P}(\hat{\mu}_i(n) - \mu_i - (\hat{\mu}_1(n) - \mu_1) \geq \mu_1 - \mu_i) \\
&= \mathbb{P}((\hat{\mu}_i(n) - \mu_i) - (\hat{\mu}_1(n) - \mu_1) \geq \Delta_i).
\end{aligned}$$

Sabemos que:

$$\hat{\mu}_i(n) = \sum_{i=1}^{n} \frac{X_i \mathbb{1}_{\{A_t = i\}}}{T_i(n)} \leq \sum_{i=1}^{n} \frac{X_i \mathbb{1}_{\{A_t = i\}}}{\lfloor n/k \rfloor},$$

$$\hat{\mu}_1(n) = \sum_{i=1}^{n} \frac{X_1 \mathbb{1}_{\{A_t = 1\}}}{T_1(n)} \leq \sum_{i=1}^{n} \frac{X_1 \mathbb{1}_{\{A_t = 1\}}}{\lfloor n/k \rfloor},$$

dessa forma, uma vez que as recompensas obtidas em cada tempo t são independentes, segue que:

$$\sum_{t=1}^{n} \mathbb{1}_{\{A_t = i\}} \frac{X_i}{\lfloor n/k \rfloor} \text{ é } \sqrt{\underbrace{\frac{1}{\lfloor n/k \rfloor^2} + \frac{1}{\lfloor n/k \rfloor^2} + \cdots + \frac{1}{\lfloor n/k \rfloor^2}}_{\lfloor n/k \rfloor \ vezes}}\text{-sub-gaussiana,}$$

$$\sum_{t=1}^{n} \mathbb{1}_{\{A_t = i\}} \frac{X_i}{\lfloor n/k \rfloor} \text{ é } \frac{1}{\sqrt{\lfloor n/k \rfloor}}\text{-sub-gaussiana.}$$

Pelo Teorema 2.2, segue que:

$$\mathbb{P}((\hat{\mu}_i(n) - \mu_i) - (\hat{\mu}_1(n) - \mu_1) \geq \Delta_i) \leq \exp\left\{ \frac{-\Delta_i^2}{2\left(\frac{1}{\sqrt{\lfloor n/k \rfloor}} + \frac{1}{\sqrt{\lfloor n/k \rfloor}}\right)} \right\} = \exp\left\{ \frac{-\Delta_i^2 \lfloor n/k \rfloor}{4} \right\}.$$

Usando a Definição 4.1, temos que:

$$\begin{aligned} R_n^{\text{SIMPLES}}(\pi,\nu) &= \sum_{i=1}^{k} \Delta_i \mathbb{P}(A_{n+1}=i) \leq \Delta + \sum_{i:\Delta_i>\Delta} \Delta_i \mathbb{P}(A_{n+1}=i) \\ &\leq \Delta + \sum_{i:\Delta_i>\Delta} \Delta_i \exp\left\{\frac{-\Delta_i^2 \lfloor n/k \rfloor}{4}\right\} \\ &\leq \min_{\Delta\geq 0} \left(\Delta + \sum_{i:\Delta_i>\Delta} \Delta_i \exp\left\{-\frac{\Delta_i^2 \lfloor n/k \rfloor}{4}\right\}\right). \end{aligned}$$

□

4.1 Identificação da melhor ação com horizonte pré-determinado

Existem duas principais abordagens para identificação da melhor ação no contexto de exploração pura, a primeira delas considera um nível de confiança fixado $\delta \in [0,1]$, a partir disso, o agente deve usar o menor número de amostras possível com a finalidade de ao final informar a ação ótima com probabilidade maior ou igual a $1-\delta$. A segunda abordagem é similar a primeira, no entanto, considera previamente a existência de uma horizonte n, sendo o objetivo final minimizar a probabilidade de seleção de uma ação sub-ótima dentro das n rodadas. O presente texto se concentrará apenas na segunda abordagem, visando apresentar resultados teóricos envolvendo-a.

A segunda abordagem supracitada é denominada identificação da melhor ação com horizonte pré-determinado, nesse cenário o agente possui um horizonte n fixado e deve escolher uma política que vise, ao final das n rodadas, informar a ação ótima, possuindo como objetivo oferecer uma resposta que minimize a probabilidade de falsa seleção, ou seja, a probabilidade da política recomendar uma ação sub-ótima.

4.1.1 Algoritmo Sequential-Halving

Ainda no contexto de minimizar a probabilidade de falsa seleção, é possível notar a existência do algoritmo *Sequential Halving*, sendo seu funcionamento descrito por meio do Algoritmo 7.

Algoritmo 7 Sequential Halving

Entrada: n, k
Defina $L = \lceil \log_2(k) \rceil$ e $\mathcal{A}_1 = [k]$
Para cada $l \in \{1, \ldots, L\}$ faça:
Defina $T_l = \left\lfloor \frac{n}{L|\mathcal{A}_l|} \right\rfloor$
Escolha cada ação em $\mathcal{A}_l$ exatamente T_l vezes
Para cada $i \in \mathcal{A}_l$ calcule $\hat{\mu}_i^l(n)$ como sendo a média empírica da ação i baseada nas últimas T_l amostras
Defina $\mathcal{A}_{l+1}$ como sendo o conjunto contendo as top $\lceil |A_l|/2 \rceil$ ações em $\mathcal{A}_l$
fim-para
Retorna: A_{n+1}, a única ação remanescente em $\mathcal{A}_{L+1}$

Conforme o próprio nome, a ideia do Algoritmo 7 é gerar T_l amostras para cada uma das ações e manter apenas as $\lceil |A_l|/2 \rceil$ ações que apresentaram maior média empírica da recompensa na rodada l. Seguindo a construção do algoritmo, devido ao conjunto de ações ser sempre reduzido pela metade a cada rodada, segue que o total de rodadas será sempre igual a $\lceil \log_2(k) \rceil$ e $\mathcal{A}_1 = [k]$, garantindo que $|\mathcal{A}_{L+1}| = 1$.

Um ponto de grande importância na construção do Algoritmo 7 é que, apesar do número de rodadas ser fixo, a escolha de quantas vezes cada ação será amostrada em cada rodada é variável, sendo esse um dos fatores cruciais para definição da estratégia de escolha adotada por parte do algoritmo. No Algoritmo 7 é definido que o número de amostras de cada ação na rodada l (T_l) é igual a $\left\lfloor \frac{n}{L|\mathcal{A}_l|} \right\rfloor$. Essa escolha aloca a mesma proporção de n para cada uma das rodadas e garante que em cada rodada as amostras para cada uma das ações possuam o mesmo tamanho.

Um resultado importante obtido para o algoritmo *Sequential Halving* é expressado por meio do Teorema 4.2, onde é possível perceber que dada a escolha de uma política que siga o algoritmo, então a probabilidade de falsa seleção, ou seja, a probabilidade de que ao final do horizonte a ação recomendada seja subótima, decai de maneira exponencial de acordo com o aumento do horizonte n.

Teorema 4.2 (Teorema 33.10 em (LATTIMORE; SZEPESVáRI, 2020)). *Se $\nu \in \mathcal{E}^k_{SGau}(1)$ tem vetor de média $\mu = \mu(\nu)$ e $\mu_1 \geq \cdots \geq \mu_k$ e π segue o algoritmo Sequential Halving, então:*

$$\mathbb{P}(\Delta_{A_{n+1}} > 0) \leq 3 \log_2(k) \exp\left(-\frac{n}{16 \mathcal{H}_2(\mu) \log_2(k)} \right),$$

onde $\mathcal{H}_2(\mu) := \max_{i:\Delta_i > 0} \frac{i}{\Delta_i^2}$.

Prova. Dado um conjunto $A \subset [k]$, definamos:

$$\mathrm{TopM}(A, m) := \left\{ i \in [k] : \sum_{j \leq i} \mathbb{1}_{\{j \in A\}} \leq m \right\}.$$

Dessa forma, definimos $\mathrm{TopM}(A, m)$ como o conjunto dos índices $i \in [k]$ tais que o número de elementos de A entre os i primeiros não exceda m. Além disso, seja $A'_l := A_l \setminus \mathrm{TopM}(A_l, \lceil |A_l|/4 \rceil)$ o subconjunto de A_l com cardinalidade três quartos da cardinalidade de A_l, contendo os braços com menores médias empíricas na rodada l. A partir disso, definimos $i_l := \min A'_l$ e

$$N_l := \sum_{i \in A'_l} \mathbb{1}_{\{\hat{\mu}_i^l \geq \hat{\mu}_1^l\}}.$$

No decorrer da prova, serão usados os seguintes resultados:

$$\mathbb{P}(\hat{\mu}_1^l(n) \leq \hat{\mu}_i^l(n) | i \in A_l, 1 \in A_l) \leq \exp\left(-\frac{T_l \Delta_i^2}{4} \right), \tag{4.2}$$

$$\mathbb{E}[N_l | A_l] \leq \frac{3|A_l|}{4} \exp\left\{ -\frac{\Delta_{i_l}^2 \cdot n}{16 i_l \log_2(k)} \right\}, \tag{4.3}$$

$$\mathbb{P}(1 \notin A_{l+1} | 1 \in A_l) \leq 3 \exp\left\{ -\frac{\Delta_{i_l}^2 \cdot n}{16 i_l \log_2(k)} \right\}. \tag{4.4}$$

Utilizando o resultado em (4.4) temos que:

$$\begin{aligned}
\mathbb{P}(\Delta_{n+1} \neq 0) &= \mathbb{P}(\exists l \in [\log_2 k] : 1 \notin A_l) = \mathbb{P}\left(\bigcup_{l=1}^{\log_2 k} \{1 \notin A_l\}\right) \\
&= \mathbb{P}\left(\bigcup_{l=1}^{\log_2 k} \{1 \notin A_{l+1}\} \cap \{1 \in A_l\}\right) \\
&\leq \sum_{l=1}^{\log_2 k} \mathbb{P}(1 \notin A_{l+1} | 1 \in A_l)\mathbb{P}(1 \in A_l) \\
&\leq \sum_{l=1}^{\log_2 k} \mathbb{P}(1 \notin A_{l+1} | 1 \in A_l) \leq \sum_{l=1}^{\log_2 k} 3\exp\left\{-\frac{\Delta_{i_l}^2 \cdot n}{16 i_l \log_2(k)}\right\} \\
&\leq \log_2 k \cdot 3 \max_{l \in \{1,\dots \log_2 k\}} \exp\left\{-\frac{\Delta_{i_l}^2 \cdot n}{16 i_l \log_2(k)}\right\} \\
&\leq \log_2 k \cdot 3 \max_{i \in [k]:\Delta_i > 0} \exp\left\{-\frac{\Delta_i^2}{i}\frac{n}{16\log_2(k)}\right\} \\
&= \log_2 k \cdot 3\exp\left\{-\frac{n}{16\mathcal{H}_2(\mu)\log_2(k)}\right\},
\end{aligned}$$

onde a última desigualdade segue do fato que $\max_{l \in \{1,\dots \log_2 n\}} \frac{i_l}{\Delta_{i_l}^2} \leq \max_{i \in [k]:\Delta_i > 0} \frac{i}{\Delta_i^2}$, que por sua vez é uma consequência direta do fato de que $\{i_l : l \in \{1, \dots \log_2 n\}\} \subset [k]$. A seguir são verificados os resultados em (4.2), (4.3), (4.4):

Em (4.2) é assumido que i é um braço sub-ótimo e que $1 \in A_l$, dessa forma temos que:

$$\begin{aligned}
\mathbb{P}(\hat{\mu}_1^l(n) \leq \hat{\mu}_i^l(n) | i \in A_l, 1 \in A_l) &= \mathbb{P}(\hat{\mu}_i^l(n) - \hat{\mu}_1^l(n) \geq 0 | i \in A_l, 1 \in A_l) \\
&= \mathbb{P}(\hat{\mu}_i^l(n) - \mu_i - (\hat{\mu}_1^l(n) - \mu_1) \geq \mu_1 - \mu_i | i \in A_l, 1 \in A_l) \\
&= \mathbb{P}(\hat{\mu}_i^l(n) - \mu_i - (\hat{\mu}_1^l(n) - \mu_1) \geq \Delta_i | i \in A_l, 1 \in A_l),
\end{aligned}$$

Utilizando a Proposição 2.1, e também o fato das variáveis aleatórias associadas às recompensas obtidas em cada tempo t serem independentes, segue que $\hat{\mu}_i^l(n)$ é $\frac{1}{\sqrt{T_l}}$-sub-gaussiana. Aplicando o Teorema 2.2 podemos concluir que:

$$\mathbb{P}(\hat{\mu}_i^l(n) - \mu_i - (\hat{\mu}_1^l(n) - \mu_1) \geq \Delta_i | i \in A_l, 1 \in A_l) \leq \exp\left\{-\frac{\Delta_i^2}{2 \cdot \frac{2}{T_l}}\right\}.$$

Para verificar (4.3) basta notar que:

$$\mathbb{E}[N_l | A_l] = \mathbb{E}\left[\sum_{i \in A_l'} \mathbb{1}_{\{\hat{\mu}_i^l \geq \hat{\mu}_1^l(n)\}} | A_l\right] = \sum_{i \in A_l'} \mathbb{P}(\hat{\mu}_i^l(n) \geq \hat{\mu}_1^l(n) | A_l),$$

uma vez que, condicionado em A_l, por definição, A_l' se torna determinístico. Além disso, aplicando (4.2) e utilizando a definição de T_l segue que:

$$\begin{aligned}
\mathbb{E}[N_l | A_l] &\leq \sum_{i \in A_l'} \exp\left\{-\frac{\Delta_i^2}{2 \cdot \frac{2}{T_l}}\right\} = \sum_{i \in A_l'} \exp\left\{-\frac{\Delta_i^2 \cdot n 2^{l-1}}{4\log_2(k)}\right\} \\
&\leq |A_l'| \max_{i \in A_l'} \exp\left\{-\frac{\Delta_i^2 \cdot n 2^{l-1}}{4\log_2(k)}\right\} \\
&\leq \frac{3|A_l|}{4} \exp\left\{-\frac{\Delta_{i_l}^2 \cdot n}{4\log_2(k) 2^{1-l}}\right\},
\end{aligned}$$

onde a última desigualdade é verificada pela definição de A'_l e pelo fato de i_l ser o menor elemento de A'_l. Retomando a definição de A'_l conjuntamente com o fato da cardinalidade de A_l ser igual a $k2^{1-l}$ temos que:

$$\begin{aligned} A'_l &= A_l \setminus \text{TopM}(A_l, \lceil |A_l|/4 \rceil) \\ &= A_l \setminus \text{TopM}(A_l, \lceil k2^{1-l}/4 \rceil), \end{aligned}$$

ou seja, todos os elementos de A'_l são maiores ou iguais a $\frac{k2^{1-l}}{4}$, em específico $i_l = \min A'_l$, portanto:

$$\mathbb{E}[N_l|A_l] \leq \frac{3|A_l|}{4} \exp\left\{-\frac{\Delta_{i_l}^2 \cdot nk}{4 \cdot 4i_l \log_2(k)}\right\} \leq \frac{3|A_l|}{4} \exp\left\{-\frac{\Delta_{i_l}^2 \cdot n}{16 i_l \log_2(k)}\right\}.$$

Para verificar (4.4), primeiramente é preciso verificar que:

$$A'_l = A_l - \text{TopM}(A_l, \lceil |A_l|/4 \rceil) = A_l - \text{TopM}(A_l, \lceil k2^{1-l}/4 \rceil)$$

Essa igualdade é verificada uma vez que 3/4 dos elementos de A_l estão em A'_l. Além disso, por definição, N_l ser maior que $1/4|A_l|$ implica que, como o vetor está ordenado, $\hat{\mu}_1$ vai estar posicionado no conjunto de braços que sucede a metade do conjunto total de braços e, portanto, não será escolhido para a próxima etapa. Por outro lado, se $1 \notin A_{l+1}$, então $\hat{\mu}_1$ estava mais afastado que metade do tamanho de A_l, e, portanto $N_l > 1/4|A_l|$. Portanto, utilizando (4.3) conjuntamente com a desigualdade de Markov, segue que:

$$\mathbb{P}(N_l > 1/4|A_l||1 \in A_l) \leq \frac{\mathbb{E}[N_l|1 \in A_l]}{1/4|A_l|} \leq \frac{\frac{3|A_l|}{4} \exp\left\{-\frac{\Delta_{i_l}^2 \cdot n}{16 i_l \log_2(k)}\right\}}{1/4|A_l|} = 3 \exp\left\{-\frac{\Delta_{i_l}^2 \cdot n}{16 i_l \log_2(k)}\right\}.$$

□

4.1.2 Algoritmo Sequential-Elimination

Outro algoritmo que surge no contexto da identificação da melhor ação com um horizonte pré-determinado é o algoritmo *Sequential Elimination*, sendo seu funcionamento apresentado no Algoritmo 8

Algoritmo 8 Sequential Elimination

Entrada: n, k

Defina $\mathcal{A}_1 = [k]$, $\overline{\log}(k) = \frac{1}{2} + \sum_{i=2}^{k} \frac{1}{i}$ e $T_0 = 0$

Para cada $l \in \{1, \ldots, k-1\}$ faça:

Defina $T_l = \left\lceil \frac{1}{\overline{\log}(k)} \frac{n-k}{|\mathcal{A}_l|} \right\rceil$

Escolha cada ação em $\mathcal{A}_l$ exatamente $T_l - T_{l-1}$ vezes

Para cada $i \in \mathcal{A}_l$ calcule $\hat{\mu}_i^l(n)$ como sendo a média empírica da ação i baseada nas últimas $T_l - T_{l-1}$ amostras

Defina $\mathcal{A}_{l+1} = \mathcal{A}_l \setminus \arg\min_{i \in \mathcal{A}_l} \hat{\mu}_i^l$

fim-para

Retorna: A_{n+1}, a única ação remanescente em $\mathcal{A}_k$

No Algoritmo 8, diferentemente do que ocorria no Algoritmo 7, não fica claro em um primeiro momento o motivo de escolher cada ação na rodada l exatamente $T_l - T_{l-1}$ vezes. Com efeito, nota-se que, assim como no Algoritmo 7, dentro de cada rodada, cada ação é amostrada o mesmo número de vezes, no entanto a definição da proporção de n que será alocada em cada rodada não é evidente. A fim de explicar melhor essa definição, será desenvolvido a seguir o raciocínio por trás dessa escolha, valendo notar que esse raciocínio pode ser estendido também para o algoritmo *Sequential Halving*. Seja m_i a parcela de n que será alocada na rodada $i, i \in [k-1]$, definido da seguinte forma:

$$m_i = \begin{cases} c \cdot n & \text{se } i = 1 \\ \dfrac{c \cdot n}{|\mathcal{A}_{i-1}|} & \text{se } i \in \{2, \ldots, k-1, \} \end{cases}$$

onde $0 < c < 1$. Como m_i representa a parcela de n alocada para cada rodada i, segue que:

$$\begin{aligned} \sum_{i=1}^{k-1} m_i = n \implies & m_1 + \sum_{i=2}^{k-1} m_i = n \\ \implies & c \cdot n + \sum_{i=2}^{k-1} \frac{c \cdot n}{|\mathcal{A}_{i-1}|} = c \cdot n \left(1 + \sum_{i=2}^{k-1} \frac{1}{k+1-(i-1)}\right) = n \\ \implies & c \left(1 + \sum_{i=2}^{k-1} \frac{1}{k+1-(i-1)}\right) = c \left(1 + \sum_{j=3}^{k} \frac{1}{j}\right) = c \left(\frac{1}{2} + \sum_{j=2}^{k} \frac{1}{j}\right) = 1 \\ \implies & c = \frac{1}{\overline{\log}(k)}. \end{aligned}$$

Vale lembrar que um dos pressupostos assumidos pelo algoritmo é que cada ação sempre é escolhida o mesmo número de vezes em cada uma das rodadas, portanto, para a rodada i temos que para todo $a \in \mathcal{A}_i$ segue que o número de vezes que cada ação será realizada é igual a $\dfrac{n}{\overline{\log}(k)|\mathcal{A}_{i-1}||\mathcal{A}_i|}$ para $i \in \{2, \ldots, k-1\}$ e $\overline{\log}(k)|\mathcal{A}_i|$ para $i = 1$.

Um resultado importante no que tange o algoritmo Sequential Elimination é o Teorema 4.3. Em sua proposição é possível notar que, assim como no Teorema 4.2, é obtido um limitante superior que decai exponencialmente a medida que o horizonte aumenta, se diferenciando apenas pelas constantes. Nesse sentido, é possível afirmar que a estratégia de eliminar metade das ações a cada rodada não contribui para o surgimento do decaimento exponencial e que a probabilidade de falsa seleção dos dois algoritmos é assintoticamente similar.

Teorema 4.3. *Se $\nu \in \mathcal{E}^k_{SGau}(\sigma)$ tem vetor de média $\mu = \mu(\nu)$ e π segue o Algoritmo 8 de eliminação sequencial então:*

$$\mathbb{P}(\Delta_{A_{n+1}} > 0) \leq \frac{k(k-1)}{2} \exp\left\{-\frac{n-k}{2\sigma^2 \overline{\log}(k) H_2(\mu)}\right\},$$

onde $H_2(\mu) := \max_{i \in \{1,\ldots,k\}} \dfrac{i}{\Delta^2_{(i)}}$.

Prova. Diferentemente do que é feito em (AUDIBERT; BUBECK, 2010), a presente dissertação apresenta um resultado geral para a probabilidade de falsa seleção, não se restringindo apenas a um ambiente $\mathcal{E}^k_{Ber}$. Para a presente prova será introduzida a notação $(i) \in \{1, \ldots, k\}$ denotando a i-ésima

ação com maior média. Assumindo sem perda de generalidade que a ação 1 é a ação ótima, segue que:

$$\begin{aligned}
\mathbb{P}(\Delta_{A_{n+1}} > 0) &= \mathbb{P}\left(\exists l \in \{1,\dots,k-1\} : \hat{\mu}_1^l(n) \leq \max_{i\in\mathcal{A}_l} \hat{\mu}_i^l(n)\right) \\
&\leq \mathbb{P}\left(\bigcup_{l\in\{1,\dots,k-1\}} \bigcup_{i\in\{k-l+1,\dots,k\}} \hat{\mu}_1^l(n) \leq \hat{\mu}_{(i)}^l(n)\right) \leq \sum_{l=1}^{k-1}\sum_{i=k-(l-1)}^{k} \mathbb{P}(\hat{\mu}_1^l(n) \leq \hat{\mu}_{(i)}^l(n)) \\
&= \sum_{l=1}^{k-1}\sum_{i=k-(l-1)}^{k} \mathbb{P}(\hat{\mu}_{(i)}^l(n) - \mu_{(i)} + \mu_1 - \hat{\mu}_1^l(n) \geq \mu_1 - \mu_{(i)}) \\
&= \sum_{l=1}^{k-1}\sum_{i=k-(l-1)}^{k} \mathbb{P}(\hat{\mu}_{(i)}^l(n) - \mu_{(i)} - (\hat{\mu}_1^l(n) - \mu_1) \geq \Delta_{(i)}).
\end{aligned}$$

Seguindo o mesmo raciocínio desenvolvido na prova do Teorema 4.2, temos que $\hat{\mu}_i^l(n)$ e $\hat{\mu}_1^l(n)$ são $\frac{\sigma}{\sqrt{T_l}}$-sub-gaussianas. Além disso, $\hat{\mu}_i^l(n)$ e $\hat{\mu}_1^l(n)$ são independentes, haja vista que as recompensas associadas a cada ação são independentes uma das outras, portanto, pela Proposição 2.1 segue que:

$$\hat{\mu}_i^l(n) - \hat{\mu}_1^l(n) \text{ é } \frac{\sqrt{2}\sigma}{\sqrt{T_l}}\text{-sub-gaussiana.}$$

Usando o Teorema 2.2:

$$\begin{aligned}
\mathbb{P}(\Delta_{A_{n+1}} > 0) \leq \sum_{l=1}^{k-1}\sum_{i=k-(l-1)}^{k} \exp\left\{-\frac{\Delta_{(i)}^2}{2\cdot\frac{2\sigma^2}{T_l}}\right\} &= \sum_{l=1}^{k-1}\sum_{i=k-(l-1)}^{k} \exp\left\{-\frac{T_l\cdot\Delta_{(i)}^2}{4\sigma^2}\right\} \\
&\leq \sum_{l=1}^{k-1} l \exp\left\{-\frac{T_l\cdot\Delta_{(k-l+1)}^2}{4\sigma^2}\right\}.
\end{aligned}$$

Utilizando a definição de T_l feita no Algoritmo 8 e o fato que pelo pressuposto de ordenação, $\Delta_{(k-l+1)} \leq \Delta_{(i)}$ para todo $i > k-l+1$, segue que:

$$T_l\cdot\Delta_{(k-l+1)}^2 \geq \frac{n-k}{\overline{\log}(k)\cdot\frac{(k-l+1)}{\Delta_{(k-l+1)}^2}} \geq \frac{n-k}{\overline{\log}(k)H_2},$$

onde $H_2 := \max_{i\in\{1,\dots k\}} i\Delta_{(i)}^{-2}$.Dessa forma, temos que:

$$\mathbb{P}(\Delta_{A_{n+1}} > 0) \leq \frac{k\cdot(k-1)}{2}\exp\left\{-\frac{n-k}{4\sigma^2\overline{\log}(k)H_2}\right\}.$$

□

Um caso específico do resultado acima é obtido ao considerar o ambiente $\mathcal{E}_{Ber}^k$, o que gera $\sigma = \frac{1}{2}$, levando ao resultado apresentado em (AUDIBERT; BUBECK, 2010). Nesse trabalho, os autores seguem um raciocínio análogo ao da prova acima, mas com a restrição ao ambiente $\mathcal{E}_{Ber}^k$, sendo uma contribuição do presente trabalho a extensão para qualquer ambiente.

Conforme visto para os algoritmos que lidam com o dilema entre exploração e explotação, é possível também encontrar limitantes inferiores para a probabilidade de falsa seleção. Uma discussão sobre limitantes inferiores para os algoritmos de exploração pura pode ser vista em (WANG; TZENG; PROUTIERE, 2023).

5 EXPERIMENTOS NUMÉRICOS

Ao longo das seções anteriores foram desenvolvidos diversos teoremas que fornecem limitantes inferiores e superiores para a medida de arrependimento ou a probabilidade de falsa seleção. No entanto, um aspecto que merece atenção é a verificação se resultados teóricos se manifestam de fato na prática, bem como a análise do comportamento de cada um dos algoritmos propostos em cenários extremos. Assim, o objetivo desta seção é apresentar os resultados obtidos com cada algoritmo, destacando as principais diferenças e características particulares de cada um deles.

Ao longo das simulações apresentadas nas próximas seções, foi considerado um horizonte de tamanho 10.000 e o numero de simulações igual a 20.000, sendo considerado para algoritmo ETC $m = 100$, para o ε-guloso $\varepsilon = 0,1$ e para o UCB $\delta = 0,05$.

Com a finalidade de avaliar o desempenho dos algoritmos, duas métricas serão utilizadas: a recompensa média e o arrependimento acumulado. A recompensa média é calculada a cada rodada, sendo obtida por meio da média de todas as recompensas obtidas até a corrente rodada, sendo, ao final de todas as simulações, feito a média ao longo das simulações. Seja $X_{j,i}$ a recompensa obtida na i-ésima rodada da j-ésima simulação, com $i \in [t]$ e $j \in [s]$, sendo $t \leq n$ e s o número de simulações. A recompensa média é dada pela seguinte equação:

$$\text{RecM}(s,t) := \frac{1}{s}\sum_{j=1}^{s}\frac{1}{t}\sum_{i=1}^{t} X_{j,i}. \tag{5.1}$$

O arrependimento acumulado é realizado de maneira análoga, a única diferença é que, ao longo das rodadas, o arrependimento é acumulado, sem a realização do calculo da média, ou seja:

$$\text{Arr}(s,t) := \frac{1}{s}\sum_{j=1}^{s}\sum_{i=1}^{t}(\mu^* - X_{j,i}). \tag{5.2}$$

onde $\mu^* = \max_{i\in[k]} \mu_i$. Nas seções subsequentes serão feitas avaliações dos algoritmos apresentados nos Capítulos 3 e 4 em cenários críticos, avaliando as métricas acima definidas.

Os experimentos foram realizados através do uso da linguagem de programação Julia, o código referente a cada um deles pode ser encontrado no seguinte endereço: <https://github.com/bastosismael/dissertacao_multi_armed_bandits>.

5.1 Primeiro cenário: Capacidade de seleção da melhor ação em um ambiente com recompensas similares

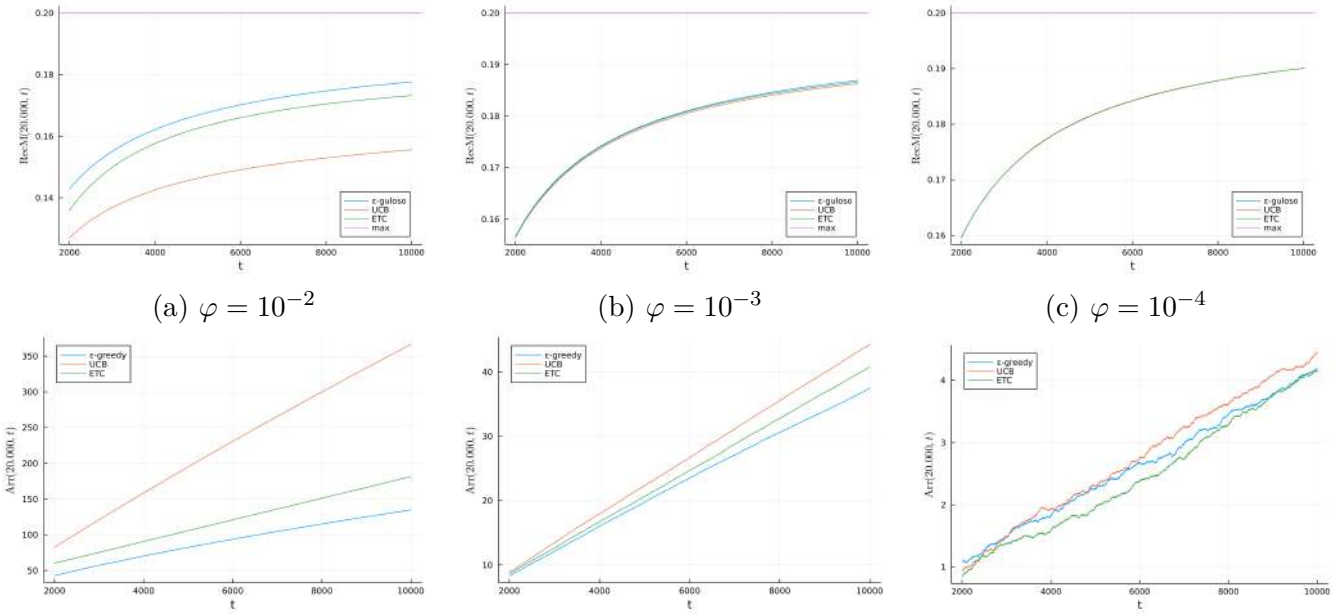
Este primeiro cenário surge a partir do questionamento sobre a capacidade dos algoritmos de identificarem a melhor ação em um ambiente onde as distribuições das recompensas de cada ação são mutuamente semelhantes. Naturalmente, espera-se que, à medida que aumente a similaridade entre as distribuições, mais difícil se torne distinguir qual ação é a mais vantajosa. Assim, o objetivo desta seção é investigar a eficácia de cada algoritmo na identificação da melhor ação considerando

um nível de similaridade φ. Dado um conjunto ordenado de ações, φ representa a diferença entre a média de duas ações consecutivas desse conjunto.

Para esse cenário, será considerado um ambiente $\mathcal{E}_{\text{Ber}}^{10}$ com cada ação $i \in [10] = \{1, 2, \ldots, 10\}$ possuindo média $0,2 - (\varphi \cdot (i-1)), \varphi \in \{10^{-2}, 10^{-3}, 10^{-4}\}$. Com efeito, a escolha da média associada ao braço ótimo ser igual a $0,2$ é arbitrária, podendo ser qualquer outro valor no intervalo $[0,1; 1]$. O importante é que, a medida que φ diminui, mais próximas se tornam as médias das ações.

As Figuras 1a, 1b e 1c apresentam uma comparação das recompensas para diferentes escolhas de φ. Observa-se que, à medida que φ diminui, as recompensas médias das ações tornam-se mais semelhantes e, consequentemente, as recompensas médias se aproximam. No entanto, ao analisar as Figuras 1a e 1b, nota-se que, de modo geral, o algoritmo ε-guloso se destacou em relação aos demais.

Sob outra perspectiva, as Figuras 1d, 1e e 1f comparam o arrependimento para os mesmos valores de φ. Os resultados obtidos estão em consonância com a análise das recompensas, indicando que o algoritmo ε-guloso apresenta o menor arrependimento. Ademais, mesmo com a adoção de 20.000 simulações, a Figura 1f ainda revela a presença de ruído.

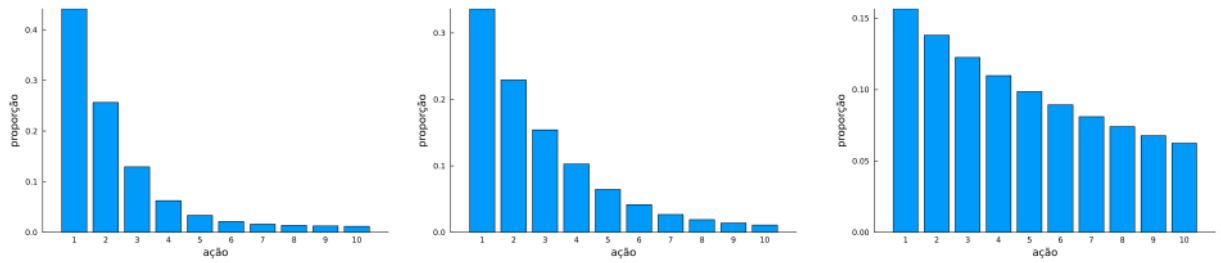


(a) $\varphi = 10^{-2}$ (b) $\varphi = 10^{-3}$ (c) $\varphi = 10^{-4}$

(d) $\varphi = 10^{-2}$ (e) $\varphi = 10^{-3}$ (f) $\varphi = 10^{-4}$

Figura 1 – Comparação entre as métricas obtidas ao longo das 20.000 simulações com diferentes valores de φ, para cada algoritmo, considerando horizonte $= 10.000$, $m = 100$ (ETC), $\varepsilon = 0{,}1$ (ε-guloso) e $\delta = 0,05$ (UCB) considerando um ambiente $\mathcal{E}_{\text{Ber}}^{10}$ com cada ação $i \in [10]$ possuindo média $0,2 - (\varphi \cdot (i-1))$.

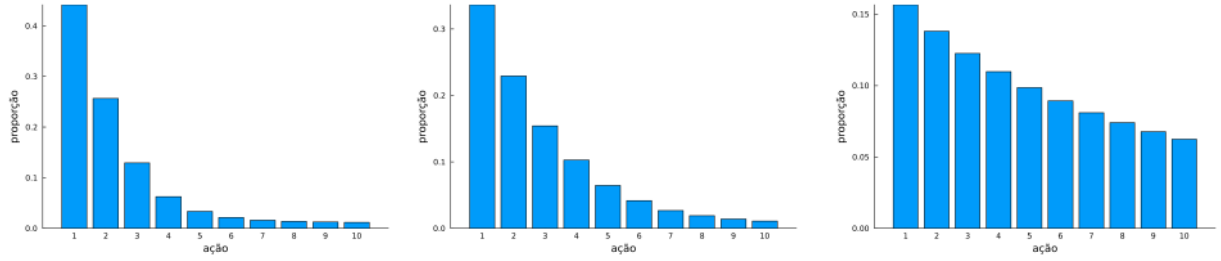
A Figura 2 apresenta os resultados obtidos para a proporção de seleção de cada um dos braços ao longo do horizonte e das simulações. É possível notar que para o caso em que $\varphi = 10^{-2}$, os algoritmos ε-guloso e o ETC parecem conseguir selecionar a ação ótima na maior parte do tempo, selecionado majoritariamente a ação ótima ou ações próximas à ótima. O algoritmo UCB, no entanto, acabou por selecionar a ações sub-ótimas um maior número de vezes quando comparado aos demais algoritmos.



(a) ε-guloso (b) ETC (c) UCB

Figura 2 – Comparação entre a proporção de seleção dos braços calculada após a execução das 20.000 simulações para cada um dos algoritmos com horizonte $= 10.000$, $\varphi = 10^{-2}$, $m = 100$ (ETC), $\varepsilon = 0,1$ (ε-guloso) e $\delta = 0,05$ (UCB) considerando um ambiente $\mathcal{E}_{\text{Ber}}^{10}$ com cada ação $i \in [10]$ possuindo média $0,2 - (\varphi \cdot (i-1))$.

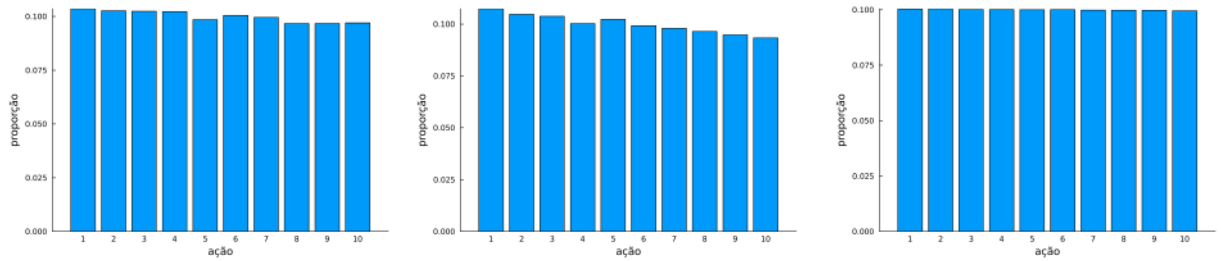
A Figura 3 apresenta os resultados obtidos para o caso em que $\varphi = 10^{-3}$. É possível notar um comportamento similar ao apresentado na Figura 2, evidenciando a dificuldade do algoritmo UCB em identificar a ação ótima quando comparado aos demais.



(a) ε-guloso (b) ETC (c) UCB

Figura 3 – Comparação entre a proporção de seleção dos preços calculada após a execulção das 20.000 simulações para cada um dos algoritmos com horizonte $= 10.000$, $\varphi = 10^{-3}$, $m = 100$ (ETC), $\varepsilon = 0,1$ (ε-guloso) e $\delta = 0,05$ (UCB) considerando um ambiente $\mathcal{E}_{\text{Ber}}^{10}$ com cada ação $i \in [10]$ possuindo média $0,2 - (\varphi \cdot (i-1))$.

A Figura 4 apresenta os resultados obtidos para o caso em que $\varphi = 10^{-4}$. Nesse caso, o algoritmo UCB seleciona cada ação exatamente na mesma proporção. Embora os algoritmos ε-guloso e ETC apresentem variações entre a proporção de seleção das ações, nota-se que os algoritmos selecionam cada ação aproximadamente o mesmo número de vezes.

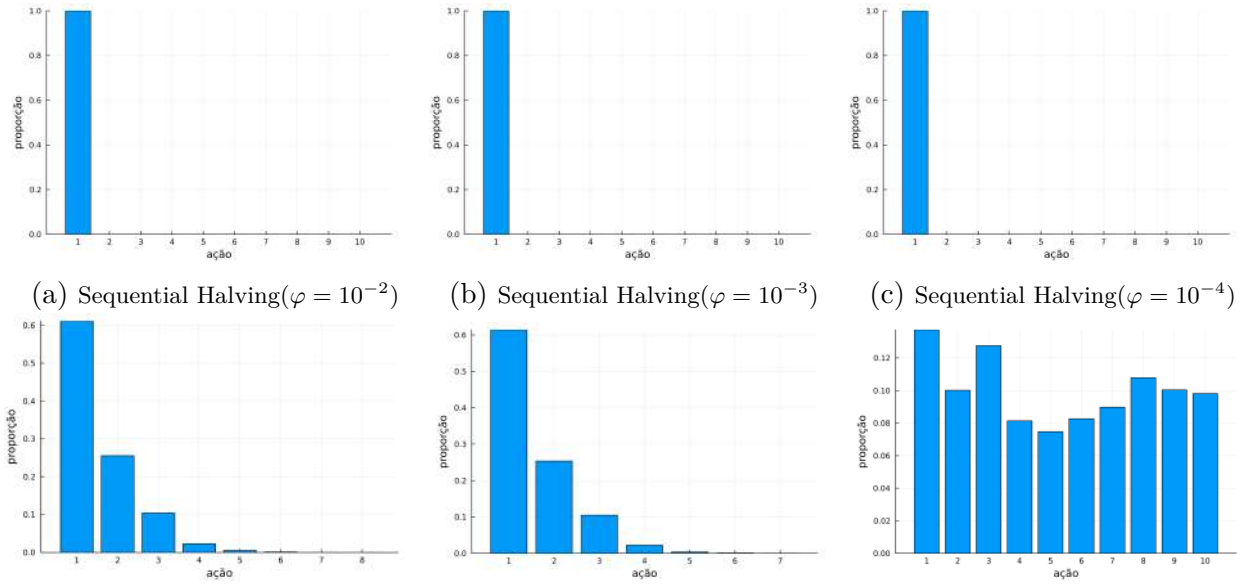


(a) ε-guloso (b) ETC (c) UCB

Figura 4 – Comparação entre a proporção de seleção dos preços calculada após a execução das 20.000 simulações para cada um dos algoritmos com horizonte $= 10.000$, $\varphi = 10^{-4}$, $m = 100$ (ETC), $\varepsilon = 0,1$ (ε-guloso) e $\delta = 0,05$ (UCB) considerando um ambiente $\mathcal{E}_{\text{Ber}}^{10}$ com cada ação $i \in [1, 10]$ possuindo média $0,2 - (\varphi \cdot (i-1))$.

De forma geral, é possível notar que o aumento da similaridade da média das recompensas associadas a cada um dos braços, faz com que se torne mais difícil identificar o braço ótimo por parte dos algoritmos testados, fazendo com que os algoritmos selecionem braços sub-ótimos um maior número de vezes. Em contrapartida, a recompensa média acaba não sendo afetada, uma vez que as recompensas são, em média, similares, então a seleção de uma ação sub-ótima não impacta significativamente na recompensa média obtida.

No contexto da exploração pura, as Figuras 5a, 5b, 5c evidenciam a proporção de seleção do algoritmo Sequantial Halving para cada um dos valores de φ. É possível perceber que mesmo com um $\varphi = 10^{-4}$ o algoritmo Sequential Halving selecionou a ação ótima em todas as 20.000 simulações, evidenciando que, mesmo em um ambiente com recompensas similares, o algoritmo consegue identificar a ação ótima. As Figuras 5d, 5e, 5f, em contrapartida, mostram que, para $\varphi = 10^{-4}$, o aumento da similaridade entre as médias das recompensas afetou negativamente a capacidade de seleção da ação ótima pelo algoritmo Sequential Elimination.



(a) Sequential Halving($\varphi = 10^{-2}$) (b) Sequential Halving($\varphi = 10^{-3}$) (c) Sequential Halving($\varphi = 10^{-4}$)

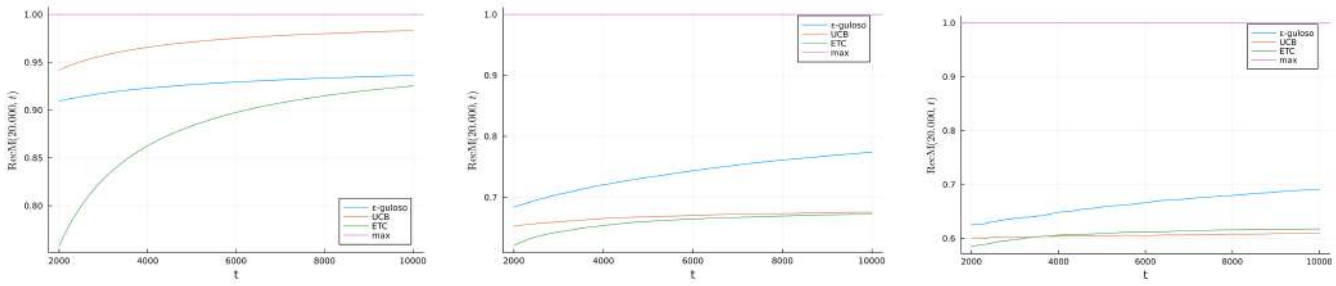
(d) Sequential Elimination($\varphi = 10^{-2}$) (e) Sequential Elimination($\varphi = 10^{-3}$) (f) Sequential Elimination($\varphi = 10^{-4}$)

Figura 5 – Comparação da proporção de vezes em que cada ação é recomendada após as 20.000 simulações com diferentes valores de φ, para cada algoritmo de exploração pura, considerando horizonte = 10.000 e um ambiente $\mathcal{E}^{10}_{\text{Ber}}$ com cada ação $i \in [1, 10]$ possuindo média $0,2 - (\varphi \cdot (i-1))$.

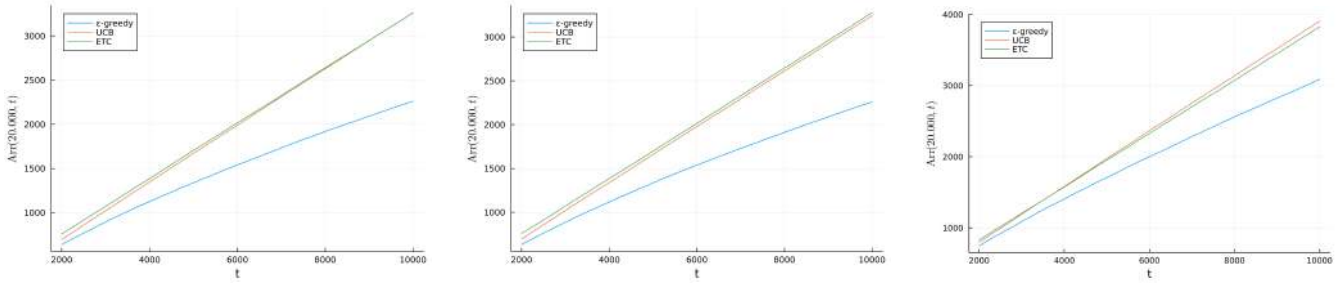
5.2 Influência da variância na seleção das ações

A fim de verificar a influência da variância na seleção das ações, na presente seção será considerado um ambiente $\mathcal{E}^{10}_{Gau}$, com cada ação possuindo média $\mu_i = 1 - (0,1 \cdot (i-1))$, com $i \in [10]$ e variância $\sigma^2 \in \{1, 10, 20\}$.

As Figuras 6a, 6b e 6c apresentam uma comparação das recompensas para diferentes escolhas de σ^2. Observa-se que, à medida que σ^2 aumenta, as recompensas médias diminuem, sendo possível destacar que, com o o aumento da variância, o algoritmo ε-guloso parece se tornar mais promissor. Ademais, as Figuras 6d, 6e e 6f indicam que o aumento da variância não provoca grandes alterações no arrependimento.



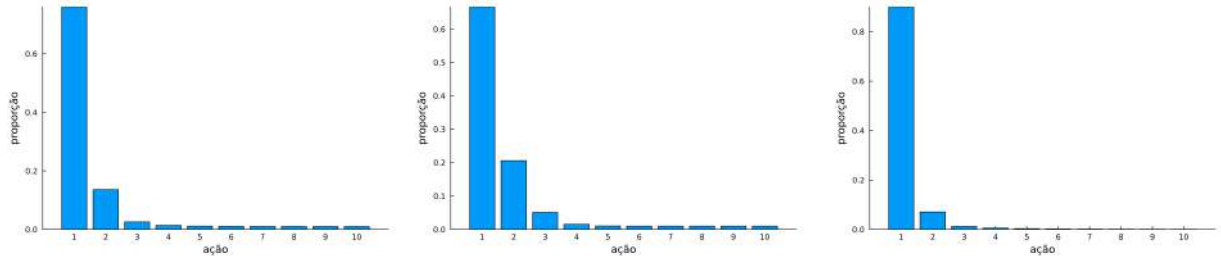
(a) $\sigma^2 = 1$ (b) $\sigma^2 = 10$ (c) $\sigma^2 = 20$



(d) $\sigma^2 = 1$ (e) $\sigma^2 = 10$ (f) $\sigma^2 = 20$

Figura 6 – Comparação entre as métricas obtidas ao longo das 20.000 simulações com diferentes valores de σ^2, para cada algoritmo, considerando horizonte = 10.000, $m = 100$ (ETC), $\varepsilon = 0{,}1$ (ε-guloso) e $\delta = 0,05$ (UCB) considerando um ambiente $\mathcal{E}^{10}_{\text{Ber}}$ com cada ação $i \in [10]$ possuindo média $1 - (0,1 \cdot (i-1))$.

A Figura 7 apresenta a proporção de seleção de cada uma das ações ao longo de 10.000 rodadas e 20.000 simulações, considerando o cenário em que as variâncias associadas às distribuições de recompensa das ações são iguais a 1. Observa-se que, de modo geral, os algoritmos conseguem identificar a ação ótima, com destaque para o algoritmo UCB, que a seleciona com maior frequência em comparação aos demais algoritmos avaliados.



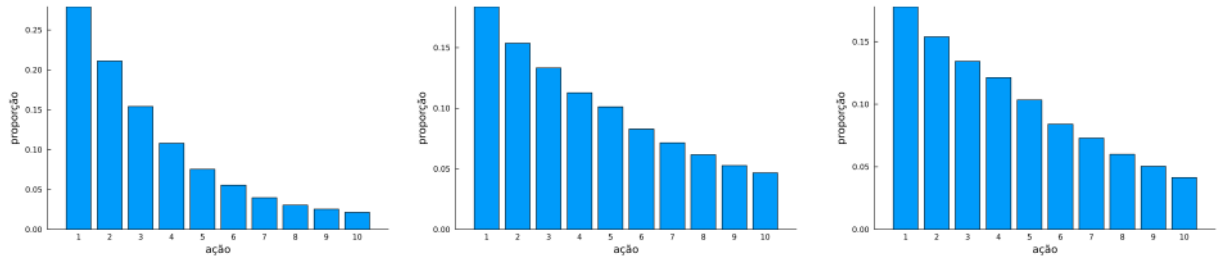
(a) ε-guloso (b) ETC (c) UCB

Figura 7 – Comparação entre a proporção de seleção das ações calculada ao final das 20.000 simulações para cada um dos algoritmos com horizonte = 10.000, $m = 100$ (ETC), $\varepsilon = 0,1$ (ε-guloso) e $\delta = 0,05$ (UCB) considerando um ambiente $\mathcal{E}^{10}_{Gau}$, com cada ação possuindo média $\mu_i = 1 - (0,1 \cdot (i-1))$ e variância igual a 1.

.

A Figura 8 apresenta a proporção de seleção de cada uma das ações ao longo do horizonte e das simulações, para cada um dos algoritmos testados, no cenário em que a variância associada à recompensa de cada braço é igual a 10. É possível notar que, em comparação com os resultados da Figura 7, a seleção de ações sub-ótimas foi mais frequente. Onde, por exemplo, o algoritmo UCB, que anteriormente selecionava a ação ótima em aproximadamente 80% das vezes, passa a selecioná-la em apenas cerca de 15% das simulações. Ademais, destaca-se que, entre os algoritmos avaliados, o

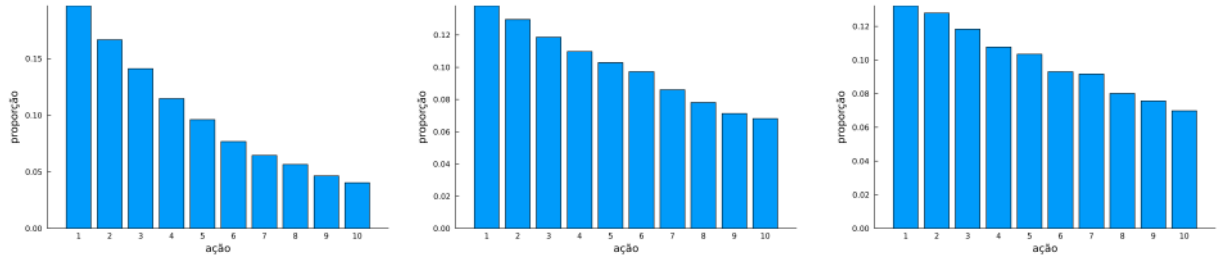
ε-guloso foi o que apresentou o melhor desempenho nesse cenário, selecionando a ação ótima em aproximadamente 25% das vezes.



(a) ε-guloso (b) ETC (c) UCB

Figura 8 – Comparação entre a proporção de seleção das ações calculada ao final das 20.000 simulações para cada um dos algoritmos com horizonte = 10.000, $m = 100$ (ETC), $\varepsilon = 0,1$ (ε-guloso) e $\delta = 0,05$ (UCB) considerando um ambiente $\mathcal{E}^{10}_{Gau}$, com cada ação possuindo média $\mu_i = 1 - (0,1 \cdot (i-1))$ e variância igual a 10.

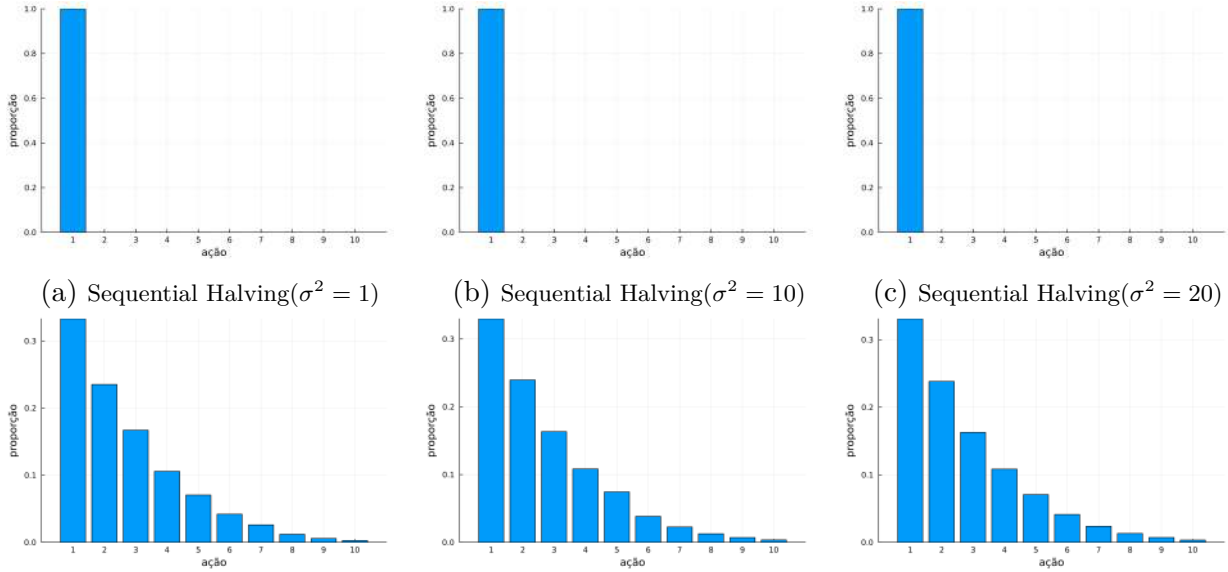
Figura 9 apresenta a proporção de seleção de cada uma das ações ao longo do horizonte e das simulações, para cada um dos algoritmos testados, no cenário em que a variância associada à recompensa de cada braço é igual a 20. Observa-se que, em comparação com os resultados da Figura 8, houve um aumento na seleção de ações sub-ótimas.



(a) ε-guloso (b) ETC (c) UCB

Figura 9 – Comparação entre a proporção de seleção das ações calculada após a execução das 20.000 simulações para cada um dos algoritmos com horizonte = 10.000, $m = 100$ (ETC), $\varepsilon = 0,1$ (ε-guloso) e $\delta = 0,05$ (UCB) considerando um ambiente $\mathcal{E}^{10}_{Gau}$, com cada ação possuindo média $\mu_i = 1 - (0,1 \cdot (i-1))$ e variância igual a 20.

No contexto da exploração pura, as Figuras 10a, 10b e 10c apresentam a proporção de seleção de cada um dos braços pelo algoritmo Sequential Halving. Observa-se que, independentemente da variância, o algoritmo selecionou a ação ótima em todas as 20.000 simulações. Por outro lado, ao analisar as Figuras 10d, 10e e 10f, nota-se que o algoritmo Sequential Elimination, embora não tenha selecionado exclusivamente a ação ótima como fez o Sequential Halving, manteve uma proporção de seleção constante entre as diferentes escolhas de σ^2, independentemente da variância.



(a) Sequential Halving($\sigma^2 = 1$) (b) Sequential Halving($\sigma^2 = 10$) (c) Sequential Halving($\sigma^2 = 20$)

(d) Sequential Elimination($\sigma^2 = 1$) (e) Sequential Elimination($\sigma^2 = 10$) (f) Sequential Elimination($\sigma^2 = 20$)

Figura 10 – Comparação da proporção de vezes em que cada ação é recomendada após as 20.000 simulações com diferentes valores de φ, para cada algoritmo de exploração pura, considerando horizonte = 10.000 e um ambiente $\mathcal{E}^{10}_{\text{Gau}}$ com cada ação $i \in [1, 10]$ possuindo média $1 - (0,1 \cdot (i-1))$.

Através das análises da presente seção, se torna evidente que a variância das recompensas impacta diretamente na recompensa média obtida, mas, no entanto, não gera grande impacto no arrependimento. Além disso, o aumento da variância parece também não gerar impacto na capacidade de seleção dos algoritmos de exploração pura.

6 PRECIFICAÇÃO DINÂMICA

De acordo com (BOER, 2015), a precificação dinâmica é o estudo da determinação de preços de venda ótimos para produtos e serviços em um cenário em que os preços podem ser ajustados com frequência. Um dos primeiros trabalhos a abordar a determinação de preços de venda ótimos é (COURNOT, 1838). No entanto, a perspectiva adotada pelo autor assume a existência de uma função contínua de demanda, que atribui um valor de demanda para cada preço de um determinado produto. Conforme discutido em (COURNOT, 1838), no âmbito da microeconomia, assume-se geralmente que a função de demanda é conhecida. Sob esta hipótese, o problema da determinação do preço ótimo reduz-se a um problema de otimização, ou seja a identificar o preço que maximiza a receita, sendo a receita definida como o produto entre a função de demanda e o preço do bem.

No entanto, em situações práticas, se torna difícil ter conhecimento prévio da função de demanda, o que complica a aplicação direta dessa estratégia. Nesse sentido, conforme será visto ao longo da presente dissertação, a adoção de algoritmos de *multi-armed bandits* oferece uma alternativa viável, uma vez que não necessita de conhecimento da função de demanda.

Relacionando o problema de precificação dinâmica ao contexto dos *multi-armed bandits*, assume-se a existência de um vendedor que deseja comercializar um produto cujo preço pode ser escolhido a partir de um conjunto finito de valores. A cada rodada, o vendedor seleciona um preço (braço) e oferece o produto a um potencial consumidor, que decide se realiza ou não a compra. Nesse caso, o vendedor não tem informação alguma acerca do comportamento do consumidor, sendo sua única fonte de informação a recompensa obtida a partir da resposta do cliente.

Nesse contexto, a decisão de compra do consumidor de um produto por um preço i é representada pela variável aleatória Y_i, onde $Y_i \sim Ber(p_i)$, com p_i representando a probabilidade de compra do produto pelo cliente ao preço i, sendo esse parâmetro desconhecido. Assim, o vendedor recebe como recompensa o preço do produto caso a compra ocorra; caso contrário, a recompensa é igual a zero. Dessa forma, a recompensa obtida pode ser expressa como $i \cdot Y_i$. É importante atentar para o fato que, no contexto da precificação dinâmica, a recompensa média, i.e., $\mu_i = ip_i$, é igual a receita média.

Uma alternativa possível consiste em modelar $Y_i \sim \text{Binomial}(n, p_i)$, onde o vendedor oferece seu produto a um grupo de n indivíduos, cada um decidindo ,de forma independente, se realiza a compra. Assim, a recompensa do vendedor é dada pelo preço do produto multiplicado pelo número total de compradores.

Outra abordagem consiste em definir Y_i como uma variável aleatória seguindo uma distribuição de Poisson composta, permitindo modelar cenários em que o número de clientes potenciais varia aleatoriamente, capturando incertezas adicionais no processo de compra. Cabe destacar, no entanto, que apesar das diferentes possibilidades de definição para Y_i, neste trabalho será considerado apenas o caso em que $Y_i \sim Ber(p_i)$.

De maneira objetiva, o vendedor deve adotar uma política, como por exemplo uma das polí-

ticas definidos no Capítulo 3, e segui-la ao longo das rodadas. Dessa forma, surge a necessidade de, conforme apresentado no início do capítulo, se estabelecer essa forma de precificação em um cenário onde os preços podem ser ajustados com frequência, haja vista que a cada rodada se faz necessário definir o preço do produto de acordo com a política adotada.

Para ilustrar o funcionamento desse processo, serão realizadas simulações da precificação dinâmica para um vendedor monopolista que comercializa um único produto, utilizando algoritmos definidos nos Capítulos 3 e 4. Para isso, define-se o conjunto de preços possíveis como $\mathcal{K} = \{2, 4, 6, 8, 10, \ldots, 40\}$, com cardinalidade $|\mathcal{K}| = 20$. O fato do vendedor ser monopolista é apenas uma simplificação do problema e evita a necessidade de se considerar a concorrência, o que eleva a complexidade do problema, uma vez que seria necessário explorar a relação do vendedor com os concorrentes[1]. Será realizada uma simulação de dois cenários envolvendo o contexto da precificação dinâmica. No primeiro cenário, assume-se a existência de um vetor de probabilidades C definido da seguinte forma:

$$C := [0,05; 0,03; 0,1; 0,02; 0,02; 0,01; 0,02; 0,01; 0,01; 0,06; 0,2; 0,02; 0,05; 0,3; 0,04; 0,06], \quad (6.1)$$

onde cada entrada i do vetor representa a probabilidade do cliente comprar o produto ao preço $2i$, ou seja:

$$p_i = C_{i/2}, \ i \in \mathcal{K}, \quad (6.2)$$

cabe ressaltar que na simulação a ser realizada, o vendedor não possui conhecimento de C.

A Figura 11 exibe a curva do valor esperado $\mu_x = x \cdot p_x$. Observa-se que o preço que maximiza o valor esperado da recompensa é $x = 28$.

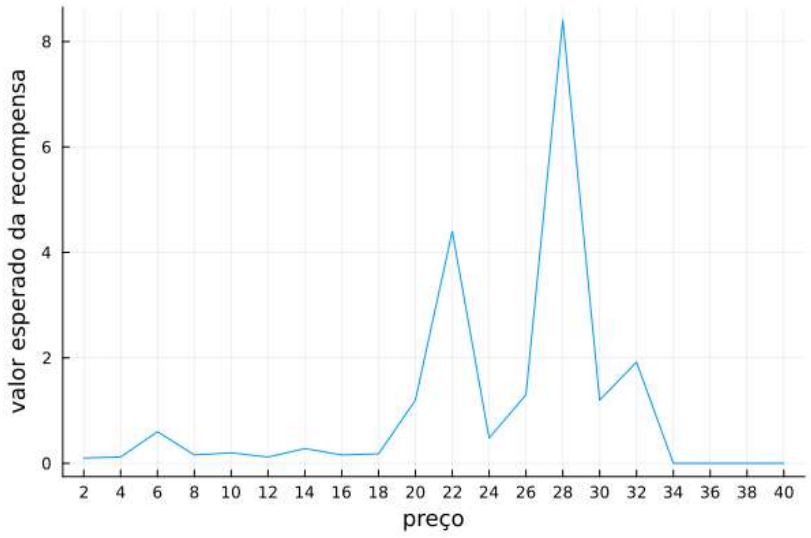


Figura 11 – Valor esperado da recompensa, i.e., $\mu_x = x \cdot p_x$, com p_x dado em (6.1).

No segundo cenário, com o objetivo de simular o comportamento do consumidor, conforme proposto em (TROVO et al., 2018), assumimos a existência de uma variável aleatória não negativa S e definimos a probabilidade de compra do produto ao preço i como:

$$p_i := \mathbb{P}(S \geq i). \quad (6.3)$$

[1] Estudos desse tipo recaem no campo dos multi-armed bandits com vários agentes (do inglês multi-agent multi-armed bandits). Esse é um tema relativamente recente (ver, por exemplo (AGARWAL; AGGARWAL; AZIZZADENESHELI, 2022)) e que não é abordado na presente dissertação.

Embora p_i seja desconhecido (a distribuição de S é desconhecida), a definição feita em (6.3) garante a propriedade de monotonicidade, ou seja, quanto maior o preço, menor a probabilidade de compra, o que é desejável e possui sentido prático.

Do ponto de vista de uma aplicação no mundo real, a distribuição da variável aleatória S é desconhecida, no entanto, para fins de simulação, assume-se que $S \sim Exp(20)$. Vale destacar que, durante a simulação, tanto o vendedor quanto os clientes desconhecem essa distribuição. Se valendo da definição de S conjuntamente com a Definição 6.3, a Figura 12 apresenta a curva do valor esperado $\mu_x = x\mathbb{P}(S \geq x) = x(1 - e^{-20x})$. Na curva apresentada na Figura 12 é possível notar que o preço que maximiza o valor esperado da recompensa é $x = 20$.

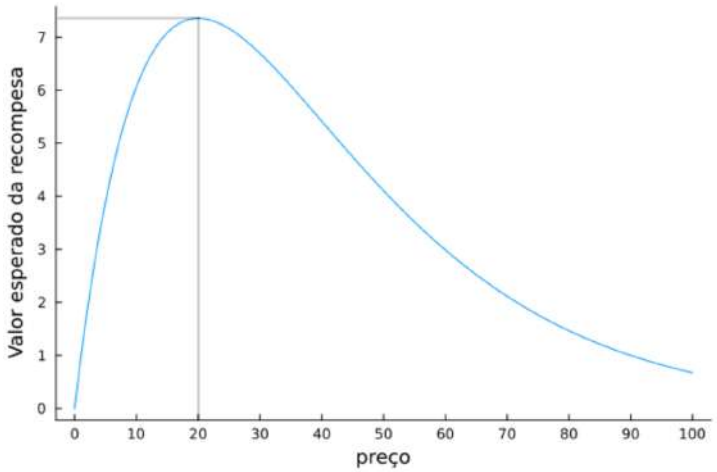
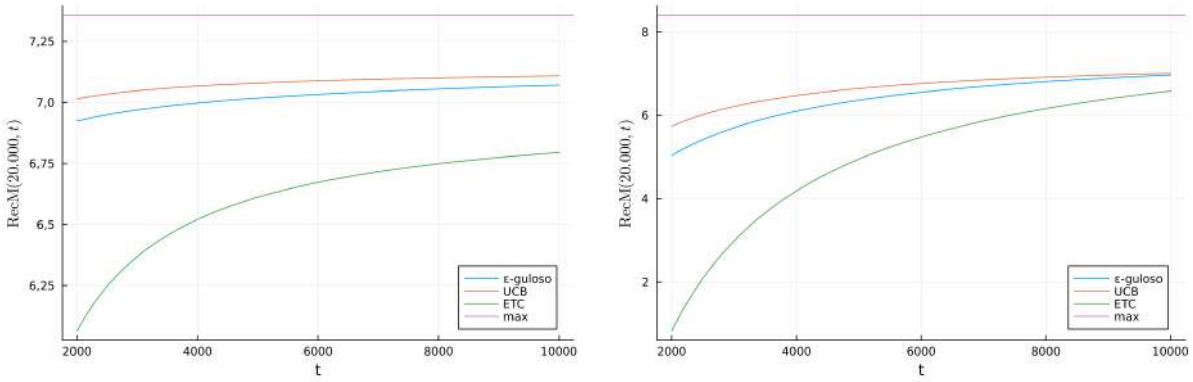


Figura 12 – Valor esperado da recompensa i.e., $\mu_x = x(1 - e^{-20x})$.

Conforme feito na Seção 5, com o objetivo de e avaliar o desempenho dos algoritmos no cenário de precificação dinâmica, serão adotas as mesma métricas adotadas na Seção 5, se valendo da Equação 5.1 para o cálculo da Recompensa Média e da Equação 5.2 para o Arrependimento Acumulado. Além disso, manteve-se a mesmo número de simulações e horizonte que o definido previamente.

6.1 Exploração vs Explotação

Para avaliar o desempenho de algoritmos que se baseiam no dilema entre exploração e explotação no contexto da precificação dinâmica, foram adotados os algoritmos UCB, ETC e ε-guloso. No caso do ETC, o parâmetro de exploração (m) foi fixado em 100. Para o ε-guloso, utilizou-se $\varepsilon = 0,1$, enquanto para o UCB, o valor de δ foi definido como $0,05$. A Figura 13 exibe os resultados da recompensa média obtida nos testes com os três algoritmos, utilizando as Definições 6.2 e 6.3.



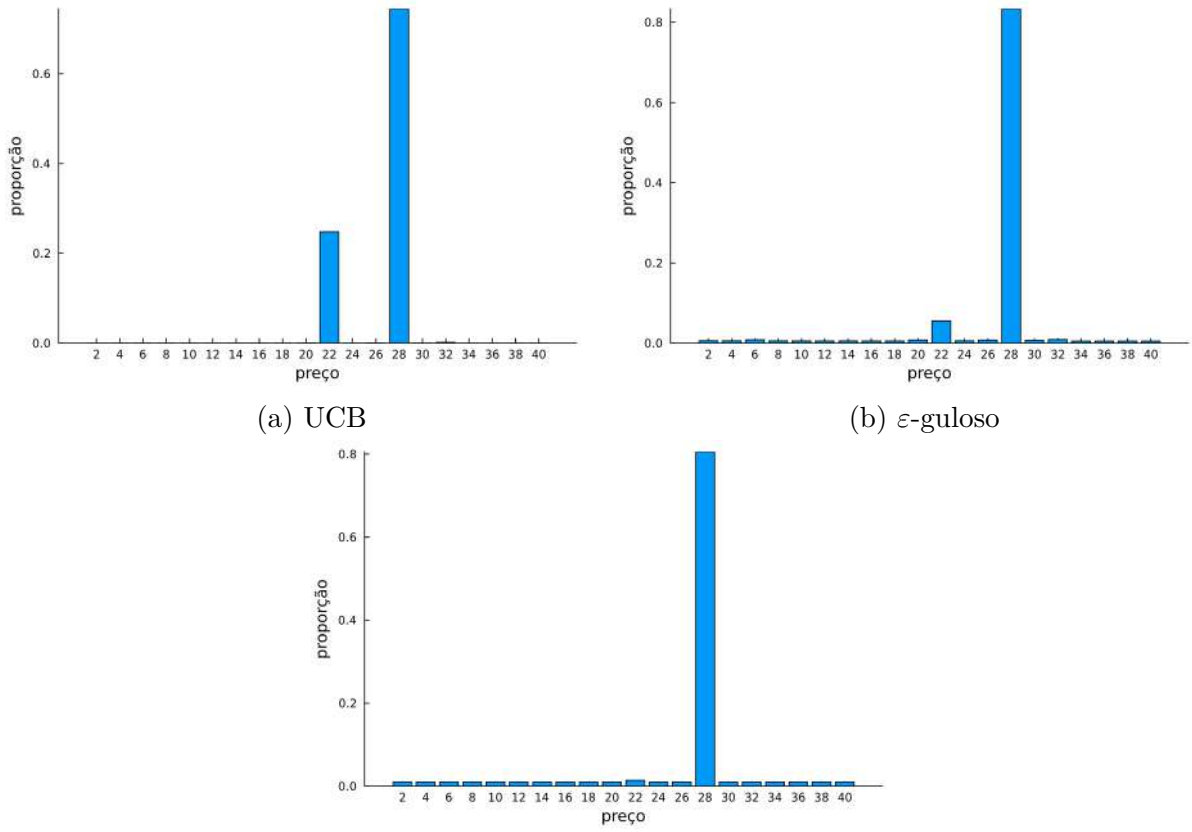
(a) $p_i = \mathbb{P}(S \geq i), S \sim \exp 20$ (b) $p_i = C_{i/2}$

Figura 13 – Recompensa média ao longo de 20.000 simulações horizonte = 10.000, $m = 100$ (ETC), $\varepsilon = 0,1$ (ε-guloso) e $\delta = 0,05$ (UCB).

Na Figura 13, é possível observar que o algoritmo UCB gerou uma recompensa média superior em comparação aos demais. `ismael:` Além disso, ainda nesta mesma figura, nota-se que, ao final das simulações, o algoritmo UCB gerou recompensa média acumulada aproximadamente 4% menor que o valor ótimo, representado no gráfico pela linha roxa. O valor ótimo é calculado pelo produto entre o preço que maximiza a recompensa esperada e a probabilidade de compra associada a esse preço, em (a) esse valor é igual a $20 \cdot \mathbb{P}(S \geq 20) \approx 7,36$ e em (b) $28 \cdot 0,3 = 8,4$.

Cabe ressaltar que, para melhor visualização dos resultados na Figura 13, foram desconsideradas as médias referente às primeiras 2.000 rodadas, uma vez que as flutuações iniciais tendiam a distorcer a visualização dos resultados. Ademais, a escolha do valor 2.000 decorre principalmente do fato de que, dada a escolha de $m = 100$ para o algoritmo ETC, as primeiras 2.000 rodadas são reservadas para a exploração, sendo a partir da rodada 2.000 que o algoritmo passa a focar na explotação.

A Figura 14 apresenta uma comparação entre a proporção de vezes que cada um dos preços é selecionado pelos algoritmos ao considerar o cenário em que p_i é definida de acordo com 6.2. É possível notar que os algoritmos se concentram no preço de \$28, enfatizando a Figura 14c, onde o algoritmo ETC se concentra quase unicamente no preço de \$28.



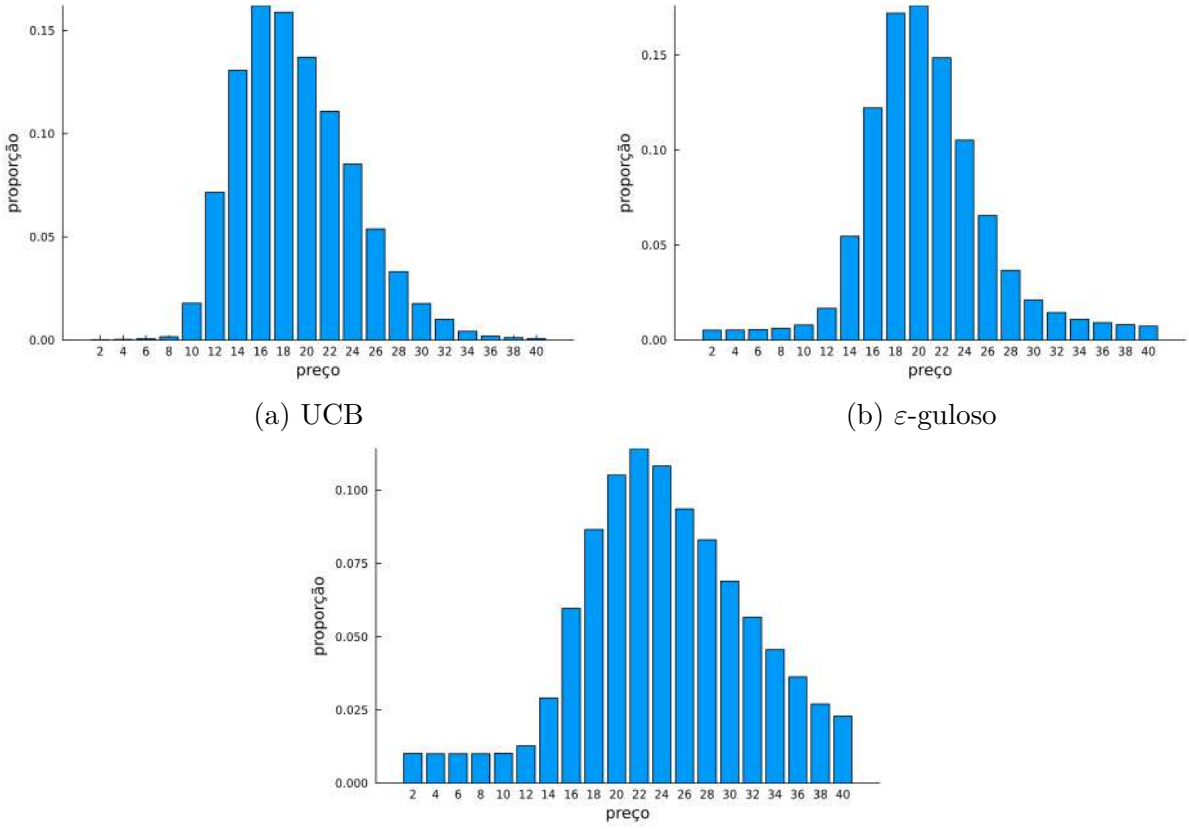
(a) UCB

(b) ε-guloso

(c) ETC

Figura 14 – Comparação entre a proporção de seleção dos preços calculada ao longo das 20.000 simulações, para cada um dos algoritmos com horizonte = 10.000, $m = 100$ (ETC), $\varepsilon = 0,1$ (ε-guloso) e $\delta = 0,05$ (UCB).

Já a Figura 15 descreve a mesma situação apresentada na Figura 14, porém se valendo da Definição 6.3. É possível perceber que os algoritmos também tendem a concentrar-se em preços próximos àquele que maximiza o valor esperado da recompensa, porém de forma menos intensa. Cabe notar que na Figura 15b os preços selecionados parecem se concentrar simetricamente ao redor do preço ótimo, enquanto na Figura 15b os preços selecionados parecem ser levemente direcionados aos preços a direita do preço ótimo, sendo essa característica presente de forma mais acentuada na Figura 15c, onde os preços selecionados se concentram majoritariamente à direita.



(a) UCB

(b) ε-guloso

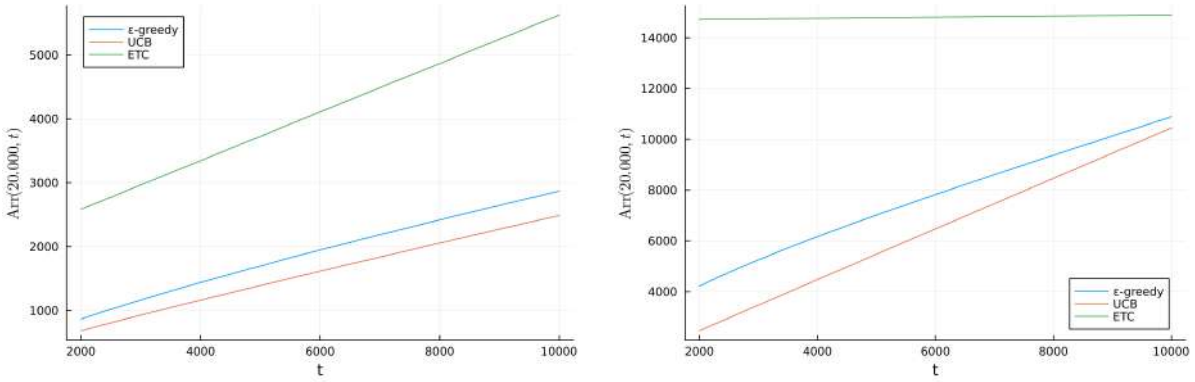
(c) ETC

Figura 15 – Comparação entre a proporção de seleção dos preços para cada um dos algoritmos com horizonte = 10.000, $m = 100$ (ETC), $\varepsilon = 0,1$ (ε-guloso) e $\delta = 0,05$ (UCB).

Analisando as Figuras 14 e 15 , nota-se que, apesar de se concentrarem no valor que maximiza a recompensa esperada, alguns algoritmos acabam escolhendo por mais tempo preços afastados do ótimo. Nesse sentido, nota-se que os resultados apresentados estão intrinsecamente ligados aos da Figuras 13, uma vez que, o fato de selecionarem preços mais afastados acaba por reduzir a recompensa média ao longo das rodadas.

Além disso, os resultados presentes na Figura 14 mostram que para o caso em que p_i é definida de acordo com 6.2, os algoritmos conseguem identificar com maior facilidades o preço ótimo quando comparado aos resultados obtidos utilizando a Definição 6.3 exibidos na Figura 15.

A Figura 16 exibe a comparação do arrependimento acumulado para os algoritmos testados. É possível notar que o algoritmo UCB apresenta arrependimento acumulado inferior aos demais, o que corrobora com o resultado obtido na Figura 13. Já na Figura 16b, é possível perceber que o algoritmo ETC apresentou arrependimento mais alto que os demais algoritmos, mas com comportamento aparentemente constante, isso se deve ao resultado obtido na Figura 14c, onde o algoritmo ETC se concentra majoritariamente no preço ótimo, sendo o arrependimento afetado fortemente pelo período de exploração.



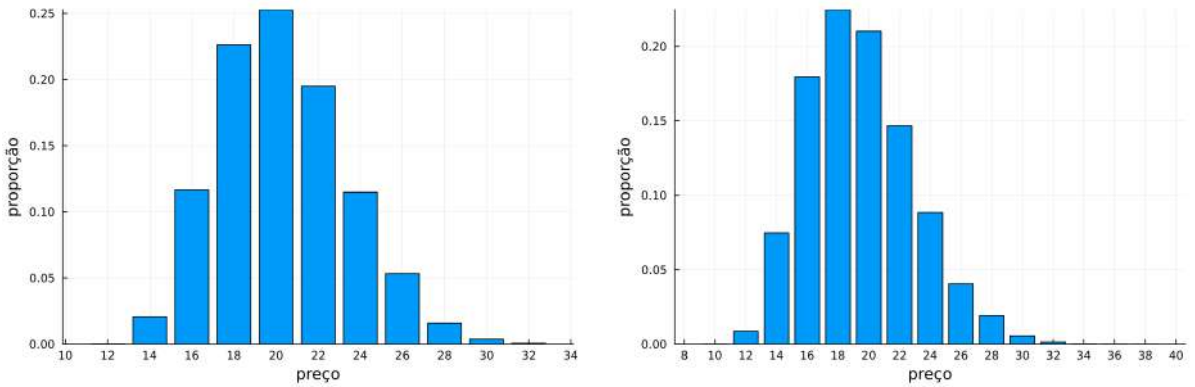
(a) $p_i = \mathbb{P}(S \geq i), S \sim \exp 20$ (b) $p_i = C_{i/2}$

Figura 16 – Arrependimento acumulado ao longo de 20.000 simulações considerando horizonte = 10.000, $m = 100$ (ETC), $\varepsilon = 0,1$ (ε-guloso) e $\delta = 0,05$ (UCB).

6.2 Exploração pura

Com o objetivo de verificar o comportamento dos algoritmos de exploração pura no cenário da precificação dinâmica, optou-se por considerar ambos os algoritmos propostos no Capítulo 4. Vale ressaltar que ao adotar algoritmos de exploração pura, o objetivo do vendedor se desloca para o de encontrar o preço ótimo, não se preocupando com a recompensa acumulada ao longo do horizonte.

A Figura 17 apresenta a proporção de vezes que cada um dos preços é selecionado pelos algoritmos Sequential Halving e Sequential Elimination. É possível perceber que ambos os algoritmos acabam por identificar o preço que maximiza o valor esperado da recompensa ou preços próximos, desde que esses preços se encontrem dentro do conjunto de possíveis preços, sendo a capacidade de identificação relacionada diretamente com o horizonte adotado, conforme evidenciado no Capítulo 4.



(a) Sequential Halving. (b) Sequential Elimination.

Figura 17 – Proporção de vezes em que acada preço é selecionado durante a execução dos algoritmos de exploração pura.

Cabe destacar que, na presente seção, são realizados testes utilizando apenas com p_i definido de acordo com 6.3, uma vez que ao adotar a Definição 6.2, dado o horizonte adotado, todos os algoritmos identificaram o preço ótimo em cada uma das simulações.

Apesar dos algoritmos de exploração pura, de maneira geral, oferecerem diretamente o preço que maximiza o valor esperado da recompensa, é importante atentar para que esse valor oferecido não

é explotado ao longo das rodadas, culminando em uma recompensa média inferior às apresentadas na Figura 13. Dessa forma, a Figura 18 evidencia a recompensa média obtida ao longo da execução dos algoritmos de exploração pura.

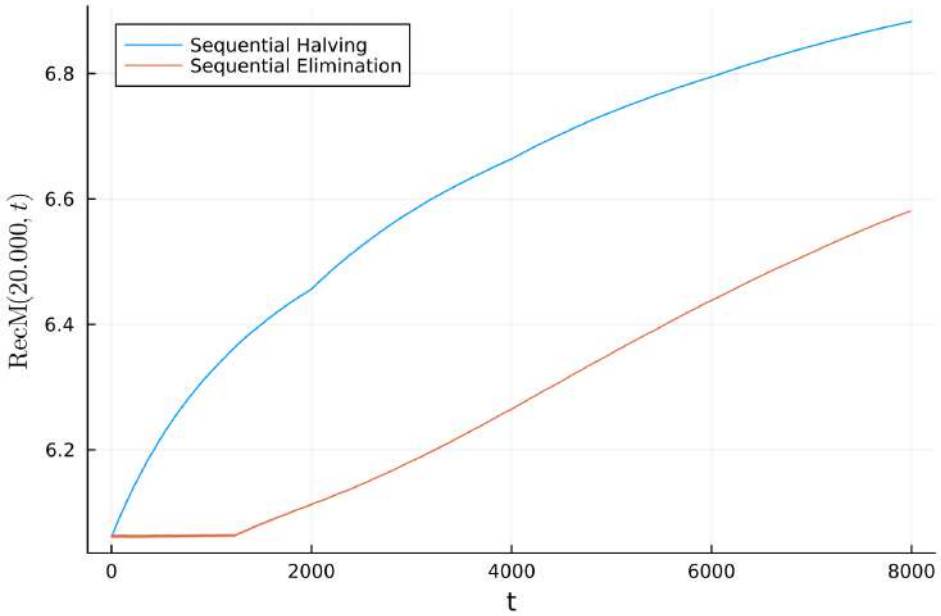


Figura 18 – Recompensa média ao longo de 20.000 simulações para os algoritmos Sequential Halving e Sequential Elimination considerando um horizonte de tamanho 10.000.

Ao analisar a Figura 18, nota-se que o algoritmo Sequential Halving apresenta recompensa média superior ao de eliminação sequencial, o que não é surpreendente, uma vez que, devido a sua construção, acaba por eliminar metade das ações potencialmente não promissoras a cada rodada, enquanto o de eliminação sequencial acaba focando em ações sub-ótimas por um maior número de rodadas.

Tendo em vista a dicotomia existente entre o problema da exploração pura e do dilema entre exploração e explotação, o presente texto apresenta uma nova abordagem que visa unir essas duas vertentes, buscando extrair os benefícios apresentados por cada um. A ideia central do algoritmo proposto é dividir o horizonte: inicialmente, aplica-se um algoritmo de exploração pura em uma parcela desse horizonte; em seguida, explora-se a ação recomendada ao longo da parte restante. Com essa estratégia, torna-se possível combinar a capacidade dos algoritmos de exploração pura em identificar a melhor ação com a fase de explotação presente nos algoritmos voltados ao dilema exploração–explotação.

Realizando um paralelo com os algoritmos definidos no Capítulo 3, a abordagem proposta se assemelha a do algoritmo ETC, porém, utilizando o algoritmo Sequential Halving durante fase de exploração. Devido a essa similaridade, o algoritmo proposto será chamado de SHTC (do inglês *Sequential Halving Then Commit*).

Um questionamento que surge em decorrência da definição do algoritmo SHTC é sobre a proporção do horizonte que deve ser alocada em cada etapa. No presente texto se considera um horizonte de exploração igual a $\alpha \cdot n$, onde $\alpha \in (0, 1)$, consequentemente o horizonte de explotação é igual a $n(1 - \alpha)$. O funcionamento do algoritmo SHTC é apresentado no Algoritmo 9.

Algoritmo 9 SHTC

Entrada: n, k, α
Defina A := Sequential-Halving($\lfloor \alpha \cdot n \rfloor, k$)
Para todo $t \in \{\lfloor \alpha \cdot n \rfloor + 1, \ldots, n\}$ escolha $A_t = \text{A}$

Ao olhar para seu funcionamento, se torna importante atentar para o fato que, diferentemente dos algoritmos que focam unicamente na exploração pura, se faz presente o dilema entre exploração e explotação, uma vez que se faz necessário definir a fração do horizonte que será alocada para o algoritmo *Sequential Halving*. Por outro lado, diferentemente dos algoritmos apresentados na Seção 3, o algoritmo SHTC também gera a necessidade de identificação da melhor ação por parte do algoritmo *Sequential Halving* durante o período de exploração, haja vista que a ação selecionada será explorada ao longo do resto do horizonte, não havendo chance para mudanças.

A Figura 19 apresenta uma comparação envolvendo o algoritmo SHTC com $\alpha = 0,5$ e os demais algoritmos testados anteriormente. É possível perceber que o algoritmo SHTC apresentou recompensa média superior ao algoritmo ETC, se mostrando uma alternativa promissora.

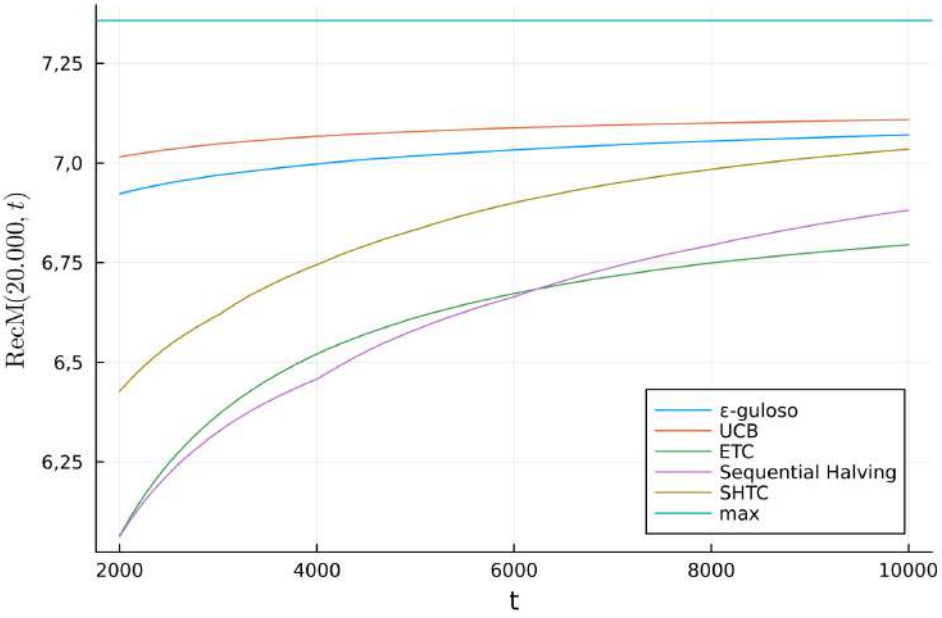
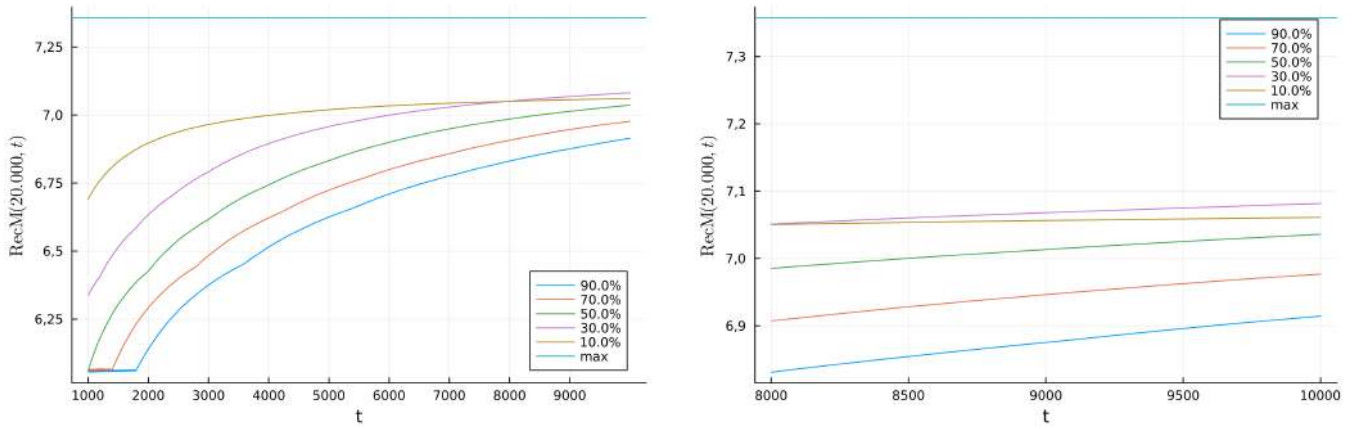


Figura 19 – Comparação da recompensa média obtida para entre o algoritmo SHTC com $\alpha = 0,5$ e os demais algoritmos.

Um questionamento que surge em decorrência da definição do algoritmo SHTC é sobre a proporção do horizonte que deve ser alocada em cada etapa. Dessa forma, a Figura 20 apresenta a recompensa média obtida utilizando diferentes proporções do horizonte para o período de exploração.



(a) Recompensa média para o algoritmo SHTC obtida ao longo das 10.000 rodadas, excluindo as 1.000 primeiras.

(b) Recompensa média obtida entre as rodadas 8.000 e 10.000.

Figura 20 – Comparação da recompensa média para o algoritmo SHTC obtida ao considerar diferentes tamanhos de horizonte durante a fase de exploração.

Com base na Figura 20a, é possível notar que reduzir a parcela do horizonte destinada ao algoritmo Sequential Halving implica no aumento da recompensa média ao longo de sua execução, evidenciando que a parcela mais impactante negativamente é a recompensa obtida ao longo da execução do algoritmo Sequential Halving. Analisando a Figura 20b, nota-se que a escolha de 30% do horizonte total como período de exploração parece ser a melhor escolha para o caso estudado.

Em consonância com a Figura 20, a Figura 21 apresenta o arrependimento acumulado ao longo das 10.000 rodadas. É possível verificar que a escolha do tamanho do período de exploração como sendo 30% gera um menor arrependimento acumulado no longo prazo. Com efeito, a escolha de 10% provoca o menor arrependimento acumulado e maior recompensa média até aproximadamente a rodada 8.000, entretanto, após esse ponto, a escolha de 30% prevalece, evidenciando a existência de uma escolha ótima para o parâmetro α.

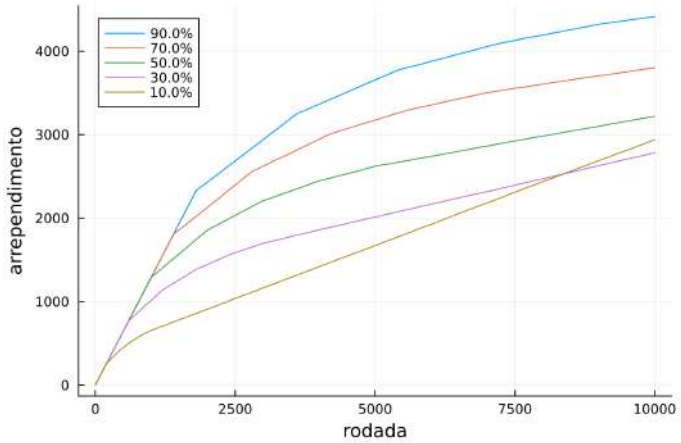
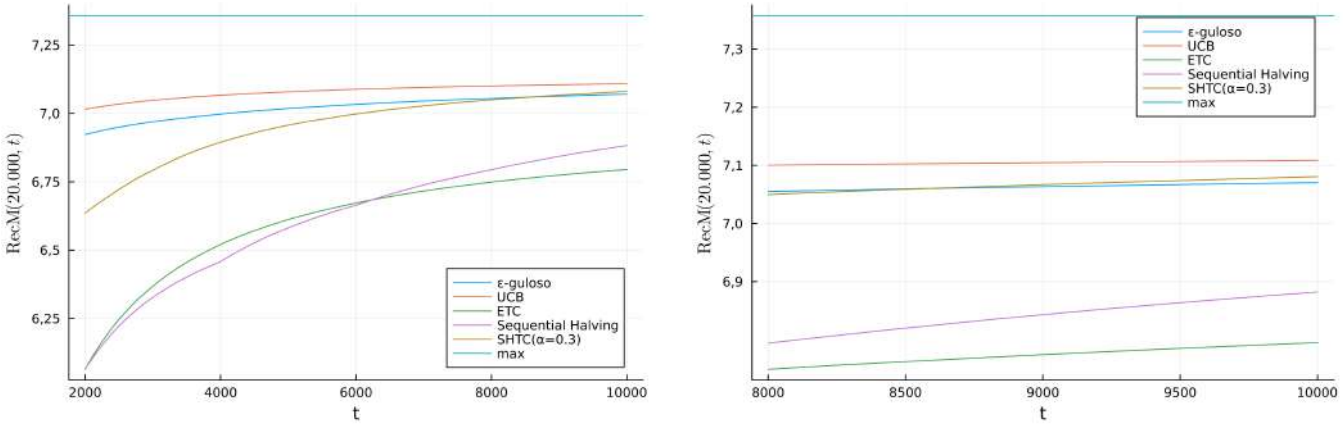


Figura 21 – Comparação do arrependimento acumulado para o algoritmo SHTC obtida ao considerar diferentes tamanhos de horizonte durante a fase de exploração.

A Figura 22 apresenta informação semelhante a apresenta na Figura 20, no entanto, considerando $\alpha = 0.3$ para o algoritmo SHTC. Ao observar a Figura 22a é possível perceber que essa escolha

de α faz com que o algoritmo SHTC apresente recompensa média superior a do algoritmo ε-guloso no longo prazo. Na Figura 22b se torna mais evidente esse fato.



(a) Recompensa média para os algoritmos testados ao longo das 10.000 rodadas, excluindo as 1.000 primeiras.

(b) Recompensa média obtida entre as rodadas 8.000 e 10.000.

Figura 22 – Comparação da recompensa média para o algoritmo SHTC obtida ao considerar diferentes tamanhos de horizonte durante a fase de exploração.

Cabe ressaltar que a presente dissertação se restringiu a adotar tamanhos de horizonte de no máximo 10.000 devido a limitações computacionais, entretanto, acredita-se que a adoção de horizontes maiores poderia ser benéfica tanto para os algoritmos de exploração pura como também para o SHTC. Ademais, por simplificação, não foi feita a exploração de outros parâmetros para os algoritmos ETC, ε-greedy e UCB.

Os testes realizados nessa seção foram realizados através do uso da linguagem de programação Julia, o código referente a cada um dos resultados apresentados pode ser encontrado no seguinte endereço: <https://github.com/bastosismael/dissertacao_multi_armed_bandits>.

7 CONCLUSÃO

Neste trabalho, foram estudados os algoritmos de multi-armed bandits como uma ferramenta para a solução de problemas de tomada de decisão sequencial. Tendo sido, ao final, apresentada uma aplicação prática desses algoritmos ao problema da precificação dinâmica, com o objetivo de explorar a adoção de algoritmos de bandits em contextos de incerteza na escolha de preços.

Com relação ao arcabouço teórico desenvolvido ao longo da dissertação, buscou-se apresentar os teoremas acompanhados de demonstrações detalhadas, com o intuito de facilitar a compreensão do raciocínio subjacente a cada resultado. Além disso, optou-se por estabelecer exclusivamente resultados teóricos no contexto de variáveis aleatórias sub-gaussianas, sendo a extensão desses resultados para outros tipos de ambientes um caminho promissor para investigações futuras.

Através dos experimentos numéricos realizados, observamos que ambientes nos quais as ações possuem recompensas com médias similares impactam negativamente a capacidade dos algoritmos que lidam com o dilema entre exploração e explotação em selecionar a ação ótima. Entretanto, a escolha de ações sub-ótimas não gerou impacto sobre a recompensa média, uma vez que as recompensas das ações são similares. No entanto, é possível perceber que, sob essas condições, o algoritmo Sequential Halving demonstrou uma robustez notável, selecionando consistentemente a ação ótima em todas as simulações realizadas, independentemente da similaridade entre as médias das recompensas.

Outro ponto relevante observado nos experimentos, foi o impacto da variância na recompensa de cada ação. Notamos que os algoritmos de exploração pura, não foram afetados pelo aumento da variância, enquanto os algoritmos que equilibram exploração e explotação, apresentaram uma redução na recompensa média conforme a variância aumentava, por outro lado, o arrependimento não se mostrou sensível a esse aumento.

No âmbito da precificação dinâmica, foi possível notar que o uso de algoritmos de bandit se mostrou uma maneira adequada para tratar o problema, onde , em um cenário de incerteza no qual não se tinha conhecimento acerca da função de demanda, foi possível obter resultados de recompensa médias que se aproximavam da recompensa ótima através do uso dos algoritmos delineados ao longo do presente texto.

Apesar de, em um contexto prático, ser mais comum que um vendedor esteja interessado em maximizar seu lucro, a presente dissertação se propôs a verificar a adoção de algoritmos de exploração pura para o problema da precificação dinâmica. Nesse sentido, foi possível verificar que tais algoritmos apresentaram recompensa média inferior quando comparados aos algoritmos que visam explorar o dilema entre exploração e explotação, o que é esperado, uma vez que no contexto da exploração pura se deseja identificar o melhor preço sem se importar com a recompensa acumulada ao longo das rodadas, entretanto, se mostram bastante promissores na identificação do preço ótimo. Ainda no contexto da exploração pura, foi possível notar que o algoritmo SHTC, proposto na presente dissertação como uma combinação dos algoritmos ETC e Sequential Halving, apresentou resultados promissores, superando os algoritmos no qual se baseia e até mesmo o algoritmo ε-greedy ao avaliar a recompensa

média acumulada no cenário da precificação dinâmica. Dessa forma, acredita-se que a combinação da exploração pura com a explotação seja um caminho promissor para futuras explorações, assim como um estudo mais aprofundado sobre a escolha do parâmetro α no algoritmo SHTC.

Cabe ressaltar que, ao analisar um cenário real de vendas, é possível identificar a existência de variáveis externas que impactam diretamente o comportamento do consumidor, como por exemplo a sazonalidade e o comportamento de possíveis competidores. Nesse sentido, enxerga-se como projeto futuro a realização de testes utilizando bandits com contexto (*contextual bandits*) onde é possível utilizar variáveis exógenas para auxiliar no funcionamento dos algoritmos aqui apresentados, ver, por exemplo, (SLIVKINS, 2011). Ademais, especificamente para o caso de presença de empresas competidoras, acredita-se que um caminho promissor seja a exploração do uso de multi armed bandits com múltiplos agentes (*multi-agent multi-armed bandits*), sendo possível modelar a interação entre o vendedor e seus competidores, ver, por exemplo, (AGARWAL; AGGARWAL; AZIZZADENESHELI, 2022).

REFERÊNCIAS

AGARWAL, M.; AGGARWAL, V.; AZIZZADENESHELI, K. Multi-agent multi-armed bandits with limited communication. *Journal of Machine Learning Research*, v. 23, n. 212, p. 1–24, 2022. Disponível em: <http://jmlr.org/papers/v23/21-138.html>.

AUDIBERT, J.-Y.; BUBECK, S. Best arm identification in multi-armed bandits. In: *COLT-23th Conference on learning theory-2010*. [S.l.: s.n.], 2010. p. 13–p.

BATHER, J.; CHERNOFF, H. Sequential decisions in the control of a spaceship. In: UNIVERSITY OF CALIFORNIA PRESS BERKELEY. *Fifth Berkeley Symposium on Mathematical Statistics and Probability*. [S.l.], 1967. v. 3, p. 181–207.

BOER, A. V. D. Dynamic pricing and learning: Historical origins, current research, and new directions. *Surveys in Operations Research and Management Science*, v. 20, n. 1, p. 1–18, 2015. ISSN 1876-7354. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1876735415000021>.

BOUNEFFOUF, D.; RISH, I.; AGGARWAL, C. Survey on applications of multi-armed and contextual bandits. In: *2020 IEEE Congress on Evolutionary Computation (CEC)*. [S.l.: s.n.], 2020. p. 1–8.

BUBECK, S.; CESA-BIANCHI, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, v. 5, n. 1, p. 1–122, 2012. ISSN 1935-8237. Disponível em: <http://dx.doi.org/10.1561/2200000024>.

BUBECK, S.; CESA-BIANCHI, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *CoRR*, abs/1204.5721, 2012. Disponível em: <http://arxiv.org/abs/1204.5721>.

BUBECK, S.; MUNOS, R.; STOLTZ, G. Pure exploration in multi-armed bandits problems. In: SPRINGER. *Algorithmic Learning Theory: 20th International Conference, ALT 2009, Porto, Portugal, October 3-5, 2009. Proceedings 20*. [S.l.], 2009. p. 23–37.

CESA-BIANCHI, N.; LUGOSI, G. *Prediction, Learning, and Games*. [S.l.]: Cambridge University Press, 2006.

COURNOT, A. Researches into the mathematical principles of the theory of wealth. In: *Forerunners of Realizable Values Accounting in Financial Reporting*. [S.l.]: Routledge, 1838.

LATTIMORE, T.; SZEPESVáRI, C. *Bandit Algorithms*. [S.l.]: Cambridge University Press, 2020.

LI, L. et al. *Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization*. 2018. Disponível em: <https://arxiv.org/abs/1603.06560>.

ROBBINS, H. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, American Mathematical Society, v. 58, n. 5, p. 527 – 535, 1952.

SLIVKINS, A. Contextual bandits with similarity information. In: JMLR WORKSHOP AND CONFERENCE PROCEEDINGS. *Proceedings of the 24th annual Conference On Learning Theory*. [S.l.], 2011. p. 679–702.

SLIVKINS, A. *Introduction to Multi-Armed Bandits*. 2024. Disponível em: <https://arxiv.org/abs/1904.07272>.

THOMPSON, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, v. 25, n. 3-4, p. 285–294, 12 1933. ISSN 0006-3444. Disponível em: <https://doi.org/10.1093/biomet/25.3-4.285>.

TROVO, F. et al. Improving multi-armed bandit algorithms for pricing. In: *16th European Conference on Multi-Agent Systems*. [S.l.: s.n.], 2018. p. 1–15.

WANG, P.-A.; TZENG, R.-C.; PROUTIERE, A. Best arm identification with fixed budget: A large deviation perspective. *Advances in Neural Information Processing Systems*, v. 36, p. 16804–16815, 2023.