VIII Encontro de Iniciação Científica e Tecnológica
VIII EnICT
ISSN: 2526-6772
IFSP – *Campus* Araraquara
19 e 20 de outubro de 2023

Análise do tempo esperado de entrada de uma palavra em um processo estocástico: comparação entre métodos de cálculo por meio de estudo de casos.

GUSTAVO MAGALHÃES[1], VITOR AMORIM[2]

[1] Graduando em Licenciatura em Matemática, IFSP *Campus* Araraquara, magalhaes.f@aluno.ifsp.edu.br.

[2] Doutor em Estatística, PIPGEs-USP-UFSCar, Docente no IFSP *Campus* Araraquara, vitoramorim@ifsp.edu.br

Área de conhecimento (Tabela CNPq): Teoria Geral e Processos Estocásticos – 1.02.01.02-5

RESUMO: Consideremos um processo estocástico $(X_m)_{m\geq 0}$ que assume valores em um conjunto finito de símbolos, chamado de alfabeto. Fixada uma sequência $A = (a_0, a_1, \cdots, a_{n-1})$ de símbolos do alfabeto, chamada de palavra, temos interesse no valor esperado da variável aleatória T_A, conhecida como tempo de entrada da palavra A. Ela é definida como a primeira coordenada do processo a partir da qual A aparece. Trabalhos anteriores mostraram que, para o caso de processos independentes e identicamente distribuídos (*i.i.d.*) com distribuição uniforme é possível obter uma fórmula exata para o tempo esperado de entrada $\mathrm{E}(T_A)$ da palavra. Para processos mais gerais, foram obtidas apenas aproximações de $\mathrm{E}(T_A)$, cuja precisão aumenta assintoticamente com o valor de n, e que dependem essencialmente da probabilidade de ocorrência da palavra e de suas características de periodicidade. Neste trabalho, fizemos a avaliação da eficiência dessas aproximações e a influência da periodicidade da palavra no tempo esperado de entrada. Como base de comparação, utilizamos a fórmula explícita do caso *i.i.d.* para obter o erro exato de aproximação, bem como foi analisado a influência da periodicidade nos resultados.

PALAVRAS-CHAVE: funções geradoras; periodicidade; poço de potencial; processos estocásticos; tempo de entrada.

INTRODUÇÃO

Em teoria das probabilidades, um processo estocástico em tempo discreto é um sequência $(X_m)_{m\geq 0}$ de variáveis aleatórias que assumem valores em um dado conjunto. Processos estocásticos possuem diversas aplicações em problemas de caráter teórico, científico e aplicações diretas, como ao modelar o movimento de uma partícula em um fluido ou no fenômeno de flutuação de preços no mercado financeiro.

O problema tratado neste trabalho está relacionado às variáveis aleatórias conhecidas como tempo de entrada e tempo de retorno, que, em essência, descrevem o tempo em que uma determinada sequência de símbolos aparece pela primeira vez ou retorna durante a realização de um processo estocástico. Investigadas há décadas por pesquisadores, a principal busca foi por fórmulas exatas ou aproximações para as suas distribuições de probabilidades e para os momentos dessas variáveis, conforme Abadi e Galves (2001).

O primeiro resultado quantitativo para o problema acima, foi encontrado por Kac (1947). Para uma classe bastante geral de processos estocásticos, ele encontrou uma fórmula exata e intuitiva para o tempo esperado de retorno. O chamado Lema de Kac afirma que o tempo esperado de retorno de uma palavra é o inverso da probabilidade de sua ocorrência. Naturalmente, pesquisadores da área começaram a busca por um resultado similar, mas desta vez relacionado ao tempo de entrada.

Essa questão obteve avanços recentemente com os trabalhos de Abadi e Vergne (2009), que obtiveram aproximações para os momentos do tempo esperado de retorno utilizando, além da probabilidade da palavra, um parâmetro intuitivo e de fácil obtenção chamado de poço de potencial, que depende das características de periodicidade (ou sobreposição) da palavra (cf. Abadi; Amorim; Gallo, 2021). Inspirado nesta ideia, Amorim

(2022) obteve, utilizando o poço de potencial, aproximações para os momentos dos tempos de entrada e retorno, considerando uma classe mais geral de processos.

Por outro lado, Flajolet e Sedgewick (2009) apresentam em seu livro um método combinatório, pautado em funções geradoras, que torna possível encontrar uma fórmula exata do tempo esperado de entrada. Porém, o método é válido apenas para processos *i.i.d.* com distribuição uniforme sobre o alfabeto, ou seja, para processos bem mais restritos do que os trabalhados apresentados pelos pesquisadores citados acima.

Neste trabalho, foram avaliadas, por meios de estudos de caso, a eficiência do método de aproximação utilizando do poço de potencial. Como base para comparação, utilizou-se da fórmula exata encontrada com o método combinatório de Flajolet e Sedgewick (2009).

FUNDAMENTAÇÃO TEÓRICA

Um processo estocástico a tempo discreto é uma sequência $(X_m)_{m\geq 0} = (X_0, X_1, \cdots, X_m, \cdots)$ de variáveis aleatórias definidas em um espaço de probabilidade $(\Omega, \mathcal{F}, \mathrm{P})$ e que assumem valores de um conjunto ao longo de um tempo contável. Neste trabalho, o processo $(X_m)_{m\geq 0}$ assume valores em um conjunto finito de m símbolos $\mathcal{A}$, chamado de alfabeto, e fixa-se uma sequência $A = (a_0, a_1, \cdots, a_{n-1}) \in \mathcal{A}^n$ de símbolos do alfabeto, chamada de palavra, que denotaremos também por a_0^{n-1}.

O primeiro resultado sobre tempos de entrada e retorno, chamado de Teorema da Recorrência de Poincaré (cf. Shields (1996)), afirma que, para uma classe bastante geral de processos estocásticos, os tempos de recorrência são quase certamente finitos, isto é $\mathrm{P}(T_A < \infty) = 1$ e $\mathrm{P}(T_A < \infty | A) = 1$..

Conforme já mencionado, o Lema de Kac traz o primeiro resultado quantitativo sobre o tema, que é bastante intuitivo: $\mathrm{E}_A(T_A) = \mathrm{P}(A)^{-1}$. Naturalmente, diversos pesquisadores buscararam resultados semelhantes para o tempo esperado de entrada $\mathrm{E}(T_A)$ bem como fórmulas para as distribuições de probabilidades dos tempos esperados de entrada e retorno. Um levantamento recente sobre os avanços obtidos nessas questões pode ser visto em Abadi, Amorim e Gallo (2021).

Na dificuldade de obter fórmulas exatas, diversos trabalhos estabeleceram a distribuição exponencial de probabilidades como uma boa aproximação para a distribuição dos tempos de entrada e retorno. Baseado no Lema de Kac, as primeiras tentativas usaram como parâmetro da exponencial o valor da probabilidade $\mathrm{P}(A)$. Todavia, Galves e Schmitt (1997) mostraram que era necessário adicionar um fator de correção $\theta(A)$, cujas primeiras tentativas de obtenção resultaram em um parâmetro de difícil compreensão intuitiva e cujo cálculo era impraticável.

Esses problemas obtiveram avanços com os esforços de Abadi e Vergne (2009), Abadi, Cardeño e Gallo (2015), Abadi, Amorim e Gallo (2021) e Amorim (2022), que desenvolveram e se aprofundaram no parâmetro $\rho(A)$, chamado de poço de potencial. Possuindo um apelo intuitivo e simples de se calcular, o parâmetro $\rho(A)$ é a probabilidade de um processo que começou com a palavra A ter seu retorno após o primeiro retorno possível $\tau(A)$, conhecido como período da palavra A. Formalmente, definimos:

$$\tau(A) := \min\left\{k \geq 1 : a_k^{n-1} = a_0^{n-k-1}\right\} \quad \text{e} \quad \rho(A) := \mathrm{P}_A(T_A > \tau(A)).$$

A notação utilizada acima para definir $\rho(A)$ significa a probabilidade do evento $\{T_A > \tau(A)\}$ dado que o processo iniciou com a palavra A. Já a notação a_k^{n-1} indica as últimas $n-k$ letras da palavra a_0^{n-1}.

O poço de potencial pode ser fisicamente interpretado como a "energia"necessária para que uma palavra escape do seu período.

O trabalho de Amorim (2022) mostrou que, para uma palavra suficientemente grande de uma classe geral de processos (chamada de ϕ-misturadores), tem-se $\mathrm{E}(T_A^k) \cong k!\,\{\rho(A)\mathrm{P}(A)\}^{-k}$.

Por outro lado, focando em resultados para processos mais simples, Flajolet e Sedgewick (2009) fornecem em seu livro sobre combinatória analítica, uma proposição que possibilita encontrar uma fórmula exata para o

cálculo do tempo esperado de entrada para processos *i.i.d.* com distribuição uniforme sobre o alfabeto. Dessa forma, podemos comparar os resultados deste caso restrito com as aproximações mais gerais. Isso pode ser feito tanto para o caso do tempo esperado de entrada, quando para a distribuição $\mathrm{P}(T_A > j)$, visto que sabemos ser esta última bem aproximada pela distribuição exponencial, cujo parâmetro é o inverso de $\mathrm{E}(T_A)$.

METODOLOGIA

A presente pesquisa pode ser classificada por um estudo exploratório teórico-experimental, pois visa familiarizar-se com o problema, aprimorando ideias e descobertas (GIL, 2002).

O trabalho envolveu um levantamento bibliográfico sobre as principais ferramentas utilizadas na teoria das probabilidades e processos estocásticos, bem como foram analisados exemplos para o auxílio da compreensão do conteúdo abordado. Realizou-se também a leitura e discussão, em reuniões semanais, de artigos e livros voltados para a teoria da recorrência de Poincaré e, em particular, para o problema em questão.

O desenvolvimento da pesquisa ocorreu por meio da exploração de exemplos e simulações, aplicando os resultados existentes, comparando as diferentes abordagens do problema e analisando as características particulares de cada caso, encaminhando para generalização dos resultados.

Os recursos utilizados foram artigos e livros disponíveis na biblioteca do IFSP - *Campus* Araraquara, bem como materiais de livre acesso na internet. Como ambiente de discussão e desenvolvimento do problema, foi utilizado o Laboratório de Ensino de Matemática (LEM) com seus diversos recursos, como projetor, mesa de trabalho, lousa e giz.

RESULTADOS E DISCUSSÃO

No que segue, utilizou-se das notações $(X_0, X_1, \cdots, X_{n-1}) = X_0^{n-1}$ para uma sequência de variáveis do processo, e, para uma palavra de tamanho n variável denotou-se por A_n. Para a probabilidade condicional, usou-se $\mathrm{P}(A|B) = \mathrm{P}_B(A)$.

Define-se a variável aleatória T_A, chamada de tempo de entrada,

$$T_A := \min\left\{k \geq 1 : X_k^{k+n-1} = a_0^{n-1}\right\}.$$

Observa-se que o tempo de entrada não é definido a partir da primeira coordenada do processo, pois, quando isto ocorre, temos interesse no tempo de retorno da palavra.

Para encontrar o tempo esperado de entrada $\mathrm{E}(T_A)$, define-se uma outra variável aleatória associada ao tempo de entrada. A variável Z_A, chamada de tempo de entrada completa, é definida como

$$Z_A := \min\left\{k \geq n : X_{k-n+1}^{k} = a_0^{n-1}\right\}.$$

Definimos esta variável por estar relacionada ao método combinatório de obtenção do tempo esperado de entrada, que utilizaremos como base de comparação. Note que $Z_A = T_A + n - 1$. Assim, pela linearidade da esperança, temos $\mathrm{E}(T_A) = \mathrm{E}(Z_A) - n + 1$.

Para as análises que faremos a seguir, consideramos um processo estocástico cuja as variáveis são independentes e identicamente distribuídas (*i.i.d.*), isto é,

$$\mathrm{P}(X_n = a|X_\ell = b) = \mathrm{P}(X_n = a) \quad \text{e} \quad \mathrm{P}(X_n = a) = \mathrm{P}(X_0 = a), \quad \forall\, a, b \in \mathcal{A}, \forall \ell, n \geq 0$$

Consiraremos ainda que as variáveis do processo têm distribuição uniforme sobre o alfabeto, ou seja, $P(X_j = a) = m^{-1}$ para todo $j \geq 0$ e $a \in \mathcal{A}$, onde $\#\mathcal{A} = m$.

Agora, com o objetivo de determinar o tempo esperado de entrada de uma palavra no processo pelo método combinatório, denotamos por s_j o número de configurações possíveis nas j primeiras coordenadas do processo

que não contém a palavra A, onde, por convenção, $s_0 = 1$. Sua função geradora é definida como

$$S(x) := \sum_{j=0}^{\infty} s_j x^j.$$

Note que o evento $\{Z_A > j\}$ ocorre quando a sequência $(X_1, \cdots, X_j)$ não apresenta a palavra A, independente do valor da variável inicial X_0. Logo, por se tratar de um processo independente, para todo $j \geq 0$:

$$P(Z_A > j) = \frac{ms_j}{m^{j+1}} = \frac{s_j}{m^j} \Longrightarrow E(Z_A) = \sum_{j=0}^{\infty} P(Z_A > j) = \sum_{j=0}^{\infty} \frac{s_j}{m^j} = S\left(\frac{1}{m}\right).$$

Por outro lado, a proposição I.4 de Flajolet e Sedgewick (2009), mostra que

$$S(x) = \frac{c(x)}{x^n + (1 - mx)c(x)},$$

onde $c(x)$ é o polinômio de autocorrelação da palavra A, definido por $c(x) := \sum_{j=0}^{n-1} c_j x^j$ e

$$c_j = \begin{cases} 1, & \text{se } a_j^{n-1} = a_0^{n-j-1} \\ 0, & \text{caso contrário} \end{cases}.$$

Assim, para um processo $(X_m)_{m\geq 0}$ *i.i.d.* com distribuição uniforme sobre o alfabeto, encontramos uma fórmula explícita para o tempo esperado de entrada T_A:

$$\mathrm{E}(Z_A) = S\left(\frac{1}{m}\right) = m^n c(1/m) \Longrightarrow \mathrm{E}(T_A) = m^n c(1/m) - n + 1.$$

Note que, como mencionado anteriormente, o método de aproximação de Amorim (2022) para processos mais gerais utiliza o parâmetro $\rho(A)$, que, assim como $c(1/m)$, depende das características de sobreposição da palavra. Ou seja, em ambos os métodos a periodicidade da palavra tem um papel fundamental no tempo de entrada.

Assim, apresentaremos alguns estudos de casos nos exemplos a seguir com os objetivos de: 1) verificar a eficiência dos métodos de aproximação; 2) Verificar a relação entre os métodos de medição das características de sobreposição.

Começamos obtendo os valores de $\mathrm{E}(T_A)$ pelo método combinatório.

Exemplo 1. *A palavra $A = abracadabra$ sobre o alfabeto $\mathcal{A} = \{a, \cdots, z\}$.*

Neste caso, $n = 11$, $m = 26$ e o polinômio de autocorrelação de A é $c(x) = 1 + x^7 + x^{10}$. Então,

$$\mathrm{E}(T_A) = 26^{11} c(1/26) - 10 \Longrightarrow \mathrm{E}(T_A) = 26^{11}(1 + \frac{1}{26^7} + \frac{1}{26^{10}}) - 10 \Longrightarrow \mathrm{E}(T_A) = 3.670.344.487.444.768$$

Exemplo 2. *As palavras $A_n = 11\cdots111$ e $B_n = 100\cdots00$ sobre o alfabeto binário $\mathcal{A} = \{0, 1\}$.*

Neste caso, $m = 2$. Para a palavra A_n, o polinômio de autocorrelação é $c(x) = 1 + x + x^2 + \cdots + x^{n-1} = \frac{1 - x^n}{1 - x}$. Para a palavra B_n, o polinômio de autorrelação é $c(x) = 1$. Então,

$$\mathrm{E}(T_{A_n}) = 2^n \left(\frac{1 - (1/2)^n}{1 - (1/2)}\right) - n + 1 \Longrightarrow \mathrm{E}(T_{A_n}) = 2^{n+1} - n - 1$$

e

$$\mathrm{E}(T_{B_n}) = 2^n - n + 1.$$

Por outro lado, o corolário 2.4.2 de Amorim (2022) garante que

$$\lim_{n\to\infty} \rho(A)\mathrm{P}(A)\mathrm{E}(T_A) = 1 \Longrightarrow \mathrm{E}(T_A) \cong \frac{1}{\rho(A)\mathrm{P}(A)}$$

.

Vale notar que $\rho(A) = \mathrm{P}_A(T_A > \tau(A)) = 1 - \mathrm{P}_A(T_A = \tau(A))$. Assim, para o caso *i.i.d.* com distribuição uniforme, um cálculo direto nos fornece $\rho(A) = 1 - m^{-\tau(A)}$. Logo,

$$\mathrm{E}(T_A) \cong \frac{1}{(1 - m^{-\tau(A)})(m^{-n})} \Longrightarrow \mathrm{E}(T_A) \cong \frac{m^{n+\tau(A)}}{m^{\tau(A)} - 1}.$$

A partir deste ponto, realizamos a comparação entre os dois métodos. Denotamos por $\mathrm{E}_1(T_A)$ o tempo esperado de entrada obtido pelo método exato para o caso *i.i.d.* uniforme e $\mathrm{E}_2(T_A)$ o tempo esperado de entrada obtido pelo método do poço de potencial.

Exemplo 3. *A palavra $A = abracadabra$ sobre o alfabeto $\mathcal{A} = \{a, \cdots, z\}$.*

Neste caso, $\mathrm{P}(A) = \dfrac{1}{26^{11}}$ e $\tau(A) = 7$. Logo,

$$\mathrm{E}_2(T_A) \cong \frac{26^{18}}{26^7 - 1} \cong 3.670.344.487.444.752.$$

Dessa forma, obtemos $\dfrac{\mathrm{E}_2(T_A)}{\mathrm{E}_1(T_A)} \cong 1$ e $\mathrm{E}_1(T_A) - \mathrm{E}_2(T_A) = 15$, o que implica uma diferença percentual da ordem de 10^{-15}.

Exemplo 4. *As palavras $A_n = 11\cdots111$ e $B_n = 100\cdots00$ sobre o alfabeto binário $\mathcal{A} = \{0, 1\}$.*

Inicialmente, temos $\mathrm{P}(A_n) = \dfrac{1}{2^n} = \mathrm{P}(B_n)$. Além disso, $\tau(A_n) = 1$ e $\tau(B_n) = n$. Logo,

$$\mathrm{E}_2(T_{A_n}) \cong 2^{n+1} \quad \text{e} \quad \mathrm{E}_2(T_{B_n}) \cong \frac{2^{2n}}{2^n - 1}.$$

Em termos de ordem de grandeza, as diferenças entre os métodos são desprezíveis quando $n \longrightarrow \infty$, pois

$$\lim_{n\to\infty} \frac{\mathrm{E}_2(T_{A_n})}{\mathrm{E}_1(T_{A_n})} = \lim_{n\to\infty} \frac{2^{n+1}}{2^{n+1} - n - 1} = 1$$

e

$$\lim_{n\to\infty} \frac{\mathrm{E}_2(T_{B_n})}{\mathrm{E}_1(T_{B_n})} = \lim_{n\to\infty} \frac{\frac{2^{2n}}{2^n-1}}{2^n - n + 1} = \lim_{n\to\infty} \frac{2^{2n}}{2^{2n} - n2^n + n - 1} = 1.$$

Note ainda que a diferença percentual converge para zero nos dois casos, pois

$$\lim_{n\to\infty} \frac{|\mathrm{E}_2(T_{A_n}) - \mathrm{E}_1(T_{A_n})|}{\mathrm{E}_1(T_{A_n})} = \lim_{n\to\infty} \left| \frac{\mathrm{E}_2(T_{A_n})}{\mathrm{E}_1(T_{A_n})} - 1 \right| = 0,$$

o que também vale para B_n.

Analisado o valor esperado do tempo de entrada pode-se investigar a respeito da distribuição de probabilidades $\mathrm{P}(T_A > j)$. Como mostram os diversos trabalhos citados em (Abadi; Amorim; Gallo, 2021), a distribuição

exponencial de probabilidades com parâmetro λ pode ser utilizada com boa precisão como aproximação para $P(T_A > j)$. Pelas propriedades da distribuição exponencial, cujo valor esperado é o inverso de seu parâmetro, uma boa aproximação deve satisfazer $\lambda = \dfrac{1}{\mathrm{E}(T_A)}$.

Assim obtemos,

$$\mathrm{P}(T_A > j) \cong \exp(-j\lambda) \Longrightarrow \mathrm{P}(T_A > j) \cong \exp\left(\frac{-j}{\mathrm{E}(T_A)}\right)$$

.

Para o caso do processo *i.i.d.* uniforme com o método exato e o método do poço de potencial, temos, respectivamente

$$\mathrm{P}_1(T_A > j) \cong \exp\left(\frac{-j}{m^n c(1/m) - n + 1}\right) \quad \text{e} \quad \mathrm{P}_2(T_A > j) \cong \exp\left(\frac{-j(m^n - 1)}{m^{n+\tau(A)}}\right)$$

CONCLUSÕES

Através das análises e comparações feitas, o poço de potencial fornece um método de aproximação do tempo esperado de entrada muito preciso para processos *i.i.d.* uniformes, além de poder ser aplicado a problemas mais gerais, em comparação ao método exato encontrado. Observa-se também que, a periodicidade (ou autocorrelação) é um parâmetro com um papel fundamental na distribuição do tempo de entrada.

REFERÊNCIAS

ABADI, M.; AMORIM, V.; GALLO, S. Potential well in poincaré recurrence. **Entropy**, MDPI, v. 23, n. 3, p. 379, 2021.

ABADI, M.; CARDEÑO, L.; GALLO, S. Potential well spectrum and hitting time in renewal processes. **Journal of Statistical Physics**, Springer, v. 159, n. 5, p. 1087–1106, 2015.

ABADI, M.; GALVES, A. Inequalities for the occurrence times of rare events in mixing processes. the state of the art. **Markov Process. Related Fields**, v. 7, n. 1, p. 97–112, 2001.

ABADI, M.; VERGNE, N. Sharp error terms for return time statistics under mixing conditions. **Journal of Theoretical Probability**, Springer, v. 22, p. 18–37, 2009.

AMORIM, V. **Poincaré recurrence times in stochastic mixing processes**. Tese (Doutorado) — PIPGEs-USP-UFSCar, 2022.

FLAJOLET, P.; SEDGEWICK, R. **Analytic combinatorics**. [S.l.]: cambridge University press, 2009.

GALVES, A.; SCHMITT, B. Inequalities for hitting times in mixing dynamical systems. **Random and Computational Dynamics**, New York, NY: Marcel Dekker, Inc., c1992-c1997., v. 5, n. 4, p. 337–348, 1997.

GIL, A. C. **Como elaborar projetos de pesquisa**. [S.l.]: Atlas São Paulo, 2002. v. 4.

KAC, M. On the notion of recurrence in discrete stochastic processes. 1947.

SHIELDS, P. C. **The ergodic theory of discrete sample paths**. [S.l.]: American Mathematical Soc., 1996. v. 13.