UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
DEPARTAMENTO DE MATEMÁTICA

GABRIELLE BELLOBRAYDIC

PROCESSOS DE DECISÕES DE MARKOV

SÃO CARLOS
2023

GABRIELLE BELLOBRAYDIC

PROCESSOS DE DECISÕES DE MARKOV

Monografia apresentada ao Curso de Licenciatura em Matemática da Universidade Federal de São Carlos.

Orientador: Prof. Dr. Rafael Kapp

SÃO CARLOS
2023

LISTA DE FIGURAS

SUMÁRIO

INTRODUÇÃO

Os Modelos de Decisão de Markov Sequencial, também conhecidos como Processos de Decisão de Markov (MDPs), constituem uma estrutura matemática poderosa para modelar situações em que um agente toma decisões sequenciais em um ambiente estocástico. Essa abordagem é particularmente útil e possui ampla utilidade em diversas áreas práticas, incluindo marketing e gerenciamento de estoques. No domínio do marketing, os MDPs podem ser empregados para otimizar estratégias de publicidade online, considerando a dinâmica das preferências do consumidor e a incerteza nas respostas aos anúncios. Isso permite que as empresas ajustem suas campanhas publicitárias de forma adaptativa, visando maximizar o retorno sobre o investimento.

No controle de estoque, os MDPs são valiosos para gerenciar a reposição de produtos de maneira eficiente. Ao modelar a incerteza na demanda e nas entregas, as empresas podem tomar decisões sequenciais sobre quando e quanto reabastecer seus estoques. Isso não apenas minimiza os custos associados ao armazenamento excessivo ou à falta de produtos, mas também contribui para a melhoria da experiência do cliente ao garantir a disponibilidade adequada dos itens desejados.

Esses exemplos ilustram como os MDPs sequenciais desempenham um papel crucial na tomada de decisões estratégicas em ambientes dinâmicos, onde a incerteza e a dependência temporal são fatores-chave a serem considerados. A flexibilidade dessa abordagem a torna uma ferramenta poderosa para enfrentar desafios práticos em uma variedade de setores.

No primeiro capítulo, adentraremos ao fascinante mundo dos conceitos fundamentais de probabilidade, proporcionando ao leitor uma base sólida para a imersão nos intricados domínios dos Processos de Markov e Cadeias de Markov. Ao explorar esses alicerces probabilísticos, buscamos não apenas fornecer uma revisão esclarecedora, mas também preparar o terreno para uma compreensão mais profunda dos temas que serão minuciosamente investigados ao longo desta pesquisa.

No segundo capítulo desta pesquisa, dedicaremos nossa atenção à análise abrangente e profunda das Cadeias de Markov, uma temática fundamental que se configura como elemento-chave na compreensão detalhada dos Processos de Markov, foco central deste trabalho. A compreensão antecipada da teoria probabilística é crucial, uma vez que constitui o alicerce não apenas das Cadeias de Markov, mas também dos Processos de Markov em sua totalidade. Essa abordagem sequencial é concebida para facilitar uma assimilação gradual de conceitos, promovendo uma progressão fluida de ideias e percepções.

Nos capítulos subsequentes desta pesquisa, nos dedicaremos a uma incursão mais profunda nos componentes fundamentais dos Processos de Markov, proporcionando uma análise mais minuciosa e especializada. Diferenciando-se da homogeneidade presente nas Cadeias de Markov, exploraremos as características não homogêneas dos Processos de Markov, onde as dinâmicas temporais, as transições de estado e a contabilização de recompensas não são constantes. Essa distinção crucial confere aos Processos de Markov uma versatilidade e adaptabilidade particular, tornando-os mais adequados para modelar uma gama diversificada de fenômenos.

Adicionalmente, nossa abordagem será enriquecida por exemplos teóricos e práticos do cotidiano, proporcionando ao leitor uma compreensão mais acessível da teoria em questão. Através de instâncias concretas e aplicadas, buscamos facilitar a assimilação dos conceitos abordados, estabelecendo uma ponte entre a teoria e sua aplicação prática.

Além disso, exploraremos e compararemos estados e conjuntos absorventes em Cadeias de Markov e MDPs. Abordaremos questões fundamentais como o problema de parada, considerando horizontes finitos, infinitos e extensos, entre outros tópicos pertinentes. Este enfoque prático permitirá ao leitor não apenas compreender os conceitos teóricos, mas também visualizar e contextualizar esses princípios em situações do dia a dia, aprimorando assim sua compreensão global do assunto previsto como foco do trabalho.

1 PROBABILIDADE

1.1 DEFINIÇÕES BÁSICAS DE PROBABILIDADE

A probabilidade é uma medida matemática que nos ajuda a quantificar a incerteza associada a eventos aleatórios. Ela está intrinsecamente relacionada à chance de ocorrência de resultados específicos em um experimento aleatório. A probabilidade é expressa numericamente em uma escala que varia entre 0 e 1, ou, em forma percentual, entre 0% e 100%, onde os valores mais próximos de 1 indicam uma maior probabilidade de ocorrência do evento, enquanto os valores mais próximos de 0 indicam uma probabilidade menor.

Espaço Amostral (Ω): O espaço amostral, representado pelo símbolo Ω, é uma estrutura fundamental na teoria da probabilidade. Ele é definido como o conjunto que compreende todos os possíveis resultados de um experimento aleatório. Em outras palavras, é a coleção completa de todas as observações ou eventos que podem ocorrer em um dado contexto probabilístico.

Exemplo 1.1. Considere o experimento de lançar uma moeda. O espaço amostral Ω nesse caso consiste nos dois resultados possíveis: "cara"(C) e "coroa"(K).

Exemplo 1.2. Para ilustrar ainda mais, suponha que estejamos interessados no espaço amostral ao lançar um dado padrão de seis lados. Nesse cenário, o espaço amostral Ω seria representado por $\{1, 2, 3, 4, 5, 6\}$, abrangendo todos os resultados individuais que podem ocorrer ao observar o resultado do lançamento do dado.

Evento (E): Um evento, denotado pelo símbolo E, é um subconjunto do espaço amostral (Ω) e descreve um resultado específico ou uma coleção de resultados em um experimento aleatório. Em outras palavras, um evento é uma ocorrência particular dentro do conjunto total de possíveis resultados.

Exemplo 1.3. Ao lançar uma moeda, “obter cara” é um evento, representado como um subconjunto de Ω.

Exemplo 1.4. Para ilustrar mais amplamente, considere o experimento de lançar um dado de seis lados. O evento E pode ser “obter um número par”, que é um subconjunto de Ω. Matematicamente, esse evento seria representado por $E = \{2, 4, 6\}$, englobando os resultados específicos de obter os números pares 2, 4 e 6 no lançamento do dado. Essa abordagem permite uma análise mais detalhada e específica das ocorrências desejadas dentro do contexto do espaço amostral.

σ-Álgebra: A σ-álgebra, é uma estrutura matemática fundamental na teoria da medida e probabilidade. Ela é composta por um conjunto de subconjuntos do espaço amostral (Ω) que satisfaz certas propriedades específicas.

Formalmente, é uma coleção de subconjuntos de Ω que inclui o próprio Ω, é fechada sob a operação de complemento (ou complementar) e é fechada sob a operação de união contável. Essas propriedades asseguram que a σ-álgebra seja uma estrutura estável e consistente para análises probabilísticas.

Adicionalmente, a operação de complemento de um evento A em relação à σ-álgebra é denotada por A^c ou $\bar{A}$ e consiste em todos os elementos de Ω que não pertencem a A. Vale ressaltar que $A^c = 1 - A$, onde 1 representa o conjunto completo Ω.

Exemplo 1.5. Suponha que estejamos trabalhando com o espaço amostral Ω referente ao lançamento de um dado justo de seis lados. A σ-álgebra associada, poderia ser construída incluindo subconjuntos como o conjunto vazio, o conjunto contendo todos os resultados possíveis e outros subconjuntos formados por combinações de resultados específicos, como $\{1, 3, 5\}$. Esta σ-álgebra proporciona a estrutura necessária para uma análise mais refinada e abrangente das probabilidades associadas aos eventos no contexto do espaço amostral, incluindo a consideração dos eventos complementares.

Probabilidade (P): A probabilidade é uma medida numérica que quantifica a chance de ocorrência de um evento em um experimento aleatório. Essa medida varia de 0, indicando que o evento quase nunca ocorrerá, a 1, indicando que o evento quase sempre ocorrerá. A probabilidade de um evento específico, denotado por $P(E)$, é expressa como um número entre 0 e 1.

Além disso, a probabilidade possui as seguintes propriedades fundamentais:

Não-Negatividade: Para qualquer evento E, a probabilidade $P(E)$ é sempre maior ou igual a zero: $P(E) \geq 0$.

Normalização: A probabilidade do espaço amostral completo é igual a 1: $P(\Omega) = 1$.

Aditividade: Para eventos mutuamente exclusivos (eventos que não podem ocorrer simultaneamente), a probabilidade da união desses eventos é igual à soma de suas probabilidades individuais: $P(E_1 \cup E_2) = P(E_1) + P(E_2)$.

Exemplo 1.6. Para ilustrar, considere a probabilidade de obter um número ímpar ao lançar um dado justo. A probabilidade do evento "obter um número ímpar"é calculada como P(Ímpar) =

$\frac{3}{6} = \frac{1}{2}$. Essa probabilidade reflete a razão entre o número de resultados favoráveis (ímpares) e o número total de resultados possíveis no espaço amostral do dado

Regra da Soma (OU): A Regra da Soma, ou Regra da União, estabelece que a probabilidade de ocorrer pelo menos um de dois eventos exclusivos é a soma das probabilidades de cada evento individual. Se A e B são eventos mutuamente exclusivos, então $P(A \text{ ou } B) = P(A) + P(B)$.

Além disso, a Regra da Soma respeita as propriedades fundamentais da probabilidade, incluindo a não-negatividade, normalização e aditividade.

Exemplo 1.7. Consideremos os eventos “obter um número par” ou “obter um número ímpar” ao lançar um dado justo. A probabilidade de obter pelo menos um desses eventos é dada por $P(\text{Par ou Ímpar}) = P(\text{Par}) + P(\text{Ímpar}) = \frac{1}{2} + \frac{1}{2} = 1$. Isso reflete a certeza de que, ao lançar o dado, o resultado será necessariamente par ou ímpar, pois esses eventos são complementares e cobrem todo o espaço amostral.

Regra do Produto (E): A Regra do Produto estabelece que a probabilidade de dois eventos independentes A e B ocorrerem juntos é o produto de suas probabilidades individuais. Se A e B são eventos independentes, então $P(A \text{ e } B) = P(A) \cdot P(B)$.

Esta regra é aplicável quando a ocorrência de um evento não afeta a probabilidade de ocorrência do outro.

Exemplo 1.8. Considere a probabilidade de obter cara em uma moeda justa e rolar um número ímpar em um dado justo. A probabilidade de “Cara e Ímpar” é dada por $P(\text{Cara e Ímpar}) = P(\text{Cara}) \cdot P(\text{Ímpar}) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. Isso reflete o fato de que as duas ocorrências são independentes uma da outra, e a probabilidade conjunta é calculada multiplicando as probabilidades individuais.

Probabilidade Condicional: A probabilidade condicional de um evento A ocorrer, dado que outro evento B ocorreu, é denotada por $P(A|B)$. Formalmente, ela é definida como:

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

onde $P(B) \neq 0$. Isso representa a probabilidade de A, considerando que sabemos que B já ocorreu.

Exemplo 1.9. A probabilidade de obter um número par, dado que já obtivemos cara em uma moeda justa, é $P(\text{Par} \mid \text{Cara}) = \frac{1}{2}$.

1.2 LEI TOTAL DA PROBABILIDADE

Um conceito fundamental na teoria da probabilidade é a Lei da Probabilidade Total, que permite calcular a probabilidade de um evento como a soma das probabilidades ponderadas de sua ocorrência, sendo expressa da seguinte forma:

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i),$$

com A um evento e $B_1, B_2, \ldots, B_n$ um conjunto de eventos que não se sobrepõem e juntos cobrem todos os resultados possíveis. Isso significa que a probabilidade do evento A é a soma das probabilidades de A ocorrem em cada uma das possibilidades $B_1, B_2, \ldots, B_n$ ponderadas pelas probabilidades de ocorrência desses cenários.

Prova: Como $B_1, B_2, \ldots, B_n$ formam uma partição*, então podemos escrever A da seguinte forma:

$$P(A) = P(A \cap B_1) \cup P(A \cap B_2) \cup P(A \cap B_3) \cup \cdots \cup P(A \cap B_n).$$

***Partição de um Espaço Amostral** é o termo utilizado para descrever uma situação em que os eventos $B_1, B_2, \ldots, B_n$ formam uma expressão do espaço amostral (Ω), obedecendo às seguintes condições:

a) Nenhum evento pode ser igual ao conjunto vazio;

b) Os eventos são disjuntos;

c) A união de todos os eventos totaliza o espaço amostral.

Temos também por Tijms, Axioma 3, que para qualquer sequência de eventos mutuamente exclusivos $A_1, A_2, \ldots, A_n$ a probabilidade da união desses eventos é dada por:

$$P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i).$$

Sabendo que $B_1, B_2, \ldots, B_n$ é disjunto, pelo Axioma 3 de Kolmogorov:

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3) + \cdots + P(A \cap B_n).$$

Finalizando com o **Teorema do produto**: Sejam A e B dois eventos quaisquer. A probabilidade de ambos os eventos A e B ocorrerem é dada pelo produto das probabilidades individuais de A e B, condicionada à ocorrência de A:

$$P(A \cap B) = P(A|B).P(B),$$

onde, $P(B)$ é a probabilidade do evento B ocorrer e $P(A|B)$ é a probabilidade do evento A ocorrer, dado que o evento B já ocorreu.

Podemos escrever a $P(A)$ como

1.3 VARIÁVEL ALEATÓRIA

Uma variável aleatória é uma função que associa um valor numérico a cada resultado possível de um experimento aleatório. Formalmente, seja Ω o espaço amostral de um experimento aleatório e X uma variável aleatória definida sobre Ω. A função $X : \Omega \to \mathbb{R}$ atribui a cada resultado $\omega \in \Omega$ um valor real $X(\omega)$.

Existem dois tipos principais de variáveis aleatórias: discretas e contínuas. Uma variável aleatória discreta assume valores isolados, enquanto uma variável aleatória contínua pode assumir qualquer valor em um intervalo contínuo.

Exemplo 1.10. Considere um experimento aleatório de lançar um dado justo. A variável aleatória X pode ser definida como o número que aparece no dado após o lançamento. Neste caso, o espaço amostral Ω é $\{1, 2, 3, 4, 5, 6\}$ e a variável aleatória X associa cada número a si mesmo.

Seja $p_X(x)$ a função de massa de probabilidade de X, dada por:

$$p_X(x) = \frac{1}{6}, \quad \text{para } x \in \{1, 2, 3, 4, 5, 6\}.$$

Essa função atribui probabilidades iguais a cada resultado possível, refletindo a justiça do dado. A variável aleatória X neste exemplo é discreta, pois assume valores distintos em $\{1, 2, 3, 4, 5, 6\}$.

1.4 ESPERANÇA VARIÁVEL DISCRETA

Esperança, no contexto de probabilidade e estatística, é um conceito que representa o valor médio ou média ponderada de uma variável aleatória, representada por

$$E(X) = \sum [x \cdot P(X = x)],$$

para todos os valores x possíveis de X, onde

a) $E(X)$: Esperança da variável aleatória X.

b) x: Representa cada valor possível da variável aleatória X.

c) $P(X = x)$: É a probabilidade de uma variável aleatória X assumir o valor x.

Exemplo 1.11. Considere uma variável aleatória X que representa o número de vezes que uma moeda honesta é lançada até obter cara. A distribuição de probabilidade de X é dada por:

$$P(X = k) = \frac{1}{2^k}, \quad \text{para } k = 1, 2, 3, \ldots .$$

A esperança de X pode ser calculada usando a fórmula:

$$E(X) = \sum_{k=1}^{\infty} [k \cdot P(X = k)].$$

Substituindo a distribuição de probabilidade, obtemos:

$$E(X) = \sum_{k=1}^{\infty} \frac{k}{2^k}.$$

Este é um exemplo prático da aplicação da fórmula de esperança para uma variável aleatória discreta.

1.5 TEOREMAS SOBRE ESPERANÇA

Teorema 1: (Constante Multiplicativa) Se c é uma constante, então a esperança de c multiplicado por uma variável aleatória X é igual a c multiplicado pela esperança de X:

$$E(cX) = cE(X);$$

Teorema 2: (Esperança da Soma de Variáveis) Se X e Y são variáveis aleatórias independentes, então a esperança da soma das variáveis é a soma das esperanças individuais:

$$E(X + Y) = E(X) + E(Y);$$

Teorema 3: (Linearidade e Monotonicidade) A esperança é linear, o que significa que a esperança da soma (ou diferença) de variáveis aleatórias é igual à soma (ou diferença) das esperanças das variáveis individuais:

$$E(aX + bY) = aE(X) + bE(Y)$$

Além disso, a esperança preserva a ordem monotônica, ou seja, se $X \leq Y$ quase certamente (quase sempre), então $E(X) \leq E(Y)$.

Teorema 4: (Esperança de uma Constante) A esperança de uma constante c é igual à própria constante:

$$E(c) = c;$$

Teorema 5: (Esperança de Variável Constante) Se X é uma variável aleatória que sempre assume um valor constante c, então a esperança de X é igual a c:

$$E(X) = c;$$

Teorema 6: (Esperança do Produto de Variáveis Independentes) Apenas se X e Y são independentes:

$$E(XY) = E(X) \cdot E(Y).$$

Esses teoremas foram extraídos do livro de Bremaud, referenciados como Teorema 1.3.13 na página 18 e Teorema 1.3.25 na página 21.

1.6 LEI BINOMIAL

A Lei Binomial é uma ferramenta poderosa na teoria das probabilidades e estatística, frequentemente utilizada para modelar experimentos com resultados dicotômicos. Nesta seção, exploraremos a definição da Lei Binomial, suas aplicações práticas e apresentaremos uma breve demonstração de seus principais resultados.

Ela descreve a probabilidade de um número fixo de sucessos em um número fixo de tentativas independentes, onde cada tentativa tem apenas dois resultados possíveis: sucesso ou fracasso. Se denotarmos por X o número de sucessos em n tentativas independentes, a distribuição binomial é dada por:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

onde:

- $\binom{n}{k}$ é o coeficiente binomial;
- p é a probabilidade de sucesso em uma tentativa;
- k é o número de sucessos desejados;
- n é o número total de tentativas.

A Lei Binomial é aplicada em diversas situações do cotidiano, como:

1. **Lançamento de Moedas:** Ao lançar uma moeda várias vezes, a probabilidade de obter um número fixo de caras em um número fixo de lançamentos segue uma distribuição binomial.
2. **Testes de Múltipla Escolha:** Em testes de múltipla escolha, onde cada questão tem duas opções (certa ou errada), a distribuição de acertos pode ser modelada pela Lei Binomial.
3. **Contagem de Sucessos em Experimentos Repetidos:** Se um experimento é repetido várias vezes, como testar a eficácia de um medicamento em diferentes pacientes, a Lei Binomial pode ser utilizada para modelar a probabilidade de um número específico de sucessos.

Exemplo 1.12. Exemplo de Distribuição Binomial

Considere um experimento no qual uma moeda justa é lançada cinco vezes. Cada vez que a moeda é lançada, ela pode resultar em cara (C) com probabilidade $p = \frac{1}{2}$ ou coroa (K) com a mesma probabilidade.

A variável aleatória X representa o número de caras em cinco lançamentos. Podemos modelar X como uma variável aleatória binomial, denotada por $X \sim B(5, \frac{1}{2})$.

A distribuição de probabilidade de X é dada por:

$$P(X = k) = \binom{5}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{5-k}$$

Cálculo da Probabilidade:

Para $k = 3$, temos:

$$P(X = 3) = \binom{5}{3} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^2$$

Utilizando a fórmula da binomial, onde $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, temos:

$$P(X = 3) = \frac{5!}{3!(5-3)!} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^2$$

$$P(X = 3) = \frac{120}{6} \cdot \frac{1}{8} \cdot \frac{1}{4}$$

$$P(X = 3) = 10 \cdot \frac{1}{8} \cdot \frac{1}{4}$$

$$P(X = 3) = \frac{10}{32} = \frac{5}{16}$$

Neste exemplo, a distribuição binomial é utilizada para calcular a probabilidade de obter exatamente três caras em cinco lançamentos de uma moeda justa.

1.7 O PROCESSO DE POISSON

Para um fundo prático, tome eventos aleatórios, como desintegração de partículas, chamadas telefônicas recebidas e quebras de cromossomos sob irradiação nociva. Todas as ocorrências são assumidas como sendo do mesmo tipo e estamos preocupados com o número total $Z(t)$ de ocorrências em um intervalo de tempo arbitrário de comprimento t, $t > 0$.

Admitimos que as forças e influências que governam o processo permanecem constantes de modo que a probabilidade de qualquer evento particular é a mesma para todos os intervalos de duração t, e é independente do desenvolvimento passado do evento.

Cada ocorrência é representada por um ponto no eixo do tempo e, portanto, estamos realmente preocupados com certas colocações aleatórias de pontos em uma linha. Em termos matemáticos, isso significa que o processo é um processo de Markov homogêneo no sentido descrito no parágrafo anterior.

Suponha $Z(0) = 0$, isto é, $P[Z(0) = 0] = 1$, e definimos

$$P_n(t) = P[Z(t) = n], \tag{1.1}$$

em que $n \in \mathbb{N}$ e quaisquer $t > 0$.

A definição apresentada na equação (1.1) representa a probabilidade do número de ocorrências possíveis. Dessa forma, temos que Processo de Poisson é um processo de contagem em que as ocorrências do evento não são simultâneas e uma vez que o processo está no estado n, a única transição possível é para o estado $n + 1$.

Além disso, considere um intervalo de tempo de comprimento 1. Este intervalo será particionado em M subintervalos de comprimento $h = 1/M$, em que M é suficientemente grande para que em cada subintervalo haja no máximo uma ocorrência do evento. Então, a probabilidade de uma ocorrência do evento em qualquer subintervalo é igual a $1 - P_0(h)$ e, supondo que os números de ocorrências do evento nos subintervalos sejam independentes, o número esperado de subintervalos contendo apenas um evento é

$$M.[1 - P_0(h)] = \frac{1 - P_0(h)}{h}, \qquad \text{pois} \qquad M = \frac{1}{h}.$$

Portanto, pode-se esperar que, quando $h \to 0$, este número convirja para o número esperado de ocorrências em um intervalo de tempo de comprimento 1, ou seja, quando $h \to 0$, existe $\lambda > 0$

tal que

$$\frac{1 - P_0(h)}{h} \to \lambda,$$

ou, ainda,

$$P_0(h) = 1 - \lambda h, \qquad \text{quando} \qquad h \to 0 \tag{1.2}$$

O parâmetro λ é denotado taxa, média ou intensidade do processo, e pode resumir, em um determinado instante, se um dado evento pode vir ou não a acontecer em um futuro próximo, levando em consideração apenas as ocorrências passadas.

Estes podem ser derivados rigorosamente de postulados simples sem apelar para teorias mais profundas.

Escolhemos uma origem de medição de tempo e dizemos que na época $t > 0$ o sistema está no estado E_n se exatamente n saltos ocorreram entre 0 e t.

Então $P_n(t)$ é igual à probabilidade do estado E_n na época t, mas $P_n(t)$ pode ser descrito também como a probabilidade de transição de um estado arbitrário E_j em uma época arbitrária t para o estado E_{i+n} na época $t + h$. Agora traduzimos nossa descrição informal do processo em propriedades das probabilidades de $P_n(t)$.

Com isso, é possível mostrar que :

$$P_n(t) = P[Z(t) = n] = \frac{(\lambda t)^n e^{-\lambda t}}{n!}, \tag{1.3}$$

ou, também,

$$P_n(t + h) = P(Z(t + h) = n) \tag{1.4}$$

Em que os eventos são independentes, $P_n(t)$ descreve a probabilidade da ocorrência de um evento no instante t, ocorrer exatos n eventos entre os instantes 0 e t ($t > 0$) e λ é a taxa do processo.

Para provar isso, assuma primeiro $n > 1$ e considere o evento que na época $t + h$ o sistema está no estado E_n. A probabilidade desse evento é igual a $P_n(t + h)$, e o evento pode ocorrer de três maneiras mutuamente excludentes.

Demonstração. Primeiro, na época t o sistema pode estar no estado E_n e nenhum salto ocorre entre t e $t + h$. A probabilidade dessa contingência é

$$P_n(t)P_0(h) = P_n(t)[1 - \lambda h] + o(h), \tag{1.5}$$

onde $P_0(h) = P[Z(0) = h] = 1 - \lambda h$, como apresentado na equação (1.2), e $o(h)$ denota uma função real de h que decresce para zero mais rápido que h, isto é,

$$\lim_{h \to 0} \frac{o(h)}{h} = 0,$$

como apresentado no livro do Tijms (2003).

A segunda possibilidade é que na época t o sistema está no estado E_{n-1} e exatamente um salto ocorre entre t e $t + h$. A probabilidade de que isso ocorra é

$$P_{n-1}(t)P_1(h) = P_{n-1}(t).\lambda h + o(h). \tag{1.6}$$

A terceira possibilidade é que o processo esteja em um estado menor que $n - 1$ no tempo t e haja duas ou mais ocorrências entre o tempo t e $t + h$, o que tem probabilidade $o(h)$. Consequentemente, devemos ter

$$P_n(t + h) = P_n(t)(1 - \lambda h) + P_{n-1}(t)\lambda h + o(h). \tag{1.7}$$

Para n = 0, substituindo em (1.7)

$$P_0(t + h) = P_0(t)(1 - \lambda h) + o(h),$$

ou ainda,

$$P_0(t + h) = P_0(t) - \lambda h P_0(t) + o(h). \tag{1.8}$$

Subtraindo $P_0(t)$ e dividindo a equação (1.8) por h, obtemos:

$$\frac{P_0(t + h) - P_0(t)}{h} = -\lambda P_0(t) + \frac{o(h)}{h} \tag{1.9}$$

Fazendo h $\rightarrow$ 0, chegamos por definição de derivada em

$$\lim_{h\to 0}\left[\frac{P_0(t+h)-P_0(t)}{h}\right] = -\lambda P_0(t) + \lim_{h\to 0}\left[\frac{o(h)}{h}\right]. \tag{1.10}$$

Pela definição da função $o(h)$, sabemos que $\lim_{h\to 0}\left[\frac{o(h)}{h}\right] = 0$. Substituindo esse resultado na equação (1.10), chegamos em

$$P_0'(t) = -\lambda P_0(t),$$

ou equivalente,

$$\frac{P_0'(t)}{P_0(t)} = -\lambda. \tag{1.11}$$

Integrando a equação (1.11) em relação a t, obtemos

$$\int \frac{P_0'(t)}{P_0(t)}\,\mathrm{d}t = -\int \lambda \mathrm{d}t.$$

Logo,

$$ln[P_0(t)] = -\lambda t + c, \tag{1.12}$$

em que c é constante.

Podemos determinar o $P_0(t)$ na equação (1.12) aplicando a função exponencial

$$e^{[ln[P_0(t)]]} = e^{(-\lambda t + c)} \rightarrow P_0(t) = e^{-\lambda t}e^{c} = ke^{-\lambda t}, \text{ com } k \text{ constante.}$$

Para $t = 0$,

$$P_0(0) = ke^{-\lambda 0} = k.1 = 1 \Rightarrow k = 1.$$

Portanto,

$$P_0(t) = e^{-\lambda t}.$$

Agora, se fizermos para $n > 0$ iremos utilizar a terceira possibilidade mostrada em (1.7):

$$P_n(t+h) = P_n(t)(1-\lambda h) + P_{n-1}(t)\lambda h + o(h),$$

ou, ainda,

$$P_n(t+h) = P_n(t) - \lambda h P_n(t) + P_{n-1}(t)\lambda h + o(h) \tag{1.13}$$

Subtraindo $P_n(t)$ e dividindo a equação (1.13) por h, obtemos:

$$\frac{P_n(t+h) - P_n(t)}{h} = -\lambda P_n(t) + \lambda P_{n-1}(t) + \frac{o(h)}{h}.$$

Fazendo $h \to 0$, chegamos por definição de derivada em

$$\lim_{h\to 0}\left[\frac{P_n(t+h) - P_n(t)}{h}\right] = -\lambda P_n(t) + \lambda P_{n-1}(t) + \lim_{h\to 0}\left[\frac{o(h)}{h}\right].$$

Como $\lim_{h\to 0}\left[\frac{o(h)}{h}\right] = 0$, chegamos em

$$P_n'(t) = -\lambda P_n(t) + \lambda P_{n-1}(t). \tag{1.14}$$

Vamos agrupar os termos $P_n(t)$ e $P_{n-1}(t)$, na equação (1.14), e depois multiplicar a mesma por $e^{\lambda t}$, de maneira a obtermos

$$e^{\lambda t}[P_n'(t) + \lambda P_n(t)] = e^{\lambda t}\lambda P_{n-1}(t).$$

Podemos observar que o lado esquerdo da equação corresponde à derivada de $e^{\lambda t}P_n(t)$, em relação a t, dessa forma reescrevemos como:

$$\frac{d}{dt}[e^{\lambda t}P_n(t)] = \lambda e^{\lambda t}P_{n-1}(t) \tag{1.15}$$

Se fizermos $n = 1$ na equação (1.15), já que encontramos que $P_0(t) = e^{-\lambda t}$, obtemos:

$$\begin{aligned}
&\frac{d}{dt}[e^{\lambda t}P_1(t)] = \lambda e^{\lambda t}P_0 \\
&\frac{d}{dt}[e^{\lambda t}P_1(t)] = \lambda e^{\lambda t}.[e^{-\lambda t}] \\
&\frac{d}{dt}[e^{\lambda t}P_1(t)] = \lambda. \qquad (1.16)
\end{aligned}$$

Integrando a equação (1.16) em relação a t, temos

$$\int \frac{d}{dt}[e^{\lambda t}P_1(t)]\mathrm{d}t = \int \lambda, \mathrm{d}t,$$

ou, ainda,

$$e^{\lambda t}P_1(t) = \lambda t + c. \qquad (1.17)$$

Isolando $P_1(t)$ na equação (1.17) :

$$P_1(t) = e^{-\lambda t}(\lambda t + c)$$

Como $P_1(0) = 0$, temos

$$P_1(0) = e^{-\lambda 0}(\lambda 0 + c) = c \qquad \Rightarrow \qquad P_1(0) = c = 0.$$

Então,

$$P_1(t) = \lambda t\, e^{-\lambda t}.$$

Fazendo $n = 2$ na equação (1.15), temos

$$\begin{aligned}
&\frac{d}{dt}[e^{\lambda t}P_2(t)] = \lambda e^{\lambda t}P_1 \\
&\frac{d}{dt}[e^{\lambda t}P_2(t)] = \lambda e^{\lambda t}.[\lambda t e^{-\lambda t}] \\
&\frac{d}{dt}[e^{\lambda t}P_2(t)] = \lambda^2 t. \qquad (1.18)
\end{aligned}$$

Integrando a equação (1.18) em relação a t, obtemos

$$\int \frac{d}{dt}[e^{\lambda t}P_2(t)]\mathrm{d}t = \int \lambda^2 t, \mathrm{d}t \qquad \Rightarrow \qquad e^{\lambda t}P_2(t) = \frac{\lambda^2 t^2}{2} + c,$$

ou assim,

$$P_2(t) = e^{-\lambda t}\left[\frac{\lambda^2 t^2}{2} + c\right].$$

Porém, $P_2(0) = 0$, então

$$P_2(0) = e^{-\lambda 0}\left[\frac{\lambda^2 0^2}{2} + c\right] \qquad \Rightarrow \qquad P_2(0) = c = 0.$$

Logo,

$$P_2(t) = \frac{\lambda^2 t^2 e^{-\lambda t}}{2} = \frac{(\lambda t)^2 e^{-\lambda t}}{2}.$$

Fazendo $n = 3$ na equação (1.15), temos

$$\begin{aligned}
&\frac{d}{dt}[e^{\lambda t}P_3(t)] = \lambda e^{\lambda t}P_2 \\
&\frac{d}{dt}[e^{\lambda t}P_3(t)] = \lambda e^{\lambda t}.\left[\frac{(\lambda t)^2 e^{-\lambda t}}{2}\right] \\
&\frac{d}{dt}[e^{\lambda t}P_3(t)] = \frac{\lambda^3 t^2}{2}.
\end{aligned} \tag{1.19}$$

Integrando a equação (1.19) em relação a t, obtemos

$$\begin{aligned}
\int \frac{d}{dt}[e^{\lambda t}P_3(t)]\mathrm{d}t &= \int \frac{\lambda^3 t^2}{2}\mathrm{d}t \\
e^{\lambda t}P_3(t) &= \frac{\lambda^3}{2}\left[\frac{t^3}{3} + c\right] \\
e^{\lambda t}P_3(t) &= \frac{(\lambda t)^3}{3.2} + c \\
e^{\lambda t}P_3(t) &= \frac{(\lambda t)^3}{3!} + c,
\end{aligned}$$

ou, ainda,

$$P_3(t) = e^{-\lambda t}\left[\frac{(\lambda t)^3}{3!} + c\right].$$

Novamente, como $P_3(0) = 0$, temos

$$P_3(0) = e^{-\lambda 0}\frac{(\lambda 0)^3}{3!} + c \qquad \Rightarrow \qquad P_3(0) = c = 0.$$

Então,

$$P_3(t) = e^{-\lambda t}\frac{(\lambda t)^3}{3!}.$$

Podemos, através do princípio de indução finita, assumir que é válido para $n-1$ em relação a t. Portanto,

$$P_n(t) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}.$$

Como queríamos demonstrar desde o inicio. ■

1.8 MODELOS DETERMINISTICOS E ESTOCÁSTICO

Primeiro precisamos definir passeio aleatório:

Passeio Aleatório

Um passeio aleatório, em modelos probabilísticos, é um processo estocástico no qual uma série de mudanças sucessivas ocorre de forma aleatória ao longo do tempo. Em outras palavras, é um tipo de trajetória que se move de maneira imprevisível, dando passos aleatórios em direções diferentes.

Formalmente, um passeio aleatório pode ser definido como uma sequência de variáveis aleatórias $X_0, X_1, X_2, ...$, onde cada X_i representa a posição do processo no tempo i, e as transições de uma posição para a próxima são determinadas por eventos aleatórios. A mudança de posição em cada etapa é geralmente caracterizada por um passo aleatório, que pode ser determinado por uma distribuição de probabilidade discreta.

Exemplo 1.13. Considere um exemplo simples de um passeio aleatório unidimensional, onde um indivíduo está caminhando em uma linha reta. A posição inicial é 0. Em cada passo, o indivíduo pode decidir se move para a direita (+1) ou para a esquerda (-1), cada um com uma probabilidade igual de $\frac{1}{2}$. A posição no tempo i é dada pela soma dos passos anteriores:

$$X_i = X_{i-1} + \epsilon_i$$

Onde ϵ_i é uma variável aleatória que toma valores +1 ou -1 com probabilidades iguais de $\frac{1}{2}$. Esse processo resulta em uma trajetória imprevisível, conforme o indivíduo dá passos aleatórios para a direita ou para a esquerda ao longo do tempo. Esse é um exemplo clássico de um passeio aleatório simples.

Modelos Determinísticos:

Modelos determinísticos são modelos matemáticos em que as relações entre as variáveis são completamente determinadas, sem aleatoriedade envolvida. Em outras palavras, para um conjunto de condições iniciais específicas, o modelo determinístico sempre produzirá o mesmo resultado. Esses modelos são frequentemente usados quando a incerteza e a variabilidade não são consideradas relevantes para o sistema em estudo.

Para ilustrar a natureza determinística, considere o exemplo simples de um modelo linear. Suponha que tenhamos uma equação linear simples:

$$Y = aX + b$$

Nesta equação, Y e X são variáveis, e a e b são parâmetros constantes. Se você souber os valores específicos de X, a, e b, então o valor de Y pode ser calculado de maneira única, sem qualquer componente aleatório.

Exemplo 1.14. Vamos considerar um exemplo mais concreto. Suponha que temos uma equação para o crescimento populacional ao longo do tempo, onde P_t é a população no tempo t:

$$P_{t+1} = r \cdot P_t$$

Nesta equação, r é uma taxa de crescimento constante. Se soubermos a população inicial P_0 e a taxa de crescimento r, podemos determinar a população em qualquer ponto futuro sem incerteza. Cada valor futuro é completamente determinado pelos valores iniciais e pelos parâmetros do modelo.

Ao contrário dos modelos estocásticos, onde há aleatoriedade e variação, os determinísticos são precisos e previsíveis, assumindo que as condições iniciais e os parâmetros são conhecidos.

Modelos Estocásticos:

Modelos estocásticos são uma classe de modelos matemáticos que lidam com a incerteza e a aleatoriedade. Eles são usados para descrever sistemas nos quais os resultados não são completamente determinísticos, mas seguem uma distribuição probabilística. Esses modelos são frequentemente aplicados em diversas áreas, como finanças, economia, engenharia, ciência da computação, medicina e muito mais.

Para entender a natureza estocástica de um modelo, considere o exemplo de um processo estocástico simples, como um passeio aleatório. Suponha que você esteja em um ponto em uma linha e, a cada passo, você pode mover para a direita ou para a esquerda com igual probabilidade. O resultado do próximo passo não pode ser previsto com certeza, pois é aleatório.

Vamos denotar X_n como a posição após n passos. A mudança na posição X_n em relação à posição anterior X_{n-1} é uma variável aleatória. Se definirmos Z_n como uma variável aleatória que representa o movimento em um único passo (1 para a direita ou -1 para a esquerda), então podemos escrever:

$$X_n = X_{n-1} + Z_n$$

Aqui, X_n é o resultado do processo estocástico após n passos. Cada Z_n é uma variável aleatória independente e identicamente distribuída (i.i.d), o que significa que a probabilidade de um passo para a direita ou para a esquerda é a mesma em cada passo.

Exemplo 1.15. Considere um passeio aleatório simples iniciado em $X_0 = 0$. Após 5 passos, o processo poderia ter evoluído assim:

$$X_5 = X_4 + Z_5 = X_3 + Z_4 + Z_5 = ...$$

Cada Z_i é uma variável aleatória, tornando o resultado final uma realização específica do processo estocástico. No entanto, se você repetir o experimento, obterá trajetórias diferentes devido à natureza estocástica do processo.

Este exemplo ilustra a essência dos modelos estocásticos, onde o resultado é incerto e segue uma distribuição probabilística. Modelos mais complexos podem envolver uma variedade de técnicas estatísticas e matemáticas para lidar com a aleatoriedade e incerteza em diferentes contextos.

2 CADEIAS DISCRETAS DE MARKOV

Uma cadeia de Markov é um sistema que evolui ao longo do tempo, onde o estado futuro é determinado diretamente pelo estado presente, sem considerar como o sistema chegou ao estado atual. Essa propriedade é chamada de propriedade de Markov, e é o que distingue as cadeias de Markov de outros modelos probabilísticos.

Através de um exemplo de previsão meteorológica, neste trabalho, mostraremos a aplicabilidade prática dos conceitos de Cadeias Discretas de Markov incluindo estados recorrentes e transitórios, a equação de equilíbrio para estabilidade, tempo de retorno, bem como estados absorvente

Cadeia de Markov Discreta: Um processo estocástico discreto, sendo o conjunto de estados possíveis e finito ou infinito enumerável tal que:

$$P[X_{n+1} = i_{n+1} | X_0 = i_0, ..., X_n = i_n] = P[X_{n+1} = i_{n+1} | X_n = i_n] \tag{2.1}$$

Dito então que, dado um estado no tempo $n + 1$, seu valor está condicionado com o estado no instante n. Certamente podemos expressar de outra forma, que uma cadeia de Markov é um processo estocástico sem memória, ou seja, o resultado de cada tentativa depende apenas do resultado imediatamente anterior, sem ser influenciado pelas tentativas anteriores. Denotaremos o conjunto de estados como S, e os estados individuais como i e j.

As probabilidades de transição de um estado i para um estado j são definidas como:

$$P[X_{n+1} = j | X_n = i] = p_{ij} \tag{2.2}$$

A matriz de transição P é definida como:

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix}$$

Satisfazem as propriedades:

$$p_{ij} \geq 0 \text{ com } i, j \in S \text{ e } \sum_j p_{ij} = 1 \text{ para cada } i \in S$$

Para construir uma cadeia de Markov e resolver problemas práticos, são necessários dois itens: selecionar as variáveis de estado e determinar as probabilidades de transição para cada estado, p_{ij}.

Chapman-Kolmogoroff:
O teorema de Chapman-Kolmogorov descreve a probabilidade de transições de estado em vários pontos no tempo, sendo um resultado fundamental da teoria de cadeias de Markov. Considerando um estado intermediário, chamado estado k, este teorema calcula a probabilidade de chegar do estado i ao estado j em $n + m$ passos, passando pelo estado k. . Essa habilidade é crucial em modelagem probabilística, oferecendo uma ferramenta essencial para prever o comportamento de sistemas dinâmicos ao longo do tempo.

Teorema 01: *Para todo n,m = 0,1,2,...,* temos:
$$p_{ij}^{(m+n)} = \sum_{k \in S} p_{ik}^{(m)} p_{kj}^{(n)} \text{ com i,j} \in S$$

Demonstração. Visto que é utilizado matematicamente, a Lei da Probabilidade Total pode ser expressa da seguinte forma:

$$P(A) = P(A|B_1).P(B_1) + P(A|B_2).P(B_2) + \cdots + P(A|B_n).P(B_n),$$

onde:

a) $P(A)$ é a probabilidade do evento A ocorrer;

b) $P(A|B_1), P(A|B_2), \ldots, P(A|B_n)$ são as probabilidades condicionais do evento A ocorrer dado que ocorreu cada um dos eventos $B_1, B_2, \ldots, B_n$, respectivamente;

c) $P(B_1), P(B_2), \ldots, P(B_n)$ são as probabilidades dos eventos $B_1, B_2, \ldots, B_n$ ocorrerem.

Dado a Lei da Probabilidade Total, temos então:

$$P[x_{m+n} = j = \sum_{k \in S} P[X_{m+n} = j | X_0 = i, X_{=}k] P[X_m = k | X_0 = i]. \tag{2.3}$$

Denotaremos por $p_{ij}(n)$ a probabilidade de uma transição do estado i para o estado j em exatamente n passos. Em outras palavras, $p_{ij}^{(n)}$ é a probabilidade condicional de estar no estado j na enésima etapa, dado o estado inicial i, sendo esta a soma das probabilidades de todos os

caminhos possíveis de comprimento n com início no estado i e terminando no estado j. Um caminho do estado i para o estado j é uma sequência de transições com probabilidades de ocorrência positivas que ligam os dois estados.

Em particular $p_{ij}^{(1)} = p_{ij}$, pois houve só um passo, que foi do estado i ao estado j. Agora para $n = 2$

$$p_{ij}^{(2)} = \sum_{k \in S} p_{ik}.p_{kj}.$$

Como houve duas etapas, então passou do estado inicial i ao estado intermediário k e depois ao estado final j.

Fazendo $n = 3$, temos

$$p_{ij}^{(3)} = \sum_{k \in S} p_{ik}^{(1)}. \sum_{l \in S} p_{kl}^{(1)}.p_{lj}^{(1)}.$$

Agora, distribuindo o produto e trocando a ordem das somatórias, temos:

$$p_{ij}^{(3)} = \sum_{k \in S \text{ e } l \in S} p_{ik}.p_{kj}^{(2)}.$$

É de fácil percepção que por indução, teremos para $n + 1$

$$p_{ij}^{(1+n)} = \sum_{k \in S} p_{ik}.p_{kj}^{(n)}.$$

Vamos supor para $n + 2$:

$$p_{ij}^{(2+n)} = \sum_{k \in S \text{ e } e \in S} p_{ie}.p_{ek}.p_{kj}^{(n)},$$

ou seja,

$$p_{ij}^{(2+n)} = \sum_{k \in S} p_{ik}^{(2)}.p_{kj}^{(n)}.$$

Hipótese de indução:

$$p_{ij}^{(m+n)} = \sum_{k \in S} P_{ik}^{(n)} P_{kj}^{(m)}.$$

Então, vamos mostrar que vale para $m + 1$

$$p_{ij}^{(m+1+n)} = \sum_{k \in S \; e \; e \in S} p_{ie}^{(m)} . p_{ek}^{(1)} . p_{kj}^{(n)}$$
$$p_{ij}^{(m+1+n)} = \sum_{k \in S} p_{ik}^{(m+1)} . p_{kj}^{(n)} .$$

Agora, isso é exatamente a relação que queremos provar

$$P_{ij}^{(m+n)} = \sum_{k \in S} P_{ik}^{(n)} P_{kj}^{(m)}.$$

Portanto, por indução, provamos que a relação é verdadeira para todos os valores de n e m. ■

Essa fórmula é útil porque permite calcular a probabilidade de transições de estados em cadeias de Markov em vários momentos diferentes sem a necessidade de reverter aos passos anteriores. Em vez disso, é possível calcular as probabilidades de transição de um estado para outro passo a passo, acumulando as probabilidades intermediárias ao longo do tempo.

2.1 EXEMPLO: PREVISÃO DO TEMPO

Dado uma cidade, é possível classificar o clima como:

Estado 1	**Estado 2**	**Estado 3**
Ensolarado	Nublado	Chuvoso

Supomos que o clima do dia seguinte, depende apenas do estado em que se encontra hoje, e não de como esteve nos dias anteriores. Caso o dia esteja ensolarado, o próximo dia estará ensolarado, nublado ou chuvoso com probabilidades respectivas de 0,70; 0,10; 0,20. Quando o dia atual está nublado, as probabilidades são 0,50; 0,25; 0,25 e no dia chuvoso as probabilidades são 0,40; 0,30; 0,30.

Ao elaborar a matriz de transição de cada estado, obtemos:

$$\begin{pmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{pmatrix} = \begin{pmatrix} 0,70 & 0,10 & 0,20 \\ 0,50 & 0,25 & 0,25 \\ 0,40 & 0,30 & 0,30 \end{pmatrix}$$

Onde as p_{ij} representa a probabilidade de transição do estado i para o estado j. Essa matriz é fundamental para modelar as mudanças de estado do clima ao longo do tempo, permitindo probabilísticas específicas para os dias seguintes com base no estado atual. Como visto caso o dia esteja no estado 1 (Ensolarado) a probabilidade de no outro dia estar no estado 3 (chuvoso) é de 20% e assim por diante.

2.2 CLASSIFICAÇÃO DE ESTADOS

Neste exemplo, identificam-se os estados como transitórios, absorventes ou recorrentes:

Um estado $i \in S$ é dito **recorrente** se , partindo do estado i, a cadeia retornará ao estado i dentro de um tempo finito (aleatório), com probabilidade 1, ou seja: $p_{i,i} := P[T_i^r < \infty | X_0 = i] = P[X_n = i$ para algum $n \geq 1 | X_0 = i) = 1$.

Isso significa que, uma vez que a cadeia entra em um estado recorrente, ela continuará a retornar a esse estado repetidamente ao longo do tempo.

Considere uma matriz de transição onde ambos os estados têm probabilidade significativa de permanecer no mesmo estado:

$$P = \begin{bmatrix} 0.4 & 0.6 \\ 0.3 & 0.7 \end{bmatrix}$$

Aqui, ambas as condições são recorrentes, pois, uma vez que o sistema sai de um estado, há uma alta probabilidade dele voltar ao mesmo estado. Essa recorrência é refletida nas probabilidades de transição da matriz P.

Um estado $i \in S$ é dito **transiente** quando não é recorrente, ou seja:

$$p_{i,i} = P[T_i^r = \infty | X_0 = i] > 0$$

Isso implica que existe uma chance positiva de que o processo de Markov nunca retorne ao estado i após ter começado nesse estado i no tempo X_0 . Portanto, um estado transiente não é visitado infinitas vezes pela cadeia de Markov.

Suponha que temos dois estados: "Chuva"(C) e "Nublado"(N). A matriz de transição de estados é dada por:

$$P = \begin{bmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{bmatrix}$$

Aqui, a probabilidade de transição varia com o tempo e esta cadeia de Markov tem estados transientes, pois há uma chance positiva de mudar de um estado para outro ao longo do tempo.

Um estado **absorvente** em uma cadeia de Markov é um estado no qual a cadeia entra e, uma vez alcançado esse estado, não há possibilidade de sair dele.

$$p_{i,j} = P[T_i^r = \infty | X_0 = i] = 0$$

No caso é um estado em que a transição para outros estados tem probabilidade zero, sendo então a probabilidade $p_{i,i} = 1$.

Consideremos uma matriz de transição onde o estado "Chuva"(C) é absorvente:

$$P = \begin{bmatrix} 1 & 0 \\ 0.1 & 0.9 \end{bmatrix}$$

Aqui, uma vez que o sistema entra no estado "Chuva", ele permanece lá indefinidamente, caracterizando-o como um estado absorvente. O estado "Nublado"ainda é transitório, pois há

uma probabilidade não nula de mudar para "Chuva".

Analisando a classificação dos estados no exemplo inicial de previsão do tempo, observamos que, de acordo com a definição formal, um estado i é considerado recorrente se, iniciando a partir do estado i, a cadeia retornar a i dentro de um tempo finito (aleatório) com probabilidade 1. Nesse contexto, nossa matriz apresenta exclusivamente estados recorrentes, não havendo a presença de estados transientes ou absorventes. Isso ocorre porque todos os estados possuem probabilidade de transição entre si, indicando uma dinâmica contínua onde a cadeia pode retornar repetidamente a qualquer estado.

2.3 TRANSIÇÃO DE ESTADOS - ESTACIONÁRIO

Qual é a probabilidade de termos um tempo ensolarado após três dias, considerando que o dia atual é chuvoso?

Para resolução iremos utilizar o **Teorema 01 de Chapman-Kolmogoroff**, obtemos:

$$p_{ij}^{(3)} = \sum_{k \in S} p_{ik}.p_{kj}^{(2)} = [P^{(3)}] \; com \; i,j \; \in S$$

Portando, $[P^{(3}] = \begin{pmatrix} 0,60115000 & 0,1682500 & 0,2302500 \\ 0,5912500 & 0,1756250 & 0,2331250 \\ 0,5855000 & 0,1797500 & 0,2347500 \end{pmatrix}$

Atráves desse resultado, chegamos que a $p_{31}^{(3)}$, ou seja, a probabilidade de sair do estado chuvoso para o estado ensolarado em 3 dias é de 0,585500.

Como ficariam as probabilidades de transição após vários dias?

Intuitivamente, esperamos uma distribuição independente do clima atual. De maneira informal, para determinarmos se uma matriz permanece constante ao longo do tempo, é necessário verificar se ela se estabiliza no infinito, ou seja, se uma distribuição limitada existe. Isso envolve elevar a matriz a potências maiores e observar se as entradas se estabilizam.

2.4 EQUAÇÃO DE EQUILÍBRIO

Outra forma de identificarmos se a matriz é estacionária, é atráves de equações de equilíbrio.

Se iniciarmos o processo de acordo com a distribuição limite, isso resultará em um processo que operará em equilíbrio, da seguinte forma:

Teorema 02: $\Pi_i = \sum_{(k \in S)} \Pi_k . p_{kj}$

Denominada Π como uma equação de equilíbrio. O teorema abordado encontra-se discutido detalhadamente no livro "A First Course in Stochastic Models"de Henk Tijms, Importante ressaltar que, neste contexto, não será apresentada uma demonstração formal do teorema nesta resposta. Contudo, é proposta uma abordagem heurística que viabiliza uma compreensão inicial mais intuitiva do teorema.

Explicação: Por um raciocínio heurístico, supomos que

$$\Pi_j = P[X_\infty = j].$$

Pela Lei total da probabilidade e a propriedade de homogeneidade no tempo de cadeias de Markov, temos

$$\Pi_j = P[X_\infty = j] = \sum_{k \in S} P[X_\infty = j | X_{\infty - 1} = k].P[X_{\infty - 1} = k].$$

Com isso,

$$\Pi_j = \sum_{k \in S} P_{kj} \Pi_k .$$

Temos que para cada estado da cadeia haverá uma equação de equilíbrio.

$$\Pi_1 = 0,70.\Pi_1 + 0,50.\Pi_2 + 0,40.\Pi_3$$

$$\Pi_2 = 0,10; \Pi_1 + 0,25.\Pi_2 + 0,30.\Pi_3$$

$$\Pi_3 = 0,20; \Pi_1 + 0,25.\Pi_2 + 0,30.\Pi_3$$

Sabemos que $\Pi_1 + \Pi_2 + \Pi_3 = 1$, sendo possível resolver o sistema linear e chegamos em:

$$\boxed{\Pi_1 = 0,5960, \Pi_2 = 0,1722, \Pi_3 = 0,2318}$$

Essas duas formas de chegar à probabilidade de equilíbrio estão relacionadas e são equivalentes em cadeias de Markov que satisfaçam a:

Suposição 1: A Cadeia de Markov possui
um estado r tal que $f_{ir} = 1$ para todo i$\in$ S e $\mu_{rr} < \infty$

A ideia por trás dessa suposição é garantir que exista um estado absorvente para que a cadeia possa eventualmente convergir e permanecer, tornando-se estacionária. A finitude do tempo de retorno garante que a cadeia não passará infinitamente ao longo do estado absorvente, garantindo que exista uma distribuição estacionária.

É importante notar que nem todas as cadeias de Markov satisfazem a Suposição 1. Cadeias que não atendem a essa condição não podem ter uma distribuição estacionária. No entanto, quando a Suposição 1 está satisfeita, ela garante a existência de uma distribuição de equilíbrio para a cadeia de Markov.

Isso ocorre de forma automática em uma cadeia de estados finitos, Essa condição implica que todos os estados estão interconectados, possibilitando transições de ida e volta entre eles, e garantindo que qualquer estado seja alcançável a partir de qualquer outro na cadeia.

Portanto, a distribuição estacionária é uma propriedade mais forte em Cadeias de Markov, pois representa um equilíbrio sustentado e estável das probabilidades de estado. A distribuição limite é uma ferramenta útil para entender o comportamento a longo prazo, mas não implica necessariamente na existência de um equilíbrio estável.

2.5 TEMPO MÉDIO DE RETORNO

Quantos dias levará para que o tempo Nublado (2) retorne, desde que tenha partido do mesmo?

Como queremos calcular o tempo de recorrência, iremos utilizar o cálculo de tempo médio de

retorno.

Teorema 03: $\mu_i(i) = 1 + \sum_{l \in S \text{ e } l \neq i} p_{i,l}.\mu_i(l)$

Demonstração. Sabemos que, matematicamente, a esperança de uma variável aleatória discreta X é calculada multiplicando cada valor possível de X pela sua respectiva probabilidade e somando esses produtos.

Seja X uma variável aleatória discreta com valores $x_1, x_2, \ldots, x_n$ e probabilidades $p_1, p_2, \ldots, p_n$, respectivamente, a esperança matemática $E(X)$ é dada por:

$$E(X) = x_1.p_1 + x_2.p_2 + \cdots + x_n.p_n.$$

Quando $X_1 = j$, teremos o tempo $T_j^r = P[X_1 = j | X_0 = i] = 1$. Ficando então, com l= estado intermediário,

$$\mu_j(i) = 1.P[X_1 = j | X_0 = i] + \sum_{l \in e l \neq j} P[X_1 = l | X_0 = i].(1 + E[T_j^r | X_0 = l])$$

Portanto, $(1 + E[T_j^r | X_0 = l])$, o 1 corresponde ao tempo de ir do estado i ao estado l em um passo, ou seja, quando $X_1 = l$, teremos o tempo $T_l^r = P[X_1 = l | X_0 = i] = 1$.

Já a $E[T_j^r | X_0 = l]$ segue da definição de $\mu_j(i)$ agora com $\mu_j(l)$ correspondendo ao tempo médio de passagem para o estado j, agora iniciando o processo no estado l intermediário.

Agora temos, então

$$\mu_j(i) = P_{i,j} + \sum_{l \in S \text{ e } l \neq j} P_{i,l}(1 + \mu_j(l)), \tag{2.4}$$

em que

$$1 + \mu_j(l)) = \sum_{l \in S \text{ e } l \neq j} P_{i,l}.\mu_j(l).$$

Substituindo na equação (2.4)

$$\mu_j(i) = P_{i,j} + \sum_{l \in S \text{ e } l \neq j} P_{i,l} + \sum_{l \in S e l \neq j} P_{i,l}.\mu_j(l). \tag{2.5}$$

Agora temos que

$$(P_{i,j} + \sum_{l \in S \text{ e } l \neq j} P_{i,l}) = 1.$$

Substituindo na equação (2.5)

$$\mu_j(i) = 1 + \sum_{l \in S e l \neq j} P_{i,l}.\mu_j(l).$$

Chegamos então que o tempo médio de primeira passagem é

$$\mu_j(i) = 1 + \sum_{l \in S e l \neq j} P_{i,l}.\mu_j(l)$$

■

Retornando ao Exemplo:

Para o estado Nublado(2):

$$\mu_2 = 1 + p_(2,1).\mu_1 + p_(2,2).\mu_2 + p_(2,3).\mu_3$$

$$\mu_2 = 1 + 0.10.\mu_1 + 0.25.\mu_2 + 0.20.\mu_3$$

Como a resolução dessa equação depende de outras duas incógnitas que não temos o valor, iremos calcularμ_1 e μ_3, afim de chegar em sistema de equações lineares que seja possível resolver.

Sistema de equações lineares:

$$\mu_1 = 1 + 0.70.\mu_1 + 0.10.\mu_2 + 0.20.\mu_3$$

$$\mu_2 = 1 + 0.10.\mu_1 + 0.25.\mu_2 + 0.20.\mu_3$$

$$\mu_3 = 1 + 0.40.\mu_1 + 0.30.\mu_2 + 0.30.\mu_3$$

Resolvendo esse sistema temos então os valores aproximados das soluções são:

$$(\mu_1, \mu_2, \mu_3) =\approx (14.519, 7.596, 12.980)$$

Logo o tempo médio de retorno ao estado Nublado(2) é aproximadamente 8 dias.

2.6 RECOMPENSA MÉDIA A LONGO PRAZO

Associando recompensas aos estados:

Estado 1	Estado 2	Estado 3
5	0	-2

Qual o cálculo da expectativa da recompensa média acumulada pelo agente por unidade de tempo?

Se uma Cadeia de Markov Discreta possui uma distribuição estacionária, então a média das recompensas ao longo do tempo se torna independente do tempo e tende a um valor constante para cada estado da cadeia.

Teorema 04: (A): A recompensa total ganha entre duas visitas da cadeia ao estado r tem uma expectativa finita.

$$\sum_{j \in S} |f(j)| \pi_j < \infty$$

(B) Para cada estado inicial $X_0 = i$ com $i \neq r$, a recompensa ganha até a primeira visita da cadeia ao estado r é finita com probabilidade 1.

Sendo assim $\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) = \sum_{j \in S}^{n} f(j)\pi_j$. com $X_0 = i$.

Quando lidamos com recompensas em uma cadeia de Markov ou qualquer outro processo estocástico, muitas vezes as recompensas ocorrem gradualmente ao longo do tempo entre as transições de estado da cadeia. Isso significa que a recompensa não necessariamente é recebida imediatamente após uma transição de estado, mas pode ser acumulada ao longo de várias etapas antes de ser realizada completamente.

Para que o **Teorema 04** seja aplicável, é necessário que a **Suposição 1.** seja cumprida. Visto que essa suposição está satisfeita, temos:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1} f(X_k) = f_{(1)}.\pi_1 + f_{(2)}.\pi_2 + f_{(3)}.\pi_3$$

$$= 5.0,5960 + 0.0,1722 + (-2).0,2318$$

$$= 2,5 \textit{ unidade de tempo.}$$

Portando, a expectativa da recompensa acumulada por unidade de tempo é 2,5.

Com base nas recompensas associadas às transições em nossa cadeia de Markov, fica evidente que o próximo capítulo, centrado nos "Processos de Markov", oferecerá uma oportunidade crucial para aprofundar nosso entendimento. As dinâmicas de recompensas indicam que o conhecimento mais aprofundado desses processos estocásticos não apenas ampliará nossa compreensão teórica, mas também maximizará as recompensas.

3 O MODELO DE DECISÃO SEQUENCIAL (MDPS)

Agora, após a exploração dos Capítulos 1 e 2, que estabeleceram os alicerces necessários para nossos estudos, então estamos prontos para adentrar no tema central desta pesquisa: os Processos de Decisão de Markov (MDPs). Nesta etapa, vamos mergulhar profundamente nos conceitos essenciais que fundamentam os MDPs, compreendendo cada termo de maneira detalhada. Como visto, os MDPs são uma ferramenta poderosa e versátil, frequentemente utilizada para modelar situações em que a tomada de decisões ocorre de maneira sequencial em ambientes estocásticos. Antes de explorarmos estratégias e otimizações em MDPs, é crucial compreender os elementos fundamentais que compõem essa estrutura.

3.1 DEFINIÇÃO DO PROBLEMA E NOTAÇÃO

Na elaboração de um Modelo de Decisão de Markov Sequencial (MDP), a definição do problema desempenha um papel crucial. Este modelo é frequentemente utilizado para representar cenários nos quais um agente é confrontado com a oportunidade de intervir em um sistema probabilístico à medida que o tempo avança. Os componentes fundamentais do MDP, como os estados que representam configurações possíveis do ambiente e as ações disponíveis ao agente em cada estado, devem ser cuidadosamente delineados. A precisão na identificação desses elementos é essencial, uma vez que eles formam a base do sistema .

O agente, impulsionado por objetivos específicos, tem a missão de escolher ações que otimizem o desempenho do sistema em relação a critérios predefinidos. Essa otimização envolve não apenas considerar as consequências imediatas de cada ação, mas também avaliar como essas decisões impactarão o sistema a médio e longo prazo. Em essência, o agente busca antecipar e moldar ativamente o comportamento futuro do sistema, enfrentando a incerteza inerente ao ambiente.

3.2 ÉPOCAS E PERÍODOS DE DECISÃO

As "épocas"e "períodos de decisão"são conceitos essenciais que ajudam a estruturar a dinâmica temporal do processo de tomada de decisões. Vamos explorar minuciosamente esses termos:

Épocas:
Épocas referem-se aos diferentes instantes ou pontos no tempo ao longo dos quais as decisões são tomadas no MDP. Cada época representa uma unidade de tempo ou um passo na evolução do sistema.

T= Conjunto de épocas de decisões.

t= Época de decisão.

Destacamos então que há dois tipos de épocas de decisões:

-Discreto: Este modelo é apropriado para situações em que as decisões ocorrem em intervalos discretos e claramente definidos, como em jogos por turnos ou processos industriais que operam em etapas distintas e regulares.

-Contínuo: Em um modelo com épocas contínuas, o tempo é tratado como uma variável contínua, sem intervalos ou divisões discretas. As decisões podem ser tomadas em qualquer ponto ao longo da escala temporal contínua.

Em muitos casos, a abordagem discreta é mais simples computacionalmente e mais fácil de interpretar, enquanto a abordagem contínua pode ser necessária para representar de forma mais precisa sistemas que operam de maneira contínua, como filas.

Períodos:
Os períodos de decisão referem-se a conjuntos consecutivos de épocas em que o agente toma decisões sequenciais antes de receber feedback do ambiente. Um período de decisão geralmente abrange várias épocas, eles representam intervalos em que o agente pode planejar e executar ações antes de observar as consequências dessas ações. Essa estruturação em períodos de decisão é particularmente útil em contextos nos quais as decisões não são imediatamente seguidas de resultados observáveis.

Já o conceito de horizonte de tempo, seja ele finito ou infinito, é fundamental para definir a duração do processo de tomada de decisões.

Horizonte finito: Um horizonte de tempo finito implica que o processo de tomada de decisões é limitado a um número específico de épocas ou períodos de decisão. O agente toma decisões ao longo de um intervalo predeterminado de tempo e, após atingir o final desse horizonte, não são mais permitidas ações e o processo termina.

Horizonte infinito: O processo de tomada de decisões se estende indefinidamente ao longo do tempo, sem um ponto final predeterminado. O agente continua tomando decisões em todas as épocas futuras, considerando o impacto de suas ações de forma contínua e sem uma limitação temporal específica.

A escolha entre horizontes finitos e infinitos depende da natureza do problema específico que está sendo modelado e das questões temporais relevantes para a análise do agente.

3.3 CONJUNTO DE ESTADOS E AÇÕES

Estados (s): Embora ambos compartilhem o termo "estado"e estejam relacionados à teoria de Markov, é importante destacar que há diferenças significativas entre o conceito de estado em Modelos de Decisão de Markov Sequencial (MDPs) e em Cadeias de Markov.

Em uma Cadeia de Markov, um estado representa uma condição ou configuração do sistema no qual o processo estocástico está atualmente. O estado em uma cadeia de Markov é definido de modo que a probabilidade de transição para o próximo estado depende apenas do estado atual, e não do histórico de estados anteriores.

Em contrapartida, um estado em MDP também representa uma configuração do sistema, mas a definição é mais abrangente, pois inclui a capacidade de tomar decisões, sendo assim, incorporam não apenas a dinâmica probabilística do sistema, mas também a capacidade de escolher ações ativamente para influenciar as transições de estado. Além disso, um estado em um MDP não necessariamente precisa obedecer à propriedade de Markov. Pode incluir informações adicionais, como variáveis ocultas, recompensas associadas e outros fatores relevantes para o processo de decisão.

Conjunto de Estados (S): O conjunto de estados representa todas as possíveis situações ou configurações em que o sistema pode se encontrar. Cada estado encapsula as informações relevantes para as decisões do agente. Por exemplo, em um jogo de tabuleiro, os estados podem representar diferentes disposições das peças no tabuleiro, refletindo as várias configurações possíveis durante o jogo.

S_t é o conjunto de estados possíveis no tempo de decisão t, ou então, representa o conjunto de estados relevantes para a decisão em um instante específico.

Ações (a): As ações podem ser determinístico ou probabilístico:

Ações Determinísticas: Uma ação determinística é uma escolha específica que o agente faz em um determinado estado, levando a uma transição de estado previsível. Ou seja, para uma combinação estado-ação, a transição para o próximo estado é certa e conhecida.

Ações Probabilísticas: Ao contrário, ações probabilísticas referem-se a escolhas nas quais a transição para o próximo estado é sujeita a incerteza. A execução da mesma ação em um mesmo estado pode levar a resultados diferentes com probabilidades associadas.

Distribuição de Probabilidades de Ações: A distribuição de probabilidades de ações indica a probabilidade associada a cada ação possível em um estado específico. Se um agente está em um estado s e tem duas ações possíveis a_1 e a_2, a distribuição de probabilidades de ações pode ser expressa como $p(a_1|s)$ e $p(a_2|s)$, representando as chances de escolher cada ação.

Em Cadeias de Markov, o conceito de ações não é aplicável da mesma forma, sendo a probabilidade transição para o próximo estado determinada apenas pelo estado atual, sem a presença de ações tomadas por um agente externo.

Conjunto de ações (A) : Cada estado $s \in S$ está associado a um conjunto específico de ações em A Isso significa que o conjunto de ações disponíveis para o agente pode variar dependendo do estado atual do sistema. A dependência do estado nas ações reflete a ideia de que as escolhas do agente podem ser contextualmente limitadas ou expandidas com base na situação em que se encontra.

Em um cenário de jogo de tabuleiro o conjunto de ações disponíveis podem incluir "mover uma peça para frente", "capturar uma peça adversária", entre outras. A dependência do estado nas ações reflete as escolhas táticas que os jogadores podem fazer com base na configuração atual do jogo.

3.4 RECOMPENSAS E PROBABILIDADES DE TRANSIÇÃO

Probabilidade de transição: É uma medida que descreve a chance de um processo de decisão transitar de um estado para outro ao realizar uma determinada ação. A probabilidade de transição é frequentemente representada por uma matriz P, onde $p(s'|s, a)$ denota a probabilidade de transição do estado s para o estado s' ao realizar a ação a.

E com isso, ao escolher ações, o agente leva em consideração as probabilidades de transição para avaliar como suas ações influenciarão a distribuição de estados futuros. Em alguns casos,

diferentes ações podem levar a diferentes distribuições de probabilidade de transição para o próximo estado.

Soma igual a 1: Para cada estado s e ação a, a soma das probabilidades de transição para todos os estados sucessores s' é igual a 1. Matematicamente, $\sum_{s'} p(s'|s,a) = 1$.

Função de Recompensa (R): A função de recompensa em MDPs atribui valores numéricos às transições de estado e ações, representando o benefício ou custo associado a cada decisão. Essa função é geralmente denotada por $R(s,a,s')$, onde s é o estado atual, a é a ação tomada, e s' é o próximo estado.

Temporalidade: As recompensas podem ser imediatas, ocorrendo imediatamente após uma ação, ou adiadas, acumulando-se ao longo do tempo. A escolha depende do horizonte de tempo e dos objetivos do problema.

Representação da Função de Recompensa: Pode ser expressa como uma fórmula matemática, relacionando estados, ações e transições de estado a valores numéricos. Por exemplo, $R(s,a,s') = -C$ para penalizar certa ação, ou $R(s,a,s') = +G$ para recompensar outra.

Recompensa média:

$$r_t(s,a) = \sum_{s' \in S} r_t(s,a,s').p_t(s'|s,a) \tag{3.1}$$

A fórmula é uma maneira de calcular a recompensa média ponderada pelas probabilidades de transição para todos os possíveis estados sucessores. Isso reflete a natureza estocástica de um ambiente em um MDP, onde diferentes ações podem levar a diferentes resultados com probabilidades associadas.

E todas as informações necessárias para tomar uma decisão no momento t são resumidas em $r_t(s,a)$ e $p_t(s'|s,a)$, sob alguns critérios é utilizado também $r_t(s,a,s')$

Recompensa Total Esperada:
O objetivo do agente em um MDP é encontrar uma política ótima que maximize a recompensa total esperada ao longo do tempo. Dado $X_t(\omega) = s_t$, ou seja, uma variavél aleatoria X no tempo t aplicada em ω é igual ao estado nesse mesmo tempo t,

$$r_t(s,a) + E_s^{\pi}(v(X_t)) = r_t(s,a) + \sum_{s' \in S} p_t(s'|s,a)v(s')$$

Onde:

- $r_t(s,a)$ é a recompensa imediata ao realizar a ação a no estado s no período de tempo t,
- $p_t(s'|s,a)$ é a probabilidade de transição para o estado s' dado que a ação a é tomada no estado s,
- $v(s')$ é o valor esperado do estado s'.

Essa fórmula combina a recompensa imediata com a expectativa do valor do próximo estado, considerando todas as possíveis transições de estado sob a política π. Vale ressaltar que, neste ponto, ainda não definimos explicitamente as políticas; no entanto, esclareceremos esse conceito mais adiante no desenvolvimento do texto.

Recompensa Terminal ($r_N(s)$):

$$r_N(s) = R_N(s)$$

Onde em $r_N(s)$, nenhuma ação é tomada no momento N e a recompensa nesse ponto do tempo é só em função do estado s. Sendo crucial para avaliar o valor de um estado ao final do horizonte de tempo, refletindo a utilidade imediata do estado no momento em que o processo de decisão atinge seu término.

Quando o agente está formulando estratégias de decisão ou aprendendo a partir de experiências passadas, a recompensa terminal influencia suas escolhas. O agente pode preferir ações que levem a estados com recompensas terminais mais altas.

Em jogos, $r_N(s)$ poderia representar a pontuação final atingida por um jogador em um determinado estado do jogo ao final de N rodadas.

3.5 REGRAS DE DECIÕES E POLÍTICAS:

Regras de decisões: As regras de decisão referem-se ao conjunto de diretrizes ou critérios que um agente segue para escolher suas ações em diferentes estados do ambiente. Essas

regras podem ser expressas matematicamente por meio de funções ou algoritmos que mapeiam estados para ações, indicando qual ação o agente deve tomar em cada situação. Sendo a função das regras $f_t : S \rightarrow A_s$, para cada $s \in S$, $f_t(s) \in A_s$.

As regras de decisão são a base para a formulação de estratégias ou políticas, que orientam o comportamento do agente.

Políticas: Uma política é uma estratégia completa que especifica como um agente deve se comportar em todos os estados ou estados-ação possíveis do MDP. A política define a relação entre os estados e as ações, representando a estratégia geral adotada pelo agente para otimizar seu desempenho ao longo do tempo.

Podendo ser :

Política Determinística: Em uma política determinística, a ação escolhida em um determinado estado é fixa. Denotada por $\pi : S \rightarrow A$, onde S é o conjunto de estados e A é o conjunto de ações. A política determinística atribui uma ação específica a cada estado.

Política Estocástica: Em uma política estocástica, a escolha de ações em um estado é probabilística, ou seja, a política atribui probabilidades às diferentes ações em um determinado estado. Denotada por $\pi : S \times A \rightarrow [0, 1]$, especificando a probabilidade de escolher cada ação em cada estado. A política estocástica permite uma maior flexibilidade ao lidar com incertezas e ambiguidades no ambiente.

Política Estacionária: Uma política é considerada estacionária se a probabilidade de escolher uma ação em um estado não depender do tempo ou do passo de decisão. Em uma política estacionária, a distribuição de probabilidade sobre as ações permanece constante ao longo do tempo. Denotada por $\pi(a|s)$, onde a é a ação e s é o estado.

Em resumo, enquanto as regras de decisão são as orientações específicas para a escolha de ações em situações individuais, as políticas são estratégias mais abrangentes que representam as escolhas do agente para todos os estados ou estados-ação possíveis. As regras de decisão são componentes que contribuem para a construção de políticas em um contexto de MDP.

Na teoria das Cadeias de Markov, o símbolo π assume o papel central de representar a distribuição de equilíbrio, um conceito essencial quando se estuda o comportamento estocástico de sistemas dinâmicos. Por outro lado, em Processos de Decisão de Markov (MDPs), o símbolo π ganha um novo significado como uma representação de políticas.

4 UM PROBLEMA DE DECISÃO DE MARKOV DE UM PERÍODO

Um objetivo em um MDP de um período é tomar a melhor decisão no momento atual para maximizar ou minimizar a recompensa imediata, dependendo do critério de desempenho estabelecido. A complexidade é reduzida, pois não é necessário considerar futuros passos de tempo, tornando o problema mais direto de resolver.

Assumimos S e A finitos, e tomamos N=2 e $T = \{1, 2\}$.

Suponha que o tomador de decisão encontre o sistema no estado s no início do estágio 1 e seu objetivo é escolher uma ação $a \in A$ para maximizar a soma da recompensa imediata e a final. Se o tomador escolher uma política determinística $pi = (f_t)$ que seleciona a ação $a \in A$ na época de decisão 1, então a recompensa total esperada será :

$$r_1(s, a) + E_s^{\pi}(v(X_2)) = r_1(s, a) + \sum_{s' \in S} p_1(s'|s, a)v(s')$$

Como A é finito, então existe pelo menos uma ação maximizadora a*, Onde uma ação maximizadora refere-se a uma escolha de ação que resulta no maior valor possível para uma determinada expressão. podendo ser a única maximizadora ou não.

$$r_1(s, a*) + \sum_{s' \in S} p_1(s'|s, a*)v(s') = \max_{a' \in A_s}[r_1(s, a') + \sum_{s' \in S} p_1(s'|s, a')v(s')]$$

Temos também que o argumento que maximiza é dado por :

$$\arg\max g(x) = [x' \in X, g(x') \geq g(x)]$$

para $x \in X$, sendo g(x) uma função e X uma variável aleatória, Portanto:

$$a*_s \in \arg\max_{a' \in A_S}[r_1(s, a') + \sum_{s' \in S} p_1(s'|s, a')v(s')]$$

Suponha X=a,b,c,d, com g(a)= 5, g(b)= 7, g(c)=3 e g(d)=7

Então, $\max_{x \in X} g(x) = 7$ e $\arg\max_{x \in X} g(x) = [b, d]$

Quando o maximo não é alcançado, buscamos um supremo

Exemplo 4.1. Suponha que X=1, 2, 3 ... e g(x)= $1 - \frac{1}{x}$.

Se observarmos a função g(x), percebemos que ela se aproxima de 1 à medida que x cresce. No entanto, ela nunca atinge exatamente 1. Como x pode assumir valores infinitos, não há um valor máximo específico para g(x), portanto, o max não existe. Da mesma forma, não há um conjunto de argumentos que maximize g(x) (argmax= $\oslash$).

4.1 EXPLORAÇÃO COM ALEATORIEDADE EM MDPS

Será que o tomador de decisões pode obter uma recompensa maior usando aleatoriedade na seleção de ações em estados?

Seja $q(a)$ a probabilidade de selecionar a ação a em um estado s. A recompensa total esperada ao introduzir aleatoriedade na seleção de ações em um Modelo de Decisão de Markov (MDP) é dada por:

$$\sum_{a \in A_s} q(a) \left[r_t(s, a) + \sum_{s' \in S} p_t(s' | s, a) v(s') \right]$$

Onde:

- $q(a)$ é a probabilidade de selecionar a ação a em um determinado estado s,
- $\sum_{a \in A_s} q(a) = 1$, Isso assegura que q(a) seja uma distribuição de probabilidade sobre as ações em A_s.
- $q(a) \geq 0$ para $a \in A_s$. As probabilidades não podem ser negativas.

A introdução de aleatoriedade (q(a)) permite explorar diferentes ações com diferentes probabilidades, contribuindo para a descoberta de estratégias que podem levar a recompensas mais elevadas, especialmente em estados iniciais ou desconhecidos.

Agora, nosso objetivo é maximizar a recompensa para encontrar a melhor possível. Essa abordagem de maximização visa identificar a estratégia ou ação que proporciona a maior recompensa total esperada em um dado contexto. Ao otimizar a recompensa, buscamos efetivamente encontrar a política de decisão mais vantajosa ou a ação específica que resultará nos resultados mais favoráveis.

Portanto, encontrar a ação que maximize diretamente a recompensa pode ser mais eficiente e interpretável em comparação com a consideração de distribuições de probabilidade completas e também a maximização da recompensa leva à identificação de políticas de decisão mais eficazes, contribuindo para a tomada de decisões.

Então a maximização da expressão matemática

$$\max_{q \in P(A_s)} [\sum_{a \in A_S} q(a)[r_t(s,a) + \sum_{s' \in S} p_t(s'|s,a)v(s')]]$$

ou, de forma equivalente,

$$\max_{a' \in A_s} [r_t(s,a) + \sum_{s' \in S} p_t(s'|s,a)v(s')]$$

Observação: Se s não for conhecido antes de determinar a ação ótima, então o tomador de decisões deve escolher uma ação para cada $s \in S$ possível e deve então especificar uma regra de decisão, ou seja, Diante da incerteza sobre qual estado s será observado, é necessário especificar uma regra de decisão que determine qual ação tomar em cada estado.

Como $r_t(s,a) + \sum_{s' \in S} p_t(s'|s,a)v(s') \leq r_t(s,a_s^*) + \sum_{s' \in S} p_t(s'|s,a_s^*)v(s')$, para ações $a \in A(s)$ segue que:

$$\sum_{a \in A_S} q(a)[r_t(s,a) + \sum_{s' \in S} p_t(s'|s,a)v(s') \leq \sum_{a \in A_S} q(a)[r_t(s,a_s^*) + \sum_{s' \in S} p_t(s'|s,a_s^*)v(s')$$

$$= [r_t(s,a_s^*) + \sum_{s' \in S} p_t(s'|s,a_s^*)v(s')]. \sum_{a \in A_S} q(a)$$

Tendo que $\sum_{a \in A_S} q(a) = 1$:

$$= [r_t(s, a_s^*) + \sum_{s' \in S} p_t(s'|s, a_s^*) v(s')]$$

Note que a desigualdade acima, qual seja:

$$\sum_{a \in A_S} q(a)[r_t(s, a) + \sum_{s' \in S} p_t(s'|s, a) v(s') \leq [r_t(s, a_s^*) + \sum_{s' \in S} p_t(s'|s, a_s^*) v(s')$$

E independe da distribuição de probabilidade q.
Logo,

$$\sup_{q \in P(A_s)} \sum_{a \in A_S} q(a)[r_t(s, a) + \sum_{s' \in S} p_t(s'|s, a) v(s') \leq max_{a* \in A_s} \{r_t(s, a_s^*) + \sum_{s' \in S} p_t(s'|s, a_s^*) v(s')\}$$

Na verdade temos uma igualdade, para ver isto basta lembrar que:

$$q^*(a) = \begin{cases} 1 & \text{se } a = a_s^*, \\ 0 & \text{se caso contrário} \end{cases}$$ É um elemento de $P(A_s)$

Essa regra de decisão ótima a_s^* para cada estado s é a ação que maximiza a recompensa total esperada, considerando a recompensa imediata e as probabilidades de transição para os estados futuros .Portanto, a escolha da ação a_s^* para cada estado s de acordo com a expressão fornecida irá resultar na maior recompensa total esperada, e essa é a base para encontrar regras de decisão ótimas em um MDP.

Ao abordar essa maximização, a introdução de elementos probabilísticos nas escolhas de ações permite uma exploração mais ampla do espaço de estados, contribuindo para a adaptabilidade e aprendizado do tomador de decisões. Entretanto, a análise crítica dessa expressão é essencial para equilibrar a exploração e a exploração de maneira eficaz, garantindo que a aleatoriedade nas escolhas de ações leve a uma otimização real do desempenho e à maximização das recompensas no contexto específico do problema em questão.

5 EXEMPLOS

Neste capítulo, apresentaremos dois exemplos ilustrativos com base nas informações que acumulamos até o momento, retirados do livro de Puterman. O primeiro exemplo será teórico, envolvendo apenas dois estados para proporcionar uma compreensão clara dos conceitos abordados. Já o segundo exemplo abordará um cenário prático, explorando o controle estocástico de inventário para um único produto. Esses exemplos fornecerão insights valiosos sobre a aplicação dos princípios de Modelos de Decisão de Markov (MDP) e estratégias estocásticas em situações reais. Ao analisar casos simples e mais complexos, buscaremos consolidar os fundamentos teóricos e sua aplicação prática, preparando para abordagens mais avançadas nos próximos capítulos.

5.1 UM PROCESSO DE DECISÃO DE MARKOV DE DOIS ESTADOS

Um Processo de Decisão de Markov (MDP) de dois estados refere-se a um modelo estocástico no qual um agente toma decisões sequenciais em um ambiente composto por dois estados distintos, s_1 ou s_2. A natureza markoviana do processo implica que a escolha de ações e as recompensas associadas dependem tanto do estado atual quanto do próximo estado, tornando o modelo sensível à dinâmica temporal.

Será apresentado um exemplo de MDP de dois estados, no qual assumimos recompensas e probabilidades de transição estacionárias, indicando que esses parâmetros não variam com o tempo. Nesse contexto, a recompensa associada dependerá tanto do próximo estado quanto do estado atual, caracterizando a situação como markoviana.

No contexto deste MDP específico, as épocas de decisão são representadas por $T = [1, 2, ..., N]$, onde $N \leq \infty$. Os estados possíveis são s_1 e s_2, com ações associadas $A_{s_1} = [a_{1,1}, a_{1,2}]$ e $A_{s_2} = [a_{2,1}]$. As recompensas associadas a cada ação e estado são definidas de forma estacionária, com $r_N(s_1) = 0$ e $r_N(s_2) = 0$ para $N < \infty$.

Além disso, as probabilidades de transição entre os estados variam de acordo com as ações tomadas. Essa variação é expressa por p_t para diferentes combinações de estado e ação, como $p_t(s_1|s_1, a_{1,1})$, $p_t(s_2|s_1, a_{1,1})$, entre outros.

Essa estrutura analítica do MDP de dois estados proporciona uma compreensão profunda e modelagem eficaz de situações onde a tomada de decisão sequencial é influenciada por fatores

estocásticos. Tal modelo se destaca em diversos contextos, será apresentado um exemplo também relacionado ao cotidiado de uma pessoa com diabetes.

A seguir, apresentamos as principais características do exemplo em termos de notação:

- Épocas de decisão: T=[1,2,...N], N$\leq \infty$
- Estados: S=[s_1, s_2]
- Ações: A_{s_1}=[$a_{1,1}$, $a_{1,2}$] , A_{s_2}=[$a_{2,1}$]
- As recompensas associadas a cada ação e estado são definidas como:
  $r_t(s_1, a_{1,1}) = 5$, $r_t(s_1, a_{1,2}) = 10$, $r_t(s_2, a_{2,1}) = -1$, $r_N(s_1) = 0$, $r_N(s_2) = 0$ com N $< \infty$.
- As probabilidades de transição são definidas como:
  $p_t(s_1|s_1, a_{1,1}) = 0,5$, $p_t(s_2|s_1, a_{1,1}) = 0,5$
  $p_t(s_1|s_1, a_{1,2}) = 0$, $p_t(s_2|s_1, a_{1,1}) = 1$
  $p_t(s_1|s_1, a_{2,1}) = 0$, $p_t(s_2|s_2, a_{2,1}) = 1$

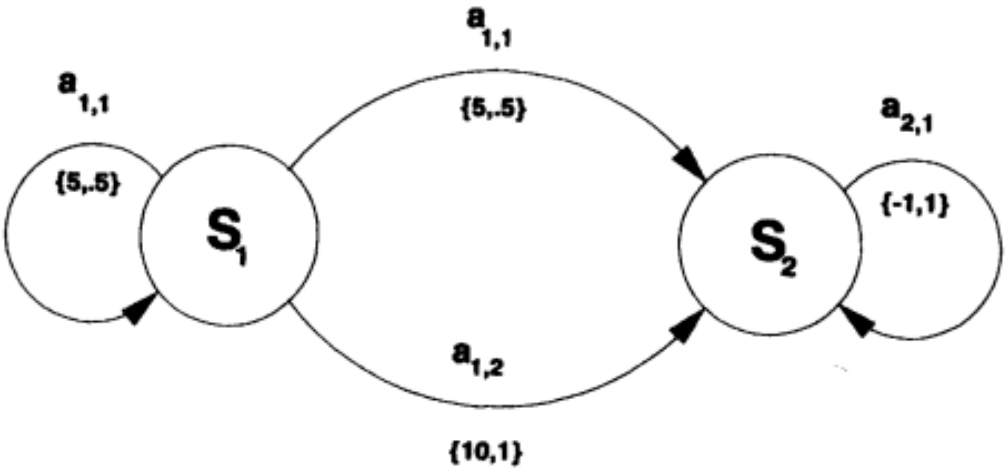

Figura 5.1 – Fonte: Markov Decision Processes : Discrete Stochasstic Dynamic Programming- Martin L. Puterman

Com este exemplo, podemos trazer para o cotidiano e relacionar a dinâmica de controle de glicose e administração de insulina com a gestão diária de pacientes com diabetes. A decisão de administrar ou não insulina em períodos regulares reflete o desafio enfrentado por indivíduos que precisam manter seus níveis de glicose dentro de uma faixa alvo.

Imagine um paciente sujeito a restrições alimentares que, em cada momento de medição de glicose, deve decidir se administrará um medicamento. Essa escolha é análoga às ações a_1)

(administrar insulina) e a_2 (não administrar insulina) no exemplo. As recompensas associadas a essas decisões refletem as consequências para o controle glicêmico do paciente.

Épocas de decisão: Cada período de medição da glicose é uma época de decisão ($T = [1, 2, \ldots, N]$).

Estados:

- s_1 representa um estado em que a glicose está dentro da faixa alvo.
- s_2 representa um estado em que a glicose está fora da faixa alvo.

Ações:

- $A_{s_1} = [a_1, a_2]$
- $A_{s_2} = [a_1]$

Recompensas:

$r_t(s_1, a_1) = -5$ (penalidade por administrar insulina quando a glicose está na faixa alvo, pode ser considerado excesso de precaução.)
$r_t(s_1, a_2) = 2$ (recompensa por não administrar insulina quando a glicose está na faixa alvo, indica uma decisão adequada)
$r_t(s_2, a_1) = 10$ (recompensa por administrar insulina quando a glicose está fora da faixa alvo, pois é uma ação benéfica para corrigir os níveis elevados de glicose.)
$r_N(s_1) = 0$ (recompensa final para s_1)
$r_N(s_2) = 0$ (recompensa final para s_2) com $N < \infty$.

Probabilidades de Transição:

$p_t(s_1|s_1, a_1) = 0.8$ (probabilidade de permanecer na faixa alvo ao administrar insulina)
$p_t(s_2|s_1, a_1) = 0.2$ (probabilidade de sair da faixa alvo ao administrar insulina)
$p_t(s_1|s_1, a_2) = 0.9$ (probabilidade de permanecer na faixa alvo sem administrar insulina)
$p_t(s_2|s_1, a_2) = 0.1$ (probabilidade de sair da faixa alvo sem administrar insulina)
$p_t(s_1|s_2, a_1) = 0.3$ (probabilidade de voltar à faixa alvo ao administrar insulina)
$p_t(s_2|s_2, a_1) = 0.7$ (probabilidade de permanecer fora da faixa alvo ao administrar insulina)

A analogia com a gestão diária de pacientes com diabetes destaca a complexidade dessas decisões e a importância de políticas de administração de insulina bem ajustadas. Pacientes e profissionais de saúde enfrentam o desafio de encontrar o equilíbrio certo entre evitar complicações associadas à hiperglicemia e minimizar os riscos relacionados ao uso excessivo de insulina.

A aplicação de conceitos de MDP a casos práticos como esse ressalta a utilidade dessa estrutura para modelar e compreender processos de decisão sequencial em ambientes estocásticos.Em suma, o exemplo destaca a importância da adaptação dinâmica de políticas de decisão para lidar com a variabilidade inerente a condições de saúde, demonstrando como os princípios dos MDPs podem ser aplicados para melhorar a tomada de decisões em situações cotidians e dinâmicas.

5.2 CONTROLE ESTOCÁSTICO DE INVENTÁRIO DE PRODUTO ÚNICO

No cenário empresarial, a gestão eficiente de estoques desempenha um papel crucial no equilíbrio entre atender à demanda do cliente e otimizar os custos associados ao armazenamento. Vamos considerar um exemplo prático em que um gerente, a cada mês, avalia o estoque atual de um determinado produto e toma a decisão estratégica de solicitar ou não a reposição desse estoque. Nesse contexto, diversos fatores entram em jogo, incluindo custos de armazenamento e a capacidade de satisfazer a demanda do cliente.

A dinâmica desse processo de decisão é modelada através de um Problema de Decisão de Markov (MDP), onde a demanda pelo produto é uma variável aleatória com uma distribuição de probabilidade conhecida. O objetivo fundamental é maximizar a medida de lucro ao longo do tempo, considerando as implicações financeiras das decisões de estoque. Este exemplo exemplifica a complexidade envolvida na gestão de inventário, onde as incertezas relacionadas à demanda e os custos de armazenamento influenciam diretamente as escolhas do gerente.

No contexto da gestão de inventário abordado, algumas suposições fundamentais são estabelecidas para simplificar e estruturar o problema. Estas suposições são cruciais para moldar o modelo de MPD que será utilizado na análise e otimização das estratégias de reposição de estoque. Vamos explorar essas suposições em detalhes:

Encomenda e Entrega:

1. A encomenda de estoque adicional é realizada no início de cada mês.

2. A entrega ocorre instantaneamente após a realização do pedido.

Atendimento de Pedidos:

Todos os pedidos são atendidos no último dia do mês. Essa sincronização simplifica o processo de gestão de estoque, permitindo uma abordagem mais uniforme nas decisões mensais.

Demanda e Acúmulo: Caso o produto não esteja disponível em estoque, os clientes procuram outras fontes, evitando a acumulação de demanda não atendida.

Estabilidade nos Parâmetros: As receitas, os custos associados e a distribuição da demanda permanecem constantes de um mês para outro. Essa estabilidade simplifica a modelagem, assumindo que as condições econômicas e operacionais permanecem inalteradas no curto prazo.

Unidades Inteiras: O produto é vendido apenas em unidades inteiras, o que facilita o controle do estoque e evita a complexidade associada a frações de produtos.

Capacidade de Armazenamento: O armazém tem uma capacidade máxima de armazenamento limitada a m unidades. Essa restrição ajuda a definir claramente os limites físicos da capacidade de estocagem.

Variáveis Temporais:

- t: Representa o mês ou a época de decisão.

Variáveis de Decisão:

- a_t: Número de unidades encomendadas no mês t.
- D_t: Demanda aleatória no mês t.
- A distribuição de probabilidade da demanda é $p_j = P\{D_t = j\}$ com $j = 0, 1, 2, ...$

Equação de Atualização do Estoque:

$$s_{t+1} = \max\{s_t + a_t - D_t, 0\} = (s_t + a_t - D_t)^+$$

A equação modela a atualização do estoque no próximo período. O termo $(s_t+a_t-D_t)^+$ representa a função de máximo, garantindo que o nível de estoque não seja negativo. Se $s_t + a_t - D_t < 0$, isso significa que a demanda excede a quantidade disponível em estoque, e o estoque na próxima época é definido como zero para evitar valores negativos.

Observação: O fato de o nível de estoque não poder ser negativo reflete a restrição física de que não é possível ter um estoque negativo. Quando a demanda é maior do que a quantidade disponível em estoque, o estoque é zerado, e não ocorre um saldo negativo.

Custo de Encomenda

$$O(u) = \begin{cases} K + X(u) & \text{se } u > 0, \\ 0 & \text{se } u = 0. \end{cases}$$

K representa o custo fixo de fazer um pedido, e C(u) é o custo variável associado à quantidade encomendada. A função O(u) modela o custo total de fazer um pedido, considerando tanto o custo fixo quanto o custo variável. Se nenhuma unidade for encomendada u = 0 o custo é zero.

Custo de Manutenção de Estoque

h(u) é uma função não decrescente ao longo do horizonte finito, e g(u) na última época.

A h(u) modela o custo de manutenção do estoque, que aumenta com a quantidade de unidades em estoque. A função g(u) é o custo de manutenção na última época. Essa estrutura permite representar custos associados ao armazenamento e gestão de estoque ao longo do tempo.

Recebimento Associado à Demanda = f(j)

assuminfo f(0) = 0.

A função f(j) modela o valor recebido pelo gestor quando a demanda é j unidades. Se nenhuma unidade for demandada j = 0 o valor recebido é zero. Essa função representa o ganho associado à satisfação da demanda do cliente.

Recompensa na Época de Decisão

$r_t(s_t, a_t, s_{t+1}) = -O(a_t) - h(s_t + a_t) + f(s_t + a_t - s_{t+1})$.

A recompensa depende do estado do sistema na época de decisão. Ela é composta pelo negativo do custo de encomenda $O(a_t)$ o negativo do custo de manutenção $h(s_t + a_t)$ e o valor recebido pelo gestor $f(s_t + a_t - s_{t+1})$. Essa formulação reflete os impactos financeiros das decisões de encomenda, manutenção de estoque e recebimento associado à demanda.

Se a quantidade disponível em estoque for maior do que a demanda, o gestor obtém uma receita igual a f(j) com uma probabilidade p_j. Isso reflete a situação em que o estoque é suficiente para atender à demanda, resultando na receita associada.

Se a demanda exceder o estoque, o valor da receita será f(u) com probabilidade $q_u = \sum_{j=u}^{\infty} P_j$.

Então $F(u) = \sum_{j=0}^{u-1} f(j)p_j + f(u)q_u$, esse é o valor da receita.

Explicação: Se a demanda for maior do que a quantidade disponível em estoque, o gestor obtém uma receita igual a f(u) com uma probabilidade q_u . Isso reflete a situação em que a demanda não pode ser totalmente atendida devido à escassez de estoque, resultando na receita associada. O termo q_u representa a probabilidade cumulativa de a demanda ser maior ou igual a u . A expressão $f(u) = \sum_{j=0}^{u-1} f(j)p_j + f(u)q_u$ representa a receita total, que é a soma ponderada das receitas para cada quantidade de demanda possível, considerando as probabilidades associadas.

Agora será destacado as principais características do exemplo em relação à sua notação:

- Épocas de decisão: T={1,2,...N}, N$\leq \infty$
- Estados: S={0, 1, 2..., M}, será a quantidade de estoque disponível no início do mês
- Ações: $A_s = \{0, 1, 2...M - s\}$, é a quantidade de unidades de produto adicional a se encomendar no mês
- A recompensa esperada será a receita esperada menos os custs de pedido e manutençaõ
  $r_t(s, a) = F(s + a) - O(a) - h(s + a)$ com $r_N(s) = g(s)$
- As probabilidades de transição são definidas como:
  $$p_s(s'|s, a) = \begin{cases} 0 & \text{se } M \geq s' > s + a, \\ p_s + a - s' & \text{se } M \geq s + a \geq s' > 0 \\ q_{s+a} & \text{se} M \geq s + a \text{e} s' = 0 \end{cases}$$
  Se $s' > s + a$, isso significa que o estoque disponível no próximo período seria insuficiente para atender à demanda, resultando em uma transição com probabilidade zero.

  Se $M \geq s + a \geq s' > 0$, a probabilidade de transição $p_s + a - s'$ indica a chance de que, após atender à demanda ($s + a$), o estoque disponível (s') seja suficiente para o próximo período.

Se $M \geq s + a$ e $s' = 0$, a probabilidade de transição q_{s+a} indica a chance de que, após atender à demanda ($s + a$), o estoque disponível seja zerado, representando um novo ciclo no sistema.

– Regra de decisão: Solicite estoque suficiente para aumentar o estoque para P unidades sempre que o nível de estoque no início de um mês for menor que σ unidades. Quando o nível de estoque no início do mês for σ ou superior, não faça pedido.

$$d_t(s) = \begin{cases} \sum -s & \text{se } s < \sigma, \\ 0 & \text{se } s \geq \sigma \end{cases}$$

Se o nível de estoque for menor que σ unidades ($s < \sigma$): Nesse caso, a decisão é solicitar estoque suficiente para aumentar o estoque para P unidades. Isso implica que, quando o estoque estiver abaixo do limiar definido por u unidades, um pedido de reposição é feito para garantir que o estoque atinja P unidades.

Se o nível de estoque for igual ou maior que u unidades ($s \geq \sigma$): Quando o nível de estoque no início do mês for igual ou superior a u unidades, a decisão é não fazer um pedido adicional. Nesse caso, a função de custo para realizar um pedido é zero.

A política de reposição de estoque, expressa por meio da função $d_t(s)$, destaca a importância de solicitar estoque de forma estratégica, visando otimizar os custos associados e garantir um equilíbrio entre a oferta e a demanda. A aplicação da política (σ, $\sum$), onde σ representa o limiar mínimo e $\sum$ é o estoque total desejado, demonstra uma abordagem criteriosa para a tomada de decisões.

Em suma, o exemplo de controle de estoque oferece insights valiosos sobre a complexidade da gestão de inventário, destacando a necessidade de estratégias bem elaboradas para minimizar custos, atender à demanda e manter um equilíbrio financeiro sustentável ao longo do tempo.

6 HORIZONTE FINITO

Neste capítulo, exploraremos a aplicação de MDPs com um foco específico em processos com leis de transição finitas e horizontes finitos. Nosso objetivo é compreender como as decisões sequenciais evoluem ao longo de um horizonte limitado, considerando as incertezas associadas às transições de estado.

Abordaremos o conceito de horizonte finito em MDPs, analisando como as decisões impactam as recompensas ao longo do tempo. Além disso, exploraremos processos com leis de transição finitas, destacando a relevância dessa abordagem em cenários específicos.

Para ilustrar esses conceitos na prática, utilizaremos um exemplo prático relacionado ao campo do marketing, extraído do livro "Dynamic Optimization: Deterministic and Stochastic Models"de Karl Hinderer, Ulrich Rieder e Michael Stieglitz. Ao longo do capítulo, examinaremos como as decisões de marketing sequenciais podem ser otimizadas considerando um horizonte de tempo limitado e transições de estado finitas. Essa abordagem permitirá uma compreensão mais aprofundada de como as estratégias de marketing podem ser ajustadas ao longo do tempo para maximizar as recompensas em contextos dinâmicos e incertos.

Exemplo 6.1. (Um problema de marketing): Uma empresa vende um determinado produto. No início de cada período, digamos um mês, a situação de mercado para o produto é avaliada e classificada como estando em algum estado s em um conjunto finito S. Em seguida, a empresa decide sobre uma ação apropriada para promover as vendas, como uma campanha publicitária. Assumimos que, com base na experiência anterior, existem estimativas razoavelmente precisas para as probabilidades $p(s'; a; s_0)$ de que o estado s será transformado no estado s_0 sob a ação a. Além disso, assumimos que (i) a recompensa r (custos de promoção deduzidos) para o período t é uma função do estado s_t e da ação a_t, e (ii) há uma recompensa terminal (por exemplo, o valor residual ao vender o equipamento de produção e os direitos de produção para outra empresa) $V_0(s_N)$ quando a empresa interrompe as vendas no tempo N no estado s_N. Qual política maximiza a recompensa total esperada para o estado inicial s_0?

Este exemplo não apenas ilustra a aplicação prática da teoria de otimização dinâmica, mas também destaca a relevância desses modelos em cenários complexos, como o ambiente volátil do marketing. Vamos agora explorar e analisar as definições :

Definição 6.1. Um processo de Markov com espaços de estados finito é uma tupla que contém S, A, D, P, p, r_s, V_o, β.Onde:

- S é o conjunto finito de estados possíveis;
- Onde D é o conjunto de todas as ações que podem ser atribuidas a todos os estados, antes tinhamos A_s que seria o conjunto de ações para um estado especifico s;
- Se D(s)=A , eliminamos S da tupla, ou seja, se as ações forem as mesma para todos os estados;
- P é uma matriz de transição de D para S,indicando as probabilidades de transição de um estado para outro, governada pela função $(s, a, s') \longrightarrow p(s, a, s')$ para D X S em $\mathbb{R}_+$ tal que

  $$\sum_{s' \in S} p(s, a, s') = 1$$

  para s,a $\in D$;
- Sendo assim p é chamada de Lei de Transição, uma função que associa a probabilidade de transição entre estados para cada ação tomada;
- r_s: D X S $\longrightarrow \mathbb{R}$, senod uma função de recompensa de 1 estágio. Se $r_s(s, a, s')$ é independente de s', substituimos r_s na tupla por r.
- V_0 pode se referir à função de valor final, representando os valores associados a cada estado no último período do horizonte finito.
- βé o parâmetro de desconto, influenciando a importância atribuída a recompensas futuras em relação às recompensas imediatas.

Essa definição esclarece de maneira abrangente os elementos essenciais de um processo de Markov, detalhando a organização e operação de ações, transições de estado, recompensas e outros parâmetros. Ao oferecer uma visão abrangente, ela estabelece uma base sólida que possibilita a compreensão e modelagem desses processos, especialmente em contextos que envolvem espaços de estados finitos.

No contexto da exploração de processos de Markov com horizonte finito, uma característica notável é a homogeneidade da matriz de transição. Ao contrário das cadeias de Markov previamente estudadas, nas quais a matriz de transição permanecia constante ao longo do tempo, nos processos com horizonte finito, essa matriz torna-se dinâmica, variando conforme o tempo avança.

A homogeneidade, nesse cenário, implica que as probabilidades de transição entre estados podem alterar-se de período para período, refletindo a dinâmica temporal do processo. Esse aspecto adiciona uma camada de complexidade e realismo aos modelos, permitindo capturar variações e adaptações nas transições de estado ao longo do tempo. Portanto, ao considerar horizontes finitos, ganhamos a flexibilidade de modelar de maneira mais precisa os processos de Markov em cenários dinâmicos e em constante mudança.

Definição 6.2. Dado que para cada $s_0 \in S$ e para cada política π_t^{N-1}, existe um MDP correspondente a uma cadeia de Markov $\zeta^N = (\zeta_t)_1^N$ em algum espaço de probabilidade $(\Omega, \mathcal{R}_{\pi_{s_0}})$ com espaço de estado S, começando em s_0, e com matriz de transição dada por $p_{\pi_t} = (p(s, \pi_t(s), s'))$ para $s, s' \in S$ e $0 \leq t \leq N-1$.

Essa expressão destaca que, para cada estado inicial s_0 em um conjunto finito de estados S e para cada política π_t^{N-1} em um horizonte de decisão finito N, existe um Processo de Decisão de Markov (MDP) correspondente a uma cadeia de Markov $\zeta^N = (\zeta_t)_1^N$. Essa cadeia está definida em algum espaço de probabilidade $(\Omega, \mathcal{R}_{\pi_{s_0}})$, onde Ω é o espaço de amostras e $\mathcal{R}_{\pi_{s_0}}$ é a σ-álgebra associada à política π_{s_0}.

A cadeia de Markov começa em um estado inicial s_0 e evolui ao longo do tempo até o horizonte de decisão N. A matriz de transição p_{π_t} descreve as probabilidades de transição entre os estados s e s' para todas as ações possíveis $\pi_t(s)$ em cada período de decisão t, onde $0 \leq t \leq N-1$.

A expressão que descreve a densidade discreta do processo de decisão ζ_N, indicando como a probabilidade de transição entre os estados evolui ao longo do tempo, é dada por:

$$\zeta_N = \prod_{t=0}^{N-1} \pi_t(s_t, s_{t+1}) p(s_t, s_{t+1}) s^N \rightarrow p_\pi(s_0, s^N) = p_{\pi_0}(s_0, s_1) \cdot p_{\pi_1}(s_1, s_2) \cdot \ldots \cdot p_{\pi_{N-1}}(s_{N-1}, s_N) \tag{6.1}$$

onde s_N representa uma sequência de estados $(s_0, s_1, \ldots, s_N)$ no processo.

A densidade discreta $\pi_t(s_t, s_{t+1}) p(s_t, s_{t+1})$ é decomposta como o produto das probabilidades de transição $\pi_t(s_t, s_{t+1})$ e $p(s_t, s_{t+1})$ para cada par de estados consecutivos (s_t, s_{t+1}), onde t varia de 0 a $N-1$. Cada termo $\pi_t(s_t, s_{t+1}) p(s_t, s_{t+1})$ representa a probabilidade de transição do estado s_t para o estado s_{t+1} no período de decisão t, sob a política π_t.

Na sequência, frequentemente denotamos a sequência de estados $s_1, s_2, \ldots, s_N$ por s^N. A razão

mais profunda pela qual pudemos resolver MDPS de forma sequencial é o fato de que, para um estado inicial fixo s_0 e uma política π, a sequência de estados $X_t = \zeta_{t\pi}^{s_0}$, para $1 \leq t \leq N$, forma uma cadeia de Markov no seguinte sentido:

Definição 6.3. Seja S finito, $s_0 \in S$, e seja $.p_t(s, s')$; $s, s' \in S$, $0 \leq t \leq N-1$, uma sequência de matrizes estocásticas $S \times S$. Uma sequência $.(X_t)_1^N$ de variáveis aleatórias com valores em S, definida em um espaço de probabilidade finito arbitrário $(\Omega; P)$, é considerada uma cadeia de Markov não homogênea de N estágios com espaço de estados S, estado inicial s_0 e matrizes de transição $(p_t)_0^{N-1}$ se a densidade discreta $s^N \mapsto P(.(X_t)_1^N = s^N)$ de $.(X_t)_1^N$ é igual a:

$$p_0(s_0, s_1) \cdot p_1(s_1, s_2) \cdot ... \cdot p_{N-1}(s_{N-1}, s_N) \quad ; \quad s_N \in S^N \tag{6.2}$$

Construímos com base no seguinte resultado simples de existência: Para um dado s_0 e $(p_t)_0^{N-1}$ sempre existe um espaço de probabilidade $(.\Omega; P)$ e, sobre ele, um vetor aleatório $.(X_t)_1^N$ que é uma cadeia de Markov com espaço de estados S, estado inicial s_0 e matrizes de transição p_t.

Essa definição introduz o conceito de uma cadeia de Markov não homogênea de N estágios em um espaço de estados finito S.

Matrizes Estocásticas ($p_t(s, s')$): Para $0 \leq t \leq N-1$, temos uma sequência de matrizes estocásticas $S \times S$, representadas por $p_t(s, s')$. Essas matrizes indicam as probabilidades de transição entre os estados s e s' no período t.

Cadeia de Markov não Homogênea de N Estágios: A sequência de variáveis aleatórias $(X_t)_1^N$, onde cada X_t é uma variável aleatória com valores em S, forma uma cadeia de Markov não homogênea de N estágios. Isso significa que a probabilidade de transição entre os estados depende do período t, conforme especificado pelas matrizes $p_t(s, s')$.

Densidade Discreta ($s^N \mapsto P((X_t)_1^N = s^N)$): A densidade discreta descreve a probabilidade de observar uma sequência específica de estados $(X_t)_1^N$ no espaço de estados S^N. A expressão $\mapsto$ denota a função que associa uma sequência de estados à sua probabilidade.

Agora, considere o problema com um espaço de estados finito S. Demonstramos facilmente que, para qualquer política $\pi = \{\pi_t\}_0^{N-1}$, a densidade discreta da sequência $(X_t)_1^N$ de estados aleatórios satisfaz (6.2) com:

$$p_{\pi t}(s, s') = P(T(s, \pi_t(s), \eta_1) = s'), 0 \leq t \leq N-1, \quad s, s' \in S \tag{6.3}$$

a expressão (6.3) está relacionada a uma perturbação, onde η_1 representa uma perturbação ou distúrbio. Essa perturbação pode ser interpretada como uma font

Construção Canônica: A definição menciona uma construção canônica para a cadeia de Markov, indicando que existe um espaço de probabilidade $(\Omega; P)$ sobre o qual é definido um vetor aleatório $(X_t)_1^N$ que segue as propriedades especificadas.

Em resumo, essa definição estabelece as bases para uma cadeia de Markov não homogênea de N estágios, especificando as matrizes de transição $p_t(s, s')$ e a densidade discreta associada a essa cadeia. Ela destaca a natureza dependente do tempo das transições entre os estados.

6.1 RECOMPENSAS EM HORIZONTE FINITO

Nesta seção, exploraremos conceitos fundamentais relacionados a processos de decisão sequencial em horizontes finitos, focando especialmente nas recompensas aleatórias e na maximização da recompensa esperada. Abordaremos a expressão de Bellman, Ao longo do estudo, discutiremos a utilidade como uma medida de valor e a interação entre recompensas e para ilustrar esses conceitos de forma prática, apresentaremos um exemplo envolvendo a venda de um produto com promoção em um horizonte finito.

Se o tomador de decisão escolher $\pi = (\pi_t)^{N-1}$, ele obtém ao iniciar em $s_0 = \zeta_0$ a recompensa aleatória de N estágios:

$$R_N^{\pi}(s_0, \zeta^N) = \sum_{t=0}^{N-1} \beta^t \cdot r_s(\zeta_t, \pi_t(\zeta_t), \zeta_{t+1}) + \beta^N \cdot V_0(\zeta_N)$$

Se (Ω, P_π, s_0) é canônico, obtemos:

$$R_N^{\pi}(s_0, s^N) = \sum_{t=0}^{N-1} \beta^t \cdot r_s(s_t, \pi_t(s_t), s_{t+1}) + \beta^N \cdot V_0(s_N)$$

A principal diferença entre essas duas formulações reside no tipo de variáveis utilizadas para representar os estados ao longo dos estágios. Na primeira expressão, os estados são tratados como variáveis aleatórias (ζ_t), o que implica em uma certa dose de incerteza ou aleatoriedade associada a cada estágio do processo. Por outro lado, na segunda expressão, os estados são considerados determinísticos (s_t), sugerindo uma ausência de aleatoriedade nos estados do

processo.

A recompensa total $R_N^{\pi}(s_0, s^N)$ é calculada a partir do estágio inicial $s_0 = s_0^N$. Essa expressão envolve uma soma que percorre cada estágio de 0 a $N-1$. Em cada estágio t, a recompensa instantânea é dada por $r_s(s_t, \pi_t(s_t), s_{t+1})$, representando a recompensa obtida ao realizar a ação π_t no estado s_t e transitar para o próximo estado s_{t+1}. Cada termo da soma é ponderado por β^t, onde β é um fator de desconto, refletindo a preferência por recompensas imediatas em relação a futuras. O último termo na soma $\beta^N \cdot V_0(s_N)$ representa a recompensa associada ao último estágio (N), onde $V_0(s_N)$ é a recompensa terminal no estágio N para o estado s_N. Essa expressão modela a acumulação de recompensas ao longo de N estágios, considerando a dinâmica das ações e estados determinados pelo processo estocástico.

Exemplo 6.2. Suponha que uma empresa possui uma máquina de produção e está tomando decisões ao longo de N períodos sobre como investir em melhorias, manutenção e eficiência da máquina. Cada estágio t envolve escolher uma estratégia de investimento π_t que pode incluir reparos, atualizações tecnológicas, entre outros.

A recompensa instantânea $r_s(s_t, \pi_t(s_t), s_{t+1})$ em cada estágio seria a contribuição específica da estratégia de investimento escolhida para o desempenho operacional da máquina. Isso poderia incluir melhorias como um aumento na produtividade, redução de custos, produção eficiente de produtos e expansão da capacidade de produção.

O valor final $V_0(s_N)$ representaria a recompensa associada à venda da máquina ao final dos N períodos, levando em consideração o sucesso operacional e o lucro acumulado ao longo do tempo.

Para cada $s_o \in S$ o problema máximo MDPNs consiste em maximizar a recompensa esperada em N-estágios:

$$V_{N\pi}(s_0) = \mathbb{E}_{\pi, s_0}\left[R_{N\pi}(s_0, s^N)\right] = \sum_{s^N \in S^N} R_{N\pi}(s_0, s^N) \cdot p_{\pi}(s_0, s^N) \tag{6.4}$$

A função $V_{N\pi}(s_0)$ representa a recompensa esperada acumulada ao longo de N estágios, iniciando no estado inicial s_0, quando seguimos uma política específica π. Em outras palavras, ela quantifica o valor médio da soma das recompensas ao longo do tempo, considerando as ações determinadas pela política e as probilidades de transições sobre a mesma política e com isso é possível avaliar a qualidade dessa política específica.

A recompensa máxima esperada no estado s_0 após N estágios é dada por:

$$V_N(s_0) = \sup\{ V_N^{\pi}(s_0) : \pi \in \mathcal{F}^N\} \tag{6.5}$$

com $-\infty < V_N(s_0) \leq \infty$.

Aqui, o conjunto $\mathcal{F}^N$ engloba todas as políticas possíveis para o referido horizonte temporal. Com isso, o operador sup busca o supremo, ou seja, o valor máximo dentre todas as recompensas esperadas $\mathcal{R}_N^{\pi}(s_0)$ para diferentes políticas π pertencentes a $\mathcal{F}^N$. A condição $-\infty < \mathcal{R}_N(s_0) \leq \infty$ estabelece que a recompensa máxima esperada está sempre limitada entre menos infinito e mais infinito, refletindo a natureza finita ou infinita das recompensas nesse contexto.

A partir desse momento, a notação s será utilizada em substituição a s_0, indicando que o processo pode iniciar a partir de qualquer estado, não necessariamente o estado inicial s_0 . Isso significa que s representará o estado inicial em contextos futuros, simplificando a expressão e tornando a leitura mais concisa, independentemente do estado inicial específico escolhido. Essa alteração não afeta o significado ou a interpretação do que foi discutido anteriormente, apenas oferece uma forma mais prática de representar o estado inicial nas formulações subsequentes, permitindo flexibilidade quanto ao ponto de partida do processo.

Relembrando (3.1):

$$r_t(s, a) = \sum_{s' \in S} r_t(s, a, s').p_t(s'|s, a)$$

Temos

$$-r_f(s) = r_f(s, f(s))$$

$$-p_f(s, s') = p(s, f(s), s')$$

com $f \in \mathcal{F}$ A notação $\mathcal{F}$ refere-se ao conjunto de todas as políticas possíveis, enquanto f indica uma sequência de decisões. As expressões $r_f(s) = r_f(s, f(s))$ e $p_f(s, s') = p(s, f(s), s')$ introduzem funções f que atuam sobre os estados. No primeiro caso, $r_f(s)$ representa a recompensa no estado s após a aplicação da função f, que pode modificar ou transformar o estado original.

Similarmente, $p_f(s, s')$ denota a probabilidade de transição de s para s' após a aplicação de f ao estado s. A sequência de decisões f pode envolver diferentes políticas ao longo do tempo, e $\mathcal{F}$ é

o conjunto que engloba todas essas políticas possíveis.

Agora usamos os operadores L, U_feU, que são definidos da seguinte maneira:

- $L_v(s,a) = r(s,a) + \beta \sum_{s' \in S} p(s' \mid s,a) \cdot v(s')$ e é usada para descrever a equação de Bellman para a função de valor v em um MDP. A expressão relaciona o valor de um estado s tomando uma ação a ao valor esperado dos estados subsequentes e recompensas.
- $U_{fv}(s) = L_v(s, f(s) = r_f(s) + \beta \sum_{s' \in S} p_f(s,s').v(s')$, reflete a utilidade cumulativa de tomar uma sequências de decisões $f(s)$ no estado s, considerando tanto a recompensa imediata quanto as recompensas futuras descontadas.
- $U_v(s) = \sup_{a \in D(s)} L_v(s,a) = \sup_{f \in \mathcal{F}} U_{fv}$, a igualdade enfatiza que a função de utilidade $U_v(s)$ é obtida maximizando sobre todas as ações possíveis $a \in D(s)$, refletindo a busca pela melhor ação para otimizar o valor acumulado. A segunda parte da igualdade introduz a ideia de que a função de utilidade também pode ser expressa como o supremo sobre todas as sequências de decisões possíveis f no conjunto $\mathcal{F}$

Na abordagem de minimização, empregamos a mesma notação L, U_feU no entanto, com algumas modificações. A função de recompensa r_s é substituída pela função de custo c_s) e o operador sup é substituído por inf. Dessa forma, a notação é adaptada para lidar com problemas de minimização, onde a ênfase está na escolha de ações ou sequências de decisões que minimizam o custo total.

Lema 6.1. *Interação de Recompensas: Em cada MDP com espaço de estados finito, a recompensa esperada de N estágios sob qualquer política pode ser calculada pela interação de recompensas (RI), dada por:*

$$V_{1f} = U_f V_0$$

$$V_{n(f,\sigma)} = U_f V_{n-1}$$

com $\sigma, n \geq 2$ *e* $(f,\sigma) \in \mathcal{F}X\mathcal{F}^{N-1}$

Essa relação expressa como a recompensa esperada no estágio n é influenciada pelas decisões da sequência f e pelas recompensas esperadas anteriores (V_{n-1}). Ou melhor, o resultado ocorre aplicando a função de utilidade nas recompensas esperadas no estágio anterior, indicando como as decisões em um estágio afetam as recompensas esperadas nos estágios subsequentes.

Demonstração. Sabendo pordefinição que:
$V_{N(f,\sigma)} = \sum_{s^N \in S^N} R_{N(f,\sigma)} p_{(f,\sigma)}(s, s^N)$
E $p_{(f,\sigma)}(s, s^N) = p_f(s, s').p_\sigma(s', s^{N-1}$
E também: $U_f V_{n-1,\sigma} = r_f(s) + \beta \sum_{s' \in S} p_f(s, s') V_{n-1,\sigma}(s')$
E $V_{n-1,\sigma}(s') = \sum_{s^N \in S^N} R_{n-1,\sigma}(s', s^N).p_\sigma(s', s^N)$ E $L_v(s, a) = r(s, a) + \beta \sum_{s' \in S} p(s' \mid s, a) \cdot v(s')$
E $U_{fv}(s) = L_v(s, f(s) = r_f(s) + \beta \sum_{s' \in S} p_f(s, s').v(s')$
E $U_v(s) = \sup_{a \in D(s)} L_v(s, a) = \sup_{f \in \mathcal{F}} U_{fv}$

Vamos mostrar o passo a passo para a relação

$$V_{N(f,\sigma)} = U_f V_{n-1,\sigma},$$

Substituindo $p_{(f,\sigma)}(s, s^N) = p_f(s, s').p_\sigma(s', s^{N-1}$

$$V_{N(f,\sigma)} = \sum_{s^N \in S^N} R_{N(f,\sigma)} p_f(s, s') p_\sigma(s', s^{N-1})$$

Agora substituindo $V_{n-1,\sigma}(s')$ em $U_f V_{n-1,\sigma} = r_f(s) + \beta \sum_{s' \in S} p_f(s, s') V_{n-1,\sigma}(s')$

$$U_f V_{n-1,\sigma} = r_f(s) + \beta \sum_{s' \in S} p_f(s, s') \sum_{s^N \in S^N} R_{n-1,\sigma}(s', s^N) p_\sigma(s', s^N)$$

Agora olhando a expressão $U_{fv}(s) = L_v(s, f(s)) = r_f(s) + \beta \sum_{s' \in S} p_f(s, s') v(s')$ e substituindo v(s')

$$U_{fv}(s) = r_f(s) + \beta \sum_{s' \in S} p_f(s, s') \sum_{s^N \in S^N} R_{n-1,\sigma}(s', s^N) p_\sigma(s', s^N)$$

E finalmente substituindo $U_f V_{n-1,\sigma}$ por $U_{fv}(s)$, conseguimos mostrar a igualdade.

$$V_n(f, \sigma) = U_f V_{n-1,\sigma} = U_{fv}(s)$$

■

Lema 6.2. *(Propriedade de L U_f e U) Se B denota qualquer um dos operadores L, U_f ou U:*

a) $Bv_1 \leq Bv_2$*;*

b) $B(v + \alpha) = Bv + B\alpha$*, para qualquer* α *real;*

c) $V_1 = UV_0$ *e* $V_n \leq UV_{n-1}$ *para* $n \in \mathbb{N}$*.*

Demonstração. Vamos provar para L. Os demais são análogos.

(a): Se $v_1(s) \leq v_2(s)$ para todos os s,
$\beta \sum_{s' \in S} p(s,a,s')v_1(s') \leq \beta \sum_{s' \in S} p(s,a,s')v_2(s')$, pois $\beta > 0$ e $p(s,a,s') \geq 0$. Com isto
$Lv_1(s,a) = r(s,a) + \beta \sum_{s' \in S} p(s,a,s')v_1(s') \leq r(s,a) + \beta \sum_{s' \in S} p(s,a,s')v_2(s') = Lv_2(s,a)$ ■

(b): $L(v+\alpha = r(s,a)+\beta \sum_{s' \in S} p(s,a,s')(v(s')+\alpha) = L(v)+\beta \sum_{s' \in S} p(s,a,s')\alpha = L(v)+\beta\alpha \sum_{s' \in S} p(s,a,s') =$
$L(v) + \beta\alpha$ ■

(c): $V_1(s_0) = \sup_{f \in \mathbb{F}} \{V_1 f(s_0)\}$.
Já foi visto no Lema 6.1 que $V_1 f = U_f V_0$, assim
$V_1(s_0) = \sup_{f \in \mathbb{F}} \{U_f V_0\}$, que por definição de U
$\sup_{f \in \mathbb{F}} \{U_f V_0\} = UV_0(s_0)$
Agora vamos tratar da outra afirmação: $V_n \leq V_{n-1}$ para $n \in \mathbb{N}$, então.
$V_n(s_0) = \sup_{\pi \in \mathbb{F}^{\aleph}} \{V_{n\pi}(s_0)\}$.
Note que cada política $\pi \in \mathbb{F}^{\aleph}$ se escreve como $\pi = (f, \sigma)$, com $f \in \mathbb{F}$ e $\sigma \in \mathbb{F}^{\aleph - \not\aleph}$, ficando
$V_n(s_0)$ como,
$V_n(s_0) = \sup_{(f,\sigma) \in \mathbb{F} X \mathbb{F}^{\aleph - \not\aleph}} \{V_{(f.\sigma)}(s_0)\}$
Novamente, pela segunda parte do lema 6.1,
$V_n(s_0) = \sup_{(f,\sigma) \in \mathbb{F} X \mathbb{F}^{\aleph - \not\aleph}} \{U_f V_{n-1,\sigma}(s_0)\}$
Usando o item (a) deste mesmo lema para U_f ,
$V_n(s_0) \leq \sup_{f \in \mathbb{F}} U_f(\sup_{\sigma \in \mathbb{F}^{\aleph - \not\aleph}} V_{n-1,\sigma}(s_0)) = \sup_{f \in \mathbb{F}} U_f V_{n-1}(s_0) = UV_{n-1}(s_0)$ ■

Teorema 6.1. *(Teorema básico p MDPs com espaços de estados finito) Dois resultados:*

a) O critéio de otimalidade (OC) diz que, se f_n é um maximizador no estado n para $1 \leq b \leq N$, então $(f_n)_N = (f_N, f_{N-1}..., f_1)$ é ideial

b) A interação de valor (VI) é mantida na forma :

$$V_n(s) = \sup_{a \in D(s)} [r(s,a) + \beta \sum_{s' \in S} p(s,a,s').V_{n-1}(s')] =$$
$$= UV_{n-1}(s) = \sup_{a \in D(s)} W_n(s,a)$$

O item (a) do teorema 6.1 nos diz que, se, em cada estágio, escolhermos a ação que maximiza a recompensa esperada para o estado atual, ao longo de N estágios, teremos uma sequência

de decisões ótima. Isso reflete a estratégia de escolher a melhor ação em cada estágio para maximizar a recompensa acumulada ao longo do tempo.

O item (b), representa como o agente atualiza suas expectativas de recompensa para cada estágio. para um determinado estado, o agente olha para todas as ações possíveis que pode tomar. Para cada ação, ele considera a recompensa imediata que receberia e adiciona a isso uma ponderação das recompensas futuras esperadas e essas recompensas futuras são obtidas a partir da função de valor do estágio anterior. O tomador de decisões então escolhe a ação que maximiza essa soma de recompensas imediatas e futuras. Portanto, esse processo é repetido em cada estágio, refinando continuamente as expectativas do agente.

Em termos práticos, a Interação de Valor é como um guia para o agente tomar decisões ao longo do tempo, considerando o impacto cumulativo de suas escolhas anteriores e a influência dessas escolhas no que está por vir.

Antes de enunciar a prova do 6.1, iremos lembrar o que é uma política maximizante no estágio n . Segundo o que aparece na página 196 de Hinderer, uma regra de decisão f é maximizante no estágio n se $f(s)$ maximiza $a \to W_n(s,a)$, para todos s, isto é, $U_f V_{n-1} = UV_{n-1}$

Demonstração. **(a):** Vamos provar por indução
(n=1) Pelo item (c) do lema 6.2, $V_1 = UV_0 = U_f 1 V_0 = V_{1,f_1}$, pois f_1 é maximizante no estágio 1. Portanto, é verdade para 1.

Suponha válida para n. Então $\sigma = (f_j)_n^1$ é ótima para MDPn. Tome $f = f_{n+1}$ temos pela página 516 do livro de Hinderer que:
$V_{n+1} \leq UV_n$, pelo lema 6.2
$= U_f V_n$, pela definição de f
$U_f V_{n\sigma}$, pois σ é ótima
$= V_{n+1,(f\ \sigma)}$, pela segunda parte do lema 6.1
$\leq V_{n+1}$
Portanto, $U_f V_n = V_{n+1}$ ■

(b): Quando há um maximizante em cada estágio, segue do item (a) acima que
$V_{n+1} = UV_n$ ■

■

Agora, aplicaremos esses conceitos a um exemplo prático envolvendo a decisão de lançar ou não

uma promoção para um produto. Essa análise será ilustrativa, utilizando gráficos e observando tendências ao longo do tempo. Este exemplo é extraído do livro "Dynamic Optimization (2016)", e nele, examinaremos como um agente, como uma empresa, pode tomar decisões ao longo do tempo para maximizar as recompensas em um cenário de marketing.

Exemplo 6.3. A cada início de mês, uma empresa que comercializa um produto avalia as condições do mercado e, posteriormente, decide uma estratégia de promoção de vendas. Nosso objetivo é obter uma compreensão mais aprofundada da solução numérica desse processo e de como ela depende de variáveis como s, n, β, e V_0. Consideramos a existência de quatro estados de negócio: bom (s = 3), médio (s = 2), ruim (s = 1), e muito ruim (s = 0), juntamente com duas ações possíveis: anunciar (a = 1) e não fazer nada (a = 0).

Ao explorar esse cenário, buscamos entender sobre como as decisões mensais de promoção influenciam os resultados ao longo do tempo, considerando o estado inicial do negócio, a duração total do processo (n), o fator de desconto (β), e o valor inicial (V_0). Essa abordagem nos permitirá entender melhor a dinâmica das ações tomadas pela empresa em resposta às diferentes condições de mercado. É feita uma suposição que V_0 é não negativo e crescente.

	p(s,uma,s')				**R(s,a)**
(s,a)	s'=0	s'=1	s'=2	s'=3	
(0,0)	1	0	0	0	0
(0,1)	0,9	0,1	0	0	-2
(1,0)	0,4	0,5	0,1	0	2
(1,1)	0,2	0,4	0,3	0,1	0
(2,0)	0,2	0,3	0,4	0,1	6
(2,1)	0,1	0,2	0,3	0,4	4
(3,0)	0	0	0,4	0,6	8
(3,1)	0	0	0,2	0,8	6

A tabela revela que escolher a=1 aumenta a probabilidade de transição para um estado mais favorável, porém, simultaneamente, reduz a recompensa devido aos custos associados à publicidade. Assim, podemos afirmar que a probabilidade condicional p(s,1,.) é estocasticamente maior do que p(s,0,.). Em outras palavras, optar por anunciar eleva a chance de uma transição para um estado mais desejável, mesmo considerando a diminuição na recompensa devido aos gastos com publicidade.

Analisando os gráficos a seguir:

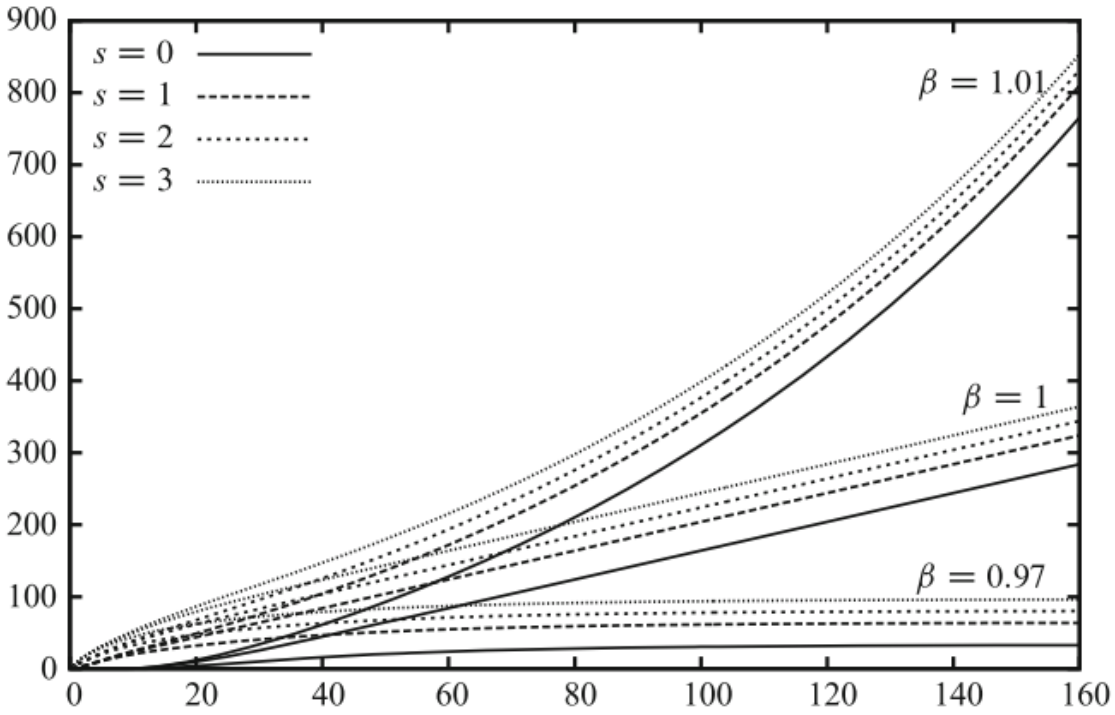

Figura 6.1 – $V_0 = 0, \beta = 0,97, \beta = 1, \beta = 1,01$.Fonte: Hinderer, K., Rieder, U., & Stieglitz, M. (2016). Dynamic Optimization: Deterministic and Stochastic Models.

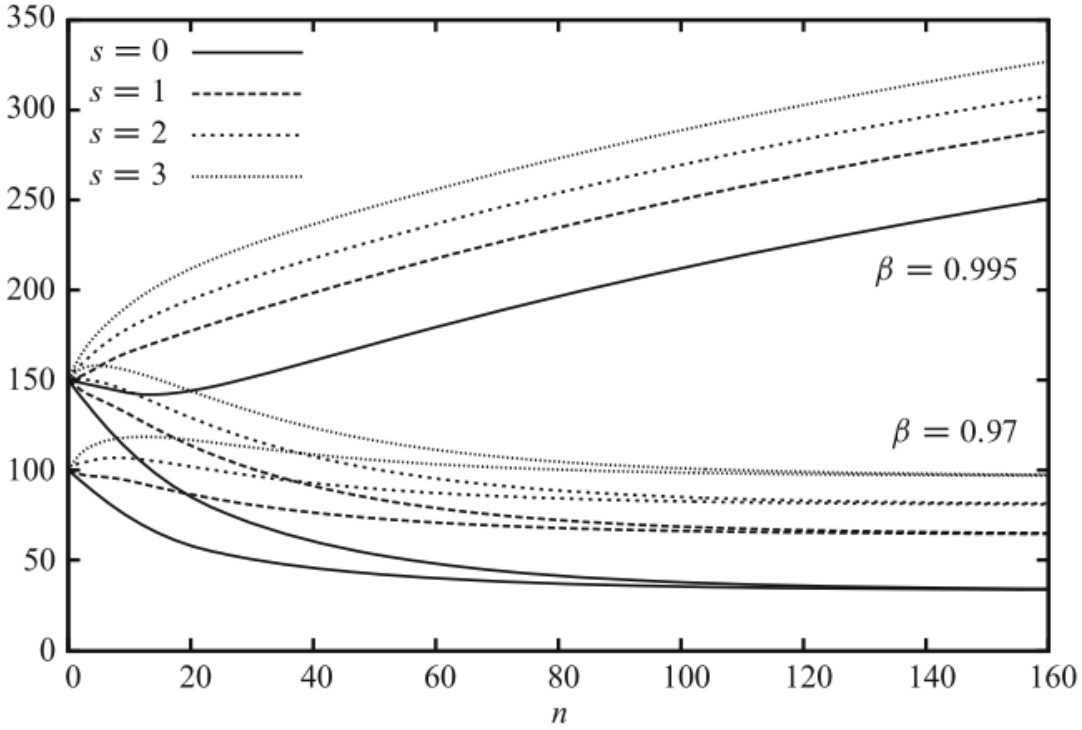

Figura 6.2 – $V_0 = 100, \beta = 0,97$ e $V_0 = 150, \beta = 0,97$, e $\beta = 0,995$ Fonte: Hinderer, K., Rieder, U., & Stieglitz, M. (2016). Dynamic Optimization: Deterministic and Stochastic Models.

(a): Monotonicidade das funções de valor em $s, n, \beta, V-_0$

(a_1) : Nas figuras 6.1 e 6.2, as funções de valor mostram que quando o estado inicial é melhor, a recompensa total esperada é maior. Se a recompensa terminal $V_0(s)$ aumenta com o aumento de s, então a função de valor $V_n(s)$ também aumenta conforme s fica maior. Em outras palavras, começar em um estado mais positivo está associado a uma trajetória de recompensas mais positiva ao longo do tempo.

(a_2) : Na Figura 6.1, a função de valor está mostrando um aumento conforme o horizonte temporal n cresce, onde $0 \leq n \leq 160$. Isso sugere que um horizonte mais longo resulta em uma recompensa total esperada mais alta. A função $V_n(s)$ está aumentando com n sempre que $V_0 = c \geq 0$ e qualquer $\beta \geq 1$ ou $c = 0$. Em outras palavras, quando a recompensa terminal V_0 é não negativa e o fator de desconto β é pelo menos 1, ou quando V_0 é igual a zero, a função de valor aumenta com um horizonte temporal mais longo.

Com $\max r(s, a) \geq 0$, temos $V_1(s) = UV_0(s) = \max r(s, a) + \beta c \geq c = V_0(s)$. Isso significa que a função de valor $V_1(s)$ no primeiro estágio é, no mínimo, a recompensa terminal $V_0(s)$.

Agora, podemos determinar facilmente, através do VI (Interação de Valor), por indução em n usando a isotonicidade de U, que $V_n \geq V_{n-1}$ para todo n. A isotonicidade de U implica que, à medida que avançamos nos estágios, a função de valor não diminui.

A Figura 8.3 também ilustra que a função de valor $V_n(s)$ não precisa ser crescente em n. Além disso, para alguns estados s, $V_n(s)$ pode aumentar ou diminuir à medida que n aumenta, enquanto para outros estados s, $V_n(s)$ pode permanecer constante. Isso destaca a complexidade das mudanças nas recompensas ao longo do tempo e como essas mudanças podem variar para diferentes estados.

(a_3) : Na Figura 6.1, a tendência de aumento de $V_n(s)$ em relação a β reflete o impacto positivo do fator de desconto no valor esperado ao longo do tempo. O fator de desconto β representa a preferência do agente por recompensas imediatas em comparação com recompensas futuras. Quando β aumenta, o agente valoriza mais as recompensas futuras, incentivando decisões que levam a estados mais vantajosos em longo prazo. Isso resulta em uma função de valor crescente em relação a β.

Além disso, para n, π e s fixos, temos que β está tendendo a $V_n^{\pi}(s, \beta)$, sendo um polinômio. Isso ocorre devido à forma como a função de valor é calculada usando o operador de Bellman, que envolve somas ponderadas dos valores nos estados futuros, e essas somas podem ser expressas como polinômios em β.

Portanto, como existem apenas um número finito de políticas, β tendendo a $V_n(s, \beta)$ é um polinômio contínuo e por partes. A expressão "contínua e por partes"refere-se à propriedade matemática de uma função que é suave em certos intervalos, mas pode apresentar descontinuidades em outros. No contexto de β tendendo a $V_n(s, \beta)$, isso implica que a função resultante, embora suave em certos intervalos de β, pode ter descontinuidades em pontos específicos.

Em outras palavras, a função pode ser contínua dentro de um intervalo de valores de β associados a uma política específica, mas pode sofrer descontinuidades quando passamos de uma política para outra. Isso permite representar a função de valor como uma combinação de polinômios associados a diferentes políticas. À medida que β aumenta ou diminui, diferentes políticas podem se sobressair, resultando em uma função global que é contínua em alguns intervalos e descontínua nos pontos de transição entre políticas.

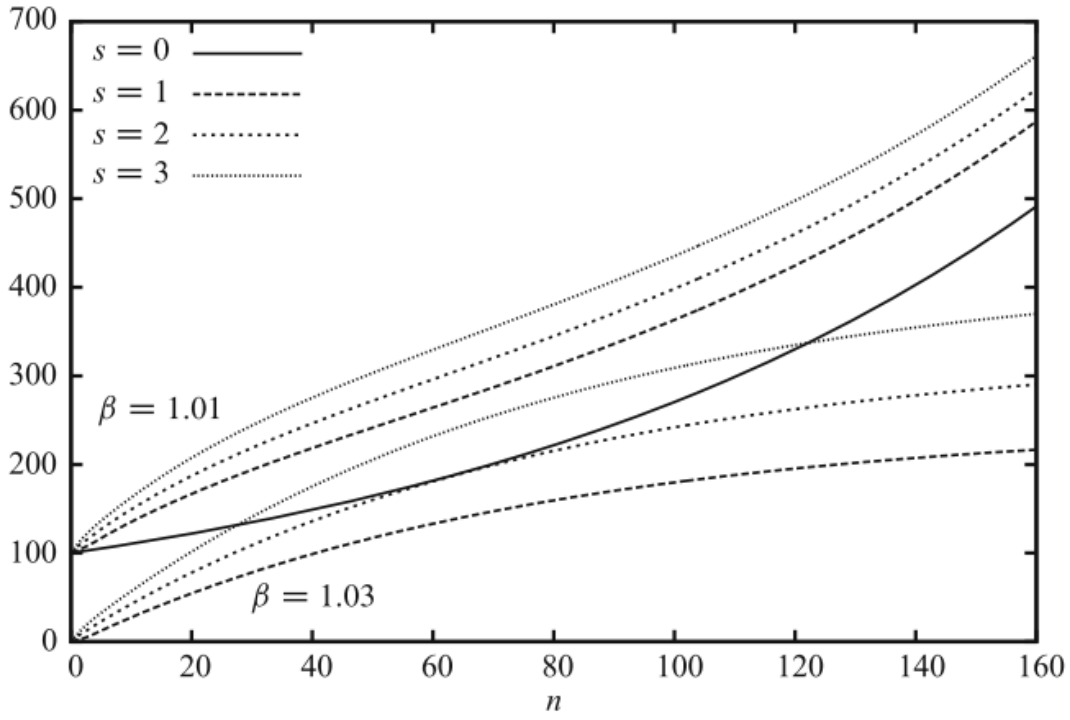

Figura 6.3 – $V_0 = 100, \beta = 0,97$ e $V_0 = 0\beta = 1,03$, e $V_0 = 100, \beta = 1,01$ Fonte: Hinderer, K., Rieder, U., & Stieglitz, M. (2016). Dynamic Optimization: Deterministic and Stochastic Models.

(b): Convergência pontual da sequência de funções de valor

(b_1) : Nas figuras 6.1 e 6.2, observamos que a convergência parece estar presente para valores de β menores que 1, enquanto a divergência para o infinito parece ocorrer para $\beta \geq 1$. Essa tendência sugere que a função de valor $V_n(s, \beta)$ pode atingir uma convergência estável para fatores de desconto β mais baixos, mas pode divergir ou crescer indefinidamente para fatores de desconto mais altos.

Por outro lado, a figura 6.3 indica que, na segunda versão com $V_0 = 0$, a convergência está presente mesmo para alguns valores de β maiores que 1, como por exemplo para $\beta = 1,03$. Essa observação pode ser explicada intuitivamente pela alta probabilidade de o processo parar no estado $s = 0$ após um número moderado de períodos. Se há uma alta probabilidade de o processo parar no estado $s = 0$ antes de atingir o horizonte N, a recompensa terminal torna-se menos relevante, e a convergência pode ocorrer mesmo para valores de β ligeiramente superiores a 1.

(b_2) : A partir da afirmação a_3, sabemos que as funções valor são crescentes em β, ou seja, à medida que β aumenta, o valor das funções também aumenta. Dessa forma, quando considera-

mos a sequência $V_n(s)_0^\infty$, ela tende ao infinito para alguns valores críticos β_0 . Isso sugere que para β superior a β_0, a função valor continua crescendo indefinidamente. No entanto, isso não contradiz a observação na figura 6.3, onde aparentemente temos convergência para $\beta = 1,03$ com $V_0 = 0$ e divergência para infinito quando $\beta = 1,01$ com $V_0 = 100$.

Ao analisar o caso em que $V_0 = 100$ e $\beta = 1,01$, observamos que $V_n(s) \geq V_0(0) = \beta^n V_0(0) = (1,01)^n \cdot 100$. Nesse caso, a sequência $V_n(s)$ cresce exponencialmente em relação a n, o que resulta em um valor que tende para infinito à medida que n aumenta.A convergência mencionada na figura 8.4 para $\beta = 1,03$ com $V_0 = 0$ pode ocorrer porque, nesse cenário específico, a influência de β na função valor é suficientemente controlada pela condição inicial $V_0 = 0$, permitindo uma convergência estável.

É importante ressaltar que a convergência na figura 6.1 ocorre para todos os estados e todos os $\beta < 1$, enquanto na figura 8.4 ocorre para todos os estados e todos os $\beta < \beta^* = 1,04$, desde que $V_0(0)$ seja adequado. Essa dependência da convergência em relação à condição inicial destaca a complexidade das interações no modelo.

(b_3) : A Figura 6.2 sugere uma observação notável: o limite $V(s)$ da sequência $V_n(s)_0^\infty$, quando esse limite existe, parece ser independente da condição inicial V_0. Em outras palavras, mesmo que começemos com diferentes valores iniciais V_0, a função valor $V(s)$ para um determinado estado s converge para um valor específico à medida que consideramos horizontes temporais mais longos.

Essa independência de V_0 destaca uma propriedade interessante do modelo de processo de decisão sequencial (MDP). Independentemente de como iniciamos o processo, com diferentes recompensas iniciais ou condições iniciais distintas, a longo prazo, a dinâmica subjacente leva à mesma avaliação da função valor $V(s)$.

A explicação da independência de V_0 na sequência $V_n(s)_0^\infty$ em um MDP pode ser comparada com a propriedade de estacionariedade em uma cadeia de Markov.

Em uma cadeia de Markov estacionária, a distribuição de probabilidade sobre os estados não muda ao longo do tempo. Essa propriedade é expressa como $\pi P = \pi$, onde π é a distribuição estacionária e P é a matriz de transição da cadeia de Markov. Isso implica que, independentemente da distribuição inicial, a distribuição de probabilidade sobre os estados convergirá para a distribuição estacionária com o tempo.

Da mesma forma, na explicação anterior sobre MDP, a independência de V_0 indica que, à medida que o horizonte temporal n aumenta, a função valor $V(s)$ convergirá para um valor

estável, independentemente da recompensa ou condição inicial específica. Essa convergência sugere uma estabilidade semelhante à distribuição estacionária em cadeias de Markov, onde as características do sistema alcançam um equilíbrio ao longo do tempo.

Horizonte de Turnipike:

Em muitos MDPs com conjuntos finitos de estados S e ações A, durante a aplicação do método de Iteração de Valor (VI), é comum observar a convergência em que uma determinada regra de decisão f emerge como um maximizador, indicando uma política ótima. Essa convergência significa que, após um número suficientemente grande de iterações, a regra de decisão f permanece como um maximizador em todos os estágios subsequentes, ou seja, ela decide precocemente a política ótima.

O menor número desses estágios para os quais essa convergência é alcançada, denotado por n_0, representa o horizonte de turnpike associado à regra de decisão f. Em termos formais, esse horizonte é definido como o ínfimo (o maior limite inferior) dos números naturais k para os quais a regra f maximiza a função de valor em todos os estágios $n \geq k$. Se o infimo é finito, então temos um horizonte de turnpike finito.

O Teorema de Turnpike é uma proposição importante neste contexto. Ele afirma que, em muitos MDPs (Processos de Decisão Markovianos) com espaços de estados e ações finitos, após um número finito de passos no processo de iteração de valor, uma regra de decisão f se torna um maximizador em todos os estágios suficientemente grandes, ou seja, para $n \geq n_0$. O horizonte de turnpike, $N^*(f)$, captura esse ponto crítico onde a regra f atinge a convergência como um maximizador em todos os estágios subsequentes. Isso é particularmente relevante para entender a estabilidade e a otimalidade de políticas de decisão em processos de decisão sequencial.

Definição 6.4. (Equivalência de modelos) Considere dois modelos, denotados como M e M^*, ambos compartilhando o mesmo conjunto S. Suponha que em M e M^* estão definidos conjuntos F e F^* de funções V_n^π e $V_n^{*\pi}$ em S para cada $n \geq 1$ e $\pi \in F^n$. Então, os modelos M e M^* são considerados equivalentes se $V_n^\pi = V_n^{*\pi}$ para $n \geq 1$ e $\pi \in F^n$.

Essa definição nos diz em outras palavaras que , se as funções de valor associadas aos dois modelos coincidem para todos os horizontes de decisão n e todas as sequências de políticas π no conjunto correspondente, então M e M^* são considerados equivalentes. Em termos mais simples, significa que, independentemente do horizonte de decisão escolhido e das sequências específicas de políticas utilizadas, os dois modelos resultam em expectativas de recompensa semelhantes ao longo do tempo.

Exemplo 6.4. (Problema do fabricante de brinquedos de Howard) Muitos dos conceitos e

resultados para MDPs podem ser ilustrados pelo seguinte exemplo simples do livro inovador de Howard (1960). É uma versão simplificada do Exemplo de marketing 6.1 com apenas dois estados: s=0 e s=1, denotando uma situação de negócios boa e ruim, respectivamente, a=0 e a=1 significam não fazer nada ou anunciar, respectivamente.Os outros dados são como na figura 8.5.

(s,a)	p(s,a,0)	p(s,a,1)	$R_s(s,a,0)$	$R_s(s,a,1)	R(s,a)
(0,0)	0,5	0,5	9	3	6
(0,1)	0,8	0,2	4	4	4
(1,0)	0,4	0,6	3	-7	-3
(1,1)	0,7	0,3	1	-19	-5

a) Para obter uma compreensão sólida do VI, calcule V_1 e V_2 para $V_0 = (105, 100)$, $\beta = 1$

Resolução:

Lembrando que $VI = V_n(s) = \sup_{a \in D(s)} \{r(s,a) + \beta \sum_{s' \in S} p(s,a,s').V_{n-1}(s')\}$

Para o cálculo de $V_1(0)$ temos:

$$V_1(0) = \sup\{r(0,0) + \beta \sum_{s'} p(0,0,s')V_0(s'),\ r(0,1) + \beta \sum_{s'} p(0,1,s')V_0(s')\}$$

$$= \sup\{r(0,0)+1.(p(0,0,0).V_0(0)+p(0,0,1).V_0(1), r(0,1)+1.(p(0,1,0).V_0(0)+p(0,1,1).V_0(1)\}$$

$$= \sup\{6 + 1 \cdot (0.5 \cdot 105 + 0.5 \cdot 100),\ 4 + 1 \cdot (0.8 \cdot 105 + 0.2 \cdot 100)\}$$

$$= \sup\{6 + 102.5,\ 4 + 82\}$$

$$= \sup\{108.5,\ 86\}$$

$$V_1(0) = 108.5.$$

Portanto, caso o sistema esteja no estado=0 e no instante 1, a melhor escolha é não anunciar.

Próximo cálculo é de $V_1(1)$:

$$V_1(1) = \sup\{r(1,0) + \beta \sum_{s'} p(1,0,s')V_0(s'),\ r(1,1) + \beta \sum_{s'} p(1,1,s')V_0(s')\}$$

$$= \sup\{r(1,0) + 1.(p(1,0,0).V_0(0) + p(1,0,1).V_0(1), r(1,1) + 1.(p(1,1,0).V_0(0) + p(1,1,1).V_0(1)\}$$

$$= \sup -3 + (0, 4.105 + 0, 6.100), -5(0, 7.105 + 0, 3.100)$$

$= \sup -3 + 102, -5 + 103, 5$

$= \sup 99, 98, 5$

$V_1(1) = 99$

Caso inicie no estado = 1 a melhor decisão no instante 1 é continuar sem aunincar também.

Agora iremos calular para o instante 2, temos $V_2(0)$, $v_2(1)$:

Para $V_2(0)$:

$$V_2(0) = \sup\{r(0,0) + \beta \sum_{s'} p(0,0,s')V_1(s'),\ r(0,1) + \beta \sum_{s'} p(0,1,s')V_1(s')\}$$

$= \sup\{r(0,0)+1.(p(0,0,0).V_1(0)+p(0,0,1).V_1(1), r(0,1)+1.(p(0,1,0).V_1(0)+p(0,1,1).V_1(1)\}$

$= \sup\{6 + 1.(0, 5.108, 5 + 0, 5.99), 4 + 1.(0, 8.108, 5 + 0, 2.99)\}$

$= \sup\{6 + 54, 25 + 49, 5, 4 + 86, 5 + 19, 8\}$

$= \sup\{109, 75, 110, 3\}$

$V_2(0) = 110, 3$

É possível perceber que no instante 2, caso o sistema esteja no estado = 0 a melhor decisão é anunciar .

Último cálculo é o de $V_2(1)$

$$V_2(1) = \sup\{r(1,0) + \beta \sum_{s'} p(1,0,s')V_1(s'),\ r(1,1) + \beta \sum_{s'} p(1,1,s')V_1(s')\}$$

$= \sup\{r(1,0)+1.(p(1,0,0).V_1(0)+p(1,0,1).V_1(1), r(1,1)+1.(p(1,1,0).V_1(0)+p(1,1,1).V_1(1)\}$

$= \sup\{-3 + 1.(0, 4.108, 5 + 0, 6.99), -5 + 1.(0, 7.108, 5 + 0, 3.99)\}$

$= \sup\{-3 + 43, 5 + 59, 4, -5 + 75, 95 + 29, 7\}$

$= \sup\{99, 9, 100, 65\}$

$V_2(1) = 110, 3$

Por fim, estando no estado=1 no instante 2, a decisão que resultará em valor máximo será a de anunciar .

Através das iterações, é possível estimar o valor esperado das recompensas ao longo do tempo, orientando assim as decisões ótimas. No entanto, conforme o horizonte temporal se estende, a complexidade do cálculo aumenta. Ao observar os minuciosos cálculos torna-se evidente que a abordagem computacional oferece vantagens, proporcionando resultados precisos e eficientes em comparação com iterações manuais extensas.

7 HORIZONTE LARGO (GRANDE)

O conceito de horizonte largo, também conhecido como horizonte grande, refere-se à capacidade de um agente de planejar e tomar decisões considerando um período extenso de tempo. Diferentemente do horizonte infinito, que considera todas as etapas futuras possíveis, e do horizonte finito, que se restringe a um número fixo de passos, o horizonte largo busca um equilíbrio, incorporando uma visão de longo prazo, mas ainda limitada em comparação ao infinito.

No contexto do cotidiano, podemos exemplificar o horizonte largo em situações como o planejamento financeiro a longo prazo, onde indivíduos consideram não apenas as despesas imediatas, mas também investimentos, aposentadoria e objetivos de vida a longo prazo. Ao tomar decisões financeiras, como investir em educação, propriedades ou planos de aposentadoria, as pessoas estão, de certa forma, aplicando o conceito de horizonte largo para maximizar seus benefícios ao longo do tempo.

Definição 7.1. O modelo é chamado de MDP com horizonte largo se as seguintes condições forem válidas.

- D, p, r_s e V_0, tem o mesmo significado que no MDP de horizonte finito;
- $A \neq 0$ e é finito;
- $\beta < 1$, essa condição indica que atribui um peso menor às recompensas e estados futuros em comparação com os imediatos.

Algumas outras definições são acrescentadas :

a) $V = \lim_{n\to\infty} V_n$, caso o limite exista nos reais,

b) A regra de decisão $f \in \mathbb{F}$, onde $\mathbb{F}$ é o conjunto de todas as possíveis f, então f é chamada assintoticamente ótima se: $V_{nf} - V_n \to 0$ para $n \to \infty$ para todos $s \in S$;

c) Para funções V a equação $V = UV$ é chamada de equação de otimalidade ou de Bellman.

Para funções V em algum conjunto finito denotamos por ||V|| a norma máxima e por $spV = maxV - minV$.

Lema 7.1. *(Estimativas para $U_v^t - U_w^t$) Para funções v e w em S e para todo $t \geq 0$ temos:*

$$\beta^t min(v - w) \leq U_v^t - U_w^t \leq \beta^t max(v - w)$$

$$||U_v^t - U_w^t|| \leq \beta^t.||v - w||$$

$$sp(U_v^t - U_w^t) \leq \beta^t sp(v - w)$$

Portanto, o lema estabelece relações entre as diferenças das funções de valor (U_v^t e U_w^t) e as diferenças das observações (v e w) ao longo do tempo, considerando um fator de desconto β. Essas estimativas são úteis para compreender como as funções de valor evoluem em relação às mudanças nas observações em um processo ao longo do tempo.

Demonstração. Começando definindo $U_v - U_w$

$$U_v - U_W \leq max_{f \in \mathbb{F}} U_{fv} - max_{f \in \mathbb{F}} U_{fw}$$

Por 6.1 temos que :

$$U_v - U_W \leq max_{f \in \mathbb{F}} U_{fv} - max_{f \in \mathbb{F}} U_{fw}$$

$$= (r_f(s) + \beta \sum_{s' \in S} p_f(s, a, s').max_{s' \in S} v(s')) - (r_f(s) + \beta \sum_{s' \in S} p_f(s, a, s').max_{s' \in S} w(s')$$

$$U_v - U_W \leq \beta \sum_{s \in S} p_f(s, a, s').max_{s' \in S}(v - w)(s')$$

$$U_v - U_W \leq \beta max_{s' \in S} \sum_{s' \in S}$$

$$U_v - U_W \leq \beta max(v - w)$$

Agora a segunda desigualdade :

A desigualdade abaixo é estabelecida utilizando a propriedade triangular da norma e a propriedade multiplicativa da norma:

$$\left\| U_v^t - U_w^t \right\| \leq \beta^t \cdot \|v - w\|$$

Usaremos a propriedade triangular da norma, que diz que $\|x + y\| \leq \|x\| + \|y\}$ para quaisquer

vetores x e y. Assim:

$$\begin{aligned}\left\|U_v^t - U_w^t\right\| &= \left\|U_v^t + (-U_w^t)\right\| \\ &\leq \left\|U_v^t\right\| + \left\|-U_w^t\right\|\end{aligned}$$

Agora, podemos usar a propriedade multiplicativa da norma (também conhecida como homogeneidade) que diz que $\|\alpha x\| = |\alpha| \cdot \|x\|$ para um escalar α e um vetor x:

$$\left\|U_v^t\right\| + \left\|-U_w^t\right\| = \left\|U_v^t\right\| + \left\|U_w^t\right\|$$

Finalmente, usando a desigualdade triangular novamente:

$$\left\|U_v^t\right\| + \left\|U_w^t\right\| \leq \left\|U_v^t - U_w^t\right\| + \beta^t \cdot \|w\|$$

Subtraindo $\beta^t \cdot \|w\|$ de ambos os lados:

$$\left\|U_v^t\right\| - \beta^t \cdot \|w\| \leq \beta^t \cdot \|v - w\|$$

Isolando $\left\|U_v^t - U_w^t\right\|$:

$$\left\|U_v^t - U_w^t\right\| \leq \beta^t \cdot \|v - w\|$$

A segunda desigualdade pode ser derivada da primeira, e é expressa como:

$$\text{sp}(U_v^t - U_w^t) \leq \beta^t \cdot \text{sp}(v - w)$$

A amplitude (ou spread) de um vetor é definida como a diferença entre seu valor máximo e mínimo. Podemos expressar isso em termos da norma máxima:

$$\text{sp}(U_v^t - U_w^t) = \left\|U_v^t - U_w^t\right\|$$

E $\text{sp}(v - w) = \|v - w\|$. Portanto, a segunda desigualdade é simplesmente a primeira desigualdade que acabamos de provar:

$$\left\|U_v^t - U_w^t\right\| \leq \beta^t \cdot \|v - w\|$$

Portanto, a segunda desigualdade também é válida. ■

Teorema 7.1. *A equação de otimalidade e regras de decisão assintoticamente ótimas para um MDP com horizonte grande.*

a) A sequência de funções de valor V_n converge para uma função limite V e $||V|| \leq$

$||r||/(1-\beta)$

b) V é a única solução finita da equação de otimalidade.

c) V_N e V satisfaz para $N \geq 0$: $V + \beta^N min(V_0 - V) \leq V_N \leq V + \beta^N max(V_0 - V)$

d) Para cada regra de decisão $f \in \mathbb{F}$ existe $V_f = lim_{n\to\infty} V_{nf}$ e V_f é o único ponto fixo finito de U_f

e) Temos $V = max_{f\in\mathbb{F}} V_f$ e $V = V_f$ se f for um maximizador de LV.

f) A regra de decisão f é assintoticamente ótima se e somente se f é um maximizador de LV e então: $0 \leq V_N - VNf \leq \beta^N sp(V - V_0)$ para $N \geq 0$

Demonstração. **(a):** Conseguimos demonstrar pelo Teorema do Ponto Fixo e contração.

Para o Teorema do Ponto Fixo de Banach ser aplicável, o espaço de funções deve ser um espaço métrico completo. Vamos considerar o espaço de funções contínuas $C(S)$ sobre o conjunto de estados S, equipado com a norma $\|V\| = \max_{s\in S} |V(s)|$.

Também definimos um operador de contração **T** que mapeia uma função V em outra função $\mathbf{T}V$ usando a equação de Bellman:

$$(\mathbf{T}V)(s) = \max_a \left\{ r(s,a) + \beta \sum_{s'} P(s'|s,a)V(s') \right\}$$

Agora, devemos verificar as condições do Teorema do Ponto Fixo de Banach:

1. Existência de um ponto fixo: Se T é um operador de contração em $C(S)$, então pelo Teorema do Ponto Fixo de Banach, existe uma função V tal que $TV = V$.

2. Unicidade do ponto fixo: Se T é uma contração, então o ponto fixo V é único.

3. Convergência da sequência: A sequência V_n converge para V na norma $\|\cdot\|$.

Vamos analisar a condição de contração:

$$\|\mathbf{T}V - \mathbf{T}W\| \leq \beta \|V - W\|$$

Isso implica que, ao longo das iterações, a norma das diferenças sucessivas das funções de valor diminui, o que é uma condição crucial para a convergência.

Sabemos por definição de Horizonte Grande que $\beta < 1$, então a condição de contração será satisfeita, e a sequência V_n converge para V. Além disso, a estimativa $\|V\| \leq \frac{\|r\|}{1-\beta}$ pode ser derivada.

(b): Utilizando novamente o Teorema do ponto Fixo.

Suponha que existam duas soluções V_1 e V_2 da equação de otimalidade, ou seja, $TV_1 = V_1$ e $TV_2 = V_2$. Queremos mostrar que $V_1 = V_2$.

Para isso, considere a norma $\|V_1 - V_2\|$. Pela condição de contração, temos:

$$\|TV_1 - TV_2\| \leq \beta\|V_1 - V_2\|$$

Substituindo $TV_1 = V_1$ e $TV_2 = V_2$, obtemos:

$$\|V_1 - V_2\| \leq \beta\|V_1 - V_2\|$$

Isso implica que $\beta < 1$, como já é definido. Portanto, a única solução finita é única, e $V_1 = V_2$.

(c): Utilizando 7.1:

$$\beta^N min(V_0 - V) \leq U_{V_0}^N - U_V^N \leq \beta^N max(V_0 - V)$$

Temos que: $U_{V_0}^N = V_N$ e $U_V^N = V$, logo

$$\beta^N min(V_0 - V) \leq V_N - V \leq \beta^N max(V_0 - V)$$

Somando V em todos os termos:

$$V + \beta^N min(V_0 - V) \leq V_N \leq V + \beta^N max(V_0 - V)$$

(d): Iremos demonstrar atráves de $U_f V_f = V_f$:

A expressão $U_f V_f = V_f$ representa a aplicação do operador U_f (operador de Bellman para uma política f) à função de valor V_f associada a essa política. Vamos entender por que essa igualdade é verdadeira.

O operador de Bellman U_f é definido da seguinte maneira para uma política f específica:

$$(U_f V)(s) = r(s, f(s)) + \beta \sum_{s'} P(s'|s, f(s)) V(s')$$

Agora, considere $U_f V_f$, onde V_f é a função de valor associada à política f. Substituímos V por V_f na definição do operador de Bellman:

$$(U_f V_f)(s) = r(s, f(s)) + \beta \sum_{s'} P(s'|s, f(s)) V_f(s')$$

Agora, observe que V_f é, por definição, a solução da equação de Bellman para a política f:

$$V_f(s) = r(s, f(s)) + \beta \sum_{s'} P(s'|s, f(s)) V_f(s')$$

Substituindo esta expressão na equação anterior, obtemos:

$$(U_f V_f)(s) = V_f(s)$$

Portanto, $U_f V_f = V_f$ porque V_f é a solução da equação de Bellman para a política f. Esse resultado mostra que V_f é uma solução fixa (ponto fixo) do operador U_f.

(e): Então temos que :
$U_f V = U_f V_f$, pois ambos são maximizados pela mesma regra de decisão f.
$U_f V_f = V_f$, pelo item d acima.
$V_f = V$, quando V atinge o máximo valor de acorod com a política ótima f
$V = UV$ pelo item (c) da definição 7
Portanto temos que :
$U_f V = U_f V_f = V_f = V = UV$.

(f): Pelo item (b) da definição 7, temos que por ser assintoticamente ótimo $V_{nf} - V_n \to 0$ quando $n \to \infty$.

Portando a primeira parte :
$0 \leq V_N - V_{Nf}$ quando $N \geq 0$

Para mostrar que $\|U_v^t - U_w^t\| \leq \beta^t \text{sp}(v - w)$, vamos usar a definição de U_v^t e U_w^t, e algumas propriedades dos operadores de Bellman.

A sequência de funções de valor U_v^t é definida como a aplicação repetida do operador de Bellman

U^t a partir da função de valor inicial V_0:

$$U_v^t = U^t V_0$$

Similarmente, U_w^t é a solução da equação de Bellman para a política w após t iterações:

$$U_w^t = U^t W$$

Agora, vamos calcular a diferença $U_v^t - U_w^t$:

$$U_v^t - U_w^t = U^t V_0 - U^t W$$

Podemos usar a propriedade de contração do operador de Bellman para obter uma estimativa dessa diferença:

$$\| U_v^t - U_w^t \| \leq \beta^t \| V_0 - W \|$$

A expressão acima é uma estimativa para a norma da diferença entre U_v^t e U_w^t, onde β^t é o fator de desconto elevado à t-ésima potência, e $\| V_0 - W \|$ é a norma da diferença entre as funções de valor inicial e final.

Além disso, a norma da diferença entre duas funções V e W é o spread dessas funções, ou seja, a diferença entre seus valores máximos e mínimos:

$$\| V_0 - W \| = \text{sp}(V_0 - W)$$

Substituindo essa expressão na nossa estimativa, obtemos:

$$\| U_v^t - U_w^t \| \leq \beta^t \text{sp}(V_0 - W)$$

Essa é a desigualdade desejada. Portanto, temos mostrado que $\| U_v^t - U_w^t \| \leq \beta^t \text{sp}(v - w)$. ■

Proposição:(Horizonte de Turnipike e otimalidade assintótica) Se f em MDP com horizonte grande tem um horizonte turnipike então f é assintóticamente ótimo.

A proposição aborda a interação entre duas características cruciais em Modelos de Decisão Markovianos (MDP) com horizonte grande: "Turnpike Horizon"e "Otimalidade Assintótica". Em um contexto de MDP, o termo "Turnpike Horizon"denota a convergência de uma política f para uma política estacionária ótima ao longo do tempo, sugerindo uma trajetória direta em direção à otimalidade. Por sua vez, a "Otimalidade Assintótica"refere-se à tendência dessa política em se tornar ótima à medida que o horizonte temporal se estende indefinidamente.

8 HORIZONTE INFINITO

Neste capítulo, abordaremos Processos de Markov com horizonte infinito, concentrando-nos na construção e compreender a estrutura desses MDPs e investigar as propriedades das políticas de decisão ao longo de horizontes temporais ilimitados. Introduziremos o Teorema de Horizonte Infinito, que fornece informações cruciais sobre as condições para políticas ótimas em contextos de horizonte infinito. Ao final, esperamos oferecer uma base sólida para a compreensão de decisões sequenciais em cenários temporais extensos.

A diferença fundamental entre horizonte finito e infinito em processos de decisão está na temporalidade das decisões. Em um horizonte finito, as decisões e as recompensas são analisadas até um número específico de estágios, enquanto em um horizonte infinito, esse número é teoricamente ilimitado. No horizonte finito, é comum avaliar uma recompensa terminal $v_0(S)$ que representa a utilidade associada ao estado final. No entanto, em um horizonte infinito, não faz sentido calcular uma recompensa terminal, pois o processo teoricamente continua indefinidamente no tempo. Em vez disso, focamos nas expectativas ao longo do tempo, considerando a soma descontada das recompensas ao longo de uma sequência infinita de estágios.

Definição 8.1. *MDPs*$_\infty$ é uma tupla (S, A, D, p , r_s e β) Para a definição da recompensa esperada em estágios infinitos, consideramos uma política de estágios infinitos $\pi = (\pi_t)_0^\infty \in \mathbb{F}^{N_0}$ e um estado inicial s. Para abordar essa situação, introduzimos uma sequência infinita de variáveis aleatórias de estado denotadas por ζ_t, onde $s \in S$ e $t \in \mathbb{N}$, definidas em um espaço de probabilidade $(\Omega, \mathbb{F}, P_\pi)$. Esse espaço de probabilidade descreve o processo de decisão ao longo de estágios infinitos.

A teoria da medida é essencial para lidar com conjuntos não triviais e possibilita a definição adequada das variáveis aleatórias e de suas propriedades em espaços de probabilidade mais gerais. Na teoria das cadeias de Markov, a medida é implicitamente incorporada ao considerar probabilidades de transição entre estados discretos. As cadeias de Markov geralmente evoluem em um espaço discreto e, portanto, a teoria da medida clássica, baseada em conjuntos contáveis, é suficiente para modelar e analisar esses processos. No entanto, ao estender o escopo para processos estocásticos com horizontes temporais infinitos, como em modelos de decisão em estágios infinitos, a teoria da medida mais avançada se torna crucial, com esse contexto, a utilização de variáveis aleatórias ζ_t em um espaço de probabilidade $(\Omega, \mathbb{F}, P_\pi)$ é crucial para modelar o processo de decisão em estágios infinitos. Essa abordagem mais rigorosa é necessária para garantir a consistência e a validade matemática ao considerar horizontes temporais ilimitados.

Assim, a construção do modelo com horizonte infinito exige a formalização desses conceitos,

incorporando a teoria da medida para garantir a fundamentação matemática sólida necessária para a análise de processos estocásticos complexos e prolongados.

Construção: A construção de MDPs com horizonte infinito envolve a definição de um espaço amostral incontável, representado por $\Omega = S^{\mathbb{N}}$. Nesse contexto, Ω é o conjunto de todas as sequências infinitas de estados possíveis, onde S é o conjunto de estados do MDP. A escolha de Ω como espaço amostral incontável é fundamental para lidar com horizontes temporais ilimitados.

Cada variável aleatória coordenada ζ_t é tomada como a t-ésima variável aleatória no processo de decisão. Isso significa que ζ_t representa o estado do sistema no tempo t. A existência de uma distribuição de probabilidade P_π em $\mathbb{F}garantidapelaescolhaapropriadadapoltica\pi$.

A distribuição de probabilidade condicional $P_{\pi s} = (\zeta_t)_1^N$ tem uma densidade por (6.1), dada a sequência de estados ζ até o instante N . Ela é definida como o produto das probabilidades condicionais p_π de transição entre estados ao longo do horizonte de tempo N.

$$P_\pi(\zeta|\zeta_0, \zeta^N) = p_\pi(\zeta_1, (\zeta_t)_2^N)$$

$$P_{\pi s}(\zeta_t = s_t) = p_\pi(\zeta_0 = s_0, \zeta_1 = s_1).p\pi(\zeta_1 = s_1, \zeta_2 = s_2) \dots p\pi(\zeta_{N-1} = s_{N-1}, \zeta_N = s_N)$$

O que indica que $\zeta = (\zeta_t)_1^\infty$ forma uma cadeia de Markov não homogênea. Aqui, ζ_0 é o estado inicial, ζ^N representa a sequência de estados até o instante N, e $p_{\pi t}$ são as matrizes de transição associadas à política π em cada estágio t.

Denote por $E_{\pi s}(.)$a expectativa em relação a $P_{\pi s}$. Desde que $\sum_{t=0}^{\infty} \beta^t.|r_s(\zeta_t, \pi_t(\zeta_t), \zeta_{t+1})| \leq ||r_s||/(1-\beta) < \infty.$, assim assegura que a expectativa é bem definida e finita, permitindo uma análise consistente de recompensas esperadas ao longo do tempo.

Definição 8.2.

- $R_{\infty\pi}(s, \zeta) = \sum_{t=0}^{\infty} \beta^t.r_s(\zeta_t, \pi_t(\zeta_t), \zeta_{t+1})$, representa a recompensa acumulada ao longo de um horizonte infinito, onde cada termo é ponderado pelo fator de desconto, é finita para políticas π e estados iniciais s, desde que o fator de desconto β seja menor que 1.
- $V_{\infty\pi}(s) = E_{\pi s}.R_{\infty\pi}(s, \zeta) = \sum_{s\in S} R_{\infty\pi}.p_\pi(s, \zeta)$, representa o valor esperado da recompensa acumulada ao longo de um horizonte infinito para um estado inicial s e uma política π A explicação de que essa expressão é finita está relacionada à finitude da

recompensa acumulada $R_{\infty\pi}(s, \zeta)$para uma política específica.

- $V_{\infty}(s) = \sup V_{\infty\pi}(s) : \pi \in \pi$, a razão pela qual buscamos o supremo é encontrar o valor máximo entre todas as possíveis políticas. Cada política π resultará em um valor esperado da recompensa acumulado e o supremo seleciona o maior desses valores.
- $\pi* \in \pi$ é chamada de ótima se $V_{\infty\pi*}(s) = V_{\infty}(s) = supV_{\infty\pi}(s) : \pi \in \pi$, Isso significa que a política $\pi*$ atinge o valor máximo em comparação com todas as outras políticas possíveis.

Na interpretação frequencial convencional de probabilidades, a expectativa $V_{\infty\pi}(s)$ da variável aleatória $R_{\infty\pi}(s, \zeta)$ é considerada uma estimativa para a média de um grande número de realizações independentes de $R_{\infty\pi}(s, \zeta)$. No entanto, vale ressaltar que o cálculo de cada realização implica aguardar um número infinito de períodos, uma prática impraticável na realidade. Para contornar essa limitação, interrompemos as observações após um número significativamente grande, porém finito, de períodos. Essa abordagem é uma maneira pragmática de lidar com a natureza infinita do horizonte em processos de decisão, permitindo uma análise eficiente e aplicável em cenários práticos.

Além disso, políticas da forma $\pi = (f)_0^{\infty} = f^{\infty}$ para alguma regra de decisão f são chamadas de estacionárias. Uma política estacionária é aquela em que a mesma regra de decisão é aplicada infinitamente, ou seja, a estratégia de decisão não muda ao longo do tempo. Em outras palavras, uma política estacionária não depende do estágio específico ou do momento no tempo; ela permanece constante ao longo de um horizonte infinito.

Ao considerar políticas estacionárias, a estratégia de decisão não muda com o passar do tempo. Em um MDP com horizonte infinito, a notação $\pi = (f, f, f, ...)$ implica que a mesma regra de decisão f é repetida infinitamente. Isso significa que, em cada estágio, a decisão é tomada de acordo com a mesma lógica, independentemente do estágio específico, a ideia é que a política escolhida é ótima e permanece válida independentemente do número de estágios.

A diferença entre horizonte finito e horizonte infinito em Processos de Decisão Markovianos (MDPs) diz respeito à duração da sequência de decisões que um agente toma ao interagir com um ambiente. Em horizonte finito, a regra de decisão pode variar de um estágio para outro. Cada estágio pode ter uma política diferente, adaptando-se a condições específicas ou a mudanças nas circunstâncias ao longo do tempo. Isso permite uma maior flexibilidade nas decisões, mas a desvantagem é que a política precisa ser definida para cada estágio separadamente, como visto $\pi = (f_1, f_2, ... , f_N$. A escolha entre essas abordagens dependerá da natureza do problema .

Teorema 8.1. *Denote por* $v_n, n \geq 0$ *a função valor em MDP que possui os mesmo dados do*

MDP_∞ e além disso alguma função de recompensa terminal v_0. Assim, $V = \lim_{n\to\infty} v_n$. Então vale o seguinte:

- *$V_\infty = V$, isso afirma que, sob certas condições, esse limite é válido, e podemos considerar V_∞ como o limite dessas aproximações. Isso é útil na prática, pois muitas vezes é mais fácil computar ou analisar funções de valor em MDPs com horizonte finito.*

- *Para cada regra de decisão f a política f^∞ é ótima se e somente se f for assintoticamente ótima, isso significa que a política continua a ser a melhor escolha à medida que o horizonte de decisão se estende indefinidamente. e se f for um maximizador de LV e seV = V_f*

- *Seja $n \geq 1$ e seja f_n um amximizador no estágio n e*
  $w_n^- = v_n + \beta.\frac{min[v_n - v_{n-1}]}{1-\beta}$
  $w_n^+ = v_n + \beta.\frac{max[v_n - v_{n-1}]}{1-\beta}$
  Temos:
  $w_n^- \leq V_{fn} \leq V_\infty \leq w_n^+$*, e os limites para V_∞ estão melhorando em n e convergindo para V_∞ para $n \to \infty$*

- $V_{fn} \geq V_\infty - \frac{\beta}{1-\beta}.\sup(V_n - V_{n-1}) \geq V_\infty - \frac{\beta}{1-\beta}.\sup(v_1 - v_0$
  A condição para que f_n seja considerada ótima é dada pela inequação acima. Se essa condição for atendida para um certo $n \geq n_0$ onde n_0 é o instante exato, então f_n é ótimo.
  A expressão $n \geq n_0 = \frac{\log[\epsilon.(1-\beta)/\sup(v_1 - v_0)]}{\log \epsilon}$, fornece uma fórmula para calcular esse instante exato da política ϵ-ótima. Isso é útil para determinar em que momento a política atinge um desempenho considerado suficientemente próximo do ótimo, conforme definido pelo parâmetro ϵ.

.

Prova:
(a): Para mostrar que $V_\infty = V$, vamos analisar o limite da função valor v_n conforme n se aproxima do infinito.

Dado que v_n é uma função valor em um MDP com os mesmos dados do MDP_∞, podemos considerar que v_n converge para V_∞ conforme n tende ao infinito, ou seja:

$$\lim_{n\to\infty} v_n = V_\infty$$

Além disso, temos que $V = \lim_{n\to\infty} v_n$, conforme teorema 7.1
Portanto, temos:

$$V = \lim_{n\to\infty} v_n = V_\infty$$

(b): 1. *f^∞ é ótima se f é assintoticamente ótima:*
Seja f uma regra de decisão assintoticamente ótima. Isso significa que, à medida que o horizonte de decisão se estende indefinidamente, a política f^∞ continua sendo ótima. Vindo do teorema 7.1 item (f):

Denotando o valor da função valor para a política f como V_f e o valor da função valor para a política f^∞ como $V_{f\infty}$, como f é assintoticamente ótima, temos $V_f = V_{f\infty}$. Portanto, f^∞ é ótima.

2. *Se f^∞ é ótima, então f é assintoticamente ótima:*

Suponha que f^∞ seja uma política ótima. Isso significa que, para qualquer horizonte de decisão finito, f^∞ é ótima. Se f^∞ é ótima para horizontes finitos, então, à medida que o horizonte se estende indefinidamente, a política f^∞ continua sendo ótima.

Isso implica que, para qualquer horizonte de decisão finito N, $V_{f\infty}(N) = V_f(N)$. À medida que N tende ao infinito, temos $V_{f\infty}(\infty) = V_f(\infty)$, o que significa que f é assintoticamente ótima.

As demonstrações dos itens (c) e (d) podem ser localizadas na página 217 do livro de (**??**), proporcionando uma fonte específica para que os leitores consultem as demonstrações completas e tenham a oportunidade de aprofundar sua compreensão sobre esses resultados.

9 MDPS COM UM CONJUNTO DE ESTADOS ABSORVENTES

O Exemplo 6.3 com $V_0(0) = 0$, mostrou a característica especial de que, sob qualquer política, o processo de decisão $(\eta_t)_1^N$ para assim que o estado s=0 é alcançado. Mais geralmente, definimos em analogia

Definição 9.1. Um subconjunto próprio não vazio J_0 do espaço dos estados S é chamado de conjunto absorvente para um MDP se tivermos:

a) $\sum_{s' \in J_0} p(s, a, s') = 1 \text{para} s \in J_0, a \in D(s)$, expressa que, a partir de qualquer estado s em J_0 sob a ação a, a probabilidade total de transição para estados em J_0 é igual a 1. Isso indica que uma vez que o processo entra em J_0 , ele permanece lá com probabilidade 1.

b) $r(s, a) = 0 \text{para todos} s \in J_0, a \in D(s)$, estabelece que as recompensas associadas às ações tomadas em estados de J_0 são nulas. Isso significa que não há recompensa adicional ao permanecer ou transitar dentro do conjunto absorvente.

c) $V_0(s) = 0 \text{para todos} s \in J_0$, essa condição implica que, ao iniciar em um estado absorvente, a recompensa terminal é zero. Em outras palavras, uma vez que o sistema atinge um estado em J_0 , não se espera mais receber recompensas adicionais, pois o processo permanecerá indefinidamente dentro desse conjunto absorvente.

Os conjuntos absorventes em MDPs e em cadeias de Markov compartilham o conceito fundamental de estados que, uma vez alcançados, garantem que o processo permaneça nesse estado indefinidamente. Em ambas as definições, a característica principal é a probabilidade de transição para fora do conjunto absorvente ser zero, tornando esse conjunto "absorvente"no sentido de que, uma vez dentro, o sistema não sairá.

No entanto, há uma distinção notável. Em MDPs, a absorção está associada não apenas a probabilidades de transição, mas também à natureza da recompensa. Em um conjunto absorvente de um MDP, a recompensa associada é sempre zero (r(s,a)=0), refletindo que, uma vez que um estado absorvente é atingido, não há mais ganhos ou perdas futuras. Em contraste, em cadeias de Markov, a absorção está relacionada apenas às probabilidades de transição, sem consideração explícita de recompensas associadas.

Sendo essa a principal distinção em MDPs, a absorção é uma combinação de propriedades probabilísticas e de recompensas, refletindo a natureza dinâmica das decisões em ambientes

estocásticos. Essa abordagem mais rica em detalhes permite modelar de maneira mais precisa situações em que a recompensa é uma consideração fundamental, o que é típico em problemas de tomada de decisão.

No contexto de Markov Decision Processes (MDPs) com conjuntos absorventes J_0, podemos simplificar a análise, concentrando-nos nos estados essenciais fora desse conjunto, representados por $J = S - J_0$ .Isso se deve ao fato de que, em J_0 o valor ótimo V_n é sempre zero para qualquer horizonte $n \in \mathbb{N}_0$

De forma mais específica, podemos afirmar o seguinte:

a) O VI assume: $V_n(s) = \sup_{a \in D(s)} \{r(s,a) + \beta \sum_{s' \in S} p(s,a,s').V_{n-1}(s')\}$ para $n \geq 1$ e s $\in J$.

b) Qualquer regra de decisão f que maximize $a \rightarrow V_n(s)$ para $s \in J$ é um maximizador no estágio n . Isso destaca a importância de escolher ações que otimizem o valor ao longo do horizonte temporal .

Em alguns exemplos com um conjunto absorvente J_0 obtem-se, além a recompensa de um estágio uma recompensa de rescisão ($r_{term}(s') \in \mathbb{R}$) assim que o processo de decisão entra em um estado $s' \in J_0$. Isso significa que, ao atingir um estado $s' \in J_0$ além da recompensa usual associada à transição no estágio atual, há uma recompensa adicional de rescisão, que é independente das ações futuras. Essa recompensa de rescisão pode refletir condições especiais ou eventos significativos associados à entrada no conjunto absorvente.

Podemos incluir este modelo no modelo anteriro usando a nova função de recompensa de um estágio $\tilde{r}_s$com$\tilde{r}_s(s,a,s') = 0$para$s \in J_0$ e

$$\tilde{r}_s(s,a,s') = r_s(s,a,s') + \beta.r_{term}(s').1J_0(s'), s' \in J_0, a \in D(s), s' \in S$$

.

Nesta expressão, a função de recompensa modificada $\tilde{r}_s$ considera o conjunto absorvente J_0. Se o processo de decisão atingir um estado $s' \in J_0$, a recompensa $\tilde{r}_s$ é composta pela recompensa usual do estágio r(s,a,s') e um termo adicional $\beta r_{term}(s')$ multiplicado por $1J_0(s')$ O indicador $1J_0(s')$é uma função que retorna 1 se s' pertence a J_0 e 0 caso contrário. Essa modificação permite considerar recompensas específicas associadas à entrada no conjunto absorvente.

Se a condição (b) da definição 9.1 para conjuntos absorventes for válida para $r(s, a)$, temos:

$$\tilde{r}(s, a) = p\tilde{r}_s(s, a) = r(s, a) + \beta. \sum_{s' \in J_0} .r_{term}(s'), s \in J_0, a \in D(s)$$

Isso reflete a influência das recompensas de rescisão associadas à entrada no conjunto absorvente.

Quando lidamos com Processos de Decisão de Markov (MDPs) que possuem um conjunto absorvente podemos reformular o problema através de um *MÎDP* expandido. Isso é feito ampliando o espaço de estados original S para $\hat{S} = S + \{\hat{S}\}$ onde$\hat{S}$ é um estado arbitrário adicionado. Além disso, definimos $\hat{D(\hat{s})} = A$, $\hat{r}(\hat{s}, a) = \hat{V}_0(\hat{s} = 0$ e ajustamos as probabilidades de transição de acordo.

As probabilidades de transição $\hat{p}(s, a, s')$ são definidas da seguinte forma:

- Se s,s' pertencem ao mesmo espaço de estados S, temos $\hat{p}(s, a, s') = p(s, a, s')$.
- Se s pertence a S e $s' = \hat{s}$, então $\hat{p}(s, a, s') = 0$, indicando que não há transição de s para $\hat{s}$.
- Se s e s' são ambos iguais a $\hat{s}$, então $\hat{p}(s, a, s') = 1$, indicando uma transição de $\hat{s}$ para si mesmo.

Essa abordagem transforma o MDP original em um *MÎDP* que possui um conjunto absorvente $\{\hat{s}\}$. Dessa forma, podemos aplicar resultados clássicos de MDPs sem conjunto absorvente aos MDPs com conjunto absorvente, proporcionando uma análise unificada que abrange cenários como estados absorventes ou terminais.

Agora, exploraremos um exemplo prático de um Processo de Decisão de Markov com um conjunto de estados absorventes. Este exemplo foi extraído do livro "Dynamic Optimization: Deterministic and Stochastic Models"(2016). A presença de estados absorventes é uma característica interessante que introduz nuances adicionais no modelo, e examinaremos como esses estados afetam as decisões e as recompensas.

Exemplo 9.1. Um cliente solicita a produção de $x \in \mathbb{N}$ peças de um produto, e estas são fabricadas em lotes. Os lotes têm tamanhos que pertencem a um subconjunto finito A de $\mathbb{N}$. A restrição é que não são permitidos mais do que N lotes, e alguns lotes podem conter peças danificadas. A probabilidade de um lote de tamanho a conter k peças boas, onde $0 \leq k \leq a$, é representada por $g(a, k)$, sujeita à condição $\sum_{k=0}^{a} g(a, k) = 1$ para todo a, em termos mais simples,

a equação assegura que, somando as probabilidades de todas as diferentes situações possíveis, a chance total de encontrar pelo menos uma peça boa no lote é completa, equivalendo a 100

Para tornar as expressões mais convenientes, definimos $g(a, k) = 0$ para $k \in \mathbb{N}_{0,a}$. O processo de produção encerra-se quando x peças boas são fabricadas ou quando N lotes são utilizados. Essa formulação permite modelar a produção de peças com diferentes tamanhos de lote e a presença de peças danificadas.

Neste contexto, consideramos custos de produção $\alpha \in \mathbb{R}_+$ por unidade e um custo de preparação de valor k, dk $\in \mathbb{R}_+$, então para cada lote de tamanho a, termos o custo de k $+\alpha.a > 0$. Além disso, se houver uma falha e você não conseguir obter exatamente x peças boas no lote, há um custo associado a essa falha. Esse custo é denotado como d.s, onde d é um número real positivo que representa o custo por peça ruim e s é a quantidade de peças que faltam para atingir o número desejado.

Ao modelar problemas relacionados à garantia de qualidade em lotes de produtos, a escolha da função de penalidade desempenha um papel crucial na representação dos custos associados à falta de peças boas. Outras abordagens podem ser consideradas para refletir diferentes aspectos do problema. Por exemplo, uma penalidade por peso pode ser adotada, onde a penalidade é proporcional ao número de peças que faltam, introduzindo um peso w associado a cada peça que faltou. Outra alternativa é utilizar uma penalidade modular, na qual a penalidade p é determinada por uma função f que varia com base na diferença entre o número real de peças boas (k) e o número desejado (x). A função f pode ser a função módulo $|x|$ ou outras funções modulares, proporcionando flexibilidade para ajustar o comportamento da penalidade conforme necessário.

Essas diferentes abordagens oferecem soluções adaptáveis a diversos contextos, permitindo que o modelo represente com precisão os custos associados à falta de comprometimento com o pedido feito pelo cliente. A escolha entre elas dependerá das características específicas e das prioridades do problema em análise.

Cada peça boa produzida além de x gera um custo de sucata de e $\in \mathbb{R}_+$. O objetivo é escolher os tamanhos dos lotes de forma a minimizar os custos esperados de produção ao longo de N estágios.

Denotamos por $s_t \in J = \{0, -1, -2, ...\}$ o número de peças que ainda devem ser produzidas no tempo t, com $s_0 = x$. Se $s_t < 0$, significa que até o momento t há um excesso de produção de peças. Portanto, $J_0 = \{0, -1, -2, ...\}$..Não há restrições quanto ao tamanho do lote, então D(s) = A para todo s. Como o processo para quando atinge J_0, temos p(s, a, s') = g(a, s - s') para todo s, a, s', à medida que s-s' aumenta, a probabilidade de transição diminui, refletindo a ideia

de que quanto maior a diferença entre o estado atual e o próximo estado desejado, menor a probabilidade de transição.

Os custos de estágio único c_1 tem a forma do $\hat{r}_s(s, a, s')$ com $-r_s(s, a, s') = (\alpha a + k).1_{J(s)}$ e $-r_{term}(s') = -es$. Aqui $1_{J(s)}$ é uma função indicadora que assume o valor 1 se $s \in J(s)$ e 0 caso contrário,sendo $J(s)$ o conjunto de estados essenciais, Dessa forma, a expressão $-r_s(s, a, s')$ representa os custos associados à produção, onde o custo é uma combinação linear do tamanho do lote e do custo de preparação. Se $s \in J(s)$, significa que há um excesso de produção de peças, e os custos são aplicados.Já o termo$-r_{term}(s') = -es$ representa os custos associados à rescisão (ou falta) de peças boas, onde s é o número de peças boas produzidas acima da quantidade desejada x, multiplicado pelo custo de penalidade por peça defeituosa e.

Os custos terminais são $C_0(s) = d.s^+$. Assumimos que não há desconto, ou seja, $\beta = 1$. Segue-se então uma versão de minimização dos custos.

$$C_n(s) \min_{a \in A}[\alpha a + k + c_1(s, a) + \sum_{j=0}^{s-1} g(a, j).c_{n-1}(s - j)$$

onde

$$c_1(s, a) = e. \sum_{s' < 0} g(a, s - s').s'$$

Sabendo que e represente o desperdício de peças feitas a mais e que $g(a, s - s')$ seja a probabilidade de ter $s - s'$ peças boas no lote, onde se s' for negativo, isso indica um estado absorvente (um estado em que não há necessidade de fazer mais peças boas). Então a expressão de c_1 calcula o custo associado à produção excessiva de peças além do necessário.

Essencialmente, o custo em cada estágio n é minimizado escolhendo adequadamente o tamanho do lote a para minimizar os custos associados à produção, penalidades por falta de peças boas e custos acumulados dos estágios anteriores.

Obviamente J_0 é absorvente, porque, uma vez que o processo atinge J_0, o mesmo permanece lá de forma permanente. Isso é evidenciado pela condição de parada do processo: quando o número de peças a serem produzidas alcança ou excede o alvo ($s_t \leq 0$), o processo termina e atinge o conjunto absorvente J_0. portanto $C_n = 0$ em J_0 para todo $n \geq 0$.

n	$f_n(s)$				$C_n(1)$	$C_n(10)$
	$s = 1$	$s = 2$	$s = 3$	$s = 4 - 10$		
1	2	2	3	5	8.240	42.500
2	2	3	3	5	8.272	35.000
3	2	3	3	5	8.273	36.042
4	2	3	3	5	8.273	36.202
5,6	2	3	3	5	8.273	36.204

Tabela de dados do exemplo 11.1. Fonte: Hinderer, K., Rieder, U., & Stieglitz, M. (2016). Dynamic Optimization: Deterministic and Stochastic Models.

A Tabela é resultado de lei de transição binominal onde cada item produzido tem a probabilidade p=0,9 de ser bom. Então g(a,.) é discreto com densidade $B_i(a, 0.9)$ e

$$g(a, j) = \binom{a}{j} 0,9^j.0,1^{a-j}, 0 \neq j \neq a$$

Para elucidar detalhadamente os cálculos apresentados na tabela, seria necessário conhecer as probabilidades específicas $g(a, k)$, que representam a chance de um lote de tamanho a conter k peças boas. Contudo, o exemplo não fornece explicitamente essas probabilidades. Os parâmetros considerados são $s_0 = x = 10$, $\alpha = 2$, $k = d = 5$ e $e = 1$. Com esses valores, a função $f_n(s)$ está aumentando em s, o que significa que, à medida que o número de peças a serem produzidas aumenta, a escolha ótima de tamanho de lote também aumenta. Isso é intuitivo, pois produzir mais peças geralmente requer lotes maiores para otimizar os custos.

No entanto, observa-se que C_n não está diminuindo em n, como evidenciado por $C_4(10) = 36,202 > C_3(10) = 36,042$. Essa observação contraintuitiva ocorre devido à restrição no espaço de ação $A = \{2, 3, 5\}$, isso significa que, em determinados estágios, a escolha ótima pode não envolver a interrupção do processo mesmo quando seria intuitivo fazê-lo, como a interrupção do processo de produção em estados baixos $s < 0$. Mesmo que a produção continue além da quantidade desejada, os custos mínimos futuros, dentro da restrição A, podem justificar essa continuação, nestes estados, os custos mínimos de produção futuros excedem os custos de penalidade associados a uma parada imediata.

Por exemplo, $C_1(1) = 8,240 > C_0(1) = 5$, o que significa que, mesmo produzindo apenas uma peça, o custo mínimo futuro de produção, considerando lotes de tamanho 2, 3 ou 5, é maior do que o custo de penalidade para uma parada imediata. Isso explica por que a escolha de A não permite interromper o processo de produção em estados baixos, resultando em custos mais

altos ao longo do tempo.

n	$f_n(s)$				$C_n(1)$	$C_n(10)$
	$s = 1$	$s = 2$	$s = 3$	$s = 4 - 10$		
1	0	2	3	5	5.000	42.500
2	0	2	3	5	5.000	35.000
3-6	0	2	3	5	5.000	34.787

Tabela de dados do exemplo 11.1. Fonte: Hinderer, K., Rieder, U., & Stieglitz, M. (2016). Dynamic Optimization: Deterministic and Stochastic Models.

A tabela fornecida agora incorpora a consideração de um tamanho de lote igual a zero. Ao introduzir um tamanho de lote igual a zero, percebemos que, no primeiro estágio ($n = 1$), o custo para todas as quantidades de peças (s) é fixado em 5.000. Isso ocorre porque, quando o tamanho do lote é zero, não há produção, e o custo é apenas o custo terminal associado à quantidade de peças desejadas. Nos estágios seguintes ($n = 3 - 6$), mesmo com o tamanho do lote zero, observamos uma diminuição nos custos para $C_n(10)$. Isso sugere que, em determinados cenários, a escolha de um tamanho de lote zero pode resultar em custos mais baixos ao longo do processo de produção. Agora, C_n está de fato diminuindo em n, conforme a indução em n como $C_1 \leq C_0$,indicando uma melhoria no desempenho em comparação com a primeira tabela.

A adoção de um tamanho de lote zero pode ser uma estratégia viável em diversos contextos do cotidiano, onde a flexibilidade na produção é crucial e as demandas do mercado são variáveis. Em indústrias onde a produção é guiada por pedidos específicos dos clientes, adotar lotes de tamanho zero evita a necessidade de manter estoques excessivos de produtos personalizados. Isso não apenas reduz custos de armazenamento, mas também atende de maneira mais eficaz às demandas únicas dos consumidores. Outro exemplo são empresas que buscam minimizar estoques e enfrentam o desafio de evitar excessos. Lotes de tamanho zero permitem uma produção alinhada com a demanda imediata, evitando acúmulos desnecessários e os custos associados à manutenção de estoques elevados.

O exemplo apresentado ilustra a importância da flexibilidade na escolha dos tamanhos de lotes em problemas de otimização de custos em processos de produção. Ao considerar o tamanho do lote a=0 como uma opção adicional, foi possível reduzir significativamente os custos mínimos de produção ao longo do tempo. A análise comparativa entre as duas versões do modelo destacou que a inclusão dessa opção permitiu uma melhor adaptação às condições específicas de produção.

10 MDPS COM ESTADO INICIAL ALEATÓRIO

Neste capítulo, adentraremos no fascinante domínio dos Processos de Decisão de Markov (MDPs) com um enfoque especial em um aspecto crucial: o estado inicial aleatório. Exploraremos as definições fundamentais associadas a MDPs com estado inicial aleatório, compreendendo a aleatoriedade do estado inicial, suas implicações e como isso impacta o desenvolvimento e a análise de políticas em um horizonte temporal específico.

Definição 10.1. Em um Processo de Decisão de Markov (MDP) com um número finito de estados e um horizonte temporal definido como $N \in \mathbb{N}$, consideramos um problema específico relacionado aos estados iniciais.

O estado inicial ζ_0 é escolhido de forma aleatória e é estocasticamente independente dos estados futuros ζ_t (isso significa que a escolha do estado inicial não depende dos estados que virão a seguir).

p_0 representa a distribuição discreta de probabilidade do estado inicial ζ_0 no conjunto de estados S.

Agora, definimos um problema específico chamado MDP_N, onde o estado inicial é escolhido aleatoriamente de acordo com a distribuição p_0. Este problema é discretizado para $\pi \in \mathbb{F}^N$ (sequências de políticas), considerando a densidade discreta conjunta das variáveis aleatórias $(\eta_t)_0^N$ no espaço amostral S^{N+1}, dado por:

$$\begin{aligned} p_\pi(s_0, s^N) &= p_0(s_0) \cdot p_{\pi 0}(s_0, s_1) \cdot \ldots \cdot p_{\pi N-1}(s_{N-1}, s_N) \\ &= p_0(s_0) \cdot p_{\pi 0}(s_0, s_1) \cdot \ldots \cdot p_{\pi N-1}(s_{N-1}, s_N) \end{aligned}$$

Isso significa que a probabilidade de uma sequência específica de estados iniciais e futuros, $(s_t)_0^N$, ocorrer, sob uma política específica π, é dada pelo produto das probabilidades de transição de um estado para o próximo, começando com a probabilidade do estado inicial s_0 de acordo com p_0.

No cenário em que p_0 atribui toda a massa a um ponto fixo (um estado específico) s_0, isso nos remete ao problema $MDP_N(s_0)$.

O que isso significa é que o estado inicial é determinístico e fixado em s_0. Em outras palavras,

não há incerteza na escolha do estado inicial; ele é pré-determinado por p_0 atribuir toda a probabilidade a s_0.

Agora, introduzimos duas funções, $V_{N\pi}$ e V_N, definidas no conjunto de estados S, como discutido anteriormente em estudos anteriores.

A recompensa esperada no estágio $V_{N\pi}(p_0)$ para a densidade discreta inicial p_0 é definida pela expressão:

$$V_{N\pi}(p_0) = p_0 V_{N\pi} = \sum_{s_0 \in S} p_0(s_0) V_{N\pi}(s_0)$$

Sendo:

- $V_{N\pi}(p_0)$: Isso representa a recompensa esperada para o estágio N sob a política π, considerando a densidade discreta inicial p_0.

- $p_0 V_{N\pi}$: É o produto de p_0 com $V_{N\pi}$, o que significa que estamos ponderando cada possível estado inicial s_0 pela probabilidade $p_0(s_0)$ atribuída a ele.

- $\sum_{s_0 \in S} p_0(s_0) V_{N\pi}(s_0)$: Esta é a soma sobre todos os estados possíveis ponderados pela probabilidade atribuída por p_0. Isso representa a recompensa total esperada para o estágio N sob a política π considerando a densidade discreta inicial p_0.

Então o problema MDPn(p0) siginifica maximizar $\pi \rightarrow V_{N\pi}(p_0)$, ou seja, encontrar:

- $V_N(p_0) = sup_{\pi \in \mathbb{F}} V_{N\pi}(p_0)$, a recompensa máxima esperada do estágio N para a distribuição de probabilidade p_0;
- Uma política $\pi^* \in \mathbb{F}^*$ que é p0-ótima no sentido de que maximiza $V_{N\pi}(p_0)$ em $\mathbb{F}^{\mathbb{N}}$.

Proposição: (Solução de MDPs com estados iniciais aleatórios) Se $\pi* \in \mathbb{F}*$ é ótimo para MDPn também é otimo p0-ótimo para MDPn(p0) para cada densidade p_0 em S e $V_N(p_0) = p_0 V_N$

Demonstração. Para cada $\pi^* \in \mathbb{F}^*$.
$V_N(p_0) \geq V_{N\pi*}(p_0)$, pois a política ótima deve, no mínimo, ser tão boa quanto qualquer outra

política possível.

$V_{N\pi*}(p_0) = \sum_{s \in S} p_0(s).V_{N\pi*}(s)$, por definição de $V_{N\pi*}(p_0)$

$\sum_{s \in S} p_0(s).V_{N\pi*}(s) \geq \sum_{s \in S} p_0(s).V_{N\pi}(s) = V_{N\pi}(p_0)$

Novamente por definição , $V_{N\pi}(p_0) = p_0 V_{N\pi}$

Então, $V_N(p_0) = p_0 V_N$ ■

Introduzimos as funções V_N e $V_{\pi N}$, definidas no conjunto de estados S, e exploramos a recompensa esperada no estágio N para a densidade discreta inicial p_0 . Formulamos a proposição de que a solução ótima para MDPN é também uma solução ótima para MDPN(p0)

Dessa forma, estabelecemos uma base sólida para abordar MDPs com estados iniciais aleatórios, considerando a incerteza na escolha do estado inicial e sua influência nas políticas e recompensas esperadas ao longo do tempo. Este capítulo nos fornece as ferramentas necessárias para abordar problemas práticos em que a aleatoriedade no estado inicial desempenha um papel crucial nas decisões sequenciais.

11 PROBLEMA DE PARADA

Abordaremos um problema relacionado ao tempo de parada em processos de decisão Markovianos (MDP). Exploraremos um exemplo específico para ilustrar a questão em foco. Em seguida, apresentaremos uma explicação detalhada, destacando as considerações relevantes e os conceitos fundamentais de MDP que serão aplicados para resolver o problema. Vamos prosseguir com a análise desse cenário, buscando compreender como políticas apropriadas podem influenciar os resultados em termos de ganhos esperados.

Exemplo 11.1. Considere uma cadeia de Markov de N estágios com espaços de estados finitos J e matriz de transição $q(s, s')$ com s e $s' \in J$. Tratamos primeiro do caso em que a cadeia começa em um determinado s_0 e depois o caso em que ela começa de acordo com uma densidade p_0 em J.

Queremos parar a cadeia em um dos N pontos de tempo $0 \leq t \leq N-1$ ou continua até N, se a cadeia ainda não estiver parada no tempo $0 \leq t \leq N-1$ e nos estados $s \in J$. Obtém-se $g(s) \in \mathbb{R}$ se estiver parada (ação=1) e $h(s) \in \mathbb{R}$ se não estiver parada (ação=0).

Se a cadeia não for parada antes do tempo N e se estiver em $s \in J$, ganha-se $g(s)$. De acordo com qual política obtemos os máximos ganho esperado $V_N(s_0)$?

Usando o seguinte MDP com conjunto absorvente $\overline{s}$ daí $S = J + \overline{s}, s_t \in J$ denota o estado no tempo t, se a cadeia ainda não foi interrompida enquanto $s_t = \overline{s}$significa que a cadeia foi interrompida antes do tempo t.

- A= D(s)=(0,1) para $s \in J$.
- $p(s, 0, s') = q(s, s'), s' \in J$ e $p(s, 1, \overline{s}) = 1$
- r(s,a)= g(s) se a=1 e h(s) se a =0
- $V_0(s) = g(s)$.

Como $\overline{s}$ é absorvente $V_n(\overline{s}) = 0$ para todo $n \geq 0$, por definição de MDP com conjunto absorvente.

Podemos trazer esse exemplo númerico para o cotidiano de forma sucinta.

Considere um agente que está gerenciando suas finanças pessoais e planejando sua aposentadoria. O conjunto de estados J representa diferentes estratégias de investimento, como ações, títulos, fundos imobiliários, etc. O conjunto absorvente $\overline{s}$ representa a decisão de se aposentar.

O agente, ao longo do tempo, decide onde investir seu dinheiro, escolhendo entre diferentes estratégias de investimento. Cada estratégia (estado em J) tem associada uma probabilidade de sucesso e uma recompensa financeira. A interrupção ocorre quando o agente decide se aposentar, indicando que atingiu seu objetivo financeiro. Uma vez aposentado ($\overline{s}$), esse estado é absorvente, e não há mais transições ou decisões relacionadas aos investimentos.

Voltando ao nosso exemplo númerico, temos que :

$$v_n(s) = qV_{n-1}(s) = \sum_{j \in J} q(s,j).V_{n-1}(j)$$

por VI implicando em J:

$$V_n = max\{h + \beta v_n, g\} \tag{11.1}$$

Além disso uma política ótima $\pi* = (f_n)_N^1$ para MDP é dada pelos maiores maximizadores f_n nos estágios n, determinados por $f_n(\overline{s}) = 1$. Isso implica que, ao longo do tempo, a política ótima busca decisões que levem ao estado absorvente .

A política ótima $\pi*$ prescreve parar no estágio n e no estado $s \in S$(desde que não tenha parado antes) se e somente se $g(s) = V_n(s)$. Essa condição reflete a ideia de que, sob a política ótima, a decisão de parar em um determinado estágio e estado é guiada pelo ganho imediato associado a essa ação

Chamamos o conjunto B_n de estados $s \in S$ onde a=1 é ótimo no estágio $n \geq 1$. Esse conjunto é denominado conjunto de parada no estágio n.

Obviamente

$$B_n = [g \geq V_n] = [g = V_n]$$

Surge naturalmente na análise devido à definição de B_n e à relação entre a função de recompensa imediata g e o valor esperado V_n. O conjunto B_n representa os estados nos quais a ação de parar no estágio n é ótima. A expressão $[g \geq V_n]$ denota o conjunto de estados onde o ganho imediato ao parar é maior ou igual ao valor esperado no estágio n, enquanto $[g = V_n]$ representa os estados onde esses valores são precisamente iguais. A igualdade entre esses conjuntos reflete a condição de que parar no estágio n é a escolha ótima, pois o ganho imediato é igual ou superior ao valor esperado nesse estágio. Portanto, a igualdade é intuitiva e resulta diretamente da relação entre a otimalidade da ação e as características das funções g e V_n no contexto do MDP.

Observe que o estado absorvente $\overline{s}$ está presente em cada conjunto de parada. A política ótima aconselha a interrupção no primeiro instante t, onde a recompensa imediata ao parar, $g(s_t)$, é pelo menos igual à recompensa imediata $h(s_t)$ de continuar, acrescida da recompensa descontada esperada para os períodos restantes ($\beta V_n(s)$) ($\beta V_n(s) \leq g(s) - h(s)$).

Essa condição reflete a busca pela maximização do momento de parada, equilibrando as recompensas imediatas com as recompensas futuras descontadas. O estado absorvente $\overline{s}$ é central nesse contexto, pois representa a possibilidade de interrupção em qualquer estágio. A equação estabelece a condição crítica para determinar o ponto ideal de parada, considerando as implicações de recompensas imediatas e futuras descontadas.

Assim, deve-se parar o sistema no momento

$$\tau_N = \tau_N((s_t)_0^{N-1}) = min\{0 \leq t \leq N-1 : s_t \in B_{N-t}\} =$$

$$min\{0 \leq t \leq N-1 : \beta v_{N-1}(s_t) \leq g(s_t) - h(s_t)\}$$

Portanto, o momento ideal para encerrar o sistema, representado por τ_N, é determinado como o primeiro instante, dentro do horizonte de decisão $N-1$, em que o estado atual s_t pertence ao conjunto B_{N-t}. Essa decisão de parar é tomada quando a recompensa descontada esperada para os períodos subsequentes ($\beta v_{N-1}(s_t)$) é menor ou igual à diferença entre as recompensas imediata ($g(s_t)$) e de continuação ($h(s_t)$). Essencialmente, τ_N é o momento em que a interrupção do sistema é ótima, equilibrando as recompensas imediatas com as expectativas de ganhos futuros descontados.

O tempo de parada ótimo τ_N no exemplo apresenta uma característica notável: para $0 \leq t \leq N-1$, o conjunto $[\tau = t]$ - definido como $\{(s)_0^{N-1} \in S^N : \tau_N(s_0^{N-1}) = t\}$ - não é influenciado pelos estados futuros $s_{t+1}, s_{t+2}, ..., s_{N-1}$. Em outras palavras, a decisão de parar ou continuar no instante t depende exclusivamente do estado presente s_t.

Isso pode ser mais claramente entendido quando comparado a uma cadeia de Markov. Em uma cadeia de Markov, a probabilidade de transição para o próximo estado depende apenas do estado atual, ou seja, é um processo "sem memória". Da mesma forma, a característica aqui ressaltada implica que a decisão de parada ótima no tempo t não considera os estados futuros, tornando-a uma escolha dependente apenas do estado presente s_t.

Essa independência em relação aos estados futuros simplifica a análise do processo de decisão, tornando-o mais gerenciável e destacando uma abordagem "sem memória"na tomada de decisões dentro do contexto do problema abordado.

Proposição: (A Solução do problema de parada)

a) $V_n = g + e_n^-, n \geq 0$, onde $e_0 = 0, e_1 = g - h - \beta.qg$ e $e_{n+1} = e_1 - \beta q e_n^-$;

b) V_n esta aumentando em n; portanto existe $V = lim_{n\to\infty} V_n$ e B_n está diminuindo em n.

c) Assume $\beta < 1$ e coloque $B = \cap_1^\infty B_n, B = [V = g]$. Então V é a solução única da equação de otimalidade $V \leq \max h/(1-\beta) + \max g < \infty$ e 1_B é assintoticamente ótimo.

d) Suponha que B_1 seja quase absorvente no sentido de que $p(s, 0, B_1) = 1$ para $s \in B_1$. Então $V_n = g$ em $B_1, n \geq 1$ e todos os conjuntos de parada B_n são iguais a B_1

Demonstração. **(a)** Por indução em n, iremos fazer para n+1, ou seja $V_{n+1} = g + e_{n+1}^-$

Sabemos por (11.1) e pela propriedade $\max\{a, b\} = b + \max\{a - b, 0\}$, *temos* : $V_{n+1} = \max\{h + \beta v_n, g\} = g + \max\{h + \beta v_n - g, 0\}$

Visto que $e_{n+1} = e_1 - \beta q e_n^-$ e $e_1 = g - h - \beta qg$, temos: $e_{n+1} = g - h - \beta qg - \beta q e_n^-$.

Sabemos também que $V_n = g + e_n^-$.

Portanto: $e_{n+1} = g - h - \beta q V_n$.

Então: $V_{n+1} = g + e_{n+1}^-$

(b): Nesse item, será necessário mostrar 3 situações:
1º situação:

Mostrar que V_n está aumentando em n

Hipótese de indução: $V_n(s) \geq V_{n-1} .... V_{n+1} \geq V_n(s)$.

Sabemos por VI em 6.1, temos:

$$V_{n+1} = \sup_{a \in D(s)} [r(s, a) + \beta \sum_{s' \in J} p(s, a, s').V_n(s').$$

Por hipótese de indução:

$$V_{n+1} \geq \sup_{a \in D(s)} [r(s,a) + \beta \sum_{s' \in J} p(s,a,s').V_{n-1}(s')$$

Logo, $V_{n+1} \geq V_n(s)$

2º situação:

Mostrar que $V = lim_{n \to \infty} V_n$

A condição $V_n \leq V_1$ para todo n implica que a sequência V_n é limitada superiormente, pois V_1 é uma cota superior para todos os termos da sequência. Portanto, a conclusão é que, devido à sequência V_n ser crescente e limitada superiormente (limitada por V_1), ela deve convergir para um limite, representado por $V = \lim_{n \to \infty} V_n$. Esse é um resultado fundamental associado a sequências que são ao mesmo tempo crescentes e limitadas. O limite V é, então, a convergência da sequência V_n à medida que n tende para o infinito.

3º situação:

Mostrar que B_n está diminuindo

$B_{n-1} \supset B_n$ = tal que
$B_{n-1} = \{s | g(s) \geq V_{n-1}\}$ e $B_n = \{s | g(s) \geq V_n\}$
Logo, $V_n \geq V_{n-1}$, então se
$s \in B_n \rightarrow g(s) \geq V_n(s) \geq V_{n-1} \rightarrow g(s) \geq V_{n-1}(s) \rightarrow s \in B_{n-1}$

(c):

(d): Iremos dividir a demonstração desse item em duas partes

1º situação:

Mostrar que em B_1 o $V_{n+1}(s) = V_1(s) = g$

É dado que $V_0 = g$ e $V_n = max\{h + \beta v_n, g\}$ e $v_n(s) = \sum + j \in Jq(s,j).V_{n-1}(j)$

Por hipótese de indução:
$V_n(s) = g$ se $s \in B_1$, ou seja, $h + \beta v_n \leq g$.

Para n+1 e substitundo v_n

$V_{n+1}(s) = max\{h + \beta \sum_{j \in B_1} q(s,j) V_n(j), g\}$

Como $V_n(s) = g$

$V_{n+1}(s) = max\{h(s) + g.\beta \sum_{j \in B_1} q(s,s'), g\}$

Dito que $p(s,0,B_1) = 1$

$V_{n+1}(s) = max\{h(s) + g.\beta, g\}$

Agora iremos comparar com V_1 se chegamos na mesma solução:

$V_1(s) = max\{h + \beta \sum_{j \in B_1} q(s,j) V_0(j)$

Sabemos que $V_0 = g$

$V_1(s) = max\{h(s) + g.\beta, g\}$

Agora pela definição de B_n, para n=1 , ou seja, definição de B_1 onde diz que $B_1 = [g \geq V_1] = [g = V_1]$

Chegamos em $V_{n+1}(s) = V_1(s) = g$

2º situação:

Mostrar que todos os conjuntos de parada B_n são iguais a B_1

Pelo item (b) temos que:

$B_1 \supset B_2 \supset B_3$...

Agora pelo item (d): $B_1 \subset B_2$, então $B_1 = B_2$

$B_1 \subset B_3$, então $B_1 = B_3$

. . . $B_1 \subset B_n$, então $B_1 = B_n$

■

Os pontos da proposição detalham a convergência e as propriedades ótimas da solução, reforçando a eficácia da abordagem "sem memória"na resolução do problema. A ideia de conjuntos de parada B_n sendo quase absorventes, especialmente quando B_1 é quase absorvente, destaca a influência significativa do estado presente em relação aos estados futuros na determinação da política ótima de parada.

Assim, a independência em relação aos estados futuros não apenas simplifica a análise do processo de decisão, tornando-o mais gerenciável, mas também ressalta a eficácia de uma

abordagem "sem memória"na tomada de decisões dentro do contexto do problema abordado. Esta característica, aliada à proposição apresentada, proporciona uma base sólida para a compreensão e resolução do problema de parada ótima.

CONCLUSÃO

Ao concluir esta pesquisa, é evidente que a jornada pelos conceitos probabilísticos fundamentais, passando pela intricada análise das Cadeias de Markov e adentrando nos Processos de Markov, proporcionou uma compreensão abrangente e aprofundada dessas estruturas estocásticas. A incorporação de exemplos teóricos e práticos do cotidiano não apenas enriqueceu a abordagem teórica, mas também facilitou a assimilação dos conceitos pelos leitores, estabelecendo uma conexão mais tangível entre a teoria e sua aplicação na prática.

Diferenciando-se da homogeneidade presente nas Cadeias de Markov, exploramos as características não homogêneas dos Processos de Markov, onde as dinâmicas temporais, as transições de estado e a contabilização de recompensas não são constantes. Essa distinção crucial confere aos Processos de Markov uma versatilidade e adaptabilidade particular, tornando-os mais adequados para modelar uma gama diversificada de fenômenos

Esta pesquisa foi estrategicamente delineada para construir um entendimento progressivo e integrado dos conceitos abordados. Iniciamos com a análise de um "Problema de Decisão de Markov de um Período", estabelecendo uma base sólida para compreensão da dinâmica fundamental dos Processos de Markov em um contexto simplificado. Em seguida, exploramos a "Exploração com Aleatoriedade em MDPs", investigando a influência da aleatoriedade nas decisões em Processos de Decisão de Markov (MDPs) e preparando o terreno para a análise de um "Processo de Decisão de Markov de Dois Estados", que introduziu dinâmicas mais complexas com a presença de múltiplos estados.

O capítulo sobre o "Controle Estocástico de Inventário de Produto Único"trouxe uma dimensão prática, ilustrando a aplicação dos Processos de Markov no cenário real do controle de inventário .No contexto dos "MDPs com um Conjunto de Estados Absorventes", analisamos como a presença desses estados impacta a dinâmica do processo. Em seguida, exploramos os "MDPs com Estado Inicial Aleatório", adicionando uma camada de complexidade ao considerar a incerteza no estado inicial do processo. Finalmente, concluímos nossa jornada abordando o "Problema de Parada", destacando a importância de compreender as condições que levam à tomada de decisão para encerrar o processo.

Essa sequência cuidadosamente elaborada teve como objetivo proporcionar aos leitores uma progressão natural de conceitos, partindo de fundamentos mais simples e evoluindo para contextos mais complexos. Ancorados em exemplos teóricos e práticos, nosso intuito foi não apenas apresentar uma variedade de cenários de aplicação, mas também possibilitar uma compreensão aprofundada das nuances inerentes aos MDPS e seus desdobramentos em diferentes domínios

12 REFERÊNCIAS

P. Bremaud, *Probability Theory and Stochastic Processes*, 2ª ed., New York: Springer, 1981,

K. Hinderer, U. Rieder, M. Stieglitz, *Dynamic Optimization: Deterministic and Stochastic Models*, Springer, 2010.

M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Nova Jersey: John Wiley e Sons, 1994.

H. Tijms, *A First Course in Stochastic Models*, Chichester: Wiley, 2003.