Marcello Nery Garcia Vidal de Barros

Causalidade, Informação, Causalidade da Informação

Belo Horizonte, MG - Brasil

Abril, 2016

Marcello Nery Garcia Vidal de Barros

Causalidade, Informação, Causalidade da Informação

Dissertação de mestrado apresentada ao programa de pós-graduação em Física da Universidade Federal de Minas Gerais como requisito parcial para obtenção do título de Mestre em Física

Universidade Federal de Minas Gerais - UFMG

Departamento de Física

Programa de Pós-Graduação

Orientador: Marcelo de Oliveira Terra Cunha

Coorientador: Rafael Luiz da Silva Rabelo

Belo Horizonte, MG - Brasil

Abril, 2016

Marcello Nery Garcia Vidal de Barros

Causalidade, Informação, Causalidade da Informação

Dissertação de mestrado apresentada ao programa de pós-graduação em Física da Universidade Federal de Minas Gerais como requisito parcial para obtenção do título de Mestre em Física

---

Marcelo de Oliveira Terra Cunha
Orientador

Rafael Luiz da Silva Rabelo
Coorientador

Carlos Henrique Monken

Reinaldo Oliveira Vianna

Belo Horizonte, MG - Brasil
Abril, 2016

"Don't ever empty the bucket of mystery.
Never let people define what you do.
It's not about zigging when you should zag.
It's not about doing something
unprecedented and unpredictable.
It's just about never being a word,
or something that is not in the process of transformation."
(Marilyn Manson)

Dedico este trabalho aos efeitos que o terão como causa comum

Agradecimentos

Os agradecimentos que farei neste trabalho valem não somente para tal, mas para todo o meu percurso. Por esse motivo, começo agradecendo aos meus pais[1] por terem me dado condições de priorizar os estudos. Ao meu amigo Arthur, que considero o irmão que escolhi, e sua mãe, tia Marcia, que considero minha segunda mãe, por todo o apoio e consideração. Agradeço também ao meu *amoro*, Thomás, por ter me dado suporte nos momentos em que eu mais precisei, me ajudando mesmo em momentos em que não percebia que estava me ajudando.

Como a carreira não é algo de natureza markoviana, agradeço também aos meus amigos e professores[2] da UFF, com quem passei loucos, divertidos e importantíssimos anos de graduação. Em especial aos amigos Caio, Vivian Maria, Wagner, Davorzin, Magnin, Dieguin, João Gnu, Gabriel Jake, Renan & Vivian afilhados, Allan & Layla[3], Gracielle, Karenn, e aos meus amigos de albergue Lar Solar: Cellõa, Filipe, Danilo, Theusão, Paulista, Vice, Pedro Pablo de *Venêssuela* e Aracelis.

Aos professores/amigos, agradeço especialmente ao Kaled Dechoum, Jorge Sá, Beatriz Boechat, Ruth Bruno e Jürgen Fritz, por serem excepcionais profissionais e gerarem muita inspiração, pelo menos para mim. Agradeço mais especialmente ainda ao meu ex-orientador, Thiago Rodrigues, por ter me inserido na área de Informação Quântica, me dado suporte com todas as minhas dificuldades[4], e me indicado realizar meu mestrado na UFMG como terráqueo[5].

Além disso, agradeço aos meus amigos da minha (quase) terra natal, Rio das Ostras: Yasmin, Marica, Paulinha, Winnie, Rômulo, Lucas Dantas, Luis Teske, Zaru, Hoogle e, claro, Natasha Safady[6].

Já na terra do pão de queijo, agradeço primeiramente ao Terra, por ter me aceitado como orientando e me dado liberdade[7] e motivação para estudar e trabalhar com qualquer coisa que eu quisesse, além de sempre vir com respostas, muitas vezes misteriosas e outras muitas milagrosas, que me ajudaram a resolver problemas no trabalho e motivaram questões a serem analisadas após esta etapa. Em segundo lugar,

1 Que tiveram um papel fundamental em não me deixar morrer de burrice na infância, coisa que certamente aconteceria comigo se seleção natural ocorresse com humanos na era atual.
2 Destaco que a interseção entre esses dois conjuntos de pessoas não é nula.
3 A partir de algum momento, é comum que um par de pessoas se acople de maneira que mencionar cada um separadamente gere perda de informação sobre o par.
4 inclusive algumas que não envolviam especificamente a minha pesquisa.
5 Nome especial dado aos orientandos do prof. Marcelo Terra Cunha.
6 Quanto a esta, assumo uma grande incerteza na distância, dado que mora em Cabo Frio, não em Rio das Ostras.
7 Até demais...

ao Rafael Rabelo, por ter me orientado nessa pesquisa, sempre alimentando minha vontade de trabalhar mais no tema e motivando constantemente meu desenvolvimento através de discussões extremamente produtivas.

Aos amigos que fiz no departamento de física: Marcantonio[8] , Eliel, Tchê, Diego, Alana, Cobra, Leonelson, Jojo, Egleito, Lucas Marçal, Paulin & Anne, etc. Aos amigos que fiz no EnLight: Ludchampz, André Luiz, Hakob, Davi, Mario, Roberto, Sheilla, Renan, Paula, Tutu, Betão, Gilberto, Pierre, Duty, Ana Paula, Roger[9], Denise. Finalmente aos amigos terráqueos, principalmente Jessga, Gláucia, Tassius, Léo, Cristhiano, Gabrielzin e Bárbara. Agradeço também a existência de uma das pessoas mais simpáticas e prestativas que tive o prazer de conhecer na vida, a querida Shirley da biblioteca de física da UFMG.

À minha psicóloga, Drª. Lilany Vieira, por ter me ajudado a manter a estabilidade, na medida do possível, para dar conta do meu trabalho sem ter surtos emocionais e panes cerebrais constantes.

Agradeço também ao Rafael Chaves (Planeta), pelo seu trabalho que despertou meu interesse por teoria de causalidade, além de todas as outras coisas que essa dissertação engloba.

Agradeço ao desenho animado *Adventure Time* por ter ensinado[10] tantas coisas sobre a vida. Finalmente, agradeço às empresas de cigarro, café e cerveja, pois o consumo desses itens é condição necessária para o meu desenvolvimento de pesquisa em física.

Agradeço à CAPES e ao CNPq pelo financiamento de minha pesquisa.

[8] Balança o cachin!
[9] Que Roger?
[10] De uma maneira muito estranha e louca, mas eficaz.

Resumo

Sistemas quânticos têm a capacidade de apresentar, entre suas partes, correlações mais fortes que as apresentadas em sistemas clássicos. Algumas dessas correlações são as denominadas *não-locais*, enquanto correlações apresentadas em sistemas clássicos são denominadas *locais*.

Correlações não-locais não implicam ações à distância, ou comunicação super luminal, pois obedecem à condição de não-sinalização. Existem correlações não-sinalizantes mais fortes que as correlações quânticas, sendo o conjunto de correlações quânticas um subconjunto das correlações não-sinalizantes.

As correlações quânticas são determinadas através da maneira com que probabilidades de obtenção de resultados condicionadas a determinadas escolhas de medição são extraídas, levando em conta um operador densidade que represente o estado quântico em questão e o conjunto de medições. Além disso, não se conhece um princípio físico capaz de determinar que as correlações quânticas são as mais não-locais que a natureza pode exibir. Por este motivo, surgiu uma importante questão: Existiria um princípio físico capaz de explicar os limites da não-localidade quântica?

Na tentativa de alcançar uma resposta positiva para essa questão, alguns princípios foram apresentados nos últimos anos, sendo o mais destacado o princípio denominado *Causalidade da Informação*. Tal princípio diz que, ao considerar um sistema com duas partes correlacionadas entre si em conjunto com o envio de uma mensagem de d bits de informação de uma parte para outra, o ganho de informação nesse processo é limitado a d bits.

A mecânica quântica satisfaz o princípio de Causalidade da Informação. Os melhores critérios conhecidos para esse princípio são capazes de excluir a maior parte das correlações não-sinalizantes mais fortes que quânticas, nos cenários mais simples. Entretanto, existem ainda correlações supra-quânticas que obedecem tais critérios.

Recentemente, uma nova abordagem tem sido utilizada para obter critérios mais fortes para Causalidade da Informação. Esta abordagem faz uso de elementos das teorias clássica e quântica da informação, além da teoria de causalidade, sendo esta última uma poderosa e interessante teoria que relaciona a observação de correlações entre eventos com relações de causa e efeito. Essa teoria tem encontrado aplicações em várias áreas do conhecimento.

Nesta dissertação apresentaremos os principais elementos das teorias da informação, da causalidade e da não-localidade, a fim de analisar o princípio da causalidade da informação e seus mais fortes critérios. Revisaremos os principais resultados obtidos e apresentaremos resultados parciais obtidos para uma generalização de um critério de

causalidade da informação, considerando cenários em que muitas cópias de recursos não locais estão disponíveis.

Abstract

Quantum systems are able to show correlations between its parties that are stronger than correlations found in classical systems. Some of those strong correlations are called *nonlocal*, as for the classical correlations, they are called *local*.

Nonlocal correlations do not imply long-distance actions or superluminal communication, because they satisfy the nonsignaling condition. There are nonsignaling correlations stronger than quantum correlations because the quantum correlation set is a strict subset of the nonsignaling correlation set.

Quantum correlations are determined by how one obtains the probabilities of geting results conditioned to the measurement choices, considering the density operator describing the quantum state and the set of measurements. Moreover, a physical principle capable of justify quantum nonlocality is unkown. Because of that, an important question appeared: Would there be a physical principle which is able to explain the quantum nonlocality limits?

Within the attempt to reach a positive answer for this question, some principles were presented in the last years. One of the most highlighted ones is the so-called *Information Causality*. This principle states that when one considers a system with two correlationed parties and the sending of a message from one party to the other containing d bits of information, the information gain within this process is limited to d bits.

Quantum correlations satisfy the Information Causality principle. The best criteria known for the principle are able to discard most of the nonsignaling correlations that are stronger than quantum correlations, in the most simple scenarios. However, there still are supra-quantum correlations that obey such criteria.

Recently, a new approach has been adopted for the achievement of a stronger criterion to represent the Information Causality principle. Such approach is based on classical and quantum information theory, and causal inference theory, the latter being a powerful and interesting theory that relates the correlations with causal relations. This theory has found applications in many areas.

This thesis presents the main elements of information theory, causal inference theory and nonlocality, aiming the analysis of the Information Causality principle and its strongest criterion. The main results obtained for the principle will be revised, and it will be presented a partial result obtained for a generalization of the more recent Information Causality criterion by considering scenarios in which multiple copies of the nonlocal resource are available.

Sumário

Introdução

Entre o final do século *XIX* e o início do século *XX*, uma revolução de descobertas em física ocorreu, principalmente em virtude do surgimento da relatividade restrita, que se propôs a tratar de sistemas envolvendo grandes escalas de velocidade, e da mecânica quântica, que trata de sistemas em escalas microscópicas. Essas descobertas trouxeram uma nova visão de mundo para a humanidade que, pouco antes destas teorias surgirem, já havia começado a acreditar que o conhecimento daquela época sobre as leis da natureza correspondia a todas as leis da natureza.

A mecânica quântica e a relatividade tiveram impactos distintos em física, gerando novas linhas de pesquisa. Em meados do século *XX*, houve o surgimento de muitos trabalhos envolvendo fenômenos peculiares que a mecânica quântica prevê que foram confirmados experimentalmente. Um desses fenômenos foi apresentado pela primeira vez na ref. [1], envolvendo o famoso *paradoxo EPR*. O paradoxo EPR surge ao tentar interpretar algumas correlações que sistemas quânticos podem exibir como realistas e locais, o que sugere que realismo e localidade são propriedades incompatíveis em muito casos.

Em 1964, o trabalho de John Bell (ref. [2]) traz uma formulação matemática por trás da hipótese de localidade, assumindo realismo nesse contexto. Nesse trabalho, Bell mostra que a mecânica quântica prevê a existência de correlações que não satisfazem certas condições, as chamadas *desigualdades de Bell*, e que isso implica que tais correlações sejam *não-locais*.

Atualmente, a existência de correlações não-locais não causa estranheza para especialistas em áreas que as envolvem, como fundamentos de mecânica quântica, informação quântica, óptica quântica, entre outras. Entretanto, não existem, até o presente momento, princípios físicos capazes de justificar a não-localidade que a mecânica quântica prevê. Por esse motivo, nos últimos anos, alguns princípios foram propostos a fim de explicar a não-localidade quântica.

Um dos objetivos deste trabalho é apresentar um destes princípios, o chamado princípio de *Causalidade da Informação* e os resultados obtidos até então com relação a esse princípio.

Outro objetivo deste trabalho é de realizar uma união entre Teoria da Informação e Teoria de Inferência Causal, com o objetivo de aplicar conceitos envolvidos nessas teorias no estudo do princípio de Causalidade da Informação, que está inserido no contexto de Não-localidade.

O capítulo 1 deste trabalho consiste de uma introdução à Teoria da Informação. Serão introduzidos conceitos como *eventos*, *espaços amostrais* que estão envolvidos em *probabilidades*. O capítulo segue com definições de *quantificadores de informação* e leis de teoria da informação envolvendo tais quantificadores. Essa última parte ilustra um pouco da natureza da Teoria da Informação.

O capítulo 2 consiste de uma introdução à Teoria de Inferência Causal, ou Teoria de Causalidade. Será estudada a diferença entre *correlações* e *causalidade*, além de serem introduzidos no capítulo, algumas ferramentas que permitem a inferência de relação causal entre dois objetos.

O capítulo 3 consiste de uma revisão de Não-localidade tentando, sempre que possível, inserir conceitos sobre causalidade na perspectiva desse estudo.

Por fim, o capítulo 4 consiste do princípio de Causalidade da Informação através dos recentes trabalhos envolvendo esse princípio. Este capítulo consistirá de uma aplicação das ferramentas e conceitos encontrados nos capítulos anteriores.

1 Teoria da Informação

Este capítulo é composto de uma introdução à Teoria da Informação. O início deste capítulo foi baseado no capítulo 1 da ref. [3] e no capítulo 2 da ref. [4]. Para maiores detalhes, consulte estas referências, ou a ref. [5]. A partir da seção 1.4, que refere-se a desigualdades de informação, utilizou-se como base o capítulo 13 e uma pequena parte do capítulo 14 da ref. [4]. Ao leitor interessado em estudar teoria quântica da informação, é indicada a ref. [6].

A teoria de probabilidade, que constitui a base de teoria da informação, é baseada em teoria de conjuntos. Iniciaremos este capítulo com uma pequena introdução à teoria de probabilidade. Conforme a necessidade, conceitos de teoria de conjuntos serão introduzidos.

1.1 Teoria de Probabilidade

1.1.1 Espaços Amostrais e Eventos

Definição 1.1.1 (Espaço Amostral)**.** Para um determinado experimento, o conjunto de todos os seus possíveis resultados é denominado *espaço amostral*.

Exemplo 1.1.1 (Lançamento de uma moeda)**.** Se o experimento consiste na observação do lançamento aleatório[1] de uma moeda, o espaço amostral contém dois resultados possíveis: Cara ou Coroa. Portanto, o espaço amostral Ω relativo a este experimento é

$$\Omega = \{\text{Cara}, \text{Coroa}\}. \tag{1.1}$$

Exemplo 1.1.2. Considerando um experimento cuja observação seja o tempo de reação a um determinado estímulo em uma pessoa, o espaço amostral pode ser, por exemplo, o conjunto de todos os números positivos (em segundos), ou seja,

$$\Omega = (0, \infty). \tag{1.2}$$

Espaços amostrais podem ser classificados em duas categorias, de acordo com o número de elementos que eles contiverem. Um espaço amostral é dito *enumerável* se os elementos deste podem ter uma correspondência um a um com um subconjunto de números inteiros. Por outro lado, será chamado de espaço amostral *não-enumerável* caso tal correspondência não seja possível. Através desta distinção, vemos que o espaço

[1] Aleatoriedade aqui significa desconhecimento por parte do observador sobre os mecanismos que determinam o resultado do lançamento da moeda. No decorrer deste trabalho, esta ideia será desconstruída.

amostral referido no exemplo 1.1.1 é enumerável, enquanto o espaço amostral do exemplo 1.1.2 é não-enumerável.

Definição 1.1.2 (Evento)**.** Um *evento* é qualquer coleção de possíveis resultados de um experimento, ou seja, algum subconjunto do espaço amostral Ω.

Se A é um evento de Ω, dizemos que o evento A ocorre se o resultado de um experimento pertence ao conjunto[2] A.

É possível impor relações de ordem entre os conjuntos de acordo com as definições a seguir.

Definição 1.1.3 (Ordenamento)**.** Dados dois eventos A e B,

$$A \subseteq B \iff x \in A \Rightarrow x \in B. \tag{1.3}$$

Ou seja, A está contido em B se, e somente se, todos os elementos de A também pertencem a B.

Definição 1.1.4 (Igualdade)**.** Dados dois eventos A e B,

$$A = B \iff A \subseteq B \text{ e } B \subseteq A. \tag{1.4}$$

Ou seja, A e B são iguais se e somente se A está contido em B e B está contido em A.

Observação. Para representar a condição de que A está estritamente contido em B, ou seja, não há a possibilidade de que $A = B$, será utilizado o símbolo "$\subset$".

Agora que sabemos como ordenar conjuntos, podemos definir um conjunto especial que será útil posteriormente. Este conjunto é o *conjunto das partes* de Ω, e será referido pelo seu nome em inglês, *powerset*.

Definição 1.1.5 (Powerset)**.** O *powerset* de um espaço amostral Ω enumerável, denotado por $\mathcal{P}(\Omega)$, é o conjunto cujos elementos são todos os eventos S_i de Ω, ou seja, cujos elementos são todos os $S_i \subseteq \Omega$.

Para um espaço amostral Ω com n elementos, seu powerset $\mathcal{P}(\Omega)$ contém 2^n elementos.

$$\mathcal{P}(\Omega) = \{S_1, \ldots, S_{2^n}\}. \tag{1.5}$$

[2] O conceito de *evento*, no contexto de teoria de probabilidade, é equivalente ao conceito de *conjunto* em teoria de conjuntos. Por este motivo, os termos *evento* e *conjunto* serão utilizados de forma equivalente, com a escolha do termo sendo feita dependendo do contexto.

Exemplo 1.1.3. Seja um espaço amostral $\Omega = \{s_1, s_2\}$, todos os seus eventos são $S_1 = \emptyset$, $S_2 = \{s_1\}$, $S_3 = \{s_2\}$ e $S_4 = \{s_1, s_2\}$. Assim, seu powerset será

$$\mathcal{P}(\Omega) = \{S_1, S_2, S_3, S_4\} = \{\emptyset, \{s_1\}, \{s_2\}, \{s_1, s_2\}\}. \tag{1.6}$$

Para quaisquer dois eventos A e B, temos as seguintes operações elementares:

Definição 1.1.6 (União)**.** A união dos eventos A e B, denotada por $A \cup B$, é o conjunto de elementos que pertencem a A, a B ou a ambos:

$$A \cup B = \{x : x \in A \text{ ou } x \in B\}. \tag{1.7}$$

Definição 1.1.7 (Interseção)**.** A interseção dos eventos A e B, denotada por $A \cap B$, é o conjunto de elementos que pertencem tanto a A quanto a B:

$$A \cap B = \{x : x \in A \text{ e } x \in B\}. \tag{1.8}$$

Definição 1.1.8 (Complementação)**.** O complemento de A, denotado por A^c, é o conjunto de todos os elementos do espaço amostral Ω que não pertençam a A:

$$A^c = \{x : x \in \Omega \text{ e } x \notin A\}. \tag{1.9}$$

Definição 1.1.9 (Diferença)**.** A diferença entre B e A, denotada por $B \setminus A$, é o conjunto de todos os elementos de B que não pertençam a A:

$$B \setminus A = \{x : x \in B \text{ e } x \notin A\}. \tag{1.10}$$

1.1.2 Probabilidade

Definição 1.1.10 (Probabilidade)**.** Para um espaço amostral Ω, uma probabilidade p é definida como

$$\begin{aligned} p : \mathcal{P}(\Omega) &\to \mathbb{R} \\ S &\mapsto p(S), \end{aligned} \tag{1.11}$$

tal que p satisfaça os seguintes axiomas:

(i) $p(S) \geq 0, \forall\, S \in \mathcal{P}(\Omega)$;

(ii) $p(\Omega) = 1$;

(iii) Para uma sequência de eventos *disjuntos* $S_1, S_2, \ldots, S_{2^n}$,
$p\left(\bigcup_i^{2^n} S_i\right) = \sum_i^{2^n} p(S_i)$.

Existem duas possíveis interpretações equivalentes de probabilidade. A primeira interpretação é a chamada *frequencista*. Esta interpretação considera probabilidades como algo tendo significado objetivo e independente do indivíduo por trás do experimento. Assim, experimentos que são realizados repetidas vezes revelarão a distribuição de probabilidade relativa à experiência, no limite em que o número de repetições da experiência seja infinito.

Definição 1.1.11 (Frequência relativa)**.** A frequência relativa f_i da ocorrência de um evento S_i contido em um espaço amostral Ω, é dada por

$$f_i = \frac{n_i}{N}, \tag{1.13}$$

sendo n_i é o número de ocorrências do evento S_i em N repetições do experimento.

Pela interpretação frequencista, a probabilidade do evento S_i é dada por

$$p(S_i) = \lim_{N\to\infty} f_i = \lim_{N\to\infty} \frac{n_i}{N}, \tag{1.14}$$

assumindo que este limite exista para todo $S_i \subseteq \Omega$.

A segunda interpretação de probabilidade é a chamada *Bayesiana*. Tal interpretação, de maneira oposta à interpretação frequencista, considera uma distribuição de probabilidade como sendo algo subjetivo e sem realidade física. A probabilidade $p(S_i)$ de que um evento S_i ocorra, representa o *grau de crença* do experimentador sobre o acontecimento do evento S_i.

A interpretação Bayesiana traz a intuição de que a probabilidade de um determinado evento $S_i \subseteq \Omega$ pode mudar caso o experimentador adquira alguma nova informação sobre o experimento. Dessa forma, a chamada *probabilidade condicional* é, para essa interpretação, a grandeza mais fundamental em teoria de probabilidades, enquanto que para a interpretação frequencista, a probabilidade conjunta de ocorrência de eventos é a grandeza que recebe esse papel. Uma probabilidade condicional, denotada por $p(S_i|S_j)$, representa a crença que o experimentador possui sobre a ocorrência do evento S_i, dado que o evento S_j ocorreu.

Observação. Para a grande maioria dos casos deste trabalho, a interpretação bayesiana de probabilidades será adotada como padrão. A interpretação frequencista será invocada explicitamente quando necessário.

Definição 1.1.12 (Regra de Bayes)**.** A probabilidade conjunta $p(S_1 \cap S_2)$ de dois eventos S_1 e S_2 ocorrerem é

$$p(S_1 \cap S_2) = p(S_1)p(S_2|S_1), \tag{1.15}$$

sendo $p(S_1)$ a probabilidade de que S_1 ocorra e $p(S_2|S_1)$ a probabilidade de que S_2 ocorra, dado que S_1 ocorreu.

Analisando a regra de Bayes através da interpretação bayesiana, essa regra relaciona a crença sobre a ocorrência conjunta de eventos com a crença sobre a ocorrência condicional de cada evento em particular.

Uma definição mais geral da regra de Bayes é:

Definição 1.1.13. A probabilidade conjunta $p(S_1 \cap \ldots \cap S_n)$ de que os eventos $S_1, \ldots, S_n$ ocorram é dada por

$$p(S_1 \cap \ldots \cap S_n) = p(S_1)p(S_2|S_1)\ldots p(S_n|S_1 \cap \ldots \cap S_{n-1}), \tag{1.16a}$$

ou, equivalentemente

$$p(S_1 \cap \ldots \cap S_n) = \prod_{i=1}^{n} p(S_i|S_1 \cap \ldots \cap S_{i-1}). \tag{1.16b}$$

Observação. A seguinte convenção será adotada neste trabalho:

$$p(S_i|\quad) = p(S_i). \tag{1.17}$$

Observação. A regra de Bayes não preferencia ordenamento, portanto não há uma maneira única de expressá-la. Por exemplo, a expressão

$$p(S_1 \cap \ldots \cap S_n) = p(S_n)p(S_{n-1}|S_n)\ldots p(S_1|S_2 \cap \ldots \cap S_n) \tag{1.18}$$

com a ordem de condicionamento invertida com relação à ordem da equação (1.16a) , também é correta. Qualquer ordem de condicionamento é válida, entretanto, a expressão final da regra de Bayes deve ser coerente com o ordenamento utilizado.

1.2 Variáveis Aleatórias e Independência

1.2.1 Variáveis Aleatórias - Representando Espaços Amostrais e Eventos

No estudo de probabilidade, diversos conceitos da teoria, como *esperança* e *variância*, exigem cálculos algébricos envolvendo resultados de estatísticas. Assim, será útil falar sobre tais resultados utilizando uma linguagem matemática, substituindo os elementos do espaço amostral Ω por números reais associados a cada elemento. Para isso, define-se *variável aleatória*.

Definição 1.2.1 (Variável aleatória)**.** Uma variável aleatória X é definida por

$$\begin{aligned} X:\ \Omega &\to \mathbb{R} \\ s &\mapsto x. \end{aligned} \tag{1.19}$$

Observação. Para todos os propósitos deste trabalho, utilizaremos variáveis aleatórias discretas, sendo considerado em geral que a variável aleatória seja uma função que leva elementos de um espaço amostral em números inteiros.

Definição 1.2.2 (Alfabeto)**.** O alfabeto $\mathcal{X}$ de uma variável aleatória X é o conjunto de possíveis valores que a variável pode assumir.

Observação. Serão utilizados neste trabalho apenas alfabetos finitos, assim, todas as provas de teoremas e proposições serão feitas levando em conta esta consideração.

Exemplo 1.2.1. Podemos definir uma variável aleatória X associada ao exemplo 1.1.1 com valores $x_1 = 0$ e $x_2 = 1$ associados a cada elemento do espaço amostral Ω:

$$\begin{aligned} X : \text{Cara} &\mapsto 0, \\ \text{Coroa} &\mapsto 1, \end{aligned} \tag{1.20}$$

assim, a variável aleatória X será representada pelo seu alfabeto como

$$\mathcal{X} = \{x_1 = 0, x_2 = 1\}. \tag{1.21}$$

Variáveis aleatórias podem representar também eventos. Além disso, muitas vezes será útil representar uma sequência de variáveis aleatórias por uma única variável aleatória. Os exemplos a seguir ilustram esses casos.

Exemplo 1.2.2. Considerando o espaço amostral do exemplo 1.1.3, é possível representar cada elemento do espaço de $\mathcal{P}(\Omega)$ por uma variável aleatória X com alfabeto $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$ tal que

$$\begin{aligned} S_1 &\mapsto x_1, \\ S_2 &\mapsto x_2, \\ S_3 &\mapsto x_3, \\ S_4 &\mapsto x_4. \end{aligned} \tag{1.22}$$

Exemplo 1.2.3. Considere uma experiência que consiste no lançamento de duas moedas. Representando os espaços amostrais correspondentes a cada moeda pelas variáveis aleatórias X_1 e X_2 de alfabetos $\mathcal{X}_1 = \{0, 1\}$ e $\mathcal{X}_2 = \{0, 1\}$ de forma equivalente ao ex. 1.2.1, podemos definir uma variável aleatória Y correspondente a $\Omega_1 \times \Omega_2$, tal que

$$\begin{aligned} Y : x_1 = 0, x_2 = 0 &\mapsto y_0, \\ x_1 = 0, x_2 = 1 &\mapsto y_1, \\ x_1 = 1, x_2 = 0 &\mapsto y_2, \\ x_1 = 1, x_2 = 1 &\mapsto y_3. \end{aligned} \tag{1.23}$$

É importante destacar que neste texto, sempre denotaremos as variáveis por letras maiúsculas (ex: X, Y) e, da mesma forma, os possíveis valores que tais variáveis possam ter serão denotados por letras minúsculas (ex: x, y).

Observação. Na transição entre um contexto baseado em conjuntos e outro baseado em variáveis aleatórias, a probabilidade conjunta de dois eventos S_1 e S_2, representados por x_1 e x_2 em uma variável aleatória, será escrita como

$$p(S_1 \cap S_2) = p(x_1, x_2).$$

Definição 1.2.3 (Suporte)**.** O suporte de uma função p de uma variável aleatória X, denotado por $\mathcal{S}_{p(X)}$, é definido como

$$\mathcal{S}_{p(X)} = \{x \in \mathcal{X} | p(x) \neq 0\}. \tag{1.24}$$

Observação. Se p for uma função de mais de uma variável, o argumento de p será removido do símbolo que representa o seu suporte, sendo então representado simplesmente como $\mathcal{S}_p$.

Supondo agora que existam duas variáveis aleatórias X e Y, cujos alfabetos são respectivamente $\mathcal{X}$ e $\mathcal{Y}$. A partir de agora, uma distribuição de probabilidade sobre variáveis aleatórias será denotada explicitamente por P. Serão discutidas relações entre as variáveis em uma distribuição de probabilidade P sobre essas variáveis.

1.2.2 Independência

Definição 1.2.4 (Independência)**.** Duas variáveis aleatórias X e Y são independentes se

$$p(x|y) = p(x) \tag{1.25a}$$

e

$$p(y|x) = p(y) \tag{1.25b}$$

para todo $x \in \mathcal{X}$ e $y \in \mathcal{Y}$.

Uma consequência importante dessa definição é que a regra de Bayes para estas variáveis será simplesmente

$$p(x, y) = p(x)p(y) \tag{1.26}$$

para todo par $(x, y) \in \mathcal{X} \times \mathcal{Y}$.
Denotamos a independência entre as variáveis X e Y por $(X \perp\!\!\!\perp Y)_P$.

Definição 1.2.5 (Independência conjunta)**.** Para $n \geq 3$, as variáveis $X_1, X_2, \ldots, X_n$ são conjuntamente independentes se

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i) \tag{1.27}$$

para todo $x_1, x_2, \ldots, x_n$.

Observação. Para que todas as variáveis sejam conjuntamente independentes, é de extrema importância que a eq. (1.27) seja válida para todos os valores das variáveis $X_1, X_2, \ldots, X_n$. Se uma probabilidade conjunta com relação a valores fixos $x_1, x_2, \ldots, x_n$ for o produto das probabilidades de cada valor $x_1, x_2, \ldots, x_n$, isso não garante que as variáveis sejam independentes. É necessário que, para todas as combinações de variáveis, as probabilidades conjuntas sobre tais combinações de variáveis sejam o produto de cada probabilidade em particular. Isso pode ser percebido ao impor que na regra de Bayes, em qualquer ordenamento escolhido, todas as probabilidades condicionais $p(x_i|x_1, \ldots, x_{i-1})$ sejam indiferentes a todos os condicionamentos, ou seja,

$$p(x_i|x_1, \ldots, x_{i-1}) = p(x_i) \; \forall \; x_i \in \mathcal{X}_i, \; i = 1, \ldots, n. \tag{1.28}$$

Definição 1.2.6 (Independência condicional)**.** Para variáveis aleatórias X, Y e Z, a variável X é independente de Y condicionada a Z, relação denotada por $(X \perp\!\!\!\perp Y|Z)_P$, se

$$p(x|y, z) = p(x|z) \tag{1.29}$$

para todo $x \in \mathcal{X}$, $y \in \mathcal{Y}$ e $z \in \mathcal{Z}$.

Observação. A relação $(X \perp\!\!\!\perp Y|Z)_P$ é equivalente à $(Y \perp\!\!\!\perp X|Z)_P$. Essa propriedade pode ser revelada pela simetria da regra de Bayes.

A independência entre X e Y condicionada a Z é intepretada como o fato de que, a crença sobre X após o conhecimento de Z não é alterada com o conhecimento de Y. Em outras palavras, Y não traz informação adicional sobre X quando Z é conhecido.

Definição 1.2.7 (Independência dois a dois)**.** Uma sequência de variáveis aleatórias X_1, X_2, $X_3, \ldots, X_n$, $n \geq 3$ são independentes dois a dois se $(X_i \perp\!\!\!\perp X_j)_P$ para $1 \leq i < j \leq n$.

Definição 1.2.8 (Cadeia de Markov)**.** Uma sequência de variáveis aleatórias X_1, X_2, $X_3, \ldots, X_n$, $n \geq 3$ forma uma cadeia de Markov se

$$p(x_1, x_2, x_3, \ldots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_2) \ldots p(x_n|x_{n-1}) \tag{1.30}$$

para todo $x_1, x_2, x_3, \ldots x_n$.

Uma cadeia de Markov é representada por $X_1 \to X_2 \to \ldots \to X_n$. De forma equivalente à def. 1.2.8, em uma cadeia de Markov,

$$p(x_i|x_1, \ldots, x_{i-1}) = p(x_i|x_{i-1}) \; \forall \; i = 3, \ldots, n. \tag{1.31}$$

Uma cadeia de Markov carrega sequências de independências condicionais do tipo $(X_k \perp\!\!\!\perp X_i|X_j)_P$ para cada trecho representado por $X_i \to X_j \to X_k$.

As relações de independência condicional em distribuições de probabilidade sobre variáveis aleatórias carregam características fundamentais sobre os tipos de informação que certas variáveis possuem sobre outras. Essas relações serão bastante exploradas ao longo desse trabalho.

1.3 Quantificadores de Informação de Shannon

1.3.1 Entropias de Shannon

Definição 1.3.1 (Entropia de Shannon)**.** A entropia de Shannon $H(X)$ de uma variável aleatória X é definida como

$$H(X) = -\sum_i p(x_i) \log p(x_i), \tag{1.32}$$

onde adotamos a convenção de que a soma é realizada sobre o suporte $\mathcal{S}_{p(X)}$, devido ao fato de que a função $p(x_i) \log p(x_i)$ em (1.32) não é definida para $p(x_i) = 0$.

A entropia de Shannon pode ser interpretada como o *grau de incerteza* do experimentador com relação à variável em questão. Ela quantifica a informação contida na distribuição de probabilidade sobre a variável.

A base do logaritmo pode ser escolhida para ser qualquer número real maior que 1. Se a base do logaritmo da entropia $H(X)$ for a *cardinalidade* $|\mathcal{X}|$ do alfabeto de X, ou seja, o número de elementos existentes no alfabeto $\mathcal{X}$, a imagem $\mathrm{Im}(H)$ da entropia de Shannon será

$$\mathrm{Im}(H) = [0, 1] \in \mathbb{R} \tag{1.33}$$

Exemplo 1.3.1 (Variável binária (*bit*))**.** Considerando duas variáveis binárias X e Y que possam assumir os valores 0 e 1, vamos calcular suas entropias de Shannon, assumindo uma distribuição de probabilidade para cada variável.
Sejam $p(x)$ e $p(y)$ as distribuições de probabilidade sobre as variáveis x e y respectivamente:

$$p(x) = \begin{cases} 0, & x = 0 \\ 1, & x = 1 \end{cases} \quad \text{e} \quad p(y) = \begin{cases} 1/2, & y = 0 \\ 1/2, & y = 1 \end{cases}. \tag{1.34}$$

Utilizando a base 2 no logaritmo da entropia, a entropia de Shannon $H(X)$ será

$$\begin{aligned} H(X) &= -\sum_i p(x_i) \log_2 p(x_i) \\ &= -p(1) \log_2 p(1) \\ &= -1 \cdot \log_2 1 = 0, \end{aligned} \tag{1.35}$$

lembrando que, como a soma é tomada apenas sobre o suporte $\mathcal{S}_{p(X)}$, apenas o valor $x = 1$ da variável foi considerado, pois é o único elemento pertencente ao suporte.

Também utilizando a base 2 no logaritmo da entropia, $H(Y)$ será

$$\begin{aligned} H(Y) &= -\sum_i p(y_i)\log_2 p(y_i) \\ &= -p(0)\log_2 p(0) - p(1)\log_2 p(1) \\ &= -\frac{1}{2}\cdot\log_2\frac{1}{2} - \frac{1}{2}\cdot\log_2\frac{1}{2} \\ &= -\frac{1}{2}\cdot(-1) - \frac{1}{2}\cdot(-1) = 1. \end{aligned} \tag{1.36}$$

Como foi mencionado no início desta seção, a entropia expressa o *grau de incerteza* sobre a variável. Como a variável X possui uma distribuição de probabilidade *determinística*, ou seja, um determinado valor da variável possui probabilidade 1 de ocorrer, enquanto o outro possui probabilidade 0, não há *incerteza* sobre a variável.

Se esta variável fosse a mesma do exemplo 1.2.1, isso representaria um experimento em que uma moeda, sempre que fosse lançada, resultaria em "Coroa". Sendo a probabilidade sobre esta variável a crença do experimentador sobre um determinado evento, isto significa que o lançador sabe que o resultado do lançamento da moeda sempre será "Coroa". Portanto, o experimentador jamais terá incerteza com relação ao resultado do experimento, e isto é representado por $H(X) = 0$.

A variável Y possui uma distribuição de probabilidade uniforme, ou seja, qualquer valor desta variável é *igualmente provável*. Novamente, se esta variável fosse a mesma do exemplo 1.2.1, isso representaria um experimento em que uma moeda, ao ser lançada, pode resultar tanto em "Cara" quanto em "Coroa", sem nenhum viés sobre qual das duas opções ocorrerá em cada jogada. Esse é o caso oposto ao caso anterior, em que o experimentador tem *máxima ignorância* sobre o resultado do lançamento da moeda. Isto está representado pelo fato de que $H(Y) = 1$. Como 1 é o valor máximo possível da entropia, devido à escolha da base do logaritmo como sendo a cardinalidade da variável, este valor da entropia representa o máximo desconhecimento do experimentador sobre um experimento, ou seja, o experimentador terá *máxima incerteza* possível com o resultado do experimento.

Dado que agora temos uma melhor intuição sobre o significado da entropia de Shannon, podemos sempre pensar nela como um quantificador da informação de alguma variável. A liberdade sobre a base do logaritmo da entropia pode ser explorada, porém isto deve ser feito com cuidado. Podemos interpretar a liberdade da base do logaritmo como uma liberdade de escala. Para tornar esta ideia clara, consideremos um exemplo com duas variáveis aleatórias:

Exemplo 1.3.2. Sejam duas variáveis aleatórias X e Y, sendo X uma variável com dois

valores possíveis e Y uma variável com quatro valores possíveis. Suponha que cada uma delas possua uma distribuição de probabilidade uniforme sobre todos os seus valores. Utilizando a cardinalidade do alfabeto de cada variável como base do logaritmo de suas entropias, temos:

$$H(X) = -\sum_i p(x_i) \log_{|\mathcal{X}|} p(x_i) = -\cancel{2}\left[\frac{1}{\cancel{2}} \log_2 \left(\frac{1}{2}\right)\right] = -[-1] = 1 \tag{1.37a}$$

e

$$H(Y) = -\sum_i p(y_i) \log_{|\mathcal{Y}|} p(y_i) = -\cancel{4}\left[\frac{1}{\cancel{4}} \log_4 \left(\frac{1}{4}\right)\right] = -[-1] = 1. \tag{1.37b}$$

Utilizando bases de logaritmos diferentes para cada entropia, encontramos o mesmo valor de entropia para variáveis cujos alfabetos possuem cardinalidades diferentes. Isso significa que, em termos de informação, não é simples comparar uma variável com a outra. Para que a comparação seja possível de ser feita diretamente, é necessário utilizar a mesma base no logaritmo da entropia para os dois casos. Utilizando como base do logaritmo a cardinalidade do alfabeto da variável X, temos

$$H(X) = -\sum_i p(x_i) \log_{|\mathcal{X}|} p(x_i) = -\cancel{2}\left[\frac{1}{\cancel{2}} \log_2 \left(\frac{1}{2}\right)\right] = -[-1] = 1 \tag{1.38a}$$

e

$$H(Y) = -\sum_i p(y_i) \log_{|\mathcal{X}|} p(y_i) = -\cancel{4}\left[\frac{1}{\cancel{4}} \log_2 \left(\frac{1}{4}\right)\right] = -[-2] = 2. \tag{1.38b}$$

Portanto, adotando a mesma base para as entropias desejadas, vemos que a variável Y carrega mais informação que a variável X. Além disso, a utilização de duas variáveis aleatórias, X_1 e X_2, binárias, independentes entre si e com distribuições de probabilidade uniformes, é equivalente a utilizar a variável aleatória Y desse exemplo.

Observação. A utilização da cardinalidade como base do logaritmo da entropia é útil para analisar grau de informação absoluto de uma variável, dado que nessa condição, o valor máximo da entropia será sempre 1. Assim, uma entropia igual a 1 significa que a variável carrega o máximo de informação que a sua capacidade permite. Além disso, como não há um valor máximo para cardinalidade, é natural usar seu menor valor interessante como uma escala, para que seja possível realizar comparações entre variáveis com alfabetos diferentes. A base 2 para o logaritmo será adotada por convenção, quando essa base não for utilizada, isso será explicitado adequadamente.

Definição 1.3.2 (Valor esperado)**.** Seja $X : \Omega \to \mathbb{R}$ uma variável aleatória, o *valor esperado*, ou *esperança*, de X é

$$E(X) = \sum_i x_i \; p(x_i), \tag{1.39}$$

onde, novamente, a soma é tomada sobre o suporte $\mathcal{S}_{p(X)}$.

A partir de agora, sempre que variáveis aleatórias forem mencionadas, será implicitamente assumida uma distribuição de probabilidade P (conjunta ou não) para tais variáveis.

Por simplicidade de notação, os índices das somas serão omitidos a partir de agora, considerando então as somas sobre todos os valores dos suportes das distribuições.

Definição 1.3.3 (Entropia conjunta de Shannon)**.** A entropia conjunta de Shannon de duas variáveis aleatórias X e Y é definida como

$$H(X,Y) = -\sum_{x,y} p(x,y)\log p(x,y) = -E(\log p(X,Y)). \tag{1.40}$$

A entropia conjunta de duas variáveis X e Y quantifica a informação total que essas variáveis carregam.

Definição 1.3.4 (Entropia condicional)**.** Para duas variáveis aleatórias X e Y, a entropia de Shannon de Y condicionada a X é definida como

$$H(Y|X) = -\sum_{x,y} p(x,y)\log p(y|x) = -E(\log p(Y|X)). \tag{1.41}$$

A entropia condicional $H(Y|X)$ quantifica a informação contida exclusivamente na variável Y, descartando a informação contida em Y que também está contida em X.

A entropia condicional da equação (1.41) pode ser escrita como

$$H(Y|X) = \sum_x p(x)\left[-\sum_y p(y|x)\log p(y|x)\right]. \tag{1.42}$$

A soma sobre y é a entropia de Y condicionada a um valor fixo $x \in \mathcal{S}_{p(X)}$. Portanto, $H(Y|X)$ também pode ser expressada como

$$H(Y|X) = \sum_x p(x)H(Y|x), \tag{1.43}$$

onde

$$H(Y|x) = -\sum_y p(y|x)\log p(y|x). \tag{1.44}$$

Assim, o lado direito das eqs. (1.32) e (1.44) possuem a mesma forma.

Proposição 1.3.1.

$$H(X,Y) = H(X) + H(Y|X) \tag{1.45a}$$

e

$$H(X,Y) = H(Y) + H(X|Y) \tag{1.45b}$$

Demonstração. Utilizando a regra de Bayes expressa na eq. (1.15) para $p(x,y)$, temos que

$$\begin{aligned} H(X,Y) &= -\sum_{x,y} p(x,y)\log p(x,y) && \text{(1.46a)}\\ &= -\sum_{x,y} p(x,y)\log[(p(x)p(y|x)] && \text{(1.46b)}\\ &= -\sum_{x,y} p(x,y)\left[\log p(x)+\log p(y|x)\right] && \text{(1.46c)}\\ &= -\left[\sum_{x,y} p(x,y)\log p(x)+\sum_{x,y} p(x,y)\log p(y|x)\right] && \text{(1.46d)}\\ &= -\sum_{x} p(x)\log p(x)-\sum_{x,y} p(x,y)\log(y|x) && \text{(1.46e)}\\ &= H(X)+H(Y|X), && \text{(1.46f)} \end{aligned}$$

provando assim a expressão 1.45a. A expressão (1.45b) pode ser provada utilizando a simetria da regra de Bayes. □

1.3.2 Informação mútua

Definição 1.3.5 (Informação mútua)**.** Para duas variáveis aleatórias X e Y, a informação mútua entre estas variáveis, denotada por $I(X;Y)$, é definida como

$$I(X;Y)=\sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}=E\left(\log\frac{p(x,y)}{p(x)p(y)}\right). \tag{1.47}$$

A informação mútua $I(X;Y)$ quantifica a informação comum entre as variáveis X e Y, eliminando a informação que somente X carrega e a informação que somente Y carrega.

Observação. A informação mútua $I(X;Y)$ é simétrica em X e Y.

Proposição 1.3.2. *A informação mútua entre uma variável X e ela mesma é simplesmente a entropia de X, ou seja,*

$$I(X;X)=H(X). \tag{1.48}$$

Demonstração.

$$\begin{aligned} I(X;X) &= \sum_{x} p(x)\log\frac{p(x)}{p(x)^2} && \text{(1.49a)}\\ &= \sum_{x} p(x)\log\frac{1}{p(x)} && \text{(1.49b)}\\ &= -\sum_{x} p(x)\log p(x) && \text{(1.49c)}\\ &= H(X) && \text{(1.49d)} \end{aligned}$$

□

Proposição 1.3.3. *A informação mútua entre duas variáveis X e Y pode ser escrita em termos das entropias destas variáveis:*

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{1.50}$$

Demonstração.

$$\begin{aligned} I(X;Y) =& \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} && \text{(1.51a)} \\ =& \sum_{x,y} p(x,y) \left[\log p(x,y) - \log p(x) - \log p(y)\right] && \text{(1.51b)} \\ =& -\sum_{x,y} p(x,y) \log p(x) - \sum_{x,y} p(x,y) \log p(y) + \sum_{x,y} p(x,y) \log p(x,y) && \text{(1.51c)} \\ =& -\sum_{x} p(x) \log p(x) - \sum_{y} p(y) \log p(y) - \left[-\sum_{x,y} p(x,y) \log p(x,y)\right] && \text{(1.51d)} \\ =& H(X) + H(Y) - H(X,Y) && \text{(1.51e)} \end{aligned}$$

□

Utilizando as equações (1.3.1), é possível também escrever a informação mútua $I(X;Y)$ em termos de entropias condicionais:

$$I(X;Y) = H(X) - H(Y|X) \tag{1.52a}$$

e

$$I(X;Y) = H(Y) - H(X|Y). \tag{1.52b}$$

Uma maneira comumente utilizada na matemática para representar conjuntos e analisar suas propriedades são Diagramas de Venn. Tais diagramas são constituídos de curvas fechadas simples desenhadas em um plano, simbolizando os conjuntos e suas propriedades. Uma maneira útil de visualizar as relações entre as diferentes quantificadores de informação de Shannon é através de um diagrama de Venn. Na figura 1 está um diagrama de Venn representando todos os quantificadores de informação de Shannon já definidos para as variáveis X e Y.

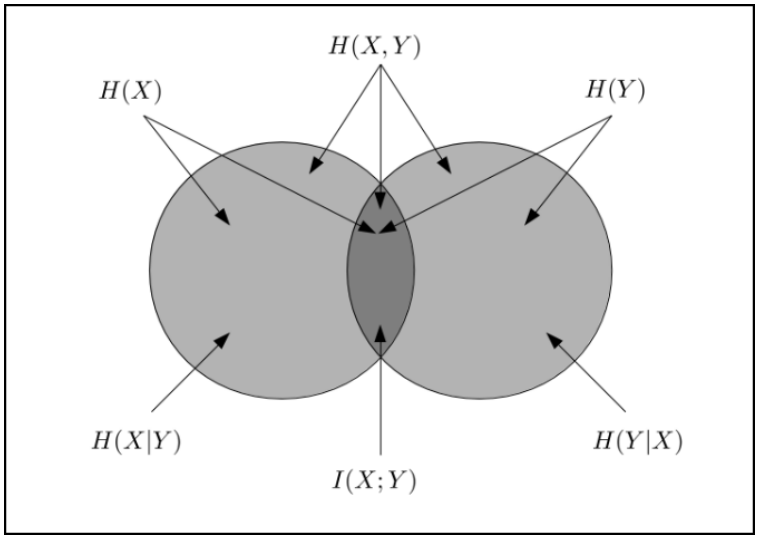


Figura 1 – Diagrama de Venn - Relações entre diferentes quantificadores de informação de Shannon para as variáveis X e Y.

Definição 1.3.6 (Informação mútua condicional)**.** Para variáveis aleatórias X, Y e Z, a informação mútua entre X e Y condicionada a Z é definida por

$$I(X;Y|Z) = \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} = E\left(\log \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}\right). \tag{1.53}$$

Proposição 1.3.4. *A informação mútua entre duas variáveis X e Y condicionada à variável Z pode ser escrita em termos das entropias das variáveis X e Y condicionadas à Z:*

$$I(X;Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z) \tag{1.54}$$

Demonstração.

$$\begin{aligned}
I(X;Y|Z) &= \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} && \text{(1.55a)}\\
&= \sum_{z} p(z) \left[\sum_{x,y} p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}\right] && \text{(1.55b)}\\
&= \sum_{z} p(z) \left[\sum_{x,y} p(x,y|z)\Big(\log p(x,y|z) - \log p(x|z) - \log p(y|z)\Big)\right] && \text{(1.55c)}\\
&= \sum_{z} p(z)\Big[-H(X,Y|z) + H(X|z) + H(Y|z)\Big] && \text{(1.55d)}\\
&= H(X|Z) + H(Y|Z) - H(X,Y|Z). && \text{(1.55e)}
\end{aligned}$$

□

Proposição 1.3.5. *Assim como na prop. 1.3.2, a informação mútua condicional entre uma variável X e ela mesma, condicionada a uma outra variável Z, é igual a entropia de X condicionada à Z.*

$$I(X;X|Z) = H(X|Z) \tag{1.56}$$

Demonstração. A demonstração dessa proposição é feita da mesma forma que a demonstração da prop. 1.3.2. □

Os quantificadores de informação de Shannon podem ser encarados como casos especiais de Informação Mútua Condicional. Por exemplo, considerando as as variáveis aleatórias X, Y, Z e uma variável Φ com uma distribuição de probabilidade determinística. Se $X = Y$ e $Z = \Phi$, a informação mútua $I(X;Y|Z) = H(X)$. Já considerando apenas que $X = Y$, $I(X;Y|Z) = H(X|Z)$. Por fim, se $Z = \Phi$, $I(X;Y|Z) = I(X;Y)$. Assim, vemos que é possível representar os quantificadores de informação de Shannon de forma geral utilizando apenas Informação Mútua Condicional.

1.3.3 Regras da Cadeia para Quantificadores de Informação de Shannon

A def. 1.1.13 fala sobre a regra de Bayes, também chamada de regra da cadeia, para uma distribuição de probabilidade sobre um conjunto de variáveis aleatórias. Mostraremos aqui que os quantificadores de informação de Shannon também possuem suas respectivas versões de regras da cadeia.

Proposição 1.3.6 (Regra da cadeia para a entropia de Shannon)**.** *Considerando n variáveis aleatórias, $X_1, \ldots, X_n$, a entropia conjunta de Shannon de todas as variáveis satisfaz*

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_{i-1}). \tag{1.57}$$

Demonstração. Como o caso em que $n = 2$ foi provado na prop. 1.3.2, podemos utilizar esse resultado com recorrência:

$$\begin{aligned} H(X_1, \ldots, X_n) =& H(X_1) + H(X_2, \ldots, X_n|X_1) && \text{(1.58a)} \\ =& H(X_1) + H(X_2|X_1) + H(X_3, \ldots, X_n|X_1, X_2) && \text{(1.58b)} \\ =& \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_{i-1}). && \text{(1.58c)} \end{aligned}$$

□

Proposição 1.3.7 (Regra da cadeia para a informação mútua condicional)**.** *Considerando as variáveis aleatórias, $X_1, \ldots, X_n, Y$ e Z a informação mútua entre as variáveis $X_1, \ldots, X_n$ e a variável Y, condicionada à Z satisfaz*

$$I(X_1, \ldots, X_n; Y|Z) = \sum_{i=1}^{n} I(X_i; Y|Z, X_1, \ldots, X_{i-1}) \tag{1.59}$$

Demonstração.

$$
\begin{aligned}
I(X_1,\ldots,X_n;Y|Z) =& H(X_1,\ldots,X_n|Z) + H(Y|Z) - H(X_1,\ldots,X_n,Y|Z) && \text{(1.60a)}\\
=& H(Z,X_1,\ldots,X_n) - H(Z) + H(Y|Z) && \text{(1.60b)}\\
& - [H(Z,X_1,\ldots,X_n,Y) - H(Z)] && \text{(1.60c)}\\
=& \cancel{H(Z)} + \sum_{i=1}^{n} H(X_i|Z,X_1,\ldots,X_{i-1}) - \cancel{H(Z)} + H(Y|Z) && \text{(1.60d)}\\
& - \left[\cancel{H(Z)} + \sum_{i=1}^{n} H(X_i,Y|Z,X_1,\ldots,X_{i-1}) - \cancel{H(Z)}\right] && \text{(1.60e)}\\
=& \sum_{i=1}^{n} I(X_i;Y|Z,X_1,\ldots,X_{i-1}) && \text{(1.60f)}
\end{aligned}
$$

□

Como os outros quantificadores de informação de Shannon são casos particulares da informação mútua condicional, as regras da cadeia são válidas também para a entropia condicional e informação mútua.

Observação. As regras da cadeia para quantificadores de informação de Shannon possuem uma forma muito semelhante à regra de Bayes da def. 1.1.13. Mesmo assim, existe uma diferença fundamental entre os dois tipos. Enquanto a regra de Bayes da def. 1.1.13 envolve um *produtório* sobre os eventos S_i, ou variáveis S_i, as regras da cadeia para quantificadores de informação de Shannon envolvem um *somatório* sobre as variáveis X_i. Esta propriedade será explorada posteriormente.

Suponha agora que existam duas distribuições de probabilidade, p e q, atuando sobre um mesmo alfabeto $\mathcal{X}$. Em certas situações, pode ser necessário saber o quanto p é diferente de q. Por este motivo, definimos *divergência informacional*.

Definição 1.3.7 (Divergência Informacional)**.** A divergência informacional, ou entropia relativa, entre duas distribuições de probabilidade p e q atuando sobre um mesmo alfabeto $\mathcal{X}$ é definida como

$$
D(p(X) \parallel q(X)) = \sum_{\mathcal{S}_{p(X)}} p(x) \log \frac{p(x)}{q(x)} = E_p \log \frac{p(X)}{q(X)}, \tag{1.61}
$$

sendo E_p a esperança com respeito à distribuição p.

É importante destacar que a Divergência Informacional é assimétrica em p e q. Vejamos um exemplo que ilustra essa propriedade.

Exemplo 1.3.3. Consideremos, novamente, um lançamento de uma moeda. Digamos que o lançamento da moeda pode satisfazer uma de duas possíveis distribuições de

probabilidade e, com isso, deseja-se determinar qual das distribuições de probabilidade é a correta. Sejam p e q tais distribuições:

$$p(x) = \begin{cases} 0, & x = 0 \\ 1, & x = 1 \end{cases} \quad e \quad q(x) = \begin{cases} 1/2, & x = 0 \\ 1/2, & x = 1. \end{cases} \tag{1.62}$$

Calculemos as Divergências Informacionais possíveis entre p e q:

$$D(p||q) = \sum_{\mathcal{S}_{p(X)}} p(x) \log \frac{p(x)}{q(x)} \tag{1.63a}$$

$$= 1 \cdot \log \frac{1}{1/2} = \log 2. \tag{1.63b}$$

$$D(q||p) = \sum_{\mathcal{S}_{p(X)}} q(x) \log \frac{q(x)}{p(x)} \tag{1.64a}$$

$$= \lim_{p(0)\to 0} \frac{1}{2} \left[\log \left(\frac{1/2}{p(0)} \right) + \log \frac{1/2}{1} \right] \tag{1.64b}$$

$$= \infty. \tag{1.64c}$$

Vejamos a interpretação desses resultados. Se o lançamento da moeda em questão for determinístico (associado à distribuição $p(x)$), ao realizar vários lançamentos da moeda, esses resultarão inúmeras vezes no mesmo valor. Essa sequência de resultados, apesar de muito improvável, também é compatível com a distribuição $q(x)$. Ao comparar a distribuição $p(x)$ com $q(x)$ dessa maneira, estamos avaliando a Divergência Informacional $D(p||q)$, cujo resultado é $\log 2$.

Já se o lançamento da moeda em questão for uniformemente aleatório (associado à distribuição $q(x)$), ao executar muitas repetições do lançamento, a moeda terá resultados que oscilam entre "Cara" e "Coroa" (associados a $x = 0$ e $x = 1$, respectivamente). Como na distribuição $p(x)$, o valor $x = 0$ não tem chance de ocorrer, a sua ocorrência na sequência de lançamentos nos leva a inferir que a distribuição correta para esse lançamento é $q(x)$. Por esse motivo, $D(q||p) = \infty$ significa que a comparação de q com p realizada dessa forma é capaz de garantir a distribuição de probabilidade correta no lançamento da moeda, enquanto que a comparação no caso $D(p||q)$ apenas sugere a probabilidade correta.

1.4 Desigualdades de Informação

Desigualdades de informação são extremamente úteis e poderosas no estudo de teoria da informação. Elas governam as impossibilidades em teoria da informação, sendo, por este motivo, muitas vezes chamadas de *leis da teoria da informação* e serão utilizadas nos principais conteúdos deste trabalho.

Definição 1.4.1 (Desigualdades de informação)**.** Uma desigualdade de informação é uma expressão da forma

$$F \geq c, \tag{1.65}$$

onde F é uma combinação linear[3] de quantificadores de informação de Shannon e c é uma constante real.

Exemplo 1.4.1. A expressão a seguir é uma desigualdade de informação

$$I(X;Y) \geq 0. \tag{1.66}$$

Essa desigualdade é válida para qualquer distribuição de probabilidade conjunta $p(x,y)$, com igualdade apenas se $(X \perp\!\!\!\perp Y)_P$.

O exemplo anterior se refere a uma desigualdade *sempre válida*, ou seja, qualquer distribuição de probabilidade conjunta sobre as variáveis envolvidas satisfaz a desigualdade. A validade desta desigualdade será provada posteriormente.

Antes de mencionar as desigualdades de informação importantes, primeiramente serão enunciados teoremas que serão necessários para demonstrar as desigualdades de informação que nos interessam.

Lema 1.4.1. *Para qualquer $a > 0 \in \mathbb{R}$,*

$$\ln a \leq a - 1. \tag{1.67}$$

Demonstração. A demonstração desse teorema será omitida, mas para realizá-la, basta utilizar ferramentas de Cálculo Diferencial.

□

Corolário 1.4.1. *Para qualquer $a > 0$,*

$$\ln a \geq 1 - \frac{1}{a}. \tag{1.68}$$

Demonstração. A demonstração é realizada através da substituição de a por $\frac{1}{a}$ na eq. (1.67). □

Teorema 1.4.2 (Desigualdade da divergência informacional)**.** *Para quaisquer duas distribuições de probabilidade p e q sobre o mesmo alfabeto $\mathcal{X}$,*

$$D(p(X) \parallel q(X)) \geq 0, \tag{1.69}$$

com igualdade apenas se $p = q$.

[3] Poderíamos considerar desigualdades compostas por combinações não lineares de quantificadores de informação de Shannon, mas essas não serão usadas nesse trabalho em nenhum momento, não sendo, portanto, necessário incluí-las na definição.

Demonstração. Se $q(x) = 0$ para algum $x \in \mathcal{S}_{p(X)}$, então $D(p \parallel q) \to +\infty$, satisfazendo o teorema neste caso. Assumindo agora que $\mathcal{S}_{p(X)} \subseteq \mathcal{S}_{q(X)}$ e utilizando o corolário 1.4.1,

$$D(p(X) \parallel q(X)) = \sum_{x \in \mathcal{S}_{p(X)}} p(x) \log \frac{p(x)}{q(x)} \tag{1.70a}$$

$$= (\log e) \sum_{x \in \mathcal{S}_{p(X)}} p(x) \ln \frac{p(x)}{q(x)} \tag{1.70b}$$

$$\geq (\log e) \sum_{x \in \mathcal{S}_{p(X)}} p(x) \left(1 - \frac{q(x)}{p(x)}\right) \tag{1.70c}$$

$$= (\log e) \left[\sum_{x \in \mathcal{S}_{p(X)}} p(x) - \sum_{x \in \mathcal{S}_{p(X)}} q(x)\right]. \tag{1.70d}$$

Como $\mathcal{S}_{q(X)} \subseteq \mathcal{S}_{p(X)}$, $\sum_{x \in \mathcal{S}_{p(X)}} q(x) \leq 1$. Sendo $\sum_{x \in \mathcal{S}_{p(X)}} p(x) = 1$, é fácil ver que

$$\log e \left[1 - \sum_{x \in \mathcal{S}_{p(X)}} q(x)\right] \geq 0. \tag{1.70e}$$

Assim sendo, $D(p(X) \parallel q(X)) = 0 \Rightarrow p(x) = q(x)\ \forall x \in \mathcal{S}_{p(X)}$. □

Teorema 1.4.3. *O condicionamento de uma variável Y em uma variável X não é capaz de aumentar a entropia de Y, isto é,*

$$H(Y|X) \leq H(Y), \tag{1.71}$$

com igualdade apenas se $(X \perp\!\!\!\perp Y)_P$.

Demonstração.

$$H(Y|X) = H(Y) - I(X;Y) \tag{1.72a}$$

$$\leq H(Y), \tag{1.72b}$$

em que a igualdade é satisfeita quando $I(X;Y) = 0$, ou seja, $(X \perp\!\!\!\perp Y)_P$. □

1.4.1 Desigualdades Básicas

As desigualdades básicas são todas aquelas que representam a não-negatividade dos quantificadores de informação de Shannon. Isso significa que, para qualquer distribuição de probabilidade conjunta sobre as variáveis envolvidas, os quantificadores de informação de Shannon são não-negativos.

Teorema 1.4.4. *Para variáveis aleatórias X,Y e Z,*

$$I(X;Y|Z) \geq 0, \tag{1.73}$$

com igualdade apenas se $(X \perp\!\!\!\perp Y|Z)_P$.

Demonstração.

$$I(X;Y|Z) = \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} \tag{1.74a}$$

$$= \sum_{z} p(z) \sum_{x,y} p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} \tag{1.74b}$$

$$= \sum_{z} p(z) D\Big(p(X,Y|z) \parallel p(X|z)p(Y|z)\Big). \tag{1.74c}$$

Como $p(X,Y|z)$ e $p(X|z)p(Y|z)$ são distribuições de probabilidade sobre o mesmo alfabeto $\mathcal{X} \times \mathcal{Y}$ condicionadas a um valor fixo z,

$$D\Big(p(X,Y|z) \parallel p(X|z)p(Y|z)\Big) \geq 0. \tag{1.74d}$$

Utilizando o teorema 1.4.2, $I(X,Y|Z) = 0$ se, e somente se, $(X \perp\!\!\!\perp Y|Z)_P$. □

Como vimos anteriormente, os quantificadores de informação de Shannon são casos particulares da informação mútua condicional, portanto, as seguintes desigualdades também são sempre válidas:

$$H(X) \geq 0 \tag{1.75a}$$

$$I(X;Y) \geq 0 \tag{1.75b}$$

$$H(X|Z) \geq 0 \tag{1.75c}$$

Assim, as desigualdades (1.73), (1.75a), (1.75b) e (1.75c) são as chamadas *desigualdades básicas*.

1.4.2 Quantificadores de Informação Nulos.

Como foi provado no teorema 1.4.4, a informação mútua condicional $I(X;Y|Z) = 0 \iff (X \perp\!\!\!\perp Y|Z)_P$.

Proposição 1.4.5. *$H(X) = 0$ se, e somente se, a distribuição de probabilidade $p(x)$ sobre a qual é calculada a entropia é determinística.*

Demonstração. Se $p(x)$ é uma distribuição de probabilidade determinística, então $\exists\, x' \in \mathcal{X}$ tal que $p(x') = 1$. Para todo outro $x \in \mathcal{X}$, $p(x) = 0$. Portanto

$$H(X) = -p(x') \log p(x') = -\log 1 = 0. \tag{1.76a}$$

Se $p(x)$ não é uma distribuição determinística, então $\exists\, x' \in \mathcal{X}$ tal que $0 < p(x') < 1$. Assim,

$$0 < -p(x') \log p(x') \leq H(X), \tag{1.76b}$$

portanto, $H(X) > 0$ para qualquer distribuição não-determinística. □

Proposição 1.4.6. *$H(Y|X) = 0$ se, e somente se, Y é uma função de X, ou seja, para cada $x \in \mathcal{X}$ existe $y \in \mathcal{Y}$ tal que $p(y|x) = 1$.*

Demonstração. Vemos na eq. (1.43) que $H(Y|X) = 0$ se, e somente se, $H(Y|x) = 0$ para todo $x \in \mathcal{S}_{p(X)}$. Sendo assim, $p(y|x)$ será uma distribuição determinística em qualquer condicionamento feito sobre X, ou seja, Y é função de X. □

Proposição 1.4.7. *$I(X;Y) = 0$ se, e somente se, $(X \perp\!\!\!\perp Y)_P$.*

Demonstração. De acordo com o teorema 1.4.4, $I(X;Y|Z) = 0 \iff (X \perp\!\!\!\perp Y|Z)_P$. Como $I(X;Y)$ é um caso particular de $I(X;Y|Z)$ ao considerar uma distribuição de probabilidade determinística sobre Z, $I(X;Y) = 0 \iff (X \perp\!\!\!\perp Y)_P$. □

1.4.3 Forma Canônica das Desigualdades de Informação

Qualquer quantificador de informação de Shannon pode ser expresso como combinação linear de entropias conjuntas utilizando as identidades abaixo:

$$H(X|Y) = H(X,Y) - H(Y) \tag{1.77a}$$
$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{1.77b}$$
$$I(X,Y|Z) = H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z) \tag{1.77c}$$

Observação. A eq. (1.77c) é a eq. (1.54) reescrita em termos de entropias conjuntas, em vez de entropias condicionais.

Assim como as expressões (1.77) são consideradas formas canônicas de representar quantificadores de informação de Shannon, diremos que uma desigualdade de informação está escrita na forma canônica se ela envolve apenas combinações lineares de entropias conjuntas.

As desigualdades de informação básicas formam o conjunto mais importante de desigualdades de informação. Todas as desigualdades de informação implicadas pelas desigualdades básicas são denomidadas *desigualdades tipo Shannon*. Da mesma forma, desigualdades de informação sempre válidas, mas que não são implicadas por desigualdades básicas, são chamadas de *desigualdades tipo não-Shannon*. Uma desigualdade tipo Shannon é *implicada* por desigualdades básicas se, usando uma manipulação algébrica apropriada, é possível obter tal desigualdade utilizando apenas as desigualdades básicas.

Apesar da categorização simples, saber se uma desigualdade é tipo Shannon ou tipo não-Shannon não é tão simples, em geral, utilizando apenas essa condição. Existe uma abordagem geométrica que permite que, dada uma desigualdade de informação, seja possível dizer se ela é tipo Shannon ou não.

Mesmo assim, ainda fica a pergunta: Existem desigualdades de informação tipo não-Shannon que compõem leis de teoria da informação? A resposta é sim. Entretanto, tais desigualdades não serão discutidas aqui. Ao leitor interessado, checar ref. [4, cap. 15].

1.4.4 Desigualdades tipo Shannon - Uma Abordagem Geométrica

Considere o conjunto $\{n\} = 0, 1, \ldots, n-1$, onde $n \geq 2$, e o conjunto

$$\Theta = \{X_i, i \in \{n\}\} \tag{1.78}$$

que representa uma coleção de variáveis aleatórias X_i. Considere também um vetor $\mathbf{H}$ com 2^n componentes pertencente ao espaço vetorial $\mathbb{R}^{2^n}$. $\mathbf{H}$ será chamado de *vetor entrópico*, ou *vetor de entropia*, se cada componente de $\mathbf{H}$ for a entropia de Shannon de um elemento de $\mathcal{P}(\Theta)$, ou seja, se cada componente de $\mathbf{H}$ for uma entropia conjunta das n variáveis de Θ.

Exemplo 1.4.2. Para $n = 3$,

$$\{n\} = \{1, 2, 3\}, \tag{1.79a}$$

$$\Theta = \{X_1, X_2, X_3\}, \tag{1.79b}$$

$$\mathcal{P}(\Theta) = \{\emptyset, \{X_1\}, \{X_2\}, \{X_3\}, \{X_1, X_2\}, \{X_1, X_3\}, \{X_2, X_3\}, \{X_1, X_2, X_3\}\}. \tag{1.79c}$$

Um vetor $\mathbf{H}$ é entrópico se puder ser escrito na forma

$$\mathbf{H} = \begin{bmatrix} H(\emptyset) \\ H(X_1) \\ H(X_2) \\ H(X_3) \\ H(X_1, X_2) \\ H(X_1, X_3) \\ H(X_2, X_3) \\ H(X_1, X_2, X_3) \end{bmatrix} \tag{1.80}$$

Por simplicidade, adotaremos a convenção de que $H(\emptyset) = 0$. Assim, voltamos nossa atenção ao espaço $\mathbb{R}^{2^n-1}$. Além disso, definimos α_i, $i = 1, \ldots, 2^n$ para denotar o i-ésimo elemento de $\mathcal{P}(\Theta)$.

De modo a caracterizar o conjunto de vetores entrópicos, serão definidas abaixo regiões de interesse do espaço $\mathbb{R}^{2^n-1}$.

Definição 1.4.2 (A região Γ_n^*)**.** Γ_n^* é a região de $\mathbb{R}^{2^n-1}$ composta por todos os pontos cujas coordenadas sejam entropias conjuntas $H(\alpha_i)$, ou seja, Γ_n^* é a região contendo apenas os vetores entrópicos $\mathbf{H}$.

A região Γ_n^* possui as seguintes propriedades:

(1) Γ_n^* contém a origem.

(2) $\overline{\Gamma_n^*}$, o fecho de Γ_n^*, é convexo.

(3) Γ_n^* pertence ao ortante[4] não negativo de $\mathbb{R}^{2^n-1}$.

Definição 1.4.3 (A região Γ_n)**.** Γ_n é a região de $\mathbb{R}^{2^n-1}$ em que as componentes $H(\alpha)$ de um vetor $\mathbf{H} \in \mathbb{R}^{2^n-1}$ satisfazem, para todo α_i e $\alpha_j \in \Theta$, as desigualdades abaixo, chamadas de *axiomas polimatroidais*:

(1) $H(\alpha_i) \geq 0$;

(2) $H(\alpha_i) \leq H(\alpha_j)$, se $\alpha_i \subset \alpha_j$;

(3) $H(\alpha_i \cup \alpha_j) + H(\alpha_i \cap \alpha_j) \leq H(\alpha_i) + H(\alpha_j)$.

Proposição 1.4.8. *Os axiomas polimatroidais são equivalentes às desigualdades básicas para todo* $\alpha \in \mathcal{P}(\Theta)$.

Demonstração. As desigualdades básicas são aquelas que representam a não-negatividade dos quantificadores de informação de Shannon para todas as variáveis em Θ. O axioma (1) já representa a não-negatividade da entropia de Shannon $H(\alpha_i)$, não tendo, portanto, que ser provado. O axioma (2) leva à não-negatividade da entropia condicional. Considerando $\alpha_k = \alpha_j \setminus \alpha_i$, o axioma (2) pode ser reescrito como:

$$\begin{aligned} & H(\alpha_i) \leq H(\alpha_k \cup \alpha_i) \\ \Rightarrow & H(\alpha_k \cup \alpha_i) - H(\alpha_i) \geq 0 \\ \Rightarrow & H(\alpha_k|\alpha_i) \geq 0. \end{aligned} \tag{1.81}$$

O axioma (3) leva à não-negatividade da informação mútua condicional. Considerando $\alpha_k = \alpha_j \setminus \alpha_i$, $\alpha_l = \alpha_i \cap \alpha_j$ e $\alpha_m = \alpha_i \setminus \alpha_j$, podemos reescrever o axioma (3) como

$$\begin{aligned} & H(\alpha_m \cup \alpha_l) + H(\alpha_k \cup \alpha_l) \geq H(\alpha_k \cup \alpha_l \cup \alpha_m) + H(\alpha_l) \\ \Rightarrow & H(\alpha_m \cup \alpha_l) + H(\alpha_k \cup \alpha_l) - H(\alpha_k \cup \alpha_l \cup \alpha_m) - H(\alpha_l) \geq 0 \\ \Rightarrow & H(\alpha_m|\alpha_l) + H(\alpha_k|\alpha_l) - H(\alpha_k \cup \alpha_m|\alpha_l) \geq 0 \\ \Rightarrow & I(\alpha_k; \alpha_m|\alpha_l) \geq 0 \end{aligned} \tag{1.82}$$

□

[4] O termo *ortante* significa uma generalização do termo *octante*, para o caso de um espaço de dimensão maior que 3.

Como todas as desigualdades de informação podem ser escritas na forma canônica, é sempre possível representar uma desigualdade por um produto interno $\mathbf{B}^{\mathsf{T}} \cdot \mathbf{H} \geq 0$, onde $\mathbf{B}^{\mathsf{T}}$ é a transposta de um vetor coluna $\in \mathbb{R}^{2^n-1}$.

Em resumo, podemos representar a hierarquia das regiões que acabamos de definir de acordo com

$$\Gamma_n^* \subseteq \overline{\Gamma_n^*} \subseteq \Gamma_n. \tag{1.83}$$

2 Causalidade

O conteúdo deste capítulo foi baseado principalmente na ref. [7]. Alguns detalhes foram consultados de maneira complementar na ref. [8], do mesmo autor. O foco desse capítulo será principalmente introduzir algumas das várias ferramentas existentes para garantir uma relação de causa e efeito entre dois objetos. Detalhes sobre como construir modelos causais, critérios para preferência de modelos, etc, podem ser encontrados na ref. [7, cap. 2].

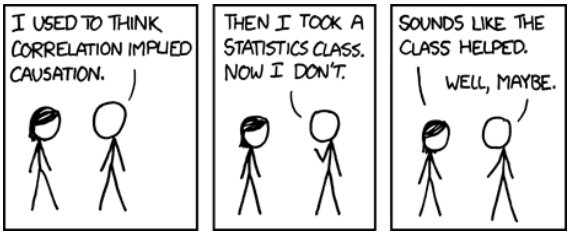
É indicado ao leitor como primeiro contato aos conceitos de teoria de Inferência Causal um *post* do blog do Michael Nielsen, ref. [9]. Outras referências possíveis para o estudo de teoria de Inferência Causal são as refs. [10, 11].

2.1 Estatística e Causalidade

A estatística é uma ciência presente na grande maioria das áreas do conhecimento. Utilizando a teoria de probabilidade como sua base, a estatística consiste de estudos sobre coleta, tratamento e interpretação de dados *observacionais*, de modo a possibilitar a criação de modelos que expliquem um determinado fenômeno e, possivelmente, fazer previsões através de tal modelo.

Por muitas vezes, o tratamento estatístico de um fenômeno considera implicitamente a interpretação *frequencista* de probabilidade[1]. Assim, considera-se que as frequências relativas geradas pela composição e análise de dados coletados para um estudo estatístico determinam, no limite em que o número de observações seja muito grande, a distribuição de probabilidade associada a tais observações.

A estatística pode trazer muita informação sobre um determinado fenômeno. Entretanto, a interpretação de resultados estatísticos pode gerar inferências errôneas, principalmente no que se refere a atribuir relações de *causalidade* entre elementos de um fenômeno estudado.



https://xkcd.com/552/

[1] A consideração é implícita pois a interpretação frequencista é intuitiva, dado que ela relaciona frequências relativas com probabilidades.

O objetivo de um estudo estatístico é descobrir e analisar *correlações* entre objetos estudados. Correlação pode ser pensada como o oposto de independência, ou seja, há correlação entre dois eventos A e B se

$$p(a,b) \neq p(a)p(b). \tag{2.1}$$

2.1.1 O Paradoxo de Simpson

O paradoxo de Simpson é um exemplo que mostra o quanto a observação de eventos puramente estatística pode ser enganosa, de modo a gerar conclusões contraditórias, dependendo da maneira com que os dados sejam tratados. Aqui, o paradoxo de Simpson será apresentado através de um exemplo.

Exemplo 2.1.1. Após a criação de um medicamento para o tratamento de uma determinada doença, este precisa passar por um teste de eficácia para, então, ser disponibilizado para o tratamento de pessoas que possuem tal doença.

Considere a realização de um teste de eficácia para tal medicamento em que 80 indivíduos foram utilizados para fazer parte. Os indivíduos são separados em dois grupos de mesmo número, os chamados *grupo de teste* e o *grupo controle*.

O grupo de teste é composto por indivíduos que fazem uso do medicamento em questão, enquanto o grupo controle é composto por indivíduos que fazem uso de *placebo*, comprimidos sem propriedades farmacológicas.

O uso do placebo é necessário pois é preciso analisar a recuperação dos indivíduos por qualquer outro possível fator desconhecido. Desta forma, a estatística do grupo controle serve para ser comparada com a estatística do grupo teste, dado que tanto no grupo teste quanto no grupo controle, os fatores desconhecidos que podem influenciar na recuperação do indivíduo são considerados existentes, sendo a única diferença entre os grupos portanto, o uso do medicamento a ser testado.

O efeito considerado no teste é a recuperação do indivíduo. Denotaremos por "e" a ocorrência do efeito, "$\neg e$" a não ocorrência do efeito, "n_e" o número de indivíduos de um grupo específico que apresentaram o efeito e "N_g" o número total de indivíduos deste mesmo grupo. O grupo de teste será representado[2] por "c", enquanto o grupo controle será representado por "$\neg c$". A taxa de recuperação de um determinado grupo é a frequência relativa $\frac{n_e}{N_g}$. A tabela a seguir apresenta os dados levantados neste estudo:

[2] As letras "c" e "e" são utilizadas para denotar *causa* e *efeito*.

	e	$\neg e$	N_g	Taxa de recuperação
c	20	20	40	$0,5$
$\neg c$	16	24	40	$0,4$
$c + \neg c$	36	44	80	$0,45$

Tabela 1 – Tabela de dados estatísticos levantados em um teste de eficácia de um medicamento.

Pela interpretação frequencista de probabilidade, ao considerar que o número de indivíduos sujeitos ao teste é suficientemente grande para assumir que as frequências relativas são aproximadamente as probabilidades sobre os eventos, as seguintes probabilidades condicionais são obtidas:

$$p(e|c) = 0,5 \tag{2.2a}$$

e

$$p(e|\neg c) = 0,4. \tag{2.2b}$$

Como $p(e|c) > p(e|\neg c)$, isso indica que uso do medicamento melhora a taxa de recuperação. Portanto, conclui-se que *o medicamento tem eficácia no combate à doença*.

Ao adicionar um novo detalhe à análise, separando os indivíduos testados por sexo, os dados desta nova estatística são representados de acordo com as tabelas a seguir:

Homens	e	$\neg e$	N_g	*Taxa de recuperação*
c	18	12	30	$0,6$
$\neg c$	7	3	10	$0,7$
$c + \neg c$	25	15	40	$0,625$

Tabela 2 – Tabela de dados estatísticos considerando apenas os homens sujeitos à experiência.

Mulheres	e	$\neg e$	N_g	*Taxa de recuperação*
c	2	8	10	$0,2$
$\neg c$	9	21	30	$0,3$
$c + \neg c$	11	29	40	$0,275$

Tabela 3 – Tabela de dados estatísticos considerando apenas as mulheres sujeitas à experiência.

Será utilizado "h" para representar os homens e "m" para representar as mulheres nessa nova análise. Fazendo a mesma consideração feita anteriormente com relação

às frequências relativas, os dados das tabelas nos levam as seguintes probabilidades condicionais:

$$p(e|c,h) = 0,6, \tag{2.3a}$$
$$p(e|\neg c,h) = 0,7 \tag{2.3b}$$
e
$$p(e|c,m) = 0,2, \tag{2.3c}$$
$$p(e|\neg c,m) = 0,3. \tag{2.3d}$$

As relações entre as probabilidades nessa nova análise são:

$$p(e|c,h) < p(e|\neg c,h), \tag{2.4a}$$
$$p(e|c,m) < p(e|\neg c,m). \tag{2.4b}$$

Essas relações levam à conclusão de que os homens têm chances maiores que as das mulheres de se recuperar da doença naturalmente, e que, tanto para homens quanto para mulheres, *o remédio acaba diminuindo as chances de recuperação dos indivíduos*.

Ou seja, inicialmente, concluiu-se que o remédio tinha eficácia no tratamento da doença. Entretando, depois de avaliar separadamente os homens e as mulheres participantes do teste, concluiu-se exatamente o oposto, levando ao paradoxo.

Neste exemplo do paradoxo de Simpson, observa-se a existência de uma correlação entre o uso do medicamento e a recuperação da doença, além de uma correlação entre o sexo dos indivíduos e a recuperação da doença. Afinal, o remédio seria responsável por fazer o indivíduo se recuperar? Ou, pelo contrário, atrapalha a recuperação do indivíduo? Estas questões nos direcionam ao nosso objetivo: Diferenciar correlações genuínas de *correlações espúrias*[3] e descobrir as relações causais por trás das correlações genuínas.

O paradoxo de Simpson ilustra o quanto um estudo estatístico pode ser enganoso. Ao longo deste capítulo, serão introduzidas diversas ferramentas da *Teoria de Inferência Causal*. O uso destas ferramentas reforça certas conclusões extraídas das estatísticas e descarta outras, tornando os resultados muito mais confiáveis.

A solução do paradoxo se dá pelo fato de que a variável "sexo" foi inclusa depois que o estudo estatístico foi elaborado. É sempre possível, a partir de dados estatísticos, manipular subpopulações dos dados, de modo a simular correlações entre coisas que não estejam inclusas na análise estatística inicialmente. Note que nas tabelas 2 e 3, o número de indivíduos no grupo teste é muito diferente do número de indivíduos

[3] Uma correlação espúria é uma correlação sem significado, quando dois objetos exibem um comportamento que se assemelha a um comportamento interdependente, mas que na verdade não é (também conhecida popularmente como "mera coincidência").

no grupo controle. Essa diferença faz com que as taxas de recuperação não façam sentido. Sendo esses números uma consequência da inclusão do sexo dos indivíduos na estatística original, qualquer inferência com relação ao sexo dos indivíduos é incorreta *nessa estatística*. A tabela 1 contém dados tratados de maneira correta, inclusive com a *intervenção* sobre o uso do medicamento. O significado de "intervenção" ficará claro no decorrer desse capítulo.

No dia a dia, é comum sermos bombardeados com informações incompatíveis, como: "Comer ovos fritos traz problemas cardíacos" e "Comer ovos fritos traz benefícios para o coração". Informações espalhadas dessa maneira costumam ser originadas pelo tratamento incorreto de estatísticas, assim como no exemplo que ilustra o paradoxo de Simpson. Muitos desses erros se originam quando, ao ser realizado um levantamento de dados para um estudo estatístico, os experimentadores inferem relações de causa e efeito após perceberem correlações em subpopulações dessa estatística, sem realizar uma nova estatística que permita avaliar se existem, de fato, tais relações causais.

A teoria de inferência causal fortalece o tratamento de dados estatísticos. Se todo estudo estatístico se utilizasse da teoria de inferência causal, as inferências seriam garantidas e as informações que nos bombardeiam no cotidiano seriam muito mais confiáveis.

2.2 Probabilidades, Grafos e Redes Bayesianas

No estudo de inferência causal, variáveis aleatórias serão utilizadas para representar os elementos de um modelo causal.

Observação. A partir de agora, adotaremos a convenção de que sempre que uma variável aleatória for utilizada, o seu alfabeto será representado pelo símbolo que identifica a variável (ex: $\mathcal{X} \to X$), pois estaremos interessados apenas no vetor de probabilidade cujas componentes são associadas aos elementos do alfabeto da variável, não nos elementos do espaço amostral.

Considere uma distribuição de probabilidade conjunta sobre o conjunto de variáveis aleatórias $X = \{X_1, \ldots, X_n\}$. Considere também um ordenamento fixo dessas variáveis $(i = 1, \ldots, n)$. *Nesse ordenamento*, a sequência de variáveis $\{X_1, \ldots, X_{i-1}\}$ contém os chamados *antecessores* de X_i. De acordo com a regra de Bayes, a probabilidade conjunta sobre todas as variáveis de X é dada por

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | x_1, \ldots, x_{i-1}). \tag{2.5}$$

As relações de independência simplificam a probabilidade conjunta $p(x_1, \ldots, x_n)$. Considerando que a probabilidade condicional de cada variável X_i não seja sensível ao

condicionamento sobre todos os seus antecessores no ordenamento especificado acima, tal distribuição será sensível a apenas um subconjunto de antecessores de cada variável X_i. Cada termo existente no produtório do lado direito da eq. (2.5) pode ser substituído por uma probabilidade condicional envolvendo apenas este pequeno subconjunto de antecessores. Tal subconjunto de antecessores será assumido como o conjunto das variáveis que exercem influência causal sobre X_i.

2.2.1 Pais Markovianos

Definição 2.2.1 (Pais markovianos)**.** Seja $X = \{X_1, \ldots, X_n\}$ um conjunto ordenado de variáveis com uma distribuição de probabilidades conjunta sobre todas essas variáveis. O conjunto $\mathrm{PA}_i = \{\mathrm{pa}_i\} \subseteq \{X_1, \ldots, X_{i-1}\}$, denominado *pais markovianos* de X_i, ou *pais* de X_j, é o conjunto mínimo de antecessores de X_i que o torna independente de todos os outros antecessores. Assim, PA_i é qualquer subconjunto de $\{X_1, \ldots, X_{i-1}\}$ que satisfaça

$$p(x_i|\mathrm{pa}_i) = p(x_i|x_1, \ldots, x_{i-1}), \tag{2.6}$$

e que nenhum subconjunto de PA_i satisfaça a eq. (2.6).

A utilização dos pais markovianos para descrever a distribuição de probabilidades conjunta sobre todas as variáveis induz a seguinte forma para a regra de Bayes:

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|\mathrm{pa}_i). \tag{2.7}$$

Observação. Como foi mencionado no capítulo anterior, a regra de Bayes não preferencia ordenamento de probabilidades condicionais. A utilização dos pais markovianos para descrever uma distribuição de probabilidade conjunta representa uma escolha específica de ordenamento de probabilidades condicionais.

Observação. O ordenamento gerado pelo uso dos pais markovianos é útil para o estudo de causalidade, pois o conjunto PA_i será composto por todas as variáveis que são possíveis *causas* da variável X_i.

2.2.2 Grafos Direcionados Acíclicos - DAGs

Na maioria dos cenários estudados nesse trabalho e em todo o estudo de teoria de inferência causal, serão utilizados *Grafos Direcionados Acíclicos* para ilustrar modelos envolvendo relações de influência causal. Para defini-los, vamos definir, primeiramente, *Grafos Direcionados*.

Definição 2.2.2 (Grafo direcionado)**.** Um grafo G consiste de um conjunto de vértices $V = \{V_1, \ldots, V_n\}$ e um conjunto de arestas $A = \{A_1, \ldots, A_m\}$, onde cada aresta $A_i \in A$

conecta um par ordenado de vértices $(V_j, V_k) \in V$. As arestas são direcionadas, ou seja, possuem um sentido de incidência bem definido.

Dois vértices conectados por uma aresta em um grafo G são chamados de *adjacentes*. Um *caminho* c é uma sequência de vértices $(V_j, \ldots, V_m) \subseteq V$ em que os vértices V_k e V_{k+1} são adjacentes para todo k. Um caminho c é dito *orientado* se para todo par de vértices adjacentes (V_k, V_{k+1}) em um caminho c, a aresta que os conecta parte de V_k e chega em V_{k+1}. Um *ciclo orientado* em G é um caminho orientado fechado, ou seja, $V_j = V_m$. Com isso, podemos agora definir a classe de grafos que será de nosso interesse.

Definição 2.2.3 (Grafo Direcionado Acíclico)**.** Um *grafo direcionado acíclico* G é um grafo direcionado livre de ciclos orientados.

Observação. Utilizaremos a sigla *DAG* para nos referir a grafos direcionados acíclicos, devido ao seu termo em inglês *directed acyclic graph*.

O grafo gerado pela remoção do direcionamento das arestas de G é denominado *esqueleto* de G.

Exemplo 2.2.1. A imagem abaixo ilustra um DAG.

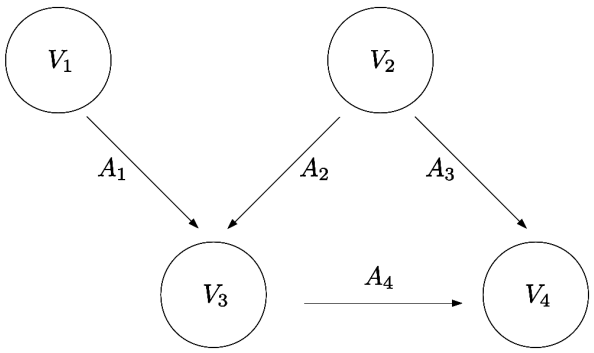


Figura 2 – Grafo direcionado acíclico com $V = \{V_1, V_2, V_3, V_4\}$ e $A = \{A_1, A_2, A_3, A_4\}$.

Definição 2.2.4 (Rede Bayesiana)**.** DAGs utilizados para representar relações temporais ou relações causais são chamados de *redes bayesianas*.

Em uma rede Bayesiana, cada vértice $V_i \in V$ representa uma variável aleatória $X_i \in X$. Os pais de uma variável X_i são todos os vértices do grafo conectados a X_i por uma aresta que parte de $X_j \in \mathrm{PA}_i$ e chega em X_i.

Exemplo 2.2.2. O grafo da fig. 2, ao representar um conjunto de variáveis aleatórias $X = \{X_1, X_2, X_3, X_4\}$ com uma distribuição de probabilidade conjunta P, exibe os

seguintes conjuntos PA_i:

$$\begin{aligned} \mathrm{PA}_1 &= \emptyset, \\ \mathrm{PA}_2 &= \emptyset, \\ \mathrm{PA}_3 &= \{X_1, X_2\}, \\ \mathrm{PA}_4 &= \{X_2, X_3\}. \end{aligned} \tag{2.8}$$

Dessa forma, o seguinte ordenamento da regra de Bayes para tais variáveis é induzido pelo grafo:

$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2)p(x_3|x_1, x_2)p(x_4|x_2, x_3). \tag{2.9}$$

Definição 2.2.5 (Compatibilidade de Markov)**.** Se uma distribuição de probabilidade P admite a fatoração da eq. (2.7) relativa a um DAG G, dizemos que G representa P, ou G e P são compatíveis, ou P é correspondente markoviano de G.

A compatibilidade entre P e G é condição necessária e suficiente para explicar, através de um DAG G, dados estatísticos capazes de gerar P.

2.3 O Critério de *d*-separação

Uma maneira de caracterizar distribuições de probabilidade compatíveis com um DAG G é através da verificação de que independências condicionais representadas em G devem ser satisfeitas por P. Existe um critério, denominado *critério de* d*-separação*, que relaciona certos tipos de estrutura de um DAG G com relações de independência condicional nas distribuições de probabilidade compatíveis com G. Esse critério é nomeado dessa forma para explicitar que a simples separação entre dois vértices V_i e V_k através de um vértice V_j conectado aos dois não implica na condição $(X_i \perp\!\!\!\perp X_k|X_j)_P$, mas que separações com direcionamento específico já são capazes de implicar tais relações. Assim, a ideia de *separação direcional* levou ao nome *d*-separação.

Para definir o critério de *d*-separação, serão considerados três conjuntos disjuntos de vértices em G: X, Y e Z. Assumiremos compatibilidade entre G e distribuições de probabilidade sobre variáveis aleatórias X, Y e Z representadas pelos conjuntos de vértices X, Y e Z em G.

Definição 2.3.1 (*d*-separação)**.** Um caminho c é *d*-separado, ou bloqueado, por um conjunto Z de vértices se, e somente se,

(i) c contenha uma *corrente* $i \to j \to k$ ou um *garfo* $i \leftarrow j \to k$, sendo $j \in Z$, ou

(ii) c contenha um *colisor* $i \to j \leftarrow k$, sendo j e todos os seus descendentes $\notin Z$.

O conjunto Z de vértices *d*-separa X de Y se, e somente se, Z bloqueia todos os caminhos que unem os conjuntos de vértices X e Y.

A *d*-separação de X e Y devido à Z será representada por $(X \perp\!\!\!\perp Y|Z)_G$. Um caminho entre X e Y conectado por Z é denominado *d*-conectado, sendo representado por $(X \not\perp\!\!\!\perp Y|Z)_G$.

Exemplo 2.3.1. O DAG a seguir será utilizado para ilustrar as três possíveis relações entre os conjuntos X, Y e Z de vértices.

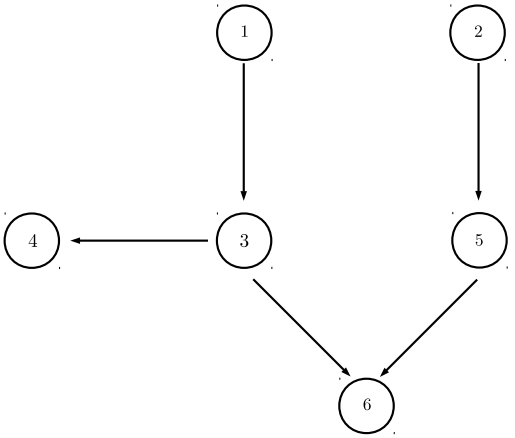
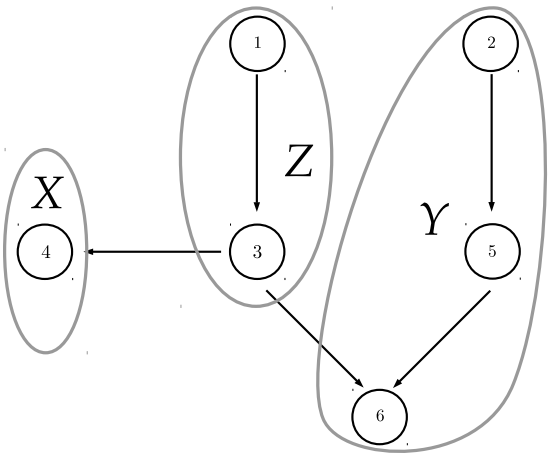
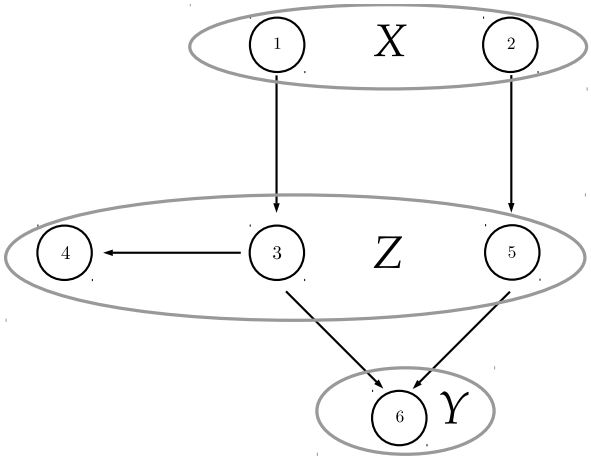


Figura 3 – DAG G com conjunto de vértices $V = \{1, 2, 3, 4, 5, 6\}$.

Representando, com as escolhas adequadas de vértices para pertencer aos conjuntos X, Y e Z, os casos em que $(X \perp\!\!\!\perp Y|Z)_G$ e o caso em que $(X \not\perp\!\!\!\perp Y|Z)_G$:



(a) $X \leftarrow Z \rightarrow Y$



(b) $X \rightarrow Z \rightarrow Y$

Figura 4 – Casos em que o conjunto Z de vértices *d*-separa os conjuntos X e Y.

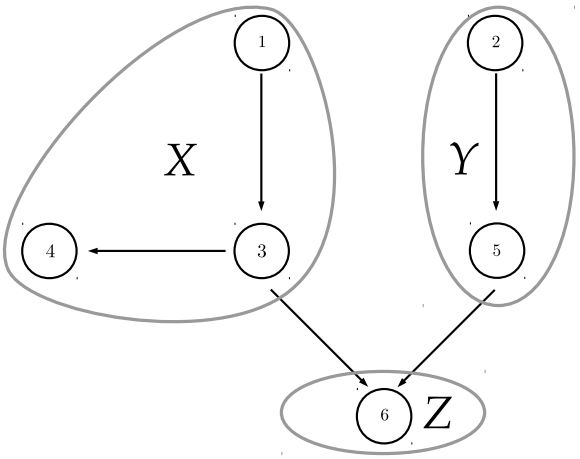


Figura 5 – Caso em que Z d-conecta X e Y ($X \to Z \leftarrow Y$)

Teorema 2.3.1 (Implicações probabilísticas de d-separação)**.** *Para quaisquer três conjuntos disjuntos (X, Y, Z) de vértices em um DAG G e, para toda distribuição de probabilidades P,*

(i) $(X \perp\!\!\!\perp Y|Z)_G \Rightarrow (X \perp\!\!\!\perp Y|Z)_P$ *sempre que G e P forem compatíveis.*

(ii) Se $(X \perp\!\!\!\perp Y|Z)_P$ *é uma relação de independência válida em* todas *as distribuições P compatíveis com G, então* $(X \perp\!\!\!\perp Y|Z)_G$.

Demonstração. A demonstração desse teorema não será realizada pois envolve ferramentas de teoria de inferência causal que não são definidas nesse trabalho. Entretanto, ela pode ser encontrada nas refs. [8] e [7, seções 2.4 e 2.9.1]. □

Observação. Como foi dito na definição, a d-separação $(X \perp\!\!\!\perp Y|Z)_G$ implica a independência *condicional* $(X \perp\!\!\!\perp Y|Z)_P$. Se na distribuição P não houver condicionamento em Z, as relações de independência são invertidas:

(i) $(X \perp\!\!\!\perp Y|Z)_G \Rightarrow (X \not\perp\!\!\!\perp Y)_P$;

(ii) $(X \not\perp\!\!\!\perp Y|Z)_G \Rightarrow (X \perp\!\!\!\perp Y)_P$.

Por exemplo, sem o condicionamento em Z, $(X \not\perp\!\!\!\perp Y)_P$ nos casos representados nas figuras 4a e 4b. Da mesma forma, $(X \perp\!\!\!\perp Y)_P$ na fig. 5.

Teorema 2.3.2 (Condição de Markov Ordenada)**.** *Uma condição necessária e suficiente para uma distribuição de probabilidade P ser compatível com um DAG G é que, condicionada em seus pais markovianos em G, cada variável seja independente de todos os seus antecessores em algum ordenamento de variáveis que concorde com as arestas de G.*

Teorema 2.3.3 (Condição de Markov Parental)**.** *Uma condição necessária e suficiente para que uma distribuição de probabilidade P seja compatível com um DAG G é que cada variável seja independente de todos os seus não-descendentes em G, condicional a seus pais markovianos.*

Os teoremas 2.3.2 e 2.3.3 seguem da definição de compatibilidade de Markov, consistindo de duas maneiras distintas de determinar a compatibilidade entre G e P.

Definição 2.3.2 (v-estruturas)**.** Uma v-estrutura de um grafo G é um caminho composto por três vértices $i, j, k \in V$, em que ocorra um colisor $i \to j \leftarrow k$.

Exemplo 2.3.2. O caminho composto pelos vértices $3, 5, 6$ no DAG da fig. 3 é uma v-estrutura.

Uma distribuição de probabilidade P pode ser compatível com dois grafos distintos G e G'. O teorema a seguir dita restrições sobre G e G' para que a distribuição P seja compatível com ambos.

Teorema 2.3.4 (Equivalência Observacional)**.** *Dois DAGs G e G' são observacionalmente equivalentes se, e somente se, eles possuem o mesmo esqueleto e o mesmo conjunto de* v*-estruturas.*

Demonstração. A demonstração desse teorema pode ser encontrada na ref. [12, pag.259] □

A equivalência observacional impõe uma restrição sobre as possíveis inferências com relação ao sentido das arestas de um grafo ao utilizar apenas uma distribuição de probabilidades P compatível com G e G'. Assim, uma distribuição de probabilidades P é capaz de garantir o sentido de algumas arestas de G. Outras hipóteses e ferramentas devem ser utilizadas para garantir o sentido das arestas restantes.

2.4 Intervenções e Redes Bayesianas Causais

2.4.1 Intervenções

Um estudo estatístico com o objetivo de determinar as relações causais entre as variáveis estudadas deve partir inicialmente de hipóteses sobre as relações de independência entre as variáveis. A partir de tais hipóteses, constrói-se um DAG G que as represente e, a partir daí, avalia-se se uma distribuição de probabilidades P obtida através da *observação* das variáveis em questão é compatível com G, ou seja, se as correlações observadas em P são representadas em G. Mesmo nesta etapa, ainda não é possível garantir relações causais entre as variáveis. Para inferir relações causais entre variáveis é necessário checar como as variáveis se comportam através de *intervenções*.

Definição 2.4.1 (Intervenção)**.** Uma intervenção, ou ação, sobre uma variável aleatória é uma realização forçada de um valor da variável. Uma intervenção elimina a influência dos pais markovianos sobre a variável.

Uma probabilidade condicionada a um valor *observado* de uma variável X é representada por $p(.|x)$. A mesma probabilidade condicionada a um valor *forçado* da mesma variável será representada por $p(.|\text{faça}(X = x))$. Uma intervenção realizada sobre uma variável faz com que seja possível checar a relação causal entre tal variável e seus descendentes.

Exemplo 2.4.1. Consideremos o seguinte DAG G:

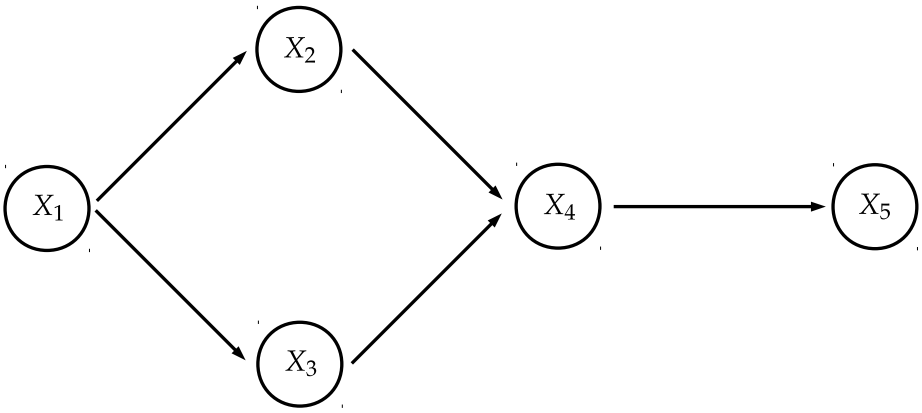


Figura 6 – DAG com $V = \{X_1, X_2, X_3, X_4, X_5\}$

Suponha que o DAG acima represente um modelo em que as váriáveis representem as seguintes observações:

(i) $X_1 =$ Estação do ano;

(ii) $X_2 =$ Ocorrência de chuva;

(iii) $X_3 =$ Estado de um regador automático de jardim;

(iv) $X_4 =$ Presença de água na calçada;

(v) $X_5 =$ Perigo de escorregamento na calçada;

X_1 é a única variável não-dicotômica:

$$\begin{aligned} X_1 : \text{Verão} &\mapsto 0 \\ \text{Outono} &\mapsto 1 \\ \text{Inverno} &\mapsto 2 \\ \text{Primavera} &\mapsto 4. \end{aligned} \tag{2.10}$$

As variáveis X_2, X_3, X_4 e X_5 são dicotômicas, assim, assume-se

$$\begin{aligned} X_i : \text{Sim} &\mapsto 0 \\ \text{Não} &\mapsto 1, \end{aligned} \tag{2.11}$$

para $i = 2, 4, 5$. Já para X_3:

$$\begin{aligned} X_3 : \text{Desligado} &\mapsto 0 \\ \text{Ligado} &\mapsto 1. \end{aligned} \tag{2.12}$$

Como mencionado anteriormente, mesmo que uma distribuição de probabilidades P obtida através da observação dessas variáveis expresse correlações entre elas, isso não é suficiente para garantir que haja uma relação causal entre tais variáveis. Suponha que deseja-se garantir a existência da aresta que liga X_3 a X_4, ou seja, deseja-se checar se o regador automático é capaz de deixar a calçada molhada ou não. Uma justificativa possível para a análise dessa relação causal é que a água lançada pelo regador é concentrada apenas sobre jardim, não chegando à calçada. Para analisar tal relação causal, uma intervenção será feita sobre X_3.

Ao realizar uma intervenção sobre X_3 do tipo "faça$(x_3 = 1)$", isso significa que o experimentador manteria o regador de jardim propositalmente ligado, fazendo uma nova estatística sobre as outras variáveis com essa condição. O mesmo deve ser feito para "faça$(x_3 = 0)$". O grafo G' gerado por esta intervenção é:

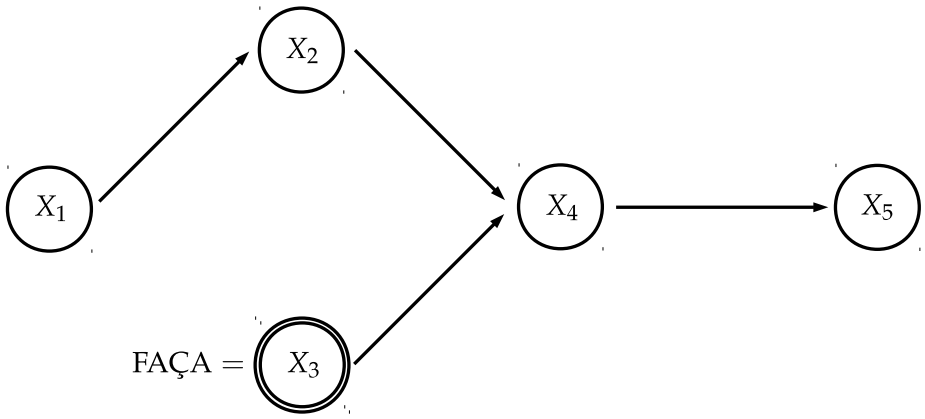


Figura 7 – Dag G' gerado pela intervenção sobre a variável X_3.

A distribuição de probabilidade resultante da operação faça$(x_3 = 1)$ é

$$p_{\text{faça}(x_3=1)}(x_1, x_2, x_4, x_5) = p(x_1)p(x_2|x_1)p(x_4|x_1, x_2, \text{faça}(x_3 = 1))p(x_5|x_4). \tag{2.13}$$

Se a partir dessa nova distribuição, forem obtida as relações

$$p(x_4 = 0|\text{faça}(x_3 = 1)) > p(x_4 = 0|\text{faça}(x_3 = 0) \tag{2.14a}$$

e

$$p(x_4 = 1|\text{faça}(x_3 = 0)) > p(x_4 = 1|\text{faça}(x_3 = 1)), \tag{2.14b}$$

isso indica que, de fato, a variável X_3 tem uma influência causal sobre X_4.

Este exemplo demonstra a diferença entre *observação* e *ação*. O efeito da observação de $X_3 = 0$ é obtido por um condicionamento bayesiano simples $p(x_1, x_2, x_4, x_5|x_3 = 0)$, enquanto que o efeito da ação "faça$(x_3 = 0)$" implica um condicionamento sobre uma modificação do grafo inicial G, eliminando as arestas que ligam PA_3 a X_3. Utilizando-se apenas da observação $x_3 = 0$, poderia existir uma tendência de inferir que X_1 corresponda a uma estação seca, portanto X_2 corresponda a falta de chuva e, por esse motivo, o regador está ligado ($x_3 = 0$), e assim por diante. Esse tipo de inferência jamais poderia ser realizado em um cenário envolvendo intervenções.

2.4.2 Redes Bayesianas Causais

Definição 2.4.2 (Rede Bayesiana Causal)**.** Considere uma distribuição de probabilidades P sobre um conjunto V de variáveis aleatórias. Definimos uma distribuição $P_{\text{faça}(x)}$ como a distribuição resultante da intervenção "faça(x)" em P, reduzindo o subconjunto X a constantes x. Seja $\mathbf{P}_{\mathbf{faça}(\mathbf{x})}$ o conjunto de todas as distribuições $P_{\text{faça}(x)}$, $X \subseteq V$, a distribuição que não envolva intervenções. Um DAG G é uma *Rede Bayesiana Causal* compatível com $\mathbf{P}_{\mathbf{faça}(\mathbf{x})}$ se, e somente se, as seguintes condições são válidas para todo $P_{\text{faça}(x)} \in \mathbf{P}_{\mathbf{faça}(\mathbf{x})}$:

(i) $P_{\text{faça}(x)}$ é compatível com G;

(ii) $p(v_i|\text{faça}(x)) = 1 \; \forall \; V_i \in X$ sempre que v_i for consistente com a ação $X = x$, ou seja, sendo V_i uma das variáveis sobre as quais há uma intervenção, a variável assume o valor v_i deterministicamente.

(iii) $p(v_i|\text{pa}_i, \text{faça}(x)) = p(v_i|\text{pa}_i) \; \forall \; V_i \notin X$ sempre que pa_i for consistente com a ação $X = x$, ou seja, cada $p(v_i|\text{pa}_i)$ seja invariante sob intervenções que não envolvam V_i.

A def. 2.4.2 impõe restrições na distribuição $\mathbf{P}_{\mathbf{faça}(\mathbf{x})}$ que possibilitam codificá-la na forma de uma única rede Bayesiana G. Tais restrições permitem escrever qualquer distribuição $P_{\text{faça}(x)}$ na forma

$$p(v|\text{faça}(x)) = \prod_{\{i|V_i \notin X\}} p(v_i|\text{pa}_i), \tag{2.15}$$

de modo que isso justifica o procedimento de arestas, como na eq. (2.13).

Redes Bayesianas Causais possuem as seguintes propriedades:

(i) Para todo i,

$$p(v_i|\text{pa}_i) = p(v_i|\text{faça}(\text{pa}_i)), \tag{2.16a}$$

(ii) Para todo i e para todo subconjunto S de variáveis disjuntas de $V_i \cup \text{PA}_i$,

$$p(v_i|\text{faça}(\text{pa}_i, s) = p(v_i|\text{faça}(\text{pa}_i)). \tag{2.16b}$$

A propriedade (i) torna cada conjunto PA_i *exógeno* relativo a seus descendentes V_i, garantindo que a probabilidade condicional $p(v_i|\text{pa}_i)$ coincida com o efeito de configurar PA_i em pa_i por controle externo. A propriedade (ii) expressa o fato de que apenas os PA_i possam influenciar seus descendentes V_i, ou seja, V_i é invariante por intervenções sobre variáveis $\notin \text{PA}_i$.

2.5 Modelos Causais Funcionais

Nessa seção, serão definidos os chamados *modelos causais funcionais*. Esses modelos são poderosos por descreverem o funcionamento de cada variável envolvida em função de seus pais markovianos, diferentemente das redes Bayesianas causais, que determinam a existência de relações causais, mas nada dizem sobre como elas se dão.

Apesar dessa descrição, os modelos causais funcionais não deixam de representar a natureza probabilística dos fenômenos estatísticos, no sentido de que se ao saber como uma variável do modelo se comporta em função de todos os seus pais markovianos e, ainda assim, os valores observados da variável não se comportam de maneira determinística, isso sugere que existam fatores não observados gerando flutuações no comportamento previsto para a variável.

Definição 2.5.1 (Modelo Causal Funcional)**.** Um modelo causal funcional consiste de um conjunto de equações da forma

$$x_i = f_i(\text{pa}_i, u_i),\ i = 1, \ldots, n, \tag{2.17}$$

expressando que uma variável aleatória é uma *função* dos seus pais markovianos PA_i e de variáveis U_i, essas últimas representando perturbações devidas a fatores omitidos no modelo e desconhecidos pelo experimentador.

Definição 2.5.2 (Modelo de Equação Estrutural Linear)**.** Um modelo de equação estrutural linear é uma relação do tipo

$$x_i = \sum_{k \neq i} \alpha_{ik} x_k + u_i,\ i = 1, \ldots, n. \tag{2.18}$$

São associados coeficientes α_{ik} nulos às variáveis $X_k \notin \text{PA}_i$, impondo a dependência de cada variável apenas aos seus pais markovianos e a perturbações U_i.

As relações causais funcionais expressam o senso comum das ciências naturais sobre o funcionamento da natureza. Tais relações podem ser interpretadas como a

maneira que a natureza determina um valor x_i para a variável X_i de acordo com todos os possíveis valores de PA_i e U_i. Apesar deste senso comum, veremos no próximo capítulo (seção 3.3.2) que a natureza, por vezes, não permite que aleatoriedade seja justificada pela falta de conhecimento sobre os fatores externos U_i, indicando que a natureza é intrinsecamente aleatória.

Um modelo causal funcional é denominado *semi-markoviano* se os valores das variáveis X são unicamente determinados pelos valores das variáveis U. Assim, a distribuição de probabilidade $p(x_1, \ldots, x_n)$ é unicamente determinada pela distribuição $p(u_1, \ldots, u_n)$ das perturbações U_i. Um diagrama causal G é denominado *markoviano* se cada variável U_i for independente de todas as outras U_j, $j \neq i$.

Observação. Apesar de uma distribuição de probabilidade ter sido assumida para as variáveis de perturbação U_i, isso não implica que essa distribuição possa ser obtida por observações, dado que tais variáveis não são observadas. Tal distribuição é assumida simplesmente para completar a distribuição do modelo e, assim, justificar o comportamento das variáveis observadas X_i.

Definição 2.5.3 (Diagrama Causal)**.** Um diagrama causal é um grafo G gerado, *a partir de um modelo causal funcional*, ao adicionar arestas partindo de cada variável em PA_i e chegando em X_i, para todas as variáveis representadas no grafo.

Teorema 2.5.1 (Condição de Markov Causal)**.** *Todo modelo funcional causal markoviano M induz uma distribuição $p(x_1, \ldots, x_n)$ que satisfaz a Condição de Markov Parental relativa ao diagrama causal G associado a M.*

Demonstração. Considerando que o conjunto $\text{PA}_i \cup U_i$ determina unicamente o valor de X_i, a distribuição de probabilidades $p(x_1, \ldots, x_n, u_1, \ldots, u_n)$ é compatível com o DAG aumentado $G(X, U)$ em que as variáveis U são representadas explicitamente. A condição de markov sobre a distribuição marginal $p(x_1, \ldots, x_n)$ segue da d-separação em $G(X, U)$. □

Em geral, serão utilizados neste trabalho apenas modelos *markovianos*. Consideraremos que se uma variável de perturbação U_i tiver alguma influência sobre outra variável que não seja a variável X_i, ela deverá ser tratada como um *covariante*.

Definição 2.5.4 (Covariante)**.** Um *covariante* é um conjunto C de vértices de um diagrama causal G que representa uma influência não-observada comum a dois outros conjuntos X e Y em G.

Diferentemente das variáveis de perturbação U_i mencionadas anteriormente, os covariantes são representados explicitamente no grafo, e como um covariante é

inacessível, ele não é representado em uma distribuição de probabilidades P compatível com G.

Exemplo 2.5.1. A imagem a seguir ilustra um covariante C em um DAG G.

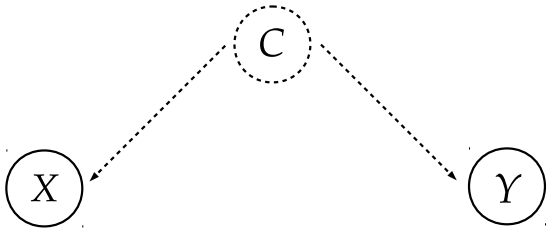


Figura 8 – Representação de um covariante C e sua influência sobre dois conjuntos X e Y em um DAG.

Para mais detalhes sobre tratamento de DAGs com covariantes podem ser encontrados na ref. [7, p.78-84].

Observação. Aos leitores familizarizados com teoria de Não-localidade, as variáveis ocultas responsáveis por definir o conjunto de correlações locais, definidas na seção 3.3.2, correspondem a uma classe dos covariantes aqui definidos.

2.6 Cálculo de Intervenções

O Cálculo de Intervenções é uma ferramenta útil para obter probabilidades condicionadas a intervenções sem que intervenções sejam, de fato, realizadas sobre as variáveis. Na prática, muitas variáveis podem não estar acessíveis para intervenções, de modo que o observador pode ser limitado a apenas observar tais variáveis. Isso pode ser visto no ex. 2.4.1, dado que as únicas variáveis sobre as quais é possível intervir são X_3, X_4 e X_5, as variáveis X_1 e X_2 representam elementos da natureza que o experimentador não controla. Por esse motivo, o Cálculo de Intervenções é útil.

Considerando três conjuntos disjuntos X, Y e Z de vértices em um DAG causal G, $G_{\overline{X}}$ denotará o grafo obtido pela remoção de todas as arestas que chegam em X em G, e $G_{\underline{X}}$ representará o grafo obtido pela remoção de todas as arestas que partem de X. A notação $p(.|\text{faça}(x))$ será substituída por $p(.|\hat{x})$ por simplicidade, dado que há, neste momento, mais intuição sobre a diferença entre observação e intervenção.

Teorema 2.6.1 (Regras do Cálculo de Intervenções)**.** *Seja G um DAG associado a um modelo causal funcional e P a distribuição de probabilidades induzida pelo modelo. Para quaisquer subconjuntos disjuntos X, Y e Z de variáveis, as seguintes regras são válidas:*

(i) Inserção/remoção de observações:

$$p(x|y,z) = p(x|y) \text{ se } (X \perp\!\!\!\perp Z|Y)_G. \tag{2.19a}$$

(ii) Permutação entre observação e ação:

$$p(x|y,\hat{z}) = p(x|y,z) \text{ se } (X \perp\!\!\!\perp Z|Y)_{G_{\underline{Z}}}. \tag{2.19b}$$

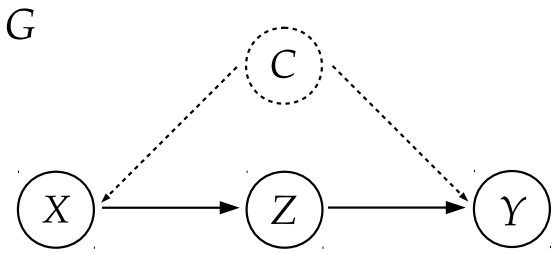
(iii) Inserção/remoção de ações:

$$p(x|y,\hat{z}) = p(x|y) \text{ se } (X \perp\!\!\!\perp Z|Y)_{G_{\overline{Z(Y)}}} \tag{2.19c}$$

Demonstração. A demonstração da validade destas regras pode ser encontrada na ref. [7, p.86]. □

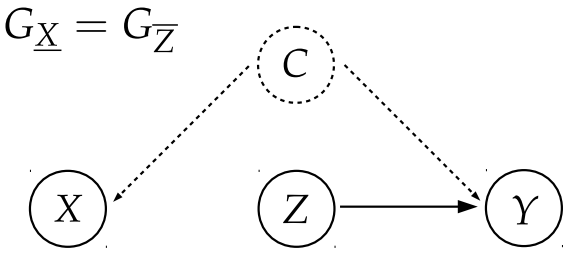
Corolário 2.6.1. *Um efeito causal $p(x_1,\ldots,x_j|\hat{z}_1,\ldots,\hat{z}_k)$ é identificável em um modelo caracterizado por um grafo G se existe uma sequência finita de transformações, cada uma de acordo com as regras do teo. 2.6.1, que reduza $p(x_1,\ldots,x_j|\hat{z}_1,\ldots,\hat{z}_k)$ a uma expressão de probabilidades que envolva apenas variáveis* observadas, *e não envolva intervenções.*

Exemplo 2.6.1 (Derivando efeitos causais por observações)**.** Considere o seguinte DAG:



Expressaremos, utilizando as regras do teorema 2.6.1, diferentes probabilidades condicionadas a intervenções através de probabilidades condicionadas a observações:

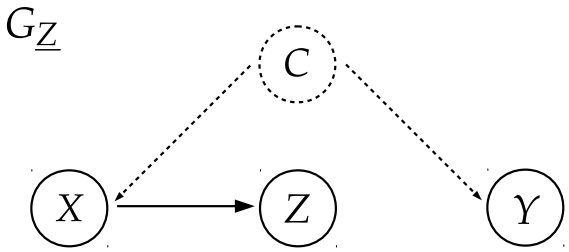
(I) Efeito causal de X sobre Z ($\mathbf{p}(\mathbf{z}|\hat{\mathbf{x}})$): Será utilizada a regra (ii) neste caso. Para isso, é necessário analisar o grafo $G_{\underline{X}}$:



Neste grafo, não há aresta partindo de X e chegando em Z. O único caminho que conecta as duas variáveis é $X \leftarrow C \rightarrow Y \leftarrow Z$ que só passa a ser d-conectado por Y ao haver condicionamento em Y. Não havendo condicionamento, o caminho é d-separado. Portanto, como $(X \perp\!\!\!\perp Z)_{G_{\underline{X}}}$,

$$p(z|\hat{x}) = p(z|x). \tag{2.20}$$

(II) Efeito causal de Z sobre Y ($\mathbf{p}(\mathbf{y}|\hat{\mathbf{z}})$): Não é possível utilizar a regra (ii) aqui. Isso pode ser percebido ao analisar o grafo $G_{\underline{Z}}$ necessário para essa regra:



O caminho que liga Y a Z é $Z \leftarrow X \leftarrow C \rightarrow Y$ que é d-conectado por não haver condicionamento em X: $(Y \not\perp\!\!\!\perp Z)_{G_{\underline{Z}}}$. Para bloquear este caminho através de X, devemos checar o condicionamento em X, incluindo-o na análise de $p(y|\hat{z})$:

$$p(y|\hat{z}) = \sum_x p(x, y|\hat{z}) = \sum_x p(x|\hat{z}) p(y|x, \hat{z}). \tag{2.21a}$$

É necessário remover agora $\hat{z}$ dos dois termos do lado direito da equação (2.21a). Utilizando a regra (iii) para o termo $p(x|\hat{z})$, remove-se o condicionamento em $\hat{z}$ pois $(X \perp\!\!\!\perp Z)_{G_{\overline{Z}}}$, como mostrado no item (I). Portanto,

$$p(x|\hat{z}) = p(x). \tag{2.21b}$$

Retomando a análise inicial de $G_{\underline{Z}}$ para modificar o termo $p(y|x, \hat{z})$ na eq (2.21a), vemos que agora o caminho que conecta Y a Z é d-separado por X: $(Y \perp\!\!\!\perp Z|X)_{G_{\underline{Z}}}$, portanto, podemos utilizar a regra (ii):

$$p(y|x, \hat{z}) = p(y|x, z). \tag{2.21c}$$

Dessa forma, reescrevemos a expressão $p(y|\hat{z})$ apenas em termos de observações:

$$p(y|\hat{z}) = \sum_x p(x) p(y|x, z). \tag{2.21d}$$

Este exemplo serve para ilustrar a força do cálculo de intervenções, reafirmando que através dele, é possível determinar efeitos de intervenções utilizando observações realizadas da maneira correta.

Vimos nesse capítulo que inferir relações causais a partir de correlações é uma tarefa não-trivial e que deve ser tratada com cautela, além disso, apresentamos diversas ferramentas que permitem garantir quando correlações são originadas por relações causais.

O próximo capítulo consiste da teoria de Não-localidade, que é baseada no estudo de tipos de correlações entre variáveis. Ter em mente a distinção entre correlações e relações causais será útil na discussão de seu conteúdo, além de poder trazer mais intuição sobre os conceitos envolvidos.

3 Não-localidade

Neste capítulo, uma breve introdução à teoria de Não-localidade será feita. O estudo de Não-localidade foi motivado pela descoberta de que a Mecânica Quântica, teoria desenvolvida para descrever o comportamento de objetos microscópicos, prevê diversos fenômenos que são difíceis de compreender sob a luz da Mecânica Clássica, reforçados por um vasto número de confirmações experimentais. Um desses fenômenos é o *emaranhamento*, um tipo de correlação que as partes de um sistema quântico podem exibir, e que não é possível de ser explicado classicamente. O emaranhamento e outras correlações que a mecânica quântica prevê, e que são incompatíveis com a Mecânica Clássica, são englobados nas chamadas *correlações não-locais*.

Ao leitor interessado em se aprofundar na teoria de Não-localidade, é indicado ler a ref. [13], na qual boa parte desse capítulo foi baseada. Para uma introdução à mecânica quântica, as refs. [14, 15]. Outra ótima referência para uma introdução à mecânica quântica, mais voltada para a área de Computação e Informação quântica é a ref. [16].

3.1 Geometria Convexa

Iniciaremos o estudo de Não-localidade através da introdução de certas ferramentas de geometria convexa, que serão a base de todo este estudo.

Definição 3.1.1 (Conjunto Convexo)**.** Um conjunto $S \subset \mathbb{R}^N$ é convexo se, para quaisquer elementos s_1 e $s_2 \in S$,

$$\alpha s_1 + (1-\alpha)s_2 \in S,\ \alpha \in [0,1]. \tag{3.1}$$

Analisando geometricamente o conjunto convexo S em um espaço Euclidiano $\mathbb{R}^N$, a sua principal característica é que, ao unir os pontos s_1 e s_2 por um segmento de reta, todos os pontos deste segmento de reta pertencerão a S.

Definição 3.1.2 (Combinação Convexa)**.** Uma *combinação convexa* s de elementos s_i pertencentes a um conjunto convexo S é também um elemento de S, sendo definido por

$$s = \sum_i \alpha_i s_i, \tag{3.2}$$

$\alpha_i \in [0,1]$ e $\sum_i \alpha_i = 1$.

Definição 3.1.3 (Elemento extremal)**.** Um elemento $s_i \in S$ é *extremal* se não puder ser escrito como combinação convexa de outros pontos pertencentes a S.

Definição 3.1.4 (Fecho convexo)**.** O *fecho convexo* $\overline{\mathcal{S}}$ de um conjunto qualquer $\mathcal{S} \subset \mathbb{R}^N$ é o menor conjunto convexo que contém $\mathcal{S}$.

Exemplo 3.1.1. Considere o conjunto de pontos $\mathcal{S} \subset \mathbb{R}^3$:

$$\mathcal{S} = \{s_1, s_2, s_3, s_4\} = \{(0,0,0), (1,0,0), (0,1,0), (0,0,1)\}, \tag{3.3}$$

o fecho convexo $\overline{\mathcal{S}}$ é o *tetraedro* de vértices s_1, s_2, s_3, s_4, representado na imagem abaixo:

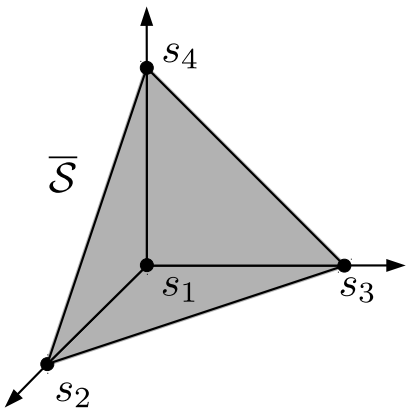


Figura 9 – Fecho convexo $\overline{\mathcal{S}}$ do conjunto de pontos $\mathcal{S}$.

Alguns conjuntos de interesse para o estudo de não-localidade são aqueles representados por fechos convexos de conjuntos *finitos* de pontos. Esses conjuntos são denominados *Politopos*.

Definição 3.1.5 (Politopo)**.** Um *politopo* é um conjunto representado pelo fecho convexo de um conjunto finito de pontos $\mathcal{S} \subset \mathbb{R}^N$.

Observação. Rigorosamente, não é necessário que um politopo seja convexo. Entretanto, como neste trabalho apenas politopos convexos serão utilizados, a definição foi restrita para esse caso.

Os vértices de um politopo são os pontos extremais desse conjunto. O conjunto representado por $\overline{\mathcal{S}}$ na fig. 9 é um politopo, pois é o fecho convexo de um conjunto finito de pontos.

Definição 3.1.6 (Hiperplano de Suporte)**.** Considere um vetor $\mathbf{a} \in \mathbb{R}^N$ e uma constante $b \in \mathbb{R}$. Um hiperplano $H = \{\mathbf{h} \in \mathbb{R}^N | \mathbf{a} \cdot \mathbf{h} = b\}$ é um hiperplano de suporte de um conjunto convexo $\mathcal{S} \subset \mathbb{R}^N$, se $H \cap \mathcal{S}$ contém pelo menos um elemento e o semiespaço H_{se} satisfaz a relação $S \subset H_{se} = \{\mathbf{h} \in \mathbb{R}^N | \mathbf{h} \cdot \mathbf{a} \leq b\}$.

Um resultado de geometria convexa, denominado *Teorema de Minkowski*, diz que um politopo $\overline{\mathcal{S}}$ pode ser representado equivalentemente de duas maneiras:

(i) *Representação $\mathcal{V}$:* Nessa representação, o politopo é descrito como o fecho convexo de um conjunto finito de pontos $\mathcal{S} = \{s_1, \ldots, s_n\}$:

$$\overline{\mathcal{S}} = \left\{ \mathbf{p} \in \mathbb{R}^N \,\middle|\, \mathbf{p} = \sum_i \alpha_i s_i,\ \alpha_i \geq 0,\ \forall\, i,\ \sum_i \alpha_i = 1 \right\}. \tag{3.4a}$$

(ii) *Representação $\mathcal{H}$:* Nessa representação, o politopo é descrito como a interseção de um número finito de semiespaços:

$$\overline{\mathcal{S}} = \left\{ \mathbf{p} \in \mathbb{R}^N \,|\, \mathbf{a}_i \cdot \mathbf{p} \geq\ b_i\ ,\ \forall\, i \in I \right\}, \tag{3.4b}$$

em que $\{(\mathbf{a}_i, b_i), i \in I\}$ denota o conjunto finito de desigualdades. As desigualdades que definem um politopo são denominadas *facetas* do politopo.

Observação. Para tais representações, assume-se implicitamente a ideia de minimalidade, no sentido de que é possível representar um determinado politopo como o fecho convexo de um número arbitrário de pontos na representação $\mathcal{V}$, ou equivalentemente, como a interseção de um número arbitrário de semiespaços na representação $\mathcal{H}$. Entretanto utilizamos sempre o número mínimo de vértices ou semiespaços capazes de caracterizar o politopo.

3.2 Abordagem Independente de Dispositivos

Neste trabalho, caracterizaremos conjuntos de correlações utilizando uma abordagem denominada *Independente de Dispositivos*. A natureza dessa abordagem é expressa pelo uso das chamadas *caixas pretas*. Uma caixa preta é um objeto matemático que representa um conjunto de funções, ou entradas, para cada qual, há um conjunto de resultados associado. Se pensarmos em uma caixa preta como a representação de um dispositivo envolvendo sistemas físicos, é possível considerar que suas entradas sejam *medições* em sistemas.

As caixas pretas recebem esse nome de modo a ilustrar a ideia de independência de dispositivos, no sentido de que há máxima ignorância sobre o funcionamento das caixas, assim, a descrição que traz o máximo de informação sobre as caixas envolve apenas distribuições de probabilidades de resultados de entradas condicionados às escolhas dessas entradas.

Observação. Não faz sentido considerar que uma caixa preta seja, em geral, uma representação de medições e resultados em um sistema físico. Entretanto, por ser sempre possível obter uma descrição de uma caixa preta a partir de um conjunto de medições em um sistema físico, faremos uma extrapolação nesse sentido, considerando que caixas pretas sejam, de fato, a representação de medições e resultados em sistemas físicos, destacando que muitos desses sistemas não são reais. A partir desta ideia, sempre chamaremos as entradas de uma caixa preta de *medições*.

Exemplo 3.2.1. Uma caixa preta que descreve um experimento em que é possível fazer alguma medição $x_i \in X = \{x_0, \ldots, x_{n-1}\}$, cada qual com possíveis resultados $A = \{a_0, \ldots, a_{n-1}\}$ pode ser representada pela figura a seguir:

Figura 10 – Caixa preta representando as possíveis medições $x_i \in X$ e possíveis resultados $a_j \in A$.

O motivo principal pelo qual a abordagem independente de dispositivos é adotada neste trabalho é estudar as correlações que a teoria quântica permite, em comparação com outros tipos de correlações. As distribuições de probabilidade que podem ser obtidas através da teoria quântica serão analisadas em conjunto com algumas hipóteses físicas que definem certos tipos de correlações. Tais hipóteses serão expressas como restrições sobre conjuntos de distribuições de probabilidade e a análise das correlações será realizada sobre *conjuntos de correlações*.

3.2.1 Caixas Pretas Simples

Definição 3.2.1 (Caixa Preta)**.** Uma caixa preta[1] **P** que descreve um conjunto de medições $\{m_0, \ldots, m_{N-1}\}$, em que cada medição possui[2] um conjunto de resultados $\{r_0, \ldots, r_{D-1}\}$, é representada por uma matriz de dimensão $N \times D$:

$$\mathbf{P} = \begin{pmatrix} p(r_0|m_0) & \ldots & p(r_0|m_{N-1}) \\ p(r_1|m_0) & \ldots & p(r_1|m_{N-1}) \\ \vdots & \ddots & \vdots \\ p(r_{D-1}|m_0) & \ldots & p(r_{D-1}|m_{N-1}) \end{pmatrix}, \tag{3.5}$$

sendo $p(r_i|m_j)$ a probabilidade de obter o resultado r_i ao realizar a medição m_j.

Cada coluna da matriz **P** contém uma distribuição de probabilidade para todos os possíveis resultados r_i condicionada à escolha de medição m_j. Portanto, a soma de

[1] Os termos "caixa preta" e "caixa" serão utilizados de forma equivalente.

[2] Aqui foi considerado que todas as medições de uma caixa preta possuem o mesmo número de resultados possíveis. Essa condição não é necessária: É possível analisar casos de caixas pretas com diferentes quantidades de possíveis resultados entre as medições, entretanto estes casos não serão de interesse para esse trabalho.

todos os elementos de cada coluna de **P** deve satisfazer a condição de normalização

$$\sum_i p(r_i|m_j) = 1, \tag{3.6a}$$

além disso, cada elemento deve satisfazer a não-negatividade

$$p(r_i|m_j) \geq 0. \tag{3.6b}$$

É conveniente introduzir, sem perda de generalidade, um vetor $\mathbf{p} \in \mathbb{R}^{N\times D}$ para representar a caixa **P**:

$$\begin{aligned}\mathbf{p} = \big[&p(r_0|m_0) \quad p(r_1|m_0) \ \ldots \ p(r_{D-1}|m_0) \quad p(r_0|m_1) \ \ldots \\ &\ldots \ p(r_{D-1}|m_1) \ \ldots \ p(r_0|m_{N-1}) \ \ldots \ p(r_{D-1}|m_{N-1})\big].\end{aligned} \tag{3.7}$$

A condição de normalização e a não-negatividade, expressos pelas eqs. (3.6), restringem os vetores $\mathbf{p}$ a um conjunto $\mathcal{B}(N, D) \subset \mathbb{R}^{N\times D}$.

Proposição 3.2.1. *O conjunto $\mathcal{B}(N, D)$ é convexo.*

Demonstração. Um vetor $\mathbf{p}$ pertence a $\mathcal{B}(N, D)$ se ele satisfaz as condições de normalização sobre todas as escolhas de medição e se $p(r_i|m_j) \geq 0$, $\forall\ p(r_i|m_j) \in \mathbf{p}$. Considerando que $\mathbf{p}$ seja um vetor obtido por combinação de caixas $\mathbf{p_1}, \mathbf{p_2} \in \mathcal{B}(N, D)$, ele é escrito como

$$\mathbf{p} = \alpha\mathbf{p_1} + (1-\alpha)\mathbf{p_2}. \tag{3.8a}$$

Cada componente de $\mathbf{p}$ é então

$$p(r_i|m_j) = \alpha p_1(r_i|m_j) + (1-\alpha)p_2(r_i|m_j), \tag{3.8b}$$

como $p_1(r_i|m_j) \geq 0$ e $p_2(r_i|m_j) \geq 0$, a combinação convexa destes dois elementos implica que $p(r_i|m_j) \geq 0\ \forall\ p(r_i|m_j) \in \mathbf{p}$, satisfazendo assim a condição de não-negatividade. A condição de normalização é satisfeita se, para todo $p(r_i|m_j)$,

$$\sum_i p(r_i|m_j) = 1. \tag{3.8c}$$

Como $\mathbf{p} = \alpha\mathbf{p_1} + (1-\alpha)\mathbf{p_2}$

$$\begin{aligned}\sum_i p(r_i|m_j) &= \sum_i \left[\alpha p_1(r_i|m_j) + (1-\alpha)p_2(r_i|m_j)\right] \\ &= \left[\alpha\left(\sum_i p_1(r_i|m_j)\right) + (1-\alpha)\left(\sum_i p_2(r_i|m_j)\right)\right] \\ &= \alpha + 1 - \alpha = 1.\end{aligned} \tag{3.8d}$$

Assim, como qualquer vetor $\mathbf{p}$ obtido por combinação convexa de dois vetores $\mathbf{p_1}, \mathbf{p_2} \in \mathcal{B}(N, D)$ também pertence a $\mathcal{B}(N, D)$, conclui-se que $\mathcal{B}(N, D)$ é um conjunto convexo. □

As chamadas *caixas determinísticas* são aquelas que possuem, para cada escolha de medição m_j,

$$p(r_i|m_j) = \begin{cases} 1, & \text{para algum } i = k \\ 0, & \forall \quad i \neq k \end{cases} . \tag{3.9}$$

As caixas determinísticas são vértices do conjunto $\mathcal{B}(N, D)$, pois não é possível obter tais caixas através de uma combinação convexa de outras caixas $\in \mathcal{B}(N, D)$. O número de caixas determinísticas de $\mathcal{B}(N, D)$ é D^N. Como essas caixas são os pontos extremais do conjunto $\mathcal{B}(N, D)$ e qualquer outro ponto desse conjunto pode ser obtido por combinação convexa de tais pontos extremais, conclui-se que $\mathcal{B}(N, D)$ é um politopo.

3.2.2 Caixas Pretas Bipartidas

Todas as análises que serão feitas sobre não-localidade se basearão em algum *cenário* fixo. Um cenário é definido pela especificação do número de partes de um sistema, número de medições que cada parte é capaz de realizar, e número de resultados possíveis de cada medição. Um cenário de Caixas Pretas Simples é determinado pelo conjunto $\mathcal{B}(N, D)$, no qual, o sistema é constituído de apenas uma parte. Já um cenário com número P de partes é determinado por um conjunto $\mathcal{B}(P, N, D)$. Este trabalho é voltado para *cenários bipartidos*, ou seja, cenários que representam um sistema composto por duas partes.

Observação. As medições e os resultados serão representados por variáveis aleatórias. Assim, cada valor da variável "medição" representa uma escolha de medição possível. Da mesma forma, cada valor da variável "resultado" representa um possível resultado.

Definição 3.2.2 (Caixa Preta Bipartida)**.** Uma *caixa preta bipartida* $\mathbf{P_{AB}}$ representa um sistema composto por duas partes A e B, em que a parte A pode realizar qualquer medição de um conjunto $X = \{x_0, \ldots, x_{N-1}\}$ de medições, e a parte B pode realizar qualquer medição de um conjunto $Y = \{y_1, \ldots, y_{N-1}\}$ de medições. Cada medição da parte A possui um conjunto de resultados possíveis $\{a_0, \ldots, a_{D-1}\}$, e cada medição da parte B possui um conjunto de resultados possíveis $\{b_0, \ldots, b_{D-1}\}$. Uma caixa preta bipartida $\mathbf{P_{AB}}$ é representada por uma matriz

$$\mathbf{P_{AB}} = \begin{bmatrix} p(a_0, b_0|x_0, y_0) & \ldots & p(a_0, b_0|x_{N-1}, y_{N-1}) \\ p(a_0, b_1|x_0, y_0) & \ldots & p(a_0, b_1|x_{N-1}, y_{N-1}) \\ \vdots & \ddots & \vdots \\ p(a_{D-1}, b_{D-1}|x_0, y_0) & \ldots & p(a_{D-1}, b_{D-1}|x_{N-1}, y_{N-1}) \end{bmatrix}, \tag{3.10}$$

sendo cada elemento $p(a_i, b_j|x_k, y_m)$ a probabilidade da parte A obter o resultado a_i após a escolha de medição x_k e a parte B de obter o resultado b_j após a escolha de medição y_m.

Observação. Na definição 3.2.2, a caixa preta bipartida foi definida de forma que tanto o número de medições quanto o número de resultados são os mesmos nas duas partes. Essa restrição não é necessária em geral, sendo adotada aqui apenas pois não nos preocuparemos com cenários em que os números de medições e resultados de cada partição sejam diferentes.

Observação. Repetiremos uma simplificação de notação adotada no cap. 1 em que os índices que identificam os valores de uma variável serão omitidos.

Observação. Por motivo de tradição na área de Informação Quântica, os responsáveis por fazer experimentos sobre os sistemas A e B serão chamados de "Alice" e "Bob", respectivamente.

De maneira similar à caixa simples, cada coluna da matriz $\mathbf{P_{AB}}$ contém a distribuição de probabilidade sobre todos os possíveis resultados para da medição conjunta de x e y. Dessa forma,

$$\sum_{a,b} p(a,b|x,y) = 1. \tag{3.11}$$

Utilizam-se probabilidades conjuntas sobre os resultados das duas partes condicionadas às escolhas de medição de cada parte pois a ideia é analisar as correlações entre os resultados das medições das duas partes.

Analogamente às caixas simples, uma caixa bipartida $\mathbf{P_{AB}}$ também pode ser representada por um vetor $\mathbf{p_{AB}} \in \mathbb{R}^{D^2 \times N^2}$. Em geral, conjuntos de caixas dependem, além do número de medições e de resultados por medição, do número de partes que compõem. Por esse motivo, o conjunto de caixas bipartidas será denotado por $\mathcal{B}(2, N, D)$.

Proposição 3.2.2. *O conjunto $\mathcal{B}(2, N, D)$ de caixas bipartidas é convexo.*

Demonstração. A demonstração pode ser feita utilizando o mesmo método utilizado para a demonstração da prop. 3.2.1. Além disso, assim como o conjunto $\mathcal{B}(N, D)$, o conjunto de caixas bipartidas $\mathcal{B}(2, N, D)$ também é um politopo. □

3.2.3 O Cenário CHSH

O cenário CHSH[3] é representado por um sistema composto por dois subsistemas, sobre os quais é possível realizar duas medições em cada, com dois possíveis resultados em cada medição. O conjunto que representa esse cenário é o $\mathcal{B}(2, 2, 2)$. A imagem abaixo ilustra o cenário CHSH.

[3] Sigla utilizada devido ao trabalho de J. F. Clauser, M. A. Horne, A. Shimony, R. A. Holt, encontrado na ref. [17].

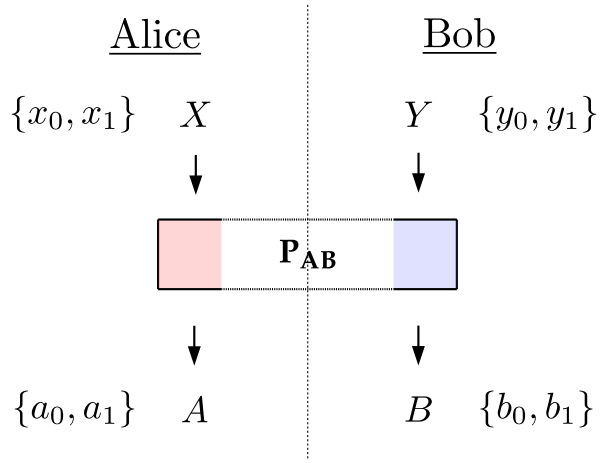


Figura 11 – Representação do cenário CHSH.

A probabilidade de obter o resultado de uma parte para uma dada escolha de medição conjunta sobre as duas partes é obtida através da marginalização sobre a probabilidade conjunta:

$$p(a|x,y) = \sum_b p(a,b|x,y), \tag{3.12a}$$

$$p(b|x,y) = \sum_a p(a,b|x,y). \tag{3.12b}$$

O número de pontos extremais do politopo $\mathcal{B}(2,2,2)$ é igual ao número de pontos extremais do politopo $\mathcal{B}(4,4)$, basta notar que a dimensão de uma matriz que representa uma caixa $\mathbf{P_{AB}} \in \mathcal{B}(2,2,2)$ é a mesma de uma matriz que representa uma caixa $\mathbf{P} \in \mathcal{B}(4,4)$. Assim, como o conjunto $\mathcal{B}(N,D)$ possui D^N pontos extremais, $\mathcal{B}(2,2,2)$ possui $4^4 = 256$ pontos extremais.

3.3 Conjuntos de Correlações

Em um cenário de não-localidade, as caixas $\mathbf{P_{AB}}$ carregam correlações entre as probabilidades de obtenção de resultados a e b condicionadas às escolhas x e y. Cada tipo de correlação será representada por um subconjunto de $\mathcal{B}(2,2,2)$.

3.3.1 Correlações Não-sinalizantes

Fisicamente, gostaríamos de nos restringir a correlações não-sinalizantes para analisar as correlações permitidas no caso em que não há informação sendo transmitida entre Alice e Bob. Tal transmissão de informação pode ocorrer quando a probabilidade do resultado da medição sobre uma parte depende da escolha de medição da outra parte, assim, nos restringiremos a casos em tal influência não ocorra. Além disso, assumimos

a hipótese de *livre escolha*, que consiste em considerar que as escolhas de medição de cada parte não dependam de outras variáveis.

Tentaremos, sempre que possível, utilizar DAGs para representar os cenários de não-localidade. Entretanto, como tais DAGs devem ser compatíveis simultaneamente com todas as distribuições de probabilidade contidas em uma caixa, isso limita a representação por DAGs, de modo que os resultados de cada parte serão representados por um único vértice que os englobe.

Definição 3.3.1 (Correlação Não-Sinalizante)**.** Uma caixa $\mathbf{P_{AB}} \in \mathcal{B}(2,2,2)$ é dita *não-sinalizante* se *todas* as probabilidades marginais sobre a parte da Alice e sobre a parte do Bob satisfazem as relações

$$p(a|x,y) = p(a|x),\ \forall y, \tag{3.13a}$$

$$p(b|x,y) = p(b|y),\ \forall x. \tag{3.13b}$$

Como dito anteriormente, a condição de não-sinalização, expressa pelas eqs. (3.13), proíbe que a escolha de medição sob cada parte influencie o resultado da outra parte. O DAG a seguir ilustra as correlações não-sinalizantes.

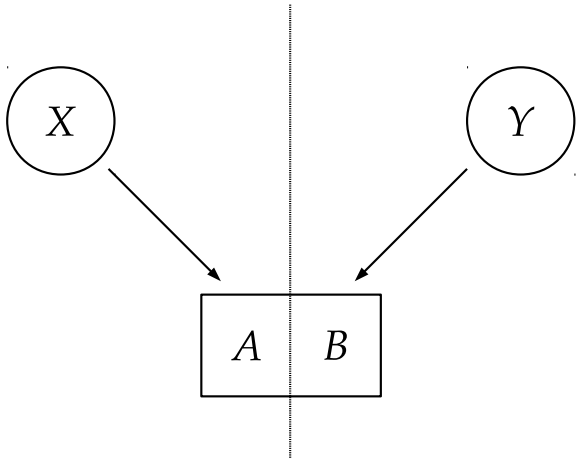


Figura 12 – Representação geral de um sistema do cenário CHSH correlacionado de forma não-sinalizante.

O conjunto de correlações não-sinalizantes no cenário CHSH é um subconjunto de $\mathcal{B}(2,2,2)$, dado que as correlações não-sinalizantes satisfazem as condições que definem o conjunto $\mathcal{B}(2,2,2)$. O conjunto de correlações não-sinalizantes será denotado por $\mathcal{NS}(2,2,2)$,

O conjunto $\mathcal{NS}(2,2,2)$ contém pontos extremais que não são pontos determinísticos, gerados pela condição de não-sinalização. Tais pontos são as denominadas *Caixas PR*. O termo "caixa PR" surgiu devido ao trabalho de Sandu Popescu e Daniel Rohrlich [18], onde foram apresentadas pela primeira vez. Essas caixas são muito importantes no estudo de Não-localidade, e uma delas será utilizada no próximo capítulo.

As caixas PR são as seguintes:

$$\mathbf{P_{PR1}} = \frac{1}{2}\begin{bmatrix} 1&1&1&0\\ 0&0&0&1\\ 0&0&0&1\\ 1&1&1&0 \end{bmatrix}, \quad \mathbf{P_{PR2}} = \frac{1}{2}\begin{bmatrix} 1&1&0&1\\ 0&0&1&0\\ 0&0&1&0\\ 1&1&0&1 \end{bmatrix}, \quad \mathbf{P_{PR3}} = \frac{1}{2}\begin{bmatrix} 1&0&1&1\\ 0&1&0&0\\ 0&1&0&0\\ 1&0&1&1 \end{bmatrix}, \quad \mathbf{P_{PR4}} = \frac{1}{2}\begin{bmatrix} 0&1&1&1\\ 1&0&0&0\\ 1&0&0&0\\ 0&1&1&1 \end{bmatrix}$$
$$\mathbf{P_{PR5}} = \frac{1}{2}\begin{bmatrix} 0&0&0&1\\ 1&1&1&0\\ 1&1&1&0\\ 0&0&0&1 \end{bmatrix}, \quad \mathbf{P_{PR6}} = \frac{1}{2}\begin{bmatrix} 0&0&1&0\\ 1&1&0&1\\ 1&1&0&1\\ 0&0&1&0 \end{bmatrix}, \quad \mathbf{P_{PR7}} = \frac{1}{2}\begin{bmatrix} 0&1&0&0\\ 1&0&1&1\\ 1&0&1&1\\ 0&1&0&0 \end{bmatrix}, \quad \mathbf{P_{PR8}} = \frac{1}{2}\begin{bmatrix} 1&0&0&0\\ 0&1&1&1\\ 0&1&1&1\\ 1&0&0&0 \end{bmatrix} \tag{3.14}$$

As caixas PR, apesar de serem pontos extremais do conjunto $\mathcal{NS}(2,2,2)$, não são pontos extremais do conjunto $\mathcal{B}(2,2,2)$ pois podem ser obtidas por combinação convexa de outros pontos do politopo $\mathcal{B}(2,2,2)$.

O conjunto de correlações não-sinalizantes é também um politopo convexo, assim como o conjunto $\mathcal{B}(2,2,2)$. Seus vértices são as oito caixas PR listadas acima e 16 dos 256 vértices de $\mathcal{B}(2,2,2)$. Esses 16 vértices de $\mathcal{B}(2,2,2)$ são os únicos pontos determinísticos que respeitam a condição de não-sinalização expressa pelas eq. (3.13). Nas refs. [19, 20] os autores apresentam justificativas para tais afirmações. Abaixo estão listadas as matrizes que representam tais vértices.

$$\mathbf{P_{L1}} = \begin{bmatrix} 1&1&1&1\\ 0&0&0&0\\ 0&0&0&0\\ 0&0&0&0 \end{bmatrix}, \quad \mathbf{P_{L2}} = \begin{bmatrix} 1&0&1&0\\ 0&1&0&1\\ 0&0&0&0\\ 0&0&0&0 \end{bmatrix}, \quad \mathbf{P_{L3}} = \begin{bmatrix} 0&1&0&1\\ 1&0&1&0\\ 0&0&0&0\\ 0&0&0&0 \end{bmatrix}, \quad \mathbf{P_{L4}} = \begin{bmatrix} 0&0&0&0\\ 1&1&1&1\\ 0&0&0&0\\ 0&0&0&0 \end{bmatrix},$$
$$\mathbf{P_{L5}} = \begin{bmatrix} 1&1&0&0\\ 0&0&0&0\\ 0&0&1&1\\ 0&0&0&0 \end{bmatrix}, \quad \mathbf{P_{L6}} = \begin{bmatrix} 1&0&0&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&0&0&1 \end{bmatrix}, \quad \mathbf{P_{L7}} = \begin{bmatrix} 0&1&0&0\\ 1&0&0&0\\ 0&0&0&1\\ 0&0&1&0 \end{bmatrix}, \quad \mathbf{P_{L8}} = \begin{bmatrix} 0&0&0&0\\ 1&1&0&0\\ 0&0&0&0\\ 0&0&1&1 \end{bmatrix},$$
$$\mathbf{P_{L9}} = \begin{bmatrix} 0&0&1&1\\ 0&0&0&0\\ 1&1&0&0\\ 0&0&0&0 \end{bmatrix}, \quad \mathbf{P_{L10}} = \begin{bmatrix} 0&0&1&0\\ 0&0&0&1\\ 1&0&0&0\\ 0&1&0&0 \end{bmatrix}, \quad \mathbf{P_{L11}} = \begin{bmatrix} 0&0&0&1\\ 0&0&1&0\\ 0&1&0&0\\ 1&0&0&0 \end{bmatrix}, \quad \mathbf{P_{L12}} = \begin{bmatrix} 0&0&0&0\\ 0&0&1&1\\ 0&0&0&0\\ 1&1&0&0 \end{bmatrix},$$
$$\mathbf{P_{L13}} = \begin{bmatrix} 0&0&0&0\\ 0&0&0&0\\ 1&1&1&1\\ 0&0&0&0 \end{bmatrix}, \quad \mathbf{P_{L14}} = \begin{bmatrix} 0&0&0&0\\ 0&0&0&0\\ 1&0&1&0\\ 0&1&0&1 \end{bmatrix}, \quad \mathbf{P_{L15}} = \begin{bmatrix} 0&0&0&0\\ 0&0&0&0\\ 0&1&0&1\\ 1&0&1&0 \end{bmatrix}, \quad \mathbf{P_{L16}} = \begin{bmatrix} 0&0&0&0\\ 0&0&0&0\\ 0&0&0&0\\ 1&1&1&1 \end{bmatrix}, \tag{3.15}$$

Em resumo, as caixas determinísticas listadas acima e as caixas PR são os pontos extremais do conjunto $\mathcal{NS}(2,2,2)$. Tal conjunto é o fecho convexo desses pontos, assim, $\mathcal{NS}(2,2,2)$ é também um politopo.

3.3.2 Correlações Locais

Definição 3.3.2 (Variável Oculta Local)**.** Uma variável oculta Λ é *local* se as medições sobre cada parte não exercem influência sobre Λ. Essa hipótese é representada por

$$p(\lambda|x,y) = p(\lambda),\ \forall\lambda \in \Lambda. \tag{3.16}$$

Definição 3.3.3 (Correlação Local)**.** Uma caixa $\mathbf{P_{AB}} \in \mathcal{NS}(2,2,2)$ é *local* se existe uma variável oculta local Λ que torna possível escrever todos os elementos $p(a,b|x,y) \in \mathbf{P_{AB}}$ como

$$p(a,b|x,y) = \sum_{\lambda} p(\lambda)p(a|x,\lambda)p(b|y,\lambda),\ \forall\, a,b \in A,B. \tag{3.17}$$

O conjunto de correlações locais é um subconjunto do conjunto não-sinalizante, dado que a eq. (3.17) representa uma nova restrição sobre o conjunto $\mathcal{NS}(2,2,2)$. Tal conjunto será denotado por $\mathcal{L}(2,2,2)$. O conjunto $\mathcal{L}(2,2,2)$ é o fecho convexo das caixas determinísticas $\mathbf{P}_{\mathbf{L}i}$, assim, $\mathcal{L}(2,2,2)$ é também um politopo. Com o conjunto de correlações locais definido, chamaremos de correlações *não-locais* todas as que pertencerem a região $\mathcal{NS}(2,2,2) \cap \mathcal{L}^c(2,2,2)$, ou seja, todas as correlações não-sinalizantes que não pertençam ao conjunto de correlações locais.

Teorema 3.3.1. *Uma caixa* $\mathbf{P_{ab}}$ *é local se, e somente se, pode ser escrita como*

$$\mathbf{P_{AB}} = \sum_{\lambda} p(\lambda)\mathbf{P_{A|\Lambda}} \otimes \mathbf{P_{B|\Lambda}}, \tag{3.18}$$

sendo $\mathbf{P_{A|\Lambda}}$ *e* $\mathbf{P_{B|\Lambda}}$ *as caixas simples correspondentes à Alice e Bob, respectivamente, condicionadas à* Λ*.*

Demonstração. Cada termo da caixa $\mathbf{P_{AB}}$ que satisfaz a eq. (3.18) é escrito como

$$p(a,b|x,y) = \sum_{\lambda} p(\lambda)p(a|x,\lambda)p(b|y,\lambda), \tag{3.19}$$

a expressão acima é exatamente a condição da definição 3.3.3, sendo portanto, $\mathbf{P_{AB}}$ uma caixa local. □

O DAG abaixo ilustra um sistema representado por uma caixa local $\mathbf{P_{AB}} \in \mathcal{L}(2,2,2)$.

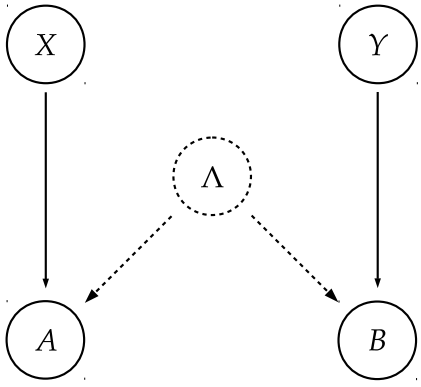


Figura 13 – DAG que representa um sistema correlacionado localmente.

O fato de que nem todas as correlações não-sinalizantes são locais implica que não é possível justificar a aleatoriedade de caixas não-locais como desconhecimento sobre uma variável oculta responsável por correlacionar as partes da caixa. Isso implica que existe aleatoriedade intrínseca às caixas.

3.3.3 Correlações Quânticas

A Mecânica Quântica é uma teoria baseada em alguns postulados. Esses postulados podem ser estudados nas refs. [14, cap. 2], [21, cap. 3]. Como nem todos os postulados da mecânica quântica serão utilizados nesse contexto, a teoria quântica será definida de maneira resumida, de modo a incluir os postulados de interesse para o estudo de correlações no cenário CHSH.

Definição 3.3.4 (Teoria Quântica)**.** Um estado quântico é representado por um operador linear $\rho : \mathcal{H}_n \to \mathcal{H}_n$, sendo $\mathcal{H}_n$ um espaço de Hilbert de dimensão n. O operador ρ satisfaz as seguintes propriedades:

(i) $\langle \mathbf{v}, \rho \mathbf{v} \rangle \geq 0, \ \forall \ \mathbf{v} \in \mathcal{H}_n$,

(ii) $\mathrm{tr}\rho = 1$.

Uma medição quântica $\mathcal{M}$ é representada por um conjunto de operadores Π_j positivos semidefinidos, ou seja, que satisfazem a propriedade (i), cuja soma sobre todos os operadores $\Pi_j \in \mathcal{M}$ deve satisfazer a equação

$$\sum_j \Pi_j = \mathbb{1}, \tag{3.20a}$$

em que $\mathbb{1}$ é o operador identidade.

A probabilidade de se obter o resultado r_j, correspondente ao elemento $\Pi_j \in \mathcal{M}$, após a realização da medição $\mathcal{M}$ sobre o estado quântico ρ é dada pela *Regra de Born*:

$$p(r_j|\mathcal{M}) = \mathrm{tr}\left[\rho \Pi_j\right]. \tag{3.20b}$$

Observação. Como frequentemente nos referiremos a várias medições, adicionaremos um índice às medições para identificá-las. Assim, uma medição será representada por $\mathcal{M}_i$ e seus respectivos operadores serão Π^i_j.

A def. 3.3.4 é geral, no sentido que o estado quântico representado por ρ pode representar um sistema composto por várias partes. As medições realizadas sobre cada parte individualmente atuarão nos subespaços de $\mathcal{H}_n$ correspondentes a cada parte. Como o nosso objetivo é o estudo das correlações quânticas no cenário CHSH, definiremos uma caixa quântica bipartida especificamente para tal cenário. Uma definição para caixa bipartida em cenários mais gerais pode ser encontrada na ref. [13, cap. 1].

Definição 3.3.5 (Caixa Quântica Bipartida)**.** Uma caixa bipartida é dita *quântica* se todo termo $p(a, b|x, y)$ da caixa satisfaz

$$p(a, b|x, y) = \mathrm{tr}\left[\rho_{AB}\left(\Pi^x_a \otimes \Pi^y_b\right)\right], \ \forall \ a \in A, \ b \in B, \ x \in X, \ y \in Y, \tag{3.21}$$

e se existirem estados quânticos $\rho_{AB} : \mathcal{H}_n \otimes \mathcal{H}_n \to \mathcal{H}_n \otimes \mathcal{H}_n$ e medições conjuntas com elementos $\Pi_a^x \otimes \Pi_b^y$, tais que a eq. (3.21) seja satisfeita.

Uma medição da forma $\{\Pi_a^x \otimes \mathbb{1}_B\}$ representa uma medição realizada apenas na parte da Alice, o mesmo pode ser dito para Bob com uma medição do tipo $\{\mathbb{1}_A \otimes \Pi_b^y\}$. Medições deste tipo são denominadas *medições locais*, tais medições realizadas sobre uma parte não influenciam a probabilidade do resultado de uma medição realizada sobre a outra parte. O conjunto de correlações quânticas no cenário CHSH será denotado por $\mathcal{Q}(2,2,2)$.

O conjunto $\mathcal{Q}(2,2,2)$ é convexo. Na ref. [13, pág. 27] encontra-se um teorema que diz que os conjuntos de correlações quânticas entre duas partes com número arbitrário de medições e resultados sobre cada parte é convexo. Para que tal resultado seja verdadeiro, é explorada a liberdade dimensional do espaço de Hilbert onde os operadores densidade atuam[4].

Teorema 3.3.2. *Uma caixa quântica bipartida* $\mathbf{P_{AB}}$ *é não-sinalizante.*

Demonstração. Considere uma caixa quântica $\mathbf{P_{AB}} \in \mathcal{Q}(2,2,2)$. Uma probabilidade marginal $p(a|x,y_0)$ pode ser obtida através de um elemento $p(a,b|x,y) \in \mathbf{P_{AB}}$ através de

$$
\begin{aligned}
p(a|x,y_0) &= \sum_b p(a,b|x,y_0) && \text{(3.22a)}\\
&= \sum_b \operatorname{tr}\left[\rho_{AB}\left(\Pi_a^x \otimes \Pi_b^{y_0}\right)\right] && \text{(3.22b)}\\
&= \operatorname{tr}\left[\rho_{AB}\left(\Pi_a^x \otimes \sum_b \Pi_b^{y_0}\right)\right] && \text{(3.22c)}\\
&= \operatorname{tr}\left[\rho_{AB}\left(\Pi_a^x \otimes \mathbb{1}_B\right)\right] && \text{(3.22d)}\\
&= p(a|x). && \text{(3.22e)}
\end{aligned}
$$

□

Lema 3.3.3. *Qualquer distribuição de probabilidade P pode ser simulada pela realização de medições sobre sistemas quânticos.*

Demonstração. Considere uma variável aleatória $A = \{a_1, \ldots, a_n\}$ representando os possíveis resultados de uma medição $\mathcal{M}$, sobre a qual existe uma distribuição de probabilidade $p(a)$. Definindo um estado quântico $\rho : \mathcal{H}_n \to \mathcal{H}_n$ e a medição $\mathcal{M} = \{\Pi_a\}$, $\Pi_a : \mathcal{H}_n \to \mathcal{H}_n$, em que os operadores Π_a sejam ortogonais entre si e cada Π_a

[4] De acordo com a ref. [22], o conjunto de caixas quânticas pode não ser convexo ao fixar a dimensão desse espaço de Hilbert.

seja associado ao resultado a após a realização da medição $\mathcal{M}$. Um estado quântico ρ que simula esta distribuição de probabilidade será

$$\rho = \sum_a p(a)\Pi_a. \tag{3.23a}$$

Recupera-se a probabilidade de obtenção de um resultado a após a medição $\mathcal{M}$ sobre o sistema quântico através de

$$p(a) = \text{tr}\left[\rho\Pi_a\right]. \tag{3.23b}$$

□

Teorema 3.3.4. $\mathcal{L}(2,2,2) \subseteq Q(2,2,2)$.

Demonstração. Suponhamos um estado quântico $\rho_{AB} : (\mathcal{H}_2^{x_0} \otimes \mathcal{H}_2^{x_1})_A \otimes (\mathcal{H}_2^{y_0} \otimes \mathcal{H}_2^{y_1})_B \to (\mathcal{H}_2^{x_0} \otimes \mathcal{H}_2^{x_1})_A \otimes (\mathcal{H}_2^{y_0} \otimes \mathcal{H}_2^{y_1})_B$, em que cada espaço de Hilbert esteja associado a uma medição no cenário (2,2,2). Supondo, também, que todas as medições sejam projetivas em seus subespaços correspondentes e possam ser escritas como $\mathcal{M} = \{|0\rangle\langle 0|, |1\rangle\langle 1|\}$, serão utilizados estados quânticos preparados em autoestados de todas as medições simultaneamente.

Utilizando, em particular, o estado quântico preparado no autoestado $|0\rangle$ de todas as possíveis medições, temos

$$\rho_{AB} = (|00\rangle\langle 00|)_A \otimes (|00\rangle\langle 00|)_B. \tag{3.24a}$$

Ao utilizar a regra de Born para obter a caixa bipartida quântica correspondente a esse estado para essa escolha de medições, obtém-se a caixa *local* $\mathbf{P_{L1}}$ encontrada na eq. (3.15).

Se explorarmos todas as possibilidades de troca de autoestados $|0\rangle$ por autoestados $|1\rangle$ na eq. (3.24a), obtemos, ao utilizar a regra de Born, estados que geram todas as caixas extremais do conjunto $\mathcal{L}(2,2,2)$, listadas na eq. (3.15). Como, de acordo com a ref. [13, pág. 27], o conjunto $\mathcal{Q}(2,2,2)$ é convexo e contém todos os pontos extremais de $\mathcal{L}(2,2,2)$, também convexo, concluímos que $\mathcal{L}(2,2,2) \subseteq \mathcal{Q}(2,2,2)$.

□

Mais especificamente, o conjunto de correlações locais é estritamente contido no conjunto de correlações quânticas, ou seja, existem correlações quânticas não-locais. Esse é um famoso resultado, conhecido como teorema de Bell, e pode ser encontrado na ref. [2]. O fato de que o conjunto de correlações locais esteja estritamente contido no conjunto de correlações quânticas implica que a teoria quântica permite a existência de correlações não-locais. Como foi discutido no final da seção 3.3.2, correlações não-locais são intrinsecamente aleatórias, ou seja, correlações quânticas podem ser intrinsecamente

aleatórias. Tais correlações não são apenas previstas pela teoria, como podem ser encontradas na natureza, e isso nos leva a concluir que existe aleatoriedade intrínseca na natureza.

De todos os conjuntos aqui introduzidos, $\mathcal{Q}(2,2,2)$ é o único que não foi perfeitamente caracterizado até então pois não é um politopo. O conjunto $\mathcal{Q}(2,2,2)$ possui um número infinito de pontos extremais, o que torna a sua caracterização mais difícil que a caracterização de politopos.

Alguns trabalhos envolvendo caracterizações do conjunto quântico são os de Boris Tsirelson [23], L. J. Landau [24] e Ll. Masanes [25].

Utilizaremos aqui o resultado de um trabalho mais recente, desenvolvido por Miguel Navascués, Stefano Pironio e Antonio Acín, encontrado na ref. [26]. Esse trabalho introduz uma hierarquia de programas semidefinidos que aproximam o conjunto quântico $\mathcal{Q}$ por conjuntos $\mathcal{Q}_i$, em que cada $\mathcal{Q}_i$ é um conjunto dessa hierarquia. Essa hierarquia é conhecida como *hierarquia NPA*, levando esse nome devido aos seus autores.

A hierarquia NPA pode ser ilustrada pela figura a seguir:

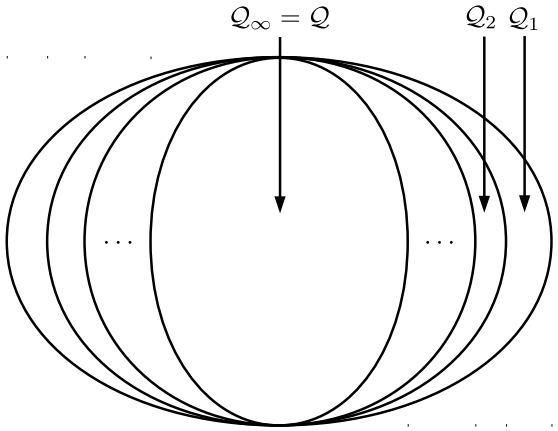


Figura 14 – Ilustração da hierarquia NPA

O pertencimento de uma caixa a um conjunto $\mathcal{Q}_i$ é uma condição necessária, mas não suficiente, para que essa seja uma caixa quântica. Entretanto, esse mesmo trabalho prova que a hieraquia converge para o conjunto de correlações quânticas, de acordo com

$$\lim_{i\to\infty} Q_i = Q. \tag{3.25}$$

Observação. Não serão introduzidos detalhes sobre esse trabalho aqui, uma introdução à hierarquia NPA um pouco mais detalhada pode ser encontrada na ref. [27, p.73].

Neste trabalho, aproximaremos o conjunto quântico $\mathcal{Q}$ pelo primeiro conjunto da hierarquia, o conjunto $\mathcal{Q}_1$. Antes de escrever a expressão correspondente a $\mathcal{Q}_1$, vamos, primeiramente, definir alguns parâmetros.

Definição 3.3.6 (Correlator)**.** Para uma caixa $\mathbf{P_{AB}}$, um correlator C_{xy} correspondente a uma das escolhas de medição conjunta xy é definido como

$$C_{xy} = \langle \mathcal{M}_x \mathcal{M}_y \rangle = \sum_{a=b} p(a,b|x,y) - \sum_{a\neq b} p(a,b|x,y). \tag{3.26}$$

Definição 3.3.7. Para uma caixa $\mathbf{P_{AB}}$, C_x correspondente à escolha de medição $\mathcal{M}_x$ pela Alice é definido como

$$C_x = \langle \mathcal{M}_x \rangle = \sum_b \left[p(0,b|x,y) - p(1,b|x,y) \right], \tag{3.27}$$

sendo y uma escolha qualquer de medição fixa sobre a parte de Bob.

Observação. Na eq. (3.27), a escolha de medição sobre a parte do Bob só é irrelevante pois nossa atenção é voltada para caixas pertencentes a $\mathcal{NS}(2,2,2)$. Se esse não fosse o caso, C_x não seria invariante sob a escolha de medição de Bob.

Analogamente, C_y correspondente à escolha de medição $\mathcal{M}_y$ por Bob é

$$C_y = \sum_a \left[p(a,0|x,y) - p(a,1|x,y) \right]. \tag{3.28}$$

Podemos agora escrever a expressão correspondente ao conjunto $\mathcal{Q}_1$.

Definição 3.3.8. O conjunto $\mathcal{Q}_1$ é o conjunto de caixas $\mathbf{P_{AB}}$ que satisfaz a condição

$$|\arcsin D_{00} + \arcsin D_{01} + \arcsin D_{10} - \arcsin D_{11}| \leq \pi, \tag{3.29}$$

em que $D_{xy} = \frac{\left(C_{xy} - C_x C_y\right)}{\sqrt{(1-C_x^2)(1-C_y^2)}}$.

3.3.4 Hierarquia das Correlações

Agora que relacionamos todos os conjuntos de interesse no cenário CHSH, podemos expressar o ordenamento entre os conjuntos de correlações:

$$\mathcal{L}(2,2,2) \subset \mathcal{Q}(2,2,2) \subset \mathcal{NS}(2,2,2) \subset \mathcal{B}(2,2,2). \tag{3.30}$$

A imagem a seguir é uma ilustração bidimensional dos conjuntos $\mathcal{NS}$, $\mathcal{Q}$ e $\mathcal{L}$:

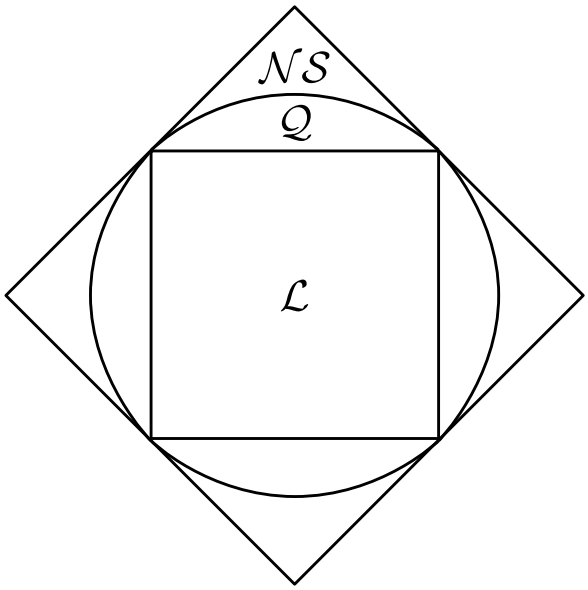


Figura 15 – Ilustração bidimensional do politopo $\mathcal{NS}(2,2,2)$ com os conjuntos $\mathcal{Q}(2,2,2)$ e $\mathcal{L}(2,2,2)$.

4 O princípio de Causalidade da Informação

No capítulo anterior, diferentes tipos de correlação no cenário CHSH foram apresentados. Sempre assumindo hipótese de livre-escolha, o conjunto de correlações não-sinalizantes contém caixas em que qualquer probabilidade de resultado de uma parte para sua dada escolha de medição não depende da escolha de medição relativa à outra parte. O conjunto de correlações locais, por sua vez, é aquele em que as partes de uma caixa são correlacionadas através de uma variável oculta, ou seja, um fator que influencia a probabilidade de obtenção de cada resultado, mas que não influencia as escolhas de medição de cada parte, visto na definição 3.3.3. Já o conjunto de correlações quânticas é definido por caixas que satisfazem a regra de Born, eq. (3.21), ou seja, a maneira com que as probabilidades de resultados para um dado conjunto de medições são extraídas de um operador densidade que represente um estado quântico.

O fato de que o conjunto de correlações quânticas contenha caixas não-locais é uma das estranhezas da natureza que a mecânica quântica veio a revelar. Tentar compreender a razão pela qual a natureza exibe não-localidade quântica, e nenhum outro tipo de não-localidade, é algo que escapa do alcance da interpretação da regra de Born. Por esse motivo, alguns princípios físicos foram sugeridos nos últimos anos na tentativa de justificar a não-localidade quântica. Alguns desses princípios são a Não-trivialidade da Complexidade de Comunicação, a Localidade Macroscópica, a Ortogonalidade Local e a Causalidade da Informação. Tais princípios podem ser encontrados respectivamente nas refs. [28, 29, 30, 31].

Neste trabalho, o princípio de Causalidade da Informação será estudado, desde a sua primeira versão, encontrada na ref. [31], até um resultado mais recente, encontrado na ref. [32]. Por fim, serão apresentados resultados parcialmente obtidos nesse trabalho de mestrado, com perspectivas de continuidade em um futuro breve.

Algumas referências interessantes para o leitor que deseja conhecer um pouco da pesquisa atual relacionada a fundamentos de mecânica quântica envolvendo teoria de inferência causal são as refs. [33, 34, 35, 36, 37, 38, 39].

4.1 O Cenário de Causalidade da Informação

Considere o seguinte cenário: Alice possui uma sequência de n bits, em que cada bit é representado por uma variável aleatória de alfabeto $Z_i = \{0, 1\}$. Sua sequência de bits, também chamada pelo termo em inglês: *bitstring*, será representada por

$$\mathbf{Z} = (Z_0, \ldots, Z_{n-1}). \tag{4.1}$$

Além disso, Alice e Bob compartilham uma caixa $\mathbf{P_{AB}}$ não-local, a qual será referida como *recurso não-local*, com a qual pode realizar alguma tarefa de comunicação com Bob que ela desejar.

Nesse cenário, o *bitstring* $\mathbf{Z}$ contém uma informação que Alice gostaria de compartilhar com Bob. Dispondo, além do recurso não-local, de um canal de comunicação clássico limitado, Alice é capaz de enviar apenas um número d de bits para Bob, em que $d < n$. Neste cenário, o princípio de Causalidade da Informação foi definido como:

Definição 4.1.1 (Causalidade da Informação)**.** O ganho de informação que Bob pode ter sobre a sequência $\mathbf{Z}$ de bits da Alice inicialmente desconhecida por ele, utilizando todos os seus recursos não-sinalizantes e uma mensagem contendo d bits enviada a ele por Alice, é de, no máximo, d bits.

Utilizando I para denotar a quantidade de informação obtida por Bob nesse cenário, a expressão abaixo resume o princípio de Causalidade da Informação.

$$I \leq d. \tag{4.2}$$

4.2 Códigos de Acesso Aleatório

Existe uma classe de protocolos, denominada *códigos de acesso aleatório*[1], que consiste de protocolos em que Alice utiliza alguma estratégia para codificar seus n bits em uma mensagem contendo d bits, $d < n$, com a intenção de que Bob consiga adquirir o máximo de informação contida nos n bits da melhor maneira possível.

Neste trabalho, estudaremos o princípio de causalidade da informação utilizando um protocolo específico. Um estudo mais geral é possível de ser realizado se todos os protocolos possíveis de serem utilizados no cenário de causalidade da informação forem testados. Essa é uma das perspectivas futuras a esse trabalho.

4.2.1 Um protocolo especial

A partir de agora, um protocolo específico da classe de códigos de acesso aleatório, que motivou o princípio de Causalidade da Informação, será adotado. Este protocolo é muito especial, pois faz com que Bob consiga ter acesso a, pelo menos, d bits codificados na mensagem. O protocolo que utilizaremos envolve uma mensagem com $d = 1$ bit. Em cada execução do protocolo, Bob escolhe um dos n bits que deseja conhecer. Tal escolha é realizada através de uma associação entre sua escolha y de medição Y e o bit desejado z_i. Isso ficará mais evidente posteriormente.

[1] do inglês *random acess codes*, muito representado na literatura apenas pela sigla "RAC".

Por exemplo, se Alice possui um *bitstring* $\mathbf{Z} = (z_0, \ldots, z_{n-1})$ e envia a Bob uma mensagem $\mathbf{M} = (m_0, \ldots, m_{d-1})$, $m_i = z_{f(i)}$ em que $\{f(i)\}$ é o conjunto de posições desejadas dos bits no *bitstring* $\mathbf{Z}$, $|\mathbf{M}| = d$, o valor da adivinhação[2] g_i de Bob sobre o bit z_i será

$$g_i = z_i \tag{4.3}$$

para todo bit z_i contido na mensagem $\mathbf{M}$. Para os bits que não estiverem contidos na mensagem, há uma probabilidade não-determinística $p(g_i = z_i|y_{f(i)})$ de que o bit z_i seja adivinhado com sucesso por Bob, condicionado à uma escolha adequada de y de acordo com o bit desejado. Essa probabilidade depende do recurso não-local utilizado e não entraremos em detalhes sobre como obtê-las pois não será necessário. A eficiência do protocolo pode ser quantificada como

$$E = \sum_{i,y} I(Z_i; G_i|y_{f(i)}). \tag{4.4}$$

Observação. Como todas as variáveis são binárias, quantificaremos a informação utilizando a base 2 para o logaritmo da informação mútua $I(Z_i; G_i|y_{f(i)})$ e para todos os quantificadores de informação de Shannon envolvidos neste capítulo.

Observação. Como cada escolha de y feita por Bob representa uma escolha de um *bit* que Bob deseja conhecer, o termo $y_{f(i)}$ será substituído por y para facilitar a notação.

O objetivo é escrever a expressão (4.2) em termos de quantificadores de informação de Shannon que facilitem o cálculo. Substituiremos, a princípio, a grandeza I da expressão (4.2) pela eficiência E do protocolo. Essa substituição será justificada após a obtenção final do critério para o princípio de Causalidade da Informação.

Retornando à eq. (4.4), cada termo $I(Z_i; G_i|y)$ obedece a relação

$$I(Z_i; G_i|y) \leq 1, \tag{4.5a}$$

com igualdade apenas se $p(z_i = g_i|y) = 1 \; \forall \; i$. Assim, o somatório da eq. (4.4) satisfaz

$$\sum_i I(Z_i; G_i|y) \leq d. \tag{4.5b}$$

Cada informação mútua $I(Z_i; G_i|y)$ pode ser reescrita como $1 - H(Z_i = G_i|G_i, y)$, em que o termo $H(Z_i = G_i|G_i, y)$ envolve a probabilidade $p(z_i = g_i|g_i, y)$, que nada mais é do que a probabilidade $p(z_i|g_i, y)$. De acordo com o teorema 1.4.3, ao remover o condicionamento em G_i dessa entropia, obtemos a relação

$$H(Z_i = G_i|G_i, y) \leq H(Z_i = G_i|y). \tag{4.5c}$$

[2] O símbolo "g" que representa a adivinhação de Bob foi escolhido devido ao termo em inglês: *guess*.

Assim,

$$I(Z_i; G_i|y) = 1 - H(Z_i = G_i|G_i, y) \geq 1 - H(Z_i = G_i|y). \tag{4.5d}$$

Isso nos permite reescrever a expressão (4.5b) como

$$N - \sum_i H(Z_i = G_i|y) \leq d. \tag{4.5e}$$

No protocolo que será introduzido na seção à seguir, a igualdade na eq. (4.5b) será sempre satisfeita por recursos locais e quânticos. Nosso interesse será voltado para os recursos que violam a desigualdade (4.5e).

A grandeza I na eq. (4.2) que quantifica o ganho de informação de Bob no processo, idealmente, deveria ser a informação mútua

$$I = I(Z_0, \ldots, Z_{n-1}; \mathbf{M}, \mathbf{P_B}), \tag{4.6}$$

em que $\mathbf{P_B}$ representa a parte de Bob do recurso não-local $\mathbf{P_{AB}}$. Nesse caso, $\mathbf{P_B}$ não é uma grandeza[3] que pode ser tratada sempre através da abordagem independente de dispositivos, não sendo possível, portanto, utilizar a informação mútua da eq. (4.6) como critério para o princípio de Causalidade da Informação.

Por este motivo, desejamos substituir a grandeza I por um objeto que seja independente de teoria. É possível mostrar que se uma informação mútua é definida de modo a obedecer certas propriedades, então,

- O princípio de Causalidade da Informação é válido;
- A grandeza I pode ser cotada superiormente por E, que é independente de teoria, ou seja

$$I \leq E. \tag{4.7}$$

No material suplementar da ref. [31] encontra-se a demonstração de tal afirmação. As relações (4.7) e (4.5d) permitem que o critério seja reescrito como

$$1 - \sum_i H(Z_i = G_i|y) \leq d. \tag{4.8}$$

4.3 Protocolo - Uma Caixa

Considerando, $n = 2$ e $d = 1$, o *bitstring* da Alice será $\mathbf{Z} = (Z_0, Z_1)$, a mensagem enviada para Bob conterá apenas 1 bit. O protocolo que será utilizado é descrito da seguinte maneira:

[3] Apesar dessa notação ter sido adotada, o caso em que $\mathbf{P_B}$ é uma caixa que represente o sistema de Bob é apenas um caso particular do que $\mathbf{P_B}$ pode representar na equação (4.6).

(1) Alice faz, em sua parte, a medição correspondente à soma módulo 2 de seus bits: $x = z_0 \oplus z_1$;

(2) Após obter o resultado a para a medição x, Alice envia para Bob a mensagem correspondente a soma módulo 2 de um de seus bits com o resultado a obtido: $m = z_0 \oplus a$;

(3) Bob realiza uma medição y em sua caixa, obtendo um resultado b;

(4) Após a obtenção do resultado b e o recebimento da mensagem m, o valor g_i, relativo ao i-ésimo bit de $\mathbf{Z}$ que Bob escolheu, é $g_i = m \oplus b$.

Uma ilustração do protocolo pode ser vista na figura a seguir:

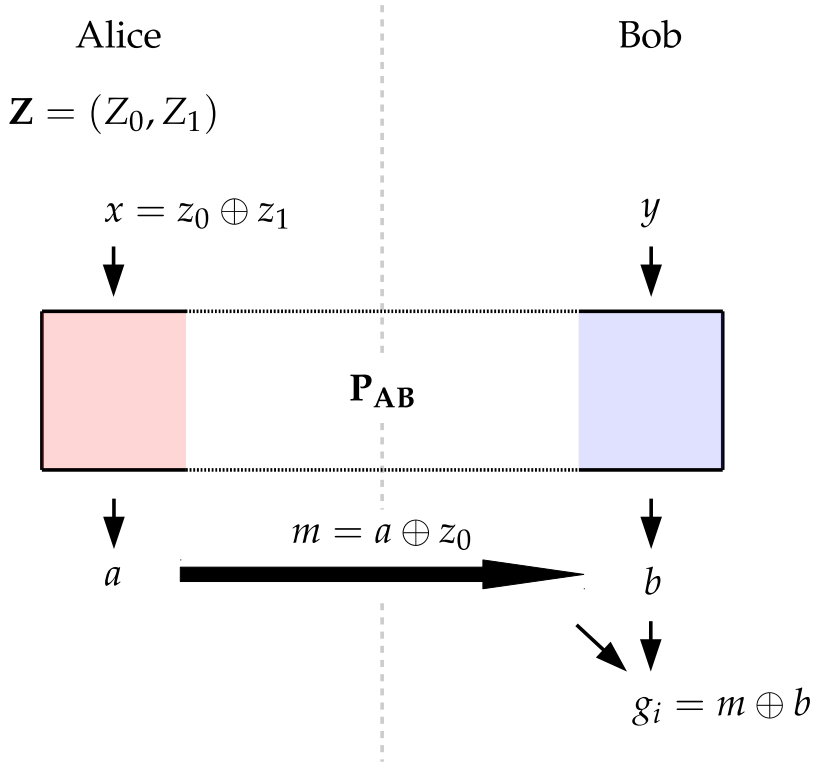


Figura 16 – Ilustração do código de acesso aleatório. g_i corresponde à tentativa de adivinhação de Bob sobre o i-ésimo bit do *bitstring* $\mathbf{Z}$ da Alice.

A escolha feita por Bob sobre o bit de $\mathbf{Z}$ que deseja descobrir é feita através da associação entre a escolha de medição $y \in \{0, 1\}$ e cada bit em $\mathbf{Z}$. Assim, se Bob deseja conhecer o bit z_0, ele escolherá a medição $y = 0$, caso contrário, ele escolherá a medição $y = 1$.

No protocolo, g_i denota a adivinhação de Bob sobre o bit z_i. O objetivo do protocolo é que Bob descubra o valor do bit z_i da Alice, ou seja, Bob terá sucesso se

$$g_i = z_i. \tag{4.9}$$

Vamos assumir que o recurso não-local compartilhado por Alice e Bob seja a caixa $\mathbf{P_{PR1}}$ expressa na seção 3.3.1. Utilizando esse recurso no protocolo, Bob é capaz de sempre acertar o valor do bit desejado, justificando a importância de tal caixa. A caixa $\mathbf{P_{PR1}}$ satisfaz a propriedade de que, para medições X, Y e resultados A, B, todos com alfabeto $\{0,1\}$, a probabilidade conjunta de resultados condicionada às escolhas de medição é sempre

$$p(a,b|x,y) = \begin{cases} 1/2 & \text{se } a \oplus b = x \cdot y, \\ 0 & \text{senão.} \end{cases} \tag{4.10}$$

Essa propriedade é facilmente verificável pela inspeção dos elementos da matriz que representa $\mathbf{P_{PR1}}$:

$a\ b$ \ $x\ y$	00	01	10	11
00	1/2	1/2	1/2	0
01	0	0	0	1/2
10	0	0	0	1/2
11	1/2	1/2	1/2	0

Tabela 4 – Tabela de elementos $p(a,b|x,y)$ referentes à caixa $\mathbf{P_{PR1}}$.

Para que Bob descubra o *bit* z_i, será necessário que ele escolha a medição $y = i$. Vejamos o que ocorre quando a caixa $\mathbf{P_{PR1}}$ é utilizada quando Bob deseja descobrir o *bit* z_i:

$$\begin{aligned} g_i =& m \oplus b \\ =& z_0 \oplus a \oplus b \\ =& z_0 \oplus y \cdot x \\ =& z_0 \oplus y \cdot (z_0 \oplus z_1). \end{aligned} \tag{4.11}$$

(i) Logo, se Bob deseja descobrir z_0:

$$\begin{aligned} g_i = g_0 =& z_0 \oplus 0 \cdot (z_0 \oplus z_1). \\ =& z_0. \end{aligned} \tag{4.12}$$

(ii) Se Bob deseja descobrir z_1:

$$\begin{aligned} g_i = g_1 =& z_0 \oplus 1 \cdot (z_0 \oplus z_1) \\ =& z_0 \oplus z_0 \oplus z_1 \\ =& z_1. \end{aligned} \tag{4.13}$$

Ou seja, se Alice e Bob compartilham uma caixa $\mathbf{P_{PR1}}$, com o envio de 1 bit de mensagem, Bob pode descobrir qualquer bit de $\mathbf{Z}$ em uma rodada do protocolo. Isso quer dizer que o Bob tem acesso aos dois bits de $\mathbf{Z}$. Como a mensagem contém apenas um único bit e é a mesma nas duas situações, o princípio de Causalidade da Informação está sendo violado.

$$I = 2 \not\leq 1. \tag{4.14}$$

O princípio de Causalidade da Informação não é violado pelo conjunto de correlações quânticas. A demonstração dessa afirmação não será feita aqui, entretanto, ela pode ser encontrada na seção IV do material suplementar da ref. [31].

4.4 Protocolo - N Cópias da caixa

O protocolo introduzido na seção anterior pode ser estendido para que seja implementado de forma recursiva, com o objetivo de potencializar o acesso de Bob à informação sobre os bits em $\mathbf{Z}$.

Para a extensão do protocolo, será necessário incluir novos detalhes[4]:

Observação. A partir de agora, a operação "soma módulo 2" aparecerá com muita frequência. Por esse motivo, o símbolo "$\oplus$" será substituído pelo símbolo "$+$". Os somatórios também serão modificados, assim, um somatório envolvendo somas módulo 2 que seria representado por "$\bigoplus_i$" será, então, representado por "$\sum_i$".

1. Execuções realizadas por Alice:

   - As cópias da caixa serão agrupadas em K níveis, representados por $k = 1, \ldots, K$. O primeiro nível será considerado como o que contém uma única cópia da caixa $\mathbf{P_{AB}}$.
   - Cada nível k conterá 2^{k-1} cópias da caixa.
   - Em um nível fixo k, cada cópia da caixa será identificada com um índice $j = 1, \ldots, 2^{k-1}$.
   - As medições e resultados de cada caixa, receberão índices que identificarão a sua caixa correspondente:

$$\begin{aligned} X \to X_j^k;\ Y \to Y_j^k; \\ A \to A_j^k;\ B \to B_j^k. \end{aligned} \tag{4.15}$$

[4] Agradeço à Jessica Bavaresco pela colaboração no desenvolvimento dessa forma do protocolo.

- Alice inicia a execução do protocolo no nível K, concluíndo-a no nível 1. Um parâmetro de recorrência é definido para ser utilizado, dessa forma, no protocolo:

  Definição 4.4.1. O parâmetro de recorrência α_i^k é definido como

  $$\alpha_i^k = \begin{cases} z_{i-1}, & \text{se } k = K, \\ a_{i-1}^{k+1} + \alpha_{2i-1}^{k+1}, & \text{se } k \neq K. \end{cases} \tag{4.16}$$

- As medições X_j^k realizadas por Alice são:

  $$x_j^k = \alpha_{2j-1}^k + \alpha_{2j}^k. \tag{4.17}$$

- A mensagem enviada por Alice para Bob será:

  $$M = a_1^1 + \alpha_1^1. \tag{4.18}$$

2. Execuções realizadas por Bob:

- A organização das cópias da caixa pelo Bob é idêntica à organização das cópias da Alice, com K níveis e 2^{k-1} caixas em cada nível k.
- A execução do protocolo pelo Bob é realizada na ordem inversa à ordem de execução da Alice, sendo iniciada no nível 1 e concluída no nível K.
- Bob não utilizará todas as caixas disponíveis, apenas uma caixa por nível.
- A caixa j que Bob utiliza para fazer a medição no nível $k+1$, é uma função do nível $k+1$. Tal caixa é relacionada com escolha de medição na caixa anterior de acordo com

  $$j_{k+1} = 2j_k + y_j^k + 1. \tag{4.19}$$

- A adivinhação de Bob sobre o bit z_i será representada por uma variável G_i, sendo, a correspondência entre a adivinhação g_i de Bob e o bit z_i da Alice, feita através do índice i. Tal índice é determinado por Bob através de suas escolhas de medições em suas caixas, de acordo com a relação

  $$i = \sum_{k=1}^{K} 2^{k-1}\, y_j^{k-1}, \tag{4.20}$$

  sendo y_j^{k-1} a medição escolhida por Bob realizada na caixa j do nível $k-1$.
- A adivinhação g_i de Bob sobre o bit z_i será

  $$g_i = m + \sum_{k=1}^{K} b_j^k, \tag{4.21}$$

  em que, a partir de $k = 2$, resultado b_j^k corresponde à caixa cujo índice j é determinado pela eq. (4.19).

Vejamos o exemplo de 3 cópias da caixa. Este é o caso mais simples depois do caso de uma única caixa. Depois disso, o processo se tornará mais intuitivo, sendo possível então, ilustrar o caso mais geral.

Exemplo 4.4.1 (Protocolo - 3 cópias)**.** Considere que o *bitstring* da Alice seja $\mathbf{Z} = (Z_0, Z_1, Z_2, Z_3)$, e que Alice e Bob compartilham três cópias de uma caixa $\mathbf{P_{AB}}$. A imagem a seguir ilustra o protocolo nesse caso.

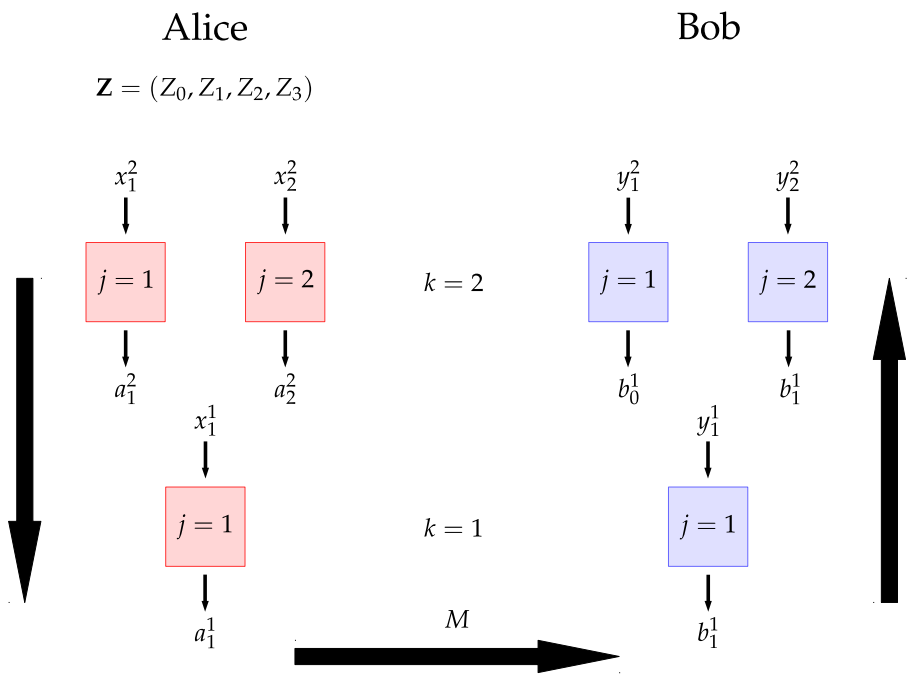


Figura 17 – Protocolo que utiliza três cópias do recurso não-local, as setas longas indicam a ordem de execução do protocolo.

1. Procedimentos executados pela Alice:

   - Parâmetros α_i^k:

     Os valores de α_i^k obtidos através da eq. (4.16) nesse cenário são representados pela tabela abaixo:

i \ k	2	1
1	z_0	$a_1^2 + z_0$
2	z_1	$a_2^2 + z_2$
3	z_2	
4	z_3	

Tabela 5 – Parâmetros α_i^k envolvidos no protocolo com 3 cópias do recurso não-local.

- Medições x_j^k:
  A tabela abaixo lista os valores, obtidos a partir da eq. (4.17), das medições realizadas em cada caixa j de cada nível k:

j \ k	2	1
1	$\alpha_1^2 + \alpha_2^2$	$\alpha_1^1 + \alpha_2^1$
2	$\alpha_3^2 + \alpha_4^2$	

Tabela 6 – Medições x_j^k em função dos parâmetros α_i^k.

- Mensagem:
  De acordo com a eq (4.18), a mensagem enviada pela Alice para Bob é

$$\begin{aligned} m &= a_1^1 + \alpha_1^1 \\ &= a_1^1 + a_1^2 + z_0. \end{aligned} \tag{4.22}$$

2. Procedimentos realizados por Bob:

- A tabela abaixo contém as escolhas de medições que Bob realiza, correspondentes a cada adivinhação G_i sobre o bit Z_i, calculadas de acordo com as eqs. (4.20) e (4.19):

g_i	y_1^1	y_1^2	y_2^2
g_0	0	0	
g_1	0	1	
g_2	1		1
g_3	1		0

Tabela 7 – Medições que Bob deve realizar para escolher realizar uma adivinhação g_i sobre o bit z_i.

- O protocolo é concluído após o cálculo da adivinhação g_i do Bob, de acordo com a eq. (4.21). Os valores de g_i, nesse caso, serão

$$g_0 = m + b_1^1 + b_1^2, \tag{4.23a}$$
$$g_1 = m + b_1^1 + b_1^2, \tag{4.23b}$$
$$g_2 = m + b_1^1 + b_2^2, \tag{4.23c}$$
$$g_3 = m + b_1^1 + b_2^2. \tag{4.23d}$$

Suponha que as caixas de Alice e Bob são cópias de uma caixa $\mathbf{P_{PR1}}$. O cálculo de G_i dependerá das escolhas de medição de Bob para cada bit Z_i desejado, conforme mostra a tabela 7. Vejamos, agora, como calcular os valores g_i para todo i.

As eqs. (4.23) mostram que a computação que Bob faz para acessar os bits da Alice é a mesma para bits que, somados, correspondem à medição x_j^K. Assim, calcularemos g_0 e g_1 separadamente de g_2 e g_3.

- Computando g_0 e g_1:

$$\begin{aligned} g_0 = g_1 &= m + b_1^1 + b_1^2 && \text{(4.24a)}\\ &= z_0 + a_1^1 + a_1^2 + b_1^1 + b_1^2 && \text{(4.24b)}\\ &= z_0 + x_1^1 \cdot y_1^1 + x_1^2 \cdot y_1^2 && \text{(4.24c)}\\ &= z_0 + y_1^1(a_1^2 + \alpha_1^2 + a_2^2 + \alpha_3^2) + y_1^2(\alpha_1^2 + \alpha_2^2) && \text{(4.24d)}\\ &= z_0 + y_1^1(a_1^2 + z_0 + a_2^2 + z_2) + y_1^2(z_0 + z_1). && \text{(4.24e)} \end{aligned}$$

A tentativa de descobrir o bit z_0 e expressa pela realização das medições $y_1^1 = 0$ e $y_1^2 = 0$ por Bob. Substituindo tais valores na eq. (4.24e),

$$g_0 = z_0. \qquad \text{(4.24f)}$$

Para descobrir o bit z_1, Bob realiza as medições $y_1^1 = 0$ e $y_1^2 = 1$, assim,

$$\begin{aligned} g_1 &= z_0 + y_1^1(a_1^2 + z_0 + a_2^2 + z_2) + y_1^2(z_0 + z_1) && \text{(4.25a)}\\ &= \cancel{z_0} + \cancel{z_0} + z_1 && \text{(4.25b)}\\ &= z_1. && \text{(4.25c)} \end{aligned}$$

- Computando g_2 e g_3:

$$\begin{aligned} g_2 = g_3 &= m + b_1^1 + b_2^2 && \text{(4.26a)}\\ &= z_0 + a_1^1 + a_1^2 + b_1^1 + b_2^2 && \text{(4.26b)}\\ &= z_0 + x_1^1 \cdot y_1^1 + a_1^2 + b_2^2 && \text{(4.26c)}\\ &= z_0 + y_1^1(a_1^2 + \alpha_1^2 + a_2^2 + \alpha_3^2) + a_1^2 + b_2^2 && \text{(4.26d)}\\ &= z_0 + y_1^1(a_1^2 + z_0 + a_2^2 + z_2) + a_1^2 + b_2^2. && \text{(4.26e)} \end{aligned}$$

Para escolher entre os bits z_2 e z_3, Bob necessariamente deve realizar, primeiramente, a medição $y_1^1 = 1$. Portanto,

$$\begin{aligned} g_2 = g_3 &= \cancel{z_0} + \cancel{a_1^2} + \cancel{z_0} + a_2^2 + z_2 + \cancel{a_1^2} + b_2^2 && \text{(4.26f)}\\ &= z_2 + a_2^2 + b_2^2 && \text{(4.26g)}\\ &= z_2 + x_2^2 \cdot y_2^2 && \text{(4.26h)}\\ &= z_2 + y_2^2(\alpha_3^2 + \alpha_4^2) && \text{(4.26i)}\\ &= z_2 + y_2^2(z_2 + z_3). && \text{(4.26j)} \end{aligned}$$

Ao tentar descobrir z_2, além de realizar a medição $y_1^1 = 1$, Bob também realiza a medição $y_2^2 = 0$. Portanto,

$$g_2 = z_2. \tag{4.26k}$$

Já ao tentar descobrir z_3, Bob realiza a medição $y_2^2 = 1$, portanto

$$g_3 = \cancel{z_2} + \cancel{z_2} + z_3 \tag{4.26l}$$

$$= z_3. \tag{4.26m}$$

Assim, se Alice e Bob dispõem de três cópias da caixa $\mathbf{P_{PR1}}$, Bob consegue acessar qualquer bit que Alice possua em seu *bitstring*, com o recebimento de uma mensagem de 1 bit fixada.

Neste exemplo, a utilização de 3 cópias de uma caixa PR faz com que a violação do princípio de Causalidade da Informação seja maior que no caso de 1 cópia, sendo

$$I = 4 \not\leq 1. \tag{4.27}$$

Vimos através do exemplo anterior que, se várias cópias de uma caixa PR formam o recurso não-sinalizante que Alice e Bob compartilham, o protocolo descrito permite com que Bob consiga adivinhar corretamente o valor de qualquer bit que Alice possua, com o envio de uma mensagem fixa de tamanho $d = 1$.

Nota. As caixas PR como recurso não-sinalizante se mostraram bastante poderosas na tarefa utilizada para o estudo do princípio de Causalidade da Informação, entretanto, esse é apenas um dos poderes dessas caixas. Na ref. [40], é apresentado o chamado *Protocolo de Van Dam*, que consiste de uma tarefa em que Alice e Bob compartilham uma caixa e desejam utilizá-la para realizar uma tarefa computacional. Essa referência mostra que, utilizando caixas PR como recurso, uma tarefa computacional de complexidade de comunicação arbitrária envolvendo as entradas das caixas se torna trivial. Essa propriedade está relacionada com o princípio de Não-trivialidade da Complexidade de Comunicação, mencionado no início desse capítulo, e apresentado na ref. [28].

4.5 Violações de Causalidade da Informação

Nesta seção, será feita uma discussão contida no material suplementar da ref. [31], a qual possibilita a caracterização de caixas não-locais através do princípio de Causalidade da Informação.

4.5.1 Caixas Não-Sinalizantes e Causalidade da Informação

Como foi visto na seção anterior, a utilização de caixas PR como recurso não-sinalizante para a tarefa determinada faz com que o acesso de Bob a todos os bits da

Alice seja perfeito. Supondo agora que o recurso não seja mais composto por cópias de uma caixa PR, mas cópias de uma caixa não-sinalizante arbitrária, isso faz com que o acesso de Bob aos bits da Alice não seja perfeito, ou seja, a probabilidade de que g_i seja igual a z_i é menor que 1. Definimos um parâmetro r que representa o número de vezes que Bob escolhe realizar a medição $y_j^k = 1$ no protocolo, assim

$$r = \sum_{k=1}^{K} y_j^k, \tag{4.28}$$

onde, novamente, o índice j depende do nível k, representando a caixa j que Bob utiliza no nível k. Definindo um parâmetro l correspondente ao número de vezes que Bob escolhe realizar a medição $y_j^k = 0$, temos que

$$l = K - r, \tag{4.29}$$

sendo n o número de bits que Alice possui. A adivinhação de Bob é computada de acordo com a eq. (4.21), que consiste da soma da mensagem com todos os resultados b_j^k obtidos.

Como cópias de uma caixa não-sinalizante diferente da caixa PR são utilizadas nesse caso, é possível que alguns resultados b_j^k sejam "errados", isto é, são diferentes dos resultados b_j^k que seriam obtidos na utilização de cópias de uma caixa PR. Entretanto, se nas K caixas utilizadas, Bob produzir um número par de erros, ele continuará determinando um valor g_i ao bit z_i corretamente, devido à propriedade da soma módulo 2.

Denotando por P a probabilidade de que a caixa gere o resultado correto, a probabilidade de Bob obter um número par de erros em r caixas é

$$\begin{aligned} Q^r_{(\text{par})} &= \sum_{k=0}^{\frac{r}{2}} \binom{r}{2k} P^{r-2k} (1-P)^{2k} \\ &= \frac{1}{2}[1 + (2P-1)^r]. \end{aligned} \tag{4.30a}$$

Da mesma forma, a probabilidade de Bob obter um número ímpar de erros em r caixas é

$$\begin{aligned} Q^r_{(\text{ímpar})} &= \sum_{k=0}^{\frac{r-1}{2}} P^{r-2k-1} (1-P)^{2k+1} = \\ &= \frac{1}{2}[1 - (2P-1)^r]. \end{aligned} \tag{4.30b}$$

Definindo os parâmetros p_{I} e p_{II} de acordo com

$$p_{\text{I}} = \frac{1}{2}[p(a+b=0|0,0) + p(a+b=0|0,1)]; \tag{4.31a}$$

$$p_{\text{II}} = \frac{1}{2}[p(a+b=0|1,1) + p(a+b=1|1,1)], \tag{4.31b}$$

é possível escrever uma expressão para a probabilidade de que Bob acerte sua adivinhação g_i sobre o bit z_i. Essa expressão é

$$\begin{aligned} p(g_i = z_i) &= Q^l_{(\text{par})}\, p_{\text{I}}\, Q^r_{(\text{par})}\, p_{\text{II}} + Q^l_{(\text{ímpar})}\, p_{\text{I}}\, Q^r_{(\text{ímpar})}\, p_{\text{II}} \\ &= \frac{1}{2}[1 + (E_{\text{I}})^l (E_{\text{II}})^r], \end{aligned} \tag{4.32}$$

sendo $E_\alpha = 2p_\alpha - 1$, $\alpha = \text{I}, \text{II}$.

Utilizando a relação $1 - h\left(\frac{1+y}{2}\right) \geq \frac{y^2}{2\ln 2}$, podemos reescrever agora, a expressão (4.8):

$$\begin{aligned} \sum_{i=0}^{n-1} [1 - H(G_i = Z_i)] &= \sum_{r=0}^{K} \binom{K}{r} \left[1 - H\left(\frac{1 + (E_{\text{I}})^l (E_{\text{II}})^r}{2}\right)\right] \\ &\geq \frac{1}{2\ln 2} \sum_{r=0}^{K} \binom{K}{r} (E_{\text{I}}^2)^l (E_{\text{II}}^2)^r \\ &= \frac{1}{2\ln 2} (E_{\text{I}}^2 + E_{\text{II}}^2)^K \leq d. \end{aligned} \tag{4.33}$$

Assim, se $E_{\text{I}}^2 + E_{\text{II}}^2 > 1$, então existe K tal que $I > 1 = d$. Ou seja, para uma caixa que satisfaz $E_{\text{I}}^2 + E_{\text{II}}^2 > 1$, é sempre possível violar o princípio de Causalidade da Informação com a utilização de um número adequado de cópias das caixas. Chamaremos esse critério de *Causalidade da Informação de Múltiplas Cópias*, no caso simples, em que $K = 1$, o critério será denominado *Causalidade da Informação de Uma Cópia*.

4.5.2 Visualização Computacional das Violações

Na ref. [41], os autores expressam o seu trabalho envolvendo um cálculo computacional do valor de I para todas as caixas em uma pequena região do politopo de não-sinalização $\mathcal{NS}(2,2,2)$. A região de interesse é a definida através da parametrização abaixo:

$$\mathbf{P_{AB}} = \alpha \mathbf{P_{PR1}} + \beta \left(\frac{1}{2}(\mathbf{P_{PR1}} + \mathbf{P}_{\mathbb{1}})\right) + (1 - \alpha - \beta)\mathbf{P_{L1}}. \tag{4.34}$$

Os parâmetros α e β são restritos às seguintes condições:

$$0 \leq \alpha \leq 1; \tag{4.35a}$$

$$0 \leq \beta \leq 1; \tag{4.35b}$$

$$\alpha + \beta \leq 1. \tag{4.35c}$$

Revisitando a fig. 15, a região do politopo definida por essa parametrização pode ser visualizada nessa seção bidimensional de acordo com a figura a seguir:

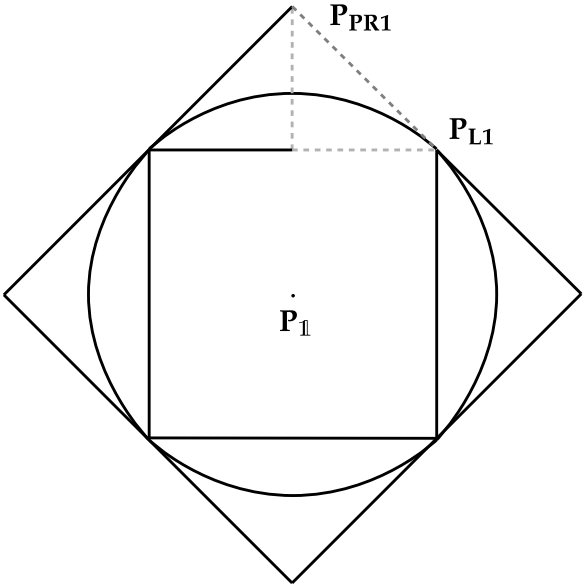


Figura 18 – Região definida pela parametrização (4.34) representada em uma seção bidimensional do politopo de não-sinalização.

O gráfico à seguir representa o resultado do cálculo computacional obtido na ref. [41]:

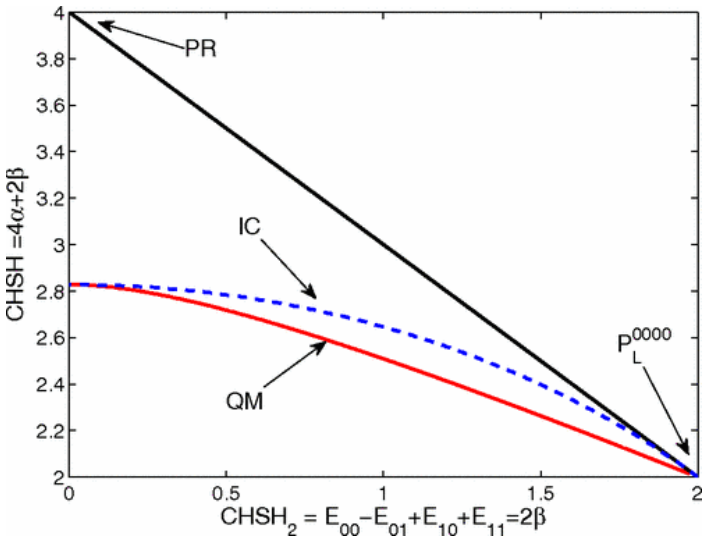


Figura 19 – Gráfico da ref. [41], a curva vermelha é a borda do conjunto $\mathcal{Q}_1$ da hierarquia NPA. A borda azul tracejada delimita a violação do princípio de Causalidade da Informação, todas as caixas acima dessa borda violam o princípio.

A borda azul tracejada representa as primeiras violações de Causalidade da Informação de Múltiplas Cópias. Esse critério é necessário, mas não suficiente, para determinar se uma caixa obedece o princípio. Assim, todas as caixas pertencentes à região superior a essa borda violam o princípio, enquanto caixas da região inferior à borda podem violar, mas não é possível saber utilizando esse critério.

Como o objetivo do estudo de Causalidade da Informação é tentar alcançar o conjunto de correlações quânticas através desse princípio, esse objetivo não é alcançado com o critério acima, portanto é necessário buscar outros critérios para tentar chegar ao objetivo.

4.6 Causalidade da Informação - Uma nova abordagem

Uma abordagem recente para o estudo de Causalidade da Informação, encontrada na ref. [32], foi desenvolvida utilizando ferramentas de teoria de causalidade. A ideia consiste na representação do cenário de Causalidade de Informação por um DAG compatível com as distribuições de probabilidade envolvendo todas as variáveis, incluindo o recurso não-local $\mathbf{P_{AB}}$ para, em seguida, utilizar desigualdades entrópicas como possíveis critérios do princípio.

Os autores representam o cenário de Causalidade da Informação de Uma Cópia através do seguinte DAG:

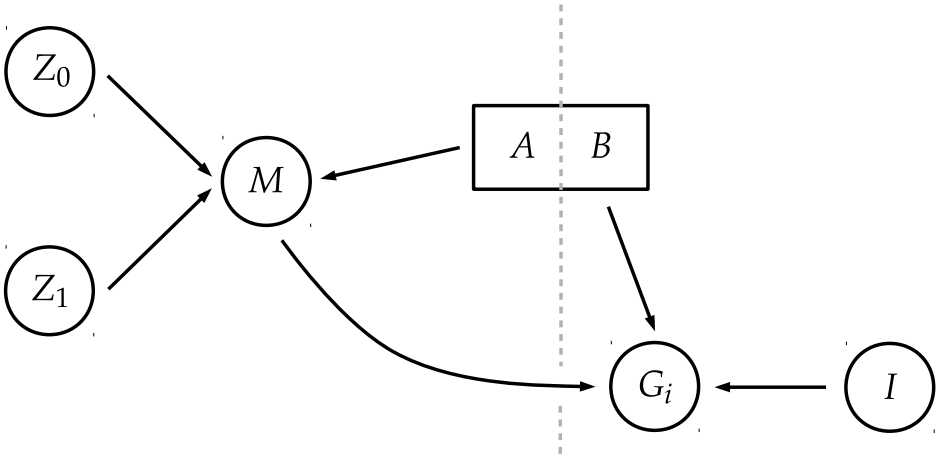


Figura 20 – Cenário de Causalidade da Informação de Uma Cópia representado por um DAG.

Nessa representação, as dependências de A e B sob as entradas X e Y são omitidas. Além disso, considera-se que a mensagem depende diretamente dos bits, o que simplifica o problema em quantidade de variáveis no cenário. A variável G_i, apesar de ser representada no DAG dessa forma, ela representa duas variáveis: G_0 e G_1, cada uma correspondendo à escolha de bit que Bob faz, através do valor i da variável I. O conjunto de todas as variáveis nesse cenário é

$$\Theta = \{Z_0, Z_1, G_0, G_1, M, AB\}. \tag{4.36}$$

Revisitando o que foi introduzido no final do cap. 1, a ideia é construir um vetor entrópico $\mathbf{H}$, em que cada entrada desse vetor é um elemento do *powerset* $\mathcal{P}(\Theta)$.

Entretanto, isso não é possível a priori, pois não é possível escrever uma distribuição de probabilidade conjunta que envolva as variáveis G_1 e G_2, assim, somos levados a utilizar o chamado *Cenário Marginal*. Mais detalhes podem ser encontrados na ref. [42].

Definição 4.6.1 (Cenário Marginal)**.** Para um conjunto Θ composto por n variáveis aleatórias, um *cenário marginal* $\mathcal{M}(\Theta)$ é definido como um conjunto $\mathcal{M} = \{\alpha_0, \ldots, \alpha_{|\mathcal{M}|-1}\}$, em que $\alpha_j \in \mathcal{P}(\Theta)$ e $|\mathcal{M}| < |\mathcal{P}(\Theta)|$.

O conjunto contendo m das n variáveis aleatórias pertencentes a Θ será denotado por M. Um vetor entrópico com componentes $H(\alpha_j)$, $\alpha_j \in \mathcal{P}(\Omega)$ pertence à região $\Gamma_n^* \in \mathbb{R}^{|\mathcal{P}|-1} = \mathbb{R}^{2^n}$. Já um vetor entrópico $\mathbf{H}$ definido em um cenário marginal $\mathcal{M}(\Theta)$ pertence ao espaço reduzido $\mathbb{R}^{|\mathcal{M}|-1} = \mathbb{R}^{2^m}$. Assim, a região Γ_m^* à qual $\mathbf{H}$ pertence, é a projeção da região Γ_n^* em $\mathbb{R}^{2^m}$.

Os autores argumentam que o cenário marginal mais geral compatível com o cenário de Causalidade da Informação envolvendo protocolos que utilizam recursos não-sinalizantes é:

$$\mathcal{M}(\Theta) = \{\alpha_i | \alpha_i \subseteq \{Z_0, Z_1, G_i, M\}\},\ i = 0, 1. \tag{4.37}$$

As independências condicionais que o DAG da fig 20 implicam são $I(Z_0, Z_1; AB) = 0$ e $I(Z_0, Z_1; G_0, G_1 | M, AB) = 0$. O primeiro caso é justificado pois M é um colisor no caminho que conecta Z_0 e Z_1 a AB, de modo que o não condicionamento em M os torna independentes. Já no segundo caso, os dois caminhos que conectam Z_0 e Z_1 a G_i são direcionados, contendo M e AB como intermediários, de modo que condicionar nessas variáveis torna Z_i independentes de G_i. Já relações do tipo $I(Z_0; G_0 | M, AB) = 0$ são obtidas a partir das desigualdades em conjunto com os axiomas polimatroidais sobre vetores entrópicos obtidos para o conjunto Θ. A partir dessa descrição, em que combinam-se os axiomas polimatroidais com as independências condicionais, é necessário eliminar todas as variáveis que não pertencem a M.

O primeiro passo é a remoção de AB através da eliminação de Fourier-Motzkin. Essa eliminação faz com que AB seja removido da sua descrição, mas suas propriedades fiquem implícitas na distribuição de probabilidade sobre as variáveis restantes. A partir daí, obtém-se o cenário marginal

$$\mathcal{M}_1 = \{\alpha_i | \alpha_i \subseteq M_1\}, \tag{4.38}$$

sendo $M_1 = \{Z_0, Z_1, G_0, G_1, M\}$.

O conjunto de desigualdades obtidas nesse cenário marginal é dado pelas desigualdades básicas definidas na seção 1.4.1, além de uma desigualdade não-trivial, obtida pelas independências condicionais:

$$H(G_0, G_1, M) + H(Z_0, Z_1) \leq H(M) + H(Z_0, Z_1, G_0, G_1, M). \tag{4.39}$$

Feito isso, eliminam-se todas as variáveis que não estejam contidas em $\mathcal{M}(\Theta)$. A descrição final obtida da região marginal Γ_m é dividida em dois grupos: O primeiro é aquele que contém todas as desigualdades válidas para $\mathcal{M}$ independentemente de suas relações causais. O segundo contém desigualdades que seguem das desigualdades básicas considerando as relações de independência condicional geradas pela estrutura causal.

O conjunto de todas essas desigualdades expressa todas as propriedades informacionais e causais obtidas no cenário de Causalidade da Informação de Uma Cópia. Uma dessas desigualdades representa o novo critério utilizado pelos autores para representar o princípio. Essa desigualdade é:

$$I(Z_0; G_0, M) + I(Z_1; G_1, M) + I(Z_0; Z_1|G_1, M) \leq H(M) + I(Z_0; Z_1). \tag{4.40}$$

O critério expresso por essa desigualdade é apenas necessário para que uma caixa não-sinalizante satisfaça o princípio de Causalidade da Informação.

Por motivos de comparação, os autores desse trabalho realizaram um cálculo computacional utilizando o critério acima, de modo a gerar um gráfico similar ao da fig. 19. A região observada é representada de acordo com a parametrização abaixo:

$$\mathbf{P_{AB}} = \gamma \mathbf{P_{PR1}} + \epsilon \mathbf{P_{L1}} + (1 - \gamma - \epsilon)\mathbf{P_{\mathbb{1}}}. \tag{4.41}$$

A seguir, encontra-se o gráfico gerado pelos autores.

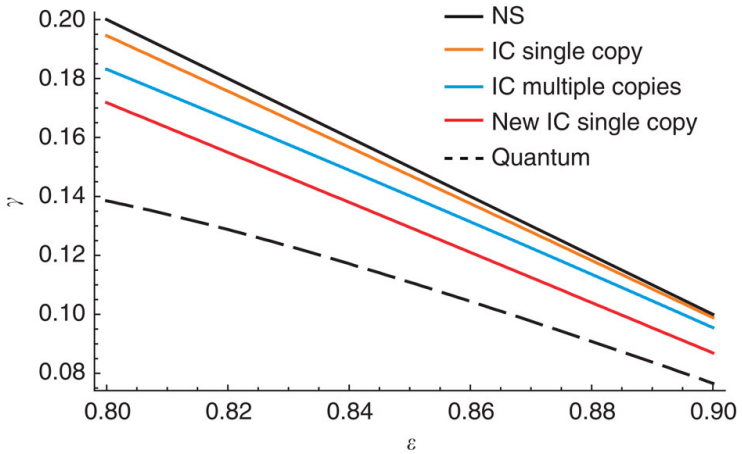


Figura 21 – A linha preta é a borda do politopo de não-sinalização, a curva laranja e a azul representam, respectivamente, os critérios de Causalidade da Informação para Uma e Múltiplas cópias. A curva vermelha representa o novo critério de Causalidade da Informação encontrado na ref. [32]. A curva tracejada representa a borda do conjunto $\mathcal{Q}_1$. Esse gráfico foi retirado da ref. [32].

Para essa região específica, o novo critério se mostra melhor que os critérios anteriores, dado que, com ele, aproxima-se mais ainda do conjunto $\mathcal{Q}_1$. Repare entretanto

que os parâmetros γ e ϵ não variam da mesma forma que os parâmetros α e β relativos à figura 19. O fato é que, no restante da região, apesar de a curva vermelha ainda continuar abaixo da curva laranja, a curva vermelha cruza a curva azul, se mostrando um critério mais fraco que o critério de Causalidade da Informação de Múltiplas Cópias utilizado na ref. [41]. Isso sugere que uma versão desse critério para Múltiplas Cópias pode se aproximar mais do conjunto $\mathcal{Q}_1$ que todos os outros.

4.7 Nova Abordagem - Múltiplas Cópias

Seguindo a linha da ref. [32], iniciei a tentativa de obter resultados utilizando uma versão de múltiplas cópias do critério introduzido na seção anterior. Nessa mesma referência, os autores argumentam que a versão de múltiplas cópias é simplesmente uma extensão do critério para uma cópia, sendo expresso pela seguinte desigualdade:

$$\sum_{i=0}^{n-1} I(Z_i; G_i, M) + \sum_{i=1}^{n-1} I(Z_1; Z_i|G_i, M) \leq H(M) + \sum_{i=0}^{n-1} H(Z_i) - H(Z_1, \ldots, Z_n). \quad (4.42)$$

Utilizando o cenário de Causalidade da Informação com três cópias do recurso não-sinalizante, reproduzimos o critério e obtivemos um gráfico para esse caso:

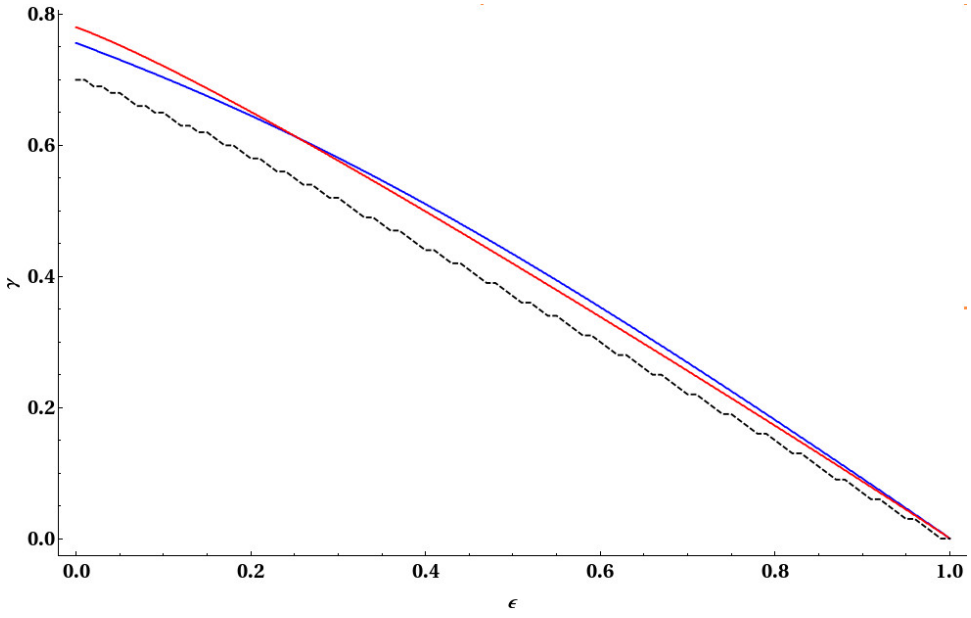


Figura 22 – Gráfico da seção do politopo definida pela eq. 4.41. A curva pontilhada representa o conjunto $\mathcal{Q}_1$, a curva vermelha representa o novo critério para uma cópia e a curva azul representa o novo critério para três cópias.

Esse resultado parcial foi obtido no final desse trabalho e carece de uma análise sobre sua validez e significado. Tal análise ocorrerá na continuação desta pesquisa.

Apesar destes avanços no desenvolvimento de critérios que representem o princípio de Causalidade da Informação, sabemos que o desenvolvimento de cenários de Causalidade da Informação considerando mais que duas partes é necessário. Isto deve-se ao resultado encontrado na ref. [43], que diz que correlações quânticas, em geral, não são possíveis de serem obtidas através de princípios envolvendo cenários bipartidos. Esse é o passo a ser dado em nossa pesquisa logo após obtermos novos resultados para o critério de Causalidade da Informação com múltiplas cópias.

Conclusão

O objetivo desse trabalho foi de tentar realizar, motivado pelo estudo do princípio de Causalidade da Informação, uma união entre a teoria da informação, a teoria de causalidade e a teoria de não-localidade.

Como o objetivo principal do trabalho foi revisitar os últimos resultados com relação a esse princípio, conclui-se que esse trabalho alcança o objetivo de mostrar, de uma forma padronizada, os últimos avanços nessa pequena parte da pesquisa sobre Não-localidade, a de buscar princípios físicos que justifiquem a não-localidade quântica.

Apesar da utilização do cenário mais simples, há muita riqueza na relação entre o princípio de Causalidade da Informação e a teoria de não-localidade. O resultado da ref.[32] aponta na direção de que esse princípio seja um ótimo candidato a ser o princípio que gere as correlações quânticas. O nosso resultado para 3 cópias do recurso não-local, expresso na fig. 22, indica que o critério pode ser melhorado de modo a se aproximar mais do conjunto $\mathcal{Q}_1$.

Acreditamos que o cruzamento entre as duas curvas nessa figura ocorra por dois possíveis motivos. O primeiro é a possibilidade de existência de vínculos, na distribuição de probabilidade sobre as variáveis do cenário, que não estejam sendo devidamente impostos pelo código utilizado para gerar o gráfico da fig. 22. O segundo motivo é que tal cruzamento esteja correto, apesar de não ser esperado, pois não é claro que, nessa nova abordagem, a utilização de muitas cópias do recurso não-local deva melhorar o critério em toda a região observada. Nesse caso, a borda azul não necessariamente deve se aproximar mais de $\mathcal{Q}_1$ que a borda vermelha. É esperado que melhore nas proximidades da caixa $\mathbf{P_{PR1}}$, de acordo com o protocolo, sendo essa melhora representada no gráfico, porém há a possibilidade de que isso ocorra apenas nessa região específica.

Caso a primeira hipótese seja a correta, há a possibilidade de realizar esse mesmo tratamento utilizando um modelo representado por um DAG que contenha mais variáveis, possivelmente gerando novas relações de independência condicional e gerando um critério melhor que o critério para uma cópia em toda a região analisada. Caso a segunda possibilidade seja a correta, isso será confirmado com novas implementações para mais cópias do recurso não-local, sendo estes os próximos passos a serem dados nessa pesquisa.

Além disso, foi utilizado um algoritmo pré-determinado aparentemente óptimo, mas que pode apresentar fraquezas no teste de violação do princípio de Causalidade da Informação. Outro problema a ser atacado é o de generalizar o critério utilizado

para testar o princípio de Causalidade da Informação utilizando todos os protocolos possíveis nesse cenário. Esperamos que isso fortaleça mais ainda o princípio, eliminando a dependência dos resultados à fixação do protocolo.

Referências

[1] Albert Einstein, Boris Podolsky, and Nathan Rosen. Can Quantum-mechanical Description of Physical Reality be Considered Complete? *Physical review*, 47(10):777, 1935. http://dx.doi.org/10.1103/PhysRev.47.777.

[2] John S Bell. On The Einstein Podolsky Rosen Paradox, 1964.

[3] George Casella and Roger L Berger. *Statistical Inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[4] Raymond W Yeung. *Information Theory and Network Coding*. Springer Science & Business Media, 2008.

[5] David JC MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge university press, 2003.

[6] Vlatko Vedral. *Introduction to Quantum Information Science (Oxford Graduate Texts)*. Oxford University Press, Inc., 2006.

[7] Judea Pearl. *Causality*. Cambridge university press, 2009.

[8] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 2014.

[9] Michael Nielsen. If Correlation Doesn't Imply Causation, then What Does?, 2012. http://www.michaelnielsen.org/ddi/if-correlation-doesnt-imply-causation-then-what-does/.

[10] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT press, 2000.

[11] Judea Pearl. On the Definition of Actual Cause, 1998.

[12] Judea Pearl and T S Verma. Equivalence and Synthesis of Causal Models. In *Proceedings of Sixth Conference on Uncertainty in Artijicial Intelligence*, pages 220–227, 1991.

[13] Marco Túlio Coelho Quintino. Black Box Correlations: Locality, Noncontextuality, and Convex Politopes. Mestrado, 2012. http://www.mat.ufmg.br/~tcunha/Disserta_MTQ.pdf.

[14] Leslie E Ballentine. *Quantum Mechanics: A Modern Development*. World scientific, 1998.

[15] David Jeffery Griffiths. *Introduction to Quantum Mechanics*. Pearson Education India, 2005.

[16] Michael A Nielsen and Isaac L Chuang. *Quantum Computation and Quantum Information*. Cambridge university press, 2010.

[17] John F. Clauser, Michael A. Horne, Abner Shimony, and Richard A. Holt. Proposed Experiment to Test Local Hidden-Variable Theories. *Phys. Rev. Lett.*, 23:880–884, Oct 1969. http://link.aps.org/doi/10.1103/PhysRevLett.23.880.

[18] Sandu Popescu and Daniel Rohrlich. Quantum Nonlocality as an Axiom. *Foundations of Physics*, 24(3):379–385, 1994. http://link.springer.com/article/10.1007/BF02058098.

[19] Jonathan Barrett, Noah Linden, Serge Massar, Stefano Pironio, Sandu Popescu, and David Roberts. Nonlocal correlations as an information-theoretic resource. *Phys. Rev. A*, 71:022101, Feb 2005. http://link.aps.org/doi/10.1103/PhysRevA.71.022101.

[20] Gláucia Murta Guimarães. Não-localidade em Sistemas Quânticos. Mestrado, 2012. www.mat.ufmg.br/~tcunha/DissGlaucia.pdf.

[21] Claude Cohen-Tannoudji, Bernard Diu, and Franck Laloë. *Quantum Mechanics*. Wiley, 1977.

[22] Károly F. Pál and Tamás Vértesi. Concavity of the set of quantum probabilities for any given dimension. *Phys. Rev. A*, 80:042114, Oct 2009. http://link.aps.org/doi/10.1103/PhysRevA.80.042114.

[23] Boris Tsirelson. Quantum Generalizations of Bell's Inequality. *Letters in Mathematical Physics*, 4(2):93–100, 1980. http://www.tau.ac.il/~tsirel/download/qbell80.pdf.

[24] Lawrence J Landau. Empirical Two-point Correlation Functions. *Foundations of Physics*, 18(4):449–460, 1988. http://link.springer.com/article/10.1007%2FBF00732549.

[25] Ll Masanes. Necessary and Sufficient Condition for Quantum-generated Correlations. *arXiv preprint quant-ph/0309137*, 2003. http://arxiv.org/abs/quant-ph/0309137v1.

[26] Miguel Navascués, Stefano Pironio, and Antonio Acín. A Convergent Hierarchy of Semidefinite Programs Characterizing the Set of Quantum Correlations. *New Journal of Physics*, 10(7):073013, 2008. http://stacks.iop.org/1367-2630/10/i=7/a=073013.

[27] Rafael Luiz Da Silva Rabelo. *On Quantum Nonlocality and the Device-independent Paradigm*. PhD thesis, 2013. https://www.quantumlah.org/media/thesis/CQT_131001_RafaelRabelo.pdf.

[28] Gilles Brassard, Harry Buhrman, Noah Linden, André Allan Méthot, Alain Tapp, and Falk Unger. Limit on Nonlocality in any World in Which Communication Complexity is not Trivial. *Physical Review Letters*, 96(25):250401, 2006. http://journals.aps.org/prl/abstract/10.1103/PhysRevLett.96.250401.

[29] Miguel Navascués and Harald Wunderlich. A Glance Beyond the Quantum Model. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, page rspa20090453. The Royal Society, 2009. http://rspa.royalsocietypublishing.org/content/466/2115/881.

[30] Tobias Fritz, Ana Belén Sainz, Remigiusz Augusiak, J Bohr Brask, Rafael Chaves, Anthony Leverrier, and Antonio Acín. Local Orthogonality as a Multipartite Principle for Quantum Correlations. *Nature communications*, 4, 2013. http://www.nature.com/ncomms/2013/130816/ncomms3263/full/ncomms3263.html.

[31] Marcin Pawłowski, Tomasz Paterek, Dagomir Kaszlikowski, Valerio Scarani, Andreas Winter, and Marek Żukowski. Information Causality as a Physical Principle. *Nature*, 461(7267):1101–1104, 2009. http://www.nature.com/nature/journal/v461/n7267/full/nature08400.html.

[32] Rafael Chaves, Christian Majenz, and David Gross. Information–theoretic Implications of Quantum Causal Structures. *Nature communications*, 6, 2015. http://www.nature.com/ncomms/2015/150106/ncomms6766/full/ncomms6766.html.

[33] Tobias Fritz. Beyond Bell's Theorem: Correlation Scenarios. *New Journal of Physics*, 14(10):103001, 2012. http://stacks.iop.org/1367-2630/14/i=10/a=103001.

[34] Tobias Fritz. Beyond Bell's Theorem II: Scenarios with Arbitrary Causal Structure. *Communications in Mathematical Physics*, 341(2):391–434, 2016. http://link.springer.com/article/10.1007%2Fs00220-015-2495-5.

[35] M. S. Leifer and Robert W. Spekkens. Towards a Formulation of Quantum Theory as a Causally Neutral Theory of Bayesian Inference. *Phys. Rev. A*, 88:052130, Nov 2013. http://link.aps.org/doi/10.1103/PhysRevA.88.052130.

[36] Christopher J Wood and Robert W Spekkens. The Lesson of Causal Discovery Algorithms for Quantum Correlations: Causal Explanations of Bell-inequality Violations Require Fine-tuning. *New Journal of Physics*, 17(3):033002, 2015. http://stacks.iop.org/1367-2630/17/i=3/a=033002.

[37] Cyril Branciard, Mateus Araújo, Adrien Feix, Fabio Costa, and Časlav Brukner. The Simplest Causal Inequalities and their Violation. *New Journal of Physics*, 18(1):013008, 2016. http://stacks.iop.org/1367-2630/18/i=1/a=013008.

[38] Časlav Brukner. Bounding Quantum Correlations with Indefinite Causal Order. *New Journal of Physics*, 17(8):083034, 2015. http://stacks.iop.org/1367-2630/17/i=8/a=083034.

[39] Mateus Araújo, Cyril Branciard, Fabio Costa, Adrien Feix, Christina Giarmatzi, and Časlav Brukner. Witnessing Causal Nonseparability. *New Journal of Physics*, 17(10):102001, 2015. http://stacks.iop.org/1367-2630/17/i=10/a=102001.

[40] Wim van Dam. Implausible Consequences of Superstrong Nonlocality. *Natural Computing*, 12(1):9–12, 2013. http://dx.doi.org/10.1007/s11047-012-9353-6.

[41] Jonathan Allcock, Nicolas Brunner, Marcin Pawlowski, and Valerio Scarani. Recovering Part of the Boundary Between Quantum and Nonquantum Correlations from Information Causality. *Physical Review A*, 80(4):040103, 2009. http://journals.aps.org/pra/abstract/10.1103/PhysRevA.80.040103.

[42] Rafael Chaves, Lukas Luft, and David Gross. Causal Structures from Entropic Information: Geometry and Novel Scenarios. *New Journal of Physics*, 16(4):043001, 2014. http://iopscience.iop.org/article/10.1088/1367-2630/16/4/043001/meta.

[43] Rodrigo Gallego, Lars Erik Würflinger, Antonio Acín, and Miguel Navascués. Quantum Correlations Require Multipartite Information Principles. *Physical Review Letters*, 107(210403), 2011. http://journals.aps.org/prl/abstract/10.1103/PhysRevLett.107.210403.