2
Teoria da Informação

A teoria da informação foi desenvolvida por Shannon nos anos 40 tendo em mente aplicações em engenharia de comunicação. O caráter inovador dessa teoria aliado a sua elegância matemática, a fez ter grande impacto não só na engenharia como também em diversas áreas tais como estatística e economia. Em especial, na área de reconhecimento de padrões, os pesquisadores têm dado crescente importância aos conceitos dessa teoria. Esta teoria apresenta dois conceitos básicos, quais sejam:

- entropia – medida de incerteza ou aleatoriedade de variáveis aleatórias individuais ou combinadas; e
- informação mútua – dependência estocástica entre variáveis aleatórias.

Este capítulo reúne fundamentos teóricos da teoria da informação, sendo expressos de maneira descritiva. Para demonstrações e explanações sobre o assunto, sugere-se, em especial, a consulta à referência Cover & Thomas (2006). As principais definições apresentadas, na seção a seguir, são baseadas em tal referência, salvo menção em contrário.

2.1
Entropia e Informação Mútua de Shannon

A incerteza caracteriza o ganho de informação que a ocorrência de um evento pode promover. Ela pode, portanto, ser traduzida através da probabilidade de ocorrência de seu evento. Um evento cuja ocorrência é certa não traz nenhum acréscimo de informação, pois toda a informação já está contida na sua certeza de ocorrência. Desta maneira, pode-se dizer que a determinação da quantidade de informação produzida pela ocorrência de um evento é determinada pela quantidade de "surpresa" que essa ocorrência traz.

Seja $E = \{e_1, e_2, ..., e_n\}$ o conjunto dos *n* resultados (eventos elementares) associados a um experimento aleatório, e seja $P = \{p_1, p_2, ..., p_n\}$

as respectivas probabilidades de ocorrência. Existe, portanto, uma incerteza quanto ao resultado. Shannon foi o primeiro a definir uma medida, quantificando essa incerteza, chamando-a de entropia (*H*), a seguir definida:

$$H(P) = -\sum_{i=1}^{n} p_i \log p_i \qquad (2.1)$$

Note-se que a entropia depende da quantidade $Q(p_i) = -\log p_i$. A medida $Q(p)$ foi proposta por Hartley (1928) como uma medida da informação produzida pela ocorrência de um evento de probabilidade p. A entropia de Shannon é, portanto, uma média ponderada das informações $Q(p_i)$.

A entropia será nula, se, e somente se, um dos resultados é certo de ocorrer, sendo, assim, toda a incerteza removida. Ela será máxima quando os *n* resultados forem equiprováveis, ou seja, $p_i = \frac{1}{n}, \forall i$. Observe-se, por fim, que quanto mais incerto for o evento e_i, maior é sua informação.

A entropia na teoria da informação corresponde, portanto, à incerteza probabilística associada a uma distribuição de probabilidade.

Definição 1 – A *entropia de Shannon H(X)* de uma variável aleatória discreta *X* , com função de massa de probabilidade $f_x(x)$, $x \in \mathbb{X}$ (conjunto domínio da variável), é definida por

$$H(X) = -\sum_{x \in \mathbb{X}} f_x(x) \log f_x(x) \qquad (2.2)$$

Representa-se também por $H(f_x)$ a quantidade acima.

Por convenção, 0 log 0 = 0 , que é facilmente justificado pela continuidade, visto que $x \log x \to 0$ quando $x \to 0$.

Decorre da própria definição que $H(X) \geq 0$.

Note-se que na definição de entropia não são considerados os valores que a variável assume, mas apenas suas probabilidades, sendo a entropia, pois, uma medida adimensional.

Com o intuito de ilustrar o conceito da entropia de Shannon, apresentar-se-á, a seguir, um exemplo simples.

Exemplo

Assuma que a variável aleatória X assuma somente dois valores com probabilidades p e $(1-p)$, respectivamente. Então a entropia de *X* ($H(X)$, ou que seja $H(p)$) será dada por: $H(p) = -p\log p - (1-p)\log(1-p)$.

A Figura 2.1 mostra o gráfico da função *H(p)*. Pode-se observar que *H(p)* é nula quando *p* é igual a 0 ou 1. Isto faz sentido, pois quando isso acontece a variável degenerou e não há mais incerteza. Por outro lado, a incerteza é máxima quando *p* = ½ (ou seja, os dois valores de *X* são equiprováveis), o que corresponde ao valor máximo da entropia.

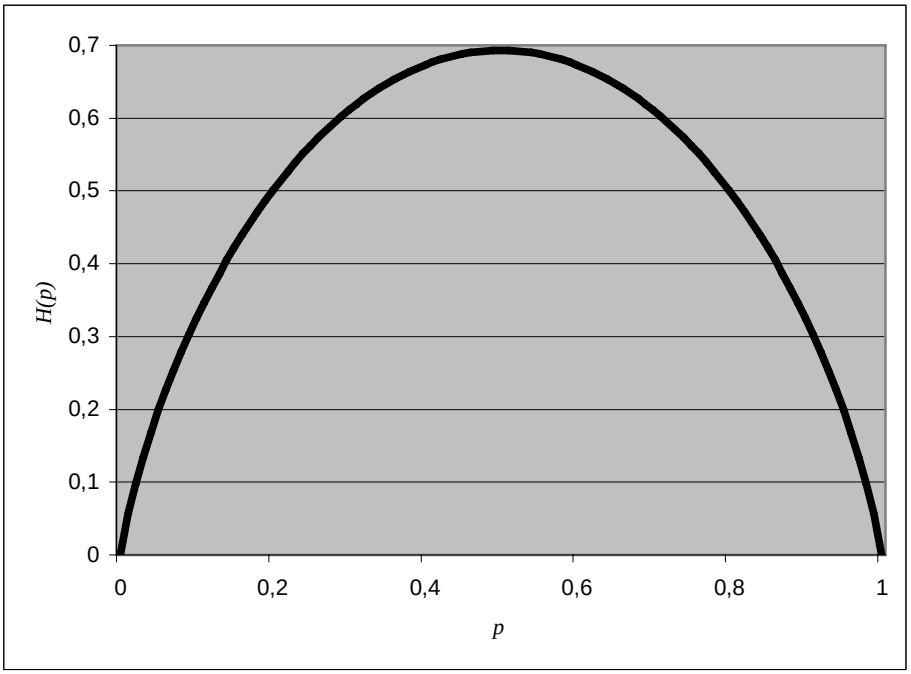

Figura 2.1 – Entropia (*H(p)*) versus *p*

Definição 2 – A *entropia conjunta (de Shannon) H(X, Y)* de um par de variáveis aleatórias discretas *(X, Y)* com distribuição de probabilidade conjunta f_{xy} é definida como

$$H(X,Y) = -\sum_{x\in\mathbb{X}}\sum_{y\in\mathbb{Y}} f_{xy}(x,y)\log f_{xy}(x,y) \tag{2.3}$$

Definição 3 – A *entropia condicional* (*de Shannon) H(Y | X)* (isto é, de *Y* dado o conhecimento de *X*) é definida como

$$\begin{aligned} H(Y\mid X) &= \sum_{x\in\mathbb{X}} f_x(x)H(Y\mid X=x) = -\sum_{x\in\mathbb{X}} f_x(x)\sum_{y\in\mathbb{Y}} f_{y|x}(y\mid x)\log f_{y|x}(y\mid x) = \\ &= -\sum_{x\in\mathbb{X}}\sum_{y\in\mathbb{Y}} f_{xy}(x,y)\log f_{y|x}(y\mid x) \end{aligned} \tag{2.4}$$

A entropia condicional mede a incerteza remanescente de *Y* dado o conhecimento de *X*.

Definição 4 – A *entropia relativa* ou *divergência (assimétrica) de Kullback–Leibler* entre duas funções de massa de probabilidade f e g é definida como

$$D_{KL}(f \parallel g) = \sum_{x\in\mathbb{X}} f(x)\log\frac{f(x)}{g(x)} = E_f\left(\log\frac{f(x)}{g(x)}\right) \tag{2.5}$$

Aqui, usa-se a convenção de que $0\log\frac{0}{0} = 0$ e a convenção (baseada nos argumentos de continuidade) de que $0\log\frac{0}{g} = 0$ e $p\log\frac{f}{0} = \infty$.

$D_{KL}(f \parallel g) \geq 0$ com igualdade se, e somente se, $f(x) = g(x)$, para todo $x \in \mathbb{X}$.

A entropia relativa (ou divergência) de Kullback–Leibler é uma medida de similaridade entre funções estritamente positivas, é também referida como "distância" entre distribuições, contudo, não é verdadeiramente uma métrica, visto que não é simétrica e não satisfaz a desigualdade triangular. Ela é muito utilizada

na comparação entre duas funções. Neste caso, a função *g* desempenha o papel de função de referência.

Em inferência estatística, a divergência de Kullback-Leibler surge como uma esperança, com respeito a *f*, do logaritmo da razão de verossimilhança.

A divergência de Kullback–Leibler é implicitamente baseada na entropia de Shannon, visto que

$$D_{KL}(f \parallel g) = -\sum_{x\in\mathbb{X}} f(x)\log g(x) - \left(-\sum_{x\in\mathbb{X}} f(x)\log f(x)\right),$$

onde "$-\sum f(x)\log f(x)$" é a entropia de Shannon com respeito a *p(x)*, e "$-\sum f(x)\log g(x)$" pode ser interpretada como a *entropia cruzada* entre *p(x)* e *q(x)* (Jenssen, 2005).

Definição 5 – A *informação mútua (de Shannon) I(X;Y)* entre duas variáveis aleatórias discretas *X* e *Y* , com função de massa de probabilidade conjunta $f_{xy}(x,y)$ e funções de massa de probabilidade marginais $f_x(x)$ e $f_y(y)$ é dada pela entropia relativa entre a distribuição conjunta e a distribuição do produto das marginais (ou seja, o quanto a primeira se diferencia da segunda):

$$I(X;Y) = \sum_{x\in\mathbb{X}}\sum_{y\in\mathbb{Y}} f_{xy}(x,y)\log\frac{f_{xy}(x,y)}{f_x(x)\,f_y(y)} = D_{KL}(f_{xy}(x,y) \parallel f_x(x)\,f_y(y)) \geq 0 \qquad (2.6)$$

com igualdade se, e somente se, $f_{xy}(x,y) = f_x(x)\ f_y(y)$ (ou seja, se *X* e *Y* são independentes).

A partir da Eq (2.6), pode-se dizer que a informação mútua é uma medida de independência estatística. Quanto maior for a informação mútua, mais relacionadas serão as variáveis.

A informação mútua pode ser vista como uma medida da quantidade de informação que uma variável aleatória tem acerca da outra.

Note-se a estreita relação com a log-verossimilhança:

$$I(X;Y) = E_{f_{xy}}\left(\log\frac{f_{xy}(x,y)}{f_x(x)\,f_y(y)}\right) = E_{f_{xy}}\left(\log f_{xy}(x,y)\right) - E_{f_{xy}}\left(\log f_x(x)\,f_y(y)\right) \qquad (2.7)$$

2.2
Principais Relações

Destacam-se os seguintes resultados:

$$\Rightarrow H(X,Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y) \quad (\textit{Regra da Cadeia}) \tag{2.8}$$

A entropia conjunta de duas variáveis aleatórias é, portanto, dada pela entropia de uma delas acrescida da entropia remanescente da outra, dado o conhecimento da primeira.

$$\Rightarrow I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) \tag{2.9}$$

A informação mútua *I(X;Y)* é, portanto, a redução da incerteza de uma variável devido ao conhecimento da outra. Assim, *X* diz tanto acerca de *Y* quanto *Y* diz acerca de *X*.

$$\Rightarrow I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y,Z) \quad (\textit{Informação Mútua Condicional}) \tag{2.10}$$

$$\Rightarrow I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{2.11}$$

$$\Rightarrow I(X;Y) = I(Y;X) \qquad (\textit{Simetria da Informação Mútua}) \tag{2.12}$$

$$\Rightarrow I(X;X) = H(X) \tag{2.13}$$

Esta é a razão de, muitas vezes, a entropia ser referida como *auto-informação*.

$$\Rightarrow H(X \mid Y) \leq H(X) \tag{2.14}$$

Com igualdade se, e somente se, X e Y são independentes.

$$\Rightarrow H(X,Y) \leq H(X) + H(Y) \tag{2.15}$$

Com igualdade se, e somente se, X e Y são independentes.

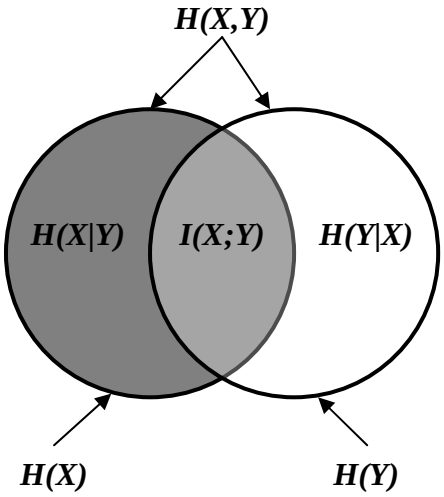

Figura 2.2 – Relação entre entropia e informação mútua

2.3
Entropia Diferencial de Shannon

A seguir, introduz-se o conceito de entropia diferencial, que é a entropia de uma variável aleatória contínua. A entropia diferencial é similar em muitas formas à entropia de uma variável aleatória discreta, mas há diferenças importantes, devendo-se ter certo cuidado no uso desse conceito. Salientando-se, aqui, que a entropia diferencial, ao contrário do caso discreto, pode ser negativa, mas a versão diferencial da informação mútua, como será visto, sempre será não-negativa.

Definição 6 – A *entropia diferencial de Shannon h(X)* de uma variável aleatória contínua X , com função de densidade de probabilidade $f_x(x)$, $x \in \mathbb{X}$ (conjunto domínio da variável) , é definida como

$$h(X) = -\int_{\mathbb{X}} f_x(x) \log f_x(x)\, dx \qquad \textit{(se existir}^{1}\textit{)} \tag{2.16}$$

Como no caso discreto, a entropia diferencial depende somente da função de densidade de probabilidade da variável aleatória, e, portanto, é, por vezes, representada por $h(f_x)$ ao invés de $H(X)$.

[1] Doravante, essa declaração fica subentendida sempre que a definição envolver uma integral.

Definição 7 – A *entropia diferencial conjunta (de Shannon) h(X,Y)* de um par de variáveis aleatórias contínuas *(X, Y)*, com densidade f_{xy} , é definida como

$$h(X,Y) = -\int_{\mathbb{Y}}\int_{\mathbb{X}} f_{xy}(x,y)\log f_{xy}(x,y)\,dx\,dy \tag{2.17}$$

Definição 8 – A *entropia diferencial condicional (de Shannon) h(X |Y)* é definida como

$$h(X\mid Y) = -\int_{\mathbb{Y}}\int_{\mathbb{X}} f_{xy}(x,y)\log f_{x|y}(x\mid y)\,dx\,dy \tag{2.18}$$

Considerando-se que *f(x | y) = f(x,y) / f(y)* , pode-se dizer que

$$h(X\mid Y) = h(X,Y) - h(Y) \tag{2.19}$$

Devendo-se ter cuidado se alguma das entropias diferenciais é infinita.

Definição 9 – A *entropia relativa ou divergência (assimétrica) de Kullback–Leibler* entre duas densidades *f* e *g* é definida por

$$D_{KL}(f \,\|\, g) = \int_{\mathbb{X}} f(x)\log\frac{f(x)}{g(x)}\,dx = E_f\left(\log\frac{f(x)}{g(x)}\right) \tag{2.20}$$

Note-se que $D_{KL}(f \,\|\, g)$ é finita somente se o conjunto suporte de *f* está contido no conjunto suporte de *g*.

Também aqui, motivado pela continuidade, toma-se $0\log\frac{0}{0} = 0$.

$D_{KL}(f \,\|\, g) \geq 0$ com igualdade se, e somente se, $f = g$ *em quase toda parte*.

Definição 10 – A *informação mútua (de Shannon) I(X;Y)* entre duas variáveis aleatórias contínuas *X* e *Y* , com densidade conjunta f_{XY} e densidades marginais f_X e f_Y é dada por:

$$I(X;Y) = \int_{\mathbb{Y}}\int_{\mathbb{X}} f(x,y)\log\frac{f_{xy}(x,y)}{f_x(x)\,f_y(y)}\,dx\,dy = D_{KL}(f_{xy}(x,y) \,\|\, f_x(x)\,f_y(y)) \geq 0 \tag{2.21}$$

com igualdade se, e somente se, $f_{xy} = f_x \ f_y$ *em quase toda parte* (ou seja, se X e Y são independentes).

Embora a entropia diferencial, ao contrário da entropia para variáveis aleatórias discretas, não possa ser interpretada como uma medida de aleatoriedade ou incerteza, a informação mútua mantém a mesma interpretação tal qual no caso discreto.

Destacam-se os seguintes resultados:

$$\Rightarrow h(X,Y) = h(X) + h(Y \mid X) = h(Y) + h(X \mid Y) \quad (2.22)$$

(*Regra da Cadeia para a Entropia Diferencial*)

$$\Rightarrow I(X;Y) = h(X) - h(X \mid Y) = h(Y) - h(Y \mid X) \quad (2.23)$$

$$\Rightarrow I(X;Y \mid Z) = h(X \mid Z) + h(X \mid Y,Z) \quad (2.24)$$

(*Informação Mútua Condicional*)

$$\Rightarrow I(X;Y) = h(X) + h(Y) - h(X,Y) \quad (2.25)$$

$$\Rightarrow I(X;Y) = I(Y;X) \quad \text{(*Simetria da Informação Mútua)*} \quad (2.26)$$

$$\Rightarrow h(X \mid Y) \leq h(X) \quad (2.27)$$

Com igualdade se, e somente se, X e Y são independentes.

$$\Rightarrow h(X,Y) \leq h(X) + h(Y) \quad (2.28)$$

Com igualdade se, e somente se, X e Y são independentes.

2.4
Entropia, Entropia Diferencial e Informação Mútua de Rényi

Definição 11 – A *entropia de Rényi* $H_{R_\alpha}(X)$ de ordem α de uma variável aleatória discreta X, com função de massa de probabilidade $f_x(x)$, $x \in \mathcal{X}$, é definida como

$$H_{R_\alpha}(X) = \frac{1}{1-\alpha} \log \sum_{x \in \mathbb{X}} f_x^{\alpha}(x), \text{ para } \alpha > 0 \text{ e } \alpha \neq 1 \tag{2.29}$$

Aqui, $0^\alpha = 0$, para todo real α.

A entropia de Shannon aparece como um caso especial da entropia de Rényi, tomando-se o limite desta, quando $\alpha \to 1$, ou seja,

$$\lim_{\alpha \to 1} H_{R_\alpha}(X) = -\sum_{x \in \mathbb{X}} f_x(x) \log f_x(x) = H(X) \tag{2.30}$$

onde H(X) é a entropia de Shannon, anteriormente, apresentada. Assim, a entropia de Shannon pode ser considerada um membro da família da entropia de Rényi.

De particular interesse, é a entropia de Rényi de ordem 2, chamada de *entropia quadrática*, definida a seguir:

Definição 11a – A *entropia quadrática de Rényi* $H_{R_2}(X)$ é dada por

$$H_{R_2}(X) = -\log \sum_{x \in \mathbb{X}} f_x^2(x) \tag{2.31}$$

Definição 12 – A *entropia diferencial de Rényi* $h_{R_\alpha}(X)$ de ordem α de uma variável aleatória contínua X, com função de densidade de probabilidade $f_x(x)$, é definida como

$$h_{R_\alpha}(X) = \frac{1}{1-\alpha}\log\left[\int_{\mathbb{X}} f_x^{\alpha}(x)dx\right], \quad \text{para } 0 < \alpha < \infty,\ \alpha \neq 1. \tag{2.32}$$

Novamente, tomando-se o limite, quando $\alpha \to 1$, obtém-se a função da entropia de Shannon:

$$h_{R_1}(X) = -\int_{\mathbb{X}} f_x(x)\log f_x(x)dx = h(X) \tag{2.33}$$

Evidencia-se, aqui, também, a entropia diferencial de Rényi de ordem 2, chamada de *entropia quadrática diferencial*, definida a seguir:

Definição 12a – A *entropia quadrática diferencial de Rényi* $h_{R_2}(X)$ é dada por

$$h_{R_2}(X) = -\log\int_{\mathbb{X}} f_x^2(x)\,dx \tag{2.34}$$

Definição 13 – A *entropia relativa* ou *divergência (assimétrica) de Rényi* entre duas funções de massa de probabilidade *f(x)* e *g(x)* é definida como

$$\begin{aligned} D_{R_\alpha}(f \,\|\, g) &= \frac{1}{\alpha-1}\log\sum_{x\in\mathbb{X}} g(x)\left(\frac{f(x)}{g(x)}\right)^{\alpha} = \\ &= \frac{1}{\alpha-1}\log\sum_{x\in\mathbb{X}} f^{\alpha}(x)\,g^{1-\alpha}(x)\ , \ \text{para } \alpha > 0 \text{ e } \alpha \neq 1 \end{aligned} \tag{2.35}$$

A divergência de Kullback–Leibler é obtida no limite, quando $\alpha \to 1$.

$$\lim_{\alpha\to 1} D_{R_\alpha}(f \,\|\, g) = \sum_{x\in\mathbb{X}} f(x)\log\frac{f(x)}{g(x)}\,dx = D_{KL}(f \,\|\, g) \tag{2.36}$$

$D_{R_\alpha}(f \,\|\, g) \geq 0$ com igualdade se, e somente se, *f(x)* = *g(x)*, para todo $x \in \mathbb{X}$.

Definição 13a – A *divergência quadrática de Rényi* entre duas funções de massa de probabilidade *f(x)* e *g(x)* , $D_{R_2}(f \parallel g)$, é dada por

$$D_{R_2}(f \parallel g) = \log \sum_{x \in \mathbb{X}} \frac{f^2(x)}{g(x)} \tag{2.37}$$

Definição 14 – A *entropia relativa ou divergência (assimétrica) de Rényi de ordem* α entre duas densidades *f* e *g* (Neemuchwala, 2005) é definida por

$$\begin{aligned} D_{R_\alpha}(f \parallel g) &= \frac{1}{\alpha - 1} \log \int_{\mathbb{X}} g(x) \left(\frac{f(x)}{g(x)} \right)^{\alpha} dx = \\ &= \frac{1}{\alpha - 1} \log \int_{\mathbb{X}} f^{\alpha}(x)\, g^{1-\alpha}(x)\, dx, \text{ para } \alpha > 0 \text{ e } \alpha \neq 1. \end{aligned} \tag{2.38}$$

A divergência de Kullback–Leibler é obtida no limite, quando $\alpha \to 1$.

$$\lim_{\alpha \to 1} D_{R_\alpha}(f \parallel g) = \int_{\mathbb{X}} f(x) \log \frac{f(x)}{g(x)}\, dx = D_{KL}(f \parallel g) \tag{2.39}$$

$D_{R_\alpha}(f \parallel g) \geq 0$ com igualdade se, e somente se, *f = g em quase toda parte*.

Definição 14a – A *divergência quadrática de Rényi* entre duas densidades *f* e *g* , $D_{R_2}(f \parallel g)$, é dada por

$$D_{R_2}(f \parallel g) = \log \int_{\mathbb{X}} \frac{f^2(x)}{g(x)} dx \tag{2.40}$$

A divergência quadrática de Rényi não está implicitamente baseada na entropia quadrática de Rényi, pois, como pode-se observar, a divergência não é quadrática em ambas as funções (Jenssen, 2005).

A divergência de Rényi também pode ser usada como uma medida de informação mútua entre variáveis aleatórias, considerando a divergência entre a distribuição conjunta e a do produto das marginais, conforme as definições subseqüentes, considerando exclusivamente a divergência quadrática, sendo simplesmente representada por $I_R(X;Y)$.

Definição 15 – A *informação mútua de Rényi* $I_R(X;Y)$ entre duas variáveis aleatórias discretas X e Y , com função de massa de probabilidade conjunta $f_{xy}(x,y)$ e funções de massa de probabilidade marginais $f_x(x)$ e $f_y(y)$ é dada pela entropia relativa entre a distribuição conjunta e a distribuição do produto das marginais:

$$I_R(X;Y) = \log \sum_{x\in\mathbb{X}} \sum_{y\in\mathbb{Y}} \frac{f_{xy}^2(x,y)}{f_x(x)\, f_y(y)} = D_{R_2}(f_{xy}(x,y) \,\|\, f_x(x)\, f_y(y)) \geq 0 \qquad (2.41)$$

com igualdade se, e somente se, $f_{xy}(x,y) = f_x(x)\ f_y(y)$ (ou seja, se X e Y são independentes).

Definição 16 – A *informação mútua de Rényi* $I_R(X;Y)$ entre duas variáveis aleatórias contínuas X e Y , com densidade conjunta $f_{xy}(x,y)$ e densidades marginais $f_x(x)$ e $f_y(y)$ é dada por:

$$I_R(X;Y) = \log \int_{\mathbb{Y}} \int_{\mathbb{X}} \frac{f_{xy}^2(x,y)}{f_x(x)\, f_y(y)}\, dx\, dy = D_{R_2}(f_{xy}(x,y) \,\|\, f_x(x)\, f_y(y)) \geq 0 \qquad (2.42)$$

com igualdade se, e somente se, $f_{xy} = f_x\, f_y$ *em quase toda parte* (ou seja, se X e Y são independentes).

A informação mútua de Rényi, ao contrário da informação mútua de Shannon, não pode ser expressa em termos das entropias (Jenssen, 2005). Entretanto, a informação mútua de Cauchy-Schwartz, apresentada na próxima seção, pode ser expressa, como será visto, através da entropia quadrática de Rényi.

2.5
Informação Mútua de Cauchy-Schwartz

Principe et al. (2000) definiram uma medida de divergência entre funções de densidade de probabilidade (ou funções de massa de probabilidade) baseada na desigualdade de Cauchy-Schwartz entre vetores.

Assim, com base na desigualdade de Cauchy-Schwartz $\|\boldsymbol{x}\|\,\|\boldsymbol{y}\| \geq \left(\boldsymbol{x}^T \boldsymbol{y}\right)$, pode-se escrever

$$-\log \frac{\boldsymbol{x}^T \boldsymbol{y}}{\sqrt{\|\boldsymbol{x}\|^2 \|\boldsymbol{y}\|^2}} \geq 0 \tag{2.43}$$

Substituindo o produto interno entre vetores na desigualdade acima pelo produto interno entre tais funções, define-se, então, tal medida de divergência.

Definição 17– A *divergência (simétrica) de Cauchy-Schwartz* entre duas funções de massa de probabilidade *f(x)* e *g(x)* é dada por

$$D_{CS}(f \,\|\, g) = -\log \frac{\sum_{x \in \mathbb{X}} f(x) g(x)}{\sqrt{\left(\sum_{x \in \mathbb{X}} f^2(x)\right)\left(\sum_{x \in \mathbb{X}} g^2(x)\right)}} \tag{2.44}$$

$D_{CS}(f \,\|\, g) \geq 0$ com igualdade se, e somente se, *f(x)* = *g(x)*, para todo $x \in \mathbb{X}$.

Desenvolvendo a equação anterior, tem-se

$$D_{CS}(f \,\|\, g) = -\log \sum_{x \in \mathbb{X}} f(x) g(x) - \frac{1}{2}\left(-\log \sum_{x \in \mathbb{X}} f^2(x)\right) - \frac{1}{2}\left(-\log \sum_{x \in \mathbb{X}} g^2(x)\right) \tag{2.45}$$

Expressando o segundo membro da Eq. (2.45), através da entropia de Rényi, obtém-se o seguinte:

$$D_{CS}(f \,\|\, g) = h_{R_2}(f \times g) - \frac{1}{2} h_{R_2}(f) - \frac{1}{2} h_{R_2}(g) \tag{2.46}$$

onde

- $h_{R_2}(f)$ é a entropia quadrática de Rényi com respeito a *f*.
- $h_{R_2}(g)$ é a entropia quadrática de Rényi com respeito a *g*.
- $h_{R_2}(f \times g)$ pode ser interpretada como a *entropia cruzada* entre *f* e *g*.

Definição 18– A *divergência (simétrica) de Cauchy-Schwartz* entre duas densidades f e g é dada por

$$D_{CS}(f \,\|\, g) = -\log \frac{\int_{\mathbb{X}} f(x)g(x)dx}{\sqrt{\left(\int_{\mathbb{X}} f^2(x)dx\right)\left(\int_{\mathbb{X}} g^2(x)dx\right)}} \tag{2.47}$$

$D_{CS}(f \,\|\, g) \geq 0$, com igualdade se, e somente se, $f = g$ *em quase toda parte*, e as integrais envolvidas são todas formas quadráticas de funções de densidade de probabilidade.

Desenvolvendo a equação anterior, tem-se

$$D_{CS}(f \,\|\, g) = -\log \int_{\mathbb{X}} f(x)g(x)dx - \frac{1}{2}\left(-\log \int_{\mathbb{X}} f^2(x)dx\right) - \frac{1}{2}\left(-\log \int_{\mathbb{X}} g^2(x)dx\right) \tag{2.48}$$

Expressando o segundo membro da Eq. (2.48), através da entropia diferencial de Rényi, obtém-se o seguinte:

$$D_{CS}(f \,\|\, g) = h_{R_2}(f \times g) - \frac{1}{2}h_{R_2}(f) - \frac{1}{2}h_{R_2}(g) \tag{2.49}$$

onde

- $h_{R_2}(f)$ é a entropia quadrática diferencial de Rényi com respeito a *f*.
- $h_{R_2}(g)$ é a entropia quadrática diferencial de Rényi com respeito a *g*.
- $h_{R_2}(f \times g)$ pode ser interpretada como a *entropia cruzada* diferencial entre *f* e *g*.

Definição 19– A *informação mútua de Cauchy-Schwartz* $I_{CS}(X;Y)$ entre duas variáveis aleatórias discretas X e Y , com função de massa de probabilidade conjunta $f_{xy}(x,y)$ e funções de massa de probabilidade marginais $f_x(x)$ e $f_y(y)$ é dada pela divergência entre a distribuição conjunta e a distribuição do produto das marginais:

$$I_{CS}(X;Y) = h_{R_2}(f_{XY} \times f_X\, f_Y) - \frac{1}{2} h_{R_2}(f_{XY}) - \frac{1}{2} h_{R_2}(f_X\, f_Y) = \\ = D_{CS}(f_{xy} \,\|\, f_x f_y) \geq 0 \tag{2.50}$$

com igualdade se, e somente se, $f_{xy}(x,y) = f_x(x)\, f_y(y)$ (ou seja, se X e Y são independentes).

Definição 20– A *informação mútua de Cauchy-Schwartz* $I_{CS}(X;Y)$ entre duas variáveis aleatórias contínuas X e Y , com densidade conjunta $f_{xy}(x,y)$ e densidades marginais $f_x(x)$ e $f_y(y)$ é dada por:

$$I_{CS}(X;Y) = h_{R_2}(f_{XY} \times f_X\, f_Y) - \frac{1}{2} h_{R_2}(f_{XY}) - \frac{1}{2} h_{R_2}(f_X\, f_Y) = \\ = D_{CS}(f_{xy} \,\|\, f_x f_y) \geq 0 \tag{2.51}$$

com igualdade se, e somente se, $f_{xy} = f_x\ f_y$ *em quase toda parte* (ou seja, se X e Y são independentes).