Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 10, n. 1, 2023.

Trabalho apresentado no XLII CNMAC, Universidade Federal de Mato Grosso do Sul - Bonito - MS, 2023

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Geometria, Estatística e Aplicações a Comunicações e Aprendizado

Henrique K. Miyamoto[1]
IMECC/Unicamp, Campinas, SP, Brasil
L2S/CentraleSupélec/Université Paris-Saclay, Gif-sur-Yvette, França

Resumo. Esta dissertação é composta por três contribuições, que têm em comum a utilização de ferramentas de geometria e/ou estatística em aplicações a comunicações e aprendizado. A primeira trata da construção de códigos esféricos a partir de um procedimento recursivo que se baseia em folheações de esferas dadas pela fibração de Hopf. Na segunda, propomos um método de compressão vetorial com perdas, formado por um quantizador adaptável aos dados, seguido de compressão dos índices de quantização com um algoritmo de árvores de contexto. A terceira consiste em usar uma função perda baseada na distância de Fisher–Rao da variedade de distribuições discretas para o treinamento de redes neurais, particularmente sob ruído de rótulo.

Palavras-chave. Aprendizado supervisionado, compressão de dados, empacotamento de esferas, geometria da informação, teoria da informação.

1 Introdução

Desde o trabalho inaugural de Shannon de 1948, a informação produzida por uma fonte e transmitida por um canal tem sido definida em termos de sua distribuição de probabilidade, através de grandezas como entropia e informação mútua. A partir de então, ferramentas de estatística têm tido papel destacado no desenvolvimento da teoria da informação. Por outro lado, abordagens geométricas têm sido usadas especialmente no desenvolvimento de códigos que buscam se aproximar dos limitantes enunciados por Shannon. Por fim, modelos estatísticos têm sido estudados de um ponto de vista de geometria diferencial na subárea de geometria da informação.

Na dissertação [8], apresentamos três contribuições ligadas à grande área de teoria da informação, e que têm em comum a utilização de ferramentas de geometria e/ou estatística em aplicações a comunicações e aprendizado. A primeira aborda a construção de códigos esféricos, que é uma versão, em superfícies de esferas, do problema clássico de empacotamento esférico. A seguir, tratamos de compressão vetorial com perdas, em que o problema passa a ser representar, de forma econômica, uma sequência de vetores, dada uma restrição de distorção. Finalmente, estudamos a geometria de famílias de distribuições de probabilidade, munidas da estrutura riemanniana dada pela matriz de Fisher; interessamo-nos particularmente pela distância geodésica induzida em tais variedades, e por uma aplicação desta no problema de classificação supervisionada.

Neste resumo expandido, apresentamos uma visão geral de cada contribuição. Maiores detalhes podem ser encontrados na dissertação [8] e nos trabalhos [9–12].

2 Construção de Códigos Esféricos por Folheações de Hopf

Um código esférico [4] $\mathcal{C}(M, n, d) \coloneqq \{x_1, x_2, \ldots, x_M\} \subset S^{n-1}$ é um conjunto de M pontos na esfera euclidiana n-dimensional $S^{n-1} \coloneqq \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$, de modo que a distância euclidiana

[1] henrique.miyamoto@centralesupelec.fr

entre cada dois pontos seja pelo menos d. O problema de construir bons códigos é um problema de *empacotamento esférico*: dada uma distância mínima d, busca-se o maior número de pontos que pode ser distribuído em S^{n-1}, respeitando a distância mútua mínima de d. Soluções ótimas em dimensão $n = 2$ são triviais (a saber, vértices de polígonos regulares), mas poucas soluções ótimas são conhecidas em dimensões mais altas. De um ponto de vista prático, o desafio é construir códigos que não só tenham boas taxas $R = (\log M)/n$, mas que também permitam realizar facilmente codificação e decodificação. Aplicações de códigos esféricos incluem transmissão de sinais por canais gaussianos (generalização da modulação PSK) e quantização vetorial. Em particular, códigos esféricos têm sido recentemente estudados para o projeto de constelações em comunicações óticas e sem-fio.

A fibração de Hopf [7] é a submersão $h : S^{2n-1} \to S^n$ dada por $(z_0, z_1) \mapsto \left(2z_0\overline{z_1}, |z_0|^2 - |z_1|^2\right)$, em que z_0 e z_1 são elementos de $\mathbb{R}$, $\mathbb{C}$, $\mathbb{H}$ ou $\mathbb{O}$, para $n \in \{1, 2, 4, 8\}$, respectivamente. Esse mapa mune a esfera S^{2n-1} de uma estrutura de fibrado vetorial $S^{n-1} \hookrightarrow S^{2n-1} \to S^n$, permitindo descrevê-la como uma folheação por variedades produto $T^{2n-2}_\eta \coloneqq S^{n-1}_{\cos\eta} \times S^{n-1}_{\sin\eta}$, $\eta \in [0, \pi/2]$. É possível estender essas folheações para qualquer $n \in \mathbb{N}^*$, a que chamamos *folheações de Hopf*.

Em [9], propomos um procedimento recursivo para construção de códigos esféricos por folheações de Hopf (SCHF) em dimensão 2^k, para qualquer distância mínima d, a partir de uma construção base em dimensão 4. Nessa dimensão, a esfera S^3 é folheada por toros planares T^2_η usando o mapa $\iota\colon (\eta; \xi_1, \xi_2) \mapsto \left(e^{i\xi_1}\cos\eta, e^{i\xi_2}\sin\eta\right) \in \mathbb{C}^2 \cong \mathbb{R}^4$, $\eta \in [0, \pi/2]$, $\xi_1, \xi_2 \in [0, 2\pi[$. Para construir códigos nessa dimensão: 1) escolhemos uma família de toros $\{T^2_\eta\}_{\eta\in H}$ mutuamente distantes de d; 2) em cada toro T^2_η, escolhemos n círculos internos (i.e., imagens por ι para η e ξ_2 fixos) separados de $\Delta\xi_2$ e m pontos separados por $\Delta\xi_1$ em cada círculo. Para construir códigos em S^{2n-1}: 1) escolhemos uma família de folhas $\{T^{2n-2}_\eta\}_{\eta\in H}$ mutuamente distantes de d; 2) em cada folha $T^{2n-2}_\eta = S^{n-1}_{\cos\eta} \times S^{n-1}_{\sin\eta} \subset S^{2n-2}$, aplicamos a distribuição da dimensão anterior a cada esfera $S^{n-1}_{\cos\eta}$, $S^{n-1}_{\sin\eta}$, com distância mínima escalada $d/\cos\eta$ e $d/\sin\eta$, respectivamente. Esse procedimento recursivo permite construir códigos em dimensões da forma $n = 2^k$.

Resultados numéricos mostram que as taxas obtidas com essa construção superam outros métodos conhecidos na literatura, em diferentes regimes de dimensão e distância mínima. Calculamos limitantes assintóticos para a densidade dos códigos SCHF, e argumentamos que nossa construção oferece um compromisso entre boas taxas e construtibilidade efetiva. Ainda, explicitamos procedimentos de baixa complexidade para codificação e decodificação dos nossos códigos. Verificamos experimentalmente que a decodificação proposta tem boa performance sob ruído gaussiano.

3 Compressão com Perdas Baseada em Árvores de Contexto

Diferentes tarefas em comunicações, como *feedback* e armazenamento podem ser modeladas como um problema de *compressão com perdas* [3], nos quais busca-se representar vetores de informação de forma econômica, i.e., com a menor quantidade de bits possível, para uma dada tolerância de distorção. Uma forma de abordar o problema é quantizar vetores de informação com um alfabeto finito e, a seguir, comprimir a sequência de índices de quantização com um compressor universal, i.e., que não depende da estatística da fonte (e.g., LZ, CTW). No entanto, a aplicação direta desses métodos apresenta inconvenientes: as sequências de saída têm comprimento variável e podem não ser imediatamente produzidas, o que não é adaptado para aplicações sensíveis a atraso. Por isso, propomos em [12] uma solução em dois passos, projetada particularmente para a compressão em tempo real de vetores de informação de estado do canal (CSI) em comunicações sem-fio.

O primeiro passo é realizar quantização (com perdas) dos vetores de CSI. Cada componente real do vetor é normalizada e quantizada separadamente com *companders* adaptados à sua distribuição. A ideia de um *compander* é aplicar uma transformação paramétrica não-linear (equivalente a uma

função distribuição acumulada) $g_\theta \colon [0,1] \to [0,1]$ que "uniformize" a distribuição de uma variável aleatória $X \in [0,1]$. Em seguida, aplicamos quantização uniforme a $g_\theta(X)$, já que esta é ótima para distribuição uniforme. Para encontrar o melhor *compander* parametrizado por $\theta \in \Theta$ resolvemos numericamente $\arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log g'_\theta(x_i)$ para uma amostra $\{x_i\}_{i=1}^n$ de componentes do vetor, o que é equivalente a minimizar a versão empírica da divergência de Kullback-Leibler entre a distribuição teórica dos dados e aquela definida pelo *compander*. Além do conhecido μ-*compander*, propomos um novo β-*compander*, inspirado pela distribuição beta.

O segundo passo trata de compressão sem perdas das sequências de índices de quantização. Para isso, supomos que essa sequência é gerada por um processo na classe das *cadeias de Markov de ordem variável* no máximo D [5]. Note que qualquer processo ergódico e estacionário pode ser aproximado por um processo nessa classe quando $D \to \infty$. Um compressor universal para essa classe consiste em usar o algoritmo CTW [14] para estimar, de modo eficiente, a distribuição de uma sequência como uma mistura de modelos dentro da classe, e usá-la como distribuição de codificação para codificação aritmética. Propomos uma modificação desse método: usando o algoritmo CTM [15] (variação do CTW), calculamos o modelo de máximo *a posteriori* dentro da classe, para a sequência que é observada. Os símbolos da sequência são então codificados com uma regra fixa a partir das probabilidades marginais estimadas pelo modelo MAP, o que limita a varição do comprimento das sequências de bits de saída.

Simulamos vetores de CSI para canais LTE, em diferentes cenários de mobilidade e correlação de antenas, e estudamos o desempenho do método proposto em termos de distorção e taxa de comunicação, em função do número de bits necessários para representar a sequência. Os resultados numéricos atestam a eficiência dos métodos de quantização e compressão propostos. É ainda importante destacar que esses métodos têm baixa complexidade computacional, podem ser implementados em tempo real e são modulares, i.e., podem ser combinados com outros quantizadores ou compressores à disposição.

4 Geometria da Informação e Aprendizado

4.1 Formas Fechadas para a Distância de Fisher–Rao

A subárea de geometria da informação [1, 2] estuda a geometria intrínseca de famílias de distribuições de probabilidade, vistas como variedades riemannianas. Um *modelo estatístico* $\mathcal{S} \coloneqq \left\{p_\xi = p(x;\xi) : \xi = (\xi^1,\ldots,\xi^n) \in \Xi \subseteq \mathbb{R}^n\right\}$ é uma família de distribuições de probabilidade parametrizadas pelo vetor $\xi = \left(\xi^1,\ldots,\xi^n\right)$, de sorte que a aplicação (parametrização) $\varphi \colon \xi \mapsto p_\xi$ seja injetiva e Ξ seja aberto em $\mathbb{R}^n$. Se $\mathcal{S}$ é suavemente parametrizado por Ξ e satisfaz certas condições de regularidade [1], torna-se uma variedade diferenciável chamada *variedade estatística*. É possível ainda munir $\mathcal{S}$ de uma estrutura riemanniana. Denotando $\partial_i \coloneqq \frac{\partial}{\partial \xi^i}$, os elementos da chamada *matriz de Fisher* $G(\xi) = [g_{ij}(\xi)]_{i,j}$ são dados por

$$g_{ij}(\xi) \coloneqq \mathbb{E}_{p_\xi}\left[\left(\partial_i \log p_\xi(x)\right)\left(\partial_j \log p_\xi(x)\right)\right]. \tag{1}$$

Como a matriz de Fisher é simétrica e definida-positiva, define uma métrica riemanniana, i.e., uma família de produtos internos que varia suavemente na variedade estatística, a chamada *métrica de Fisher*. Aplicar a métrica de Fisher g_ξ a dois vetores $v_1 = \mathrm{d}\varphi(\xi_1)$, $v_2 = \mathrm{d}\varphi(\xi_2)$ no espaço tangente $T_{p_\xi}\mathcal{S}$ equivale a calcular um produto interno mediado pela matriz $G(\xi)$ entre os vetores em coordenadas locais: $\langle v_1, v_2\rangle_{G(\xi)} \coloneqq g_\xi(v_1,v_2) = \xi_1^{\mathsf{T}} G(\xi)\xi_2$. A métrica de Fisher não só é invariante por reparametrização do espaço amostral e covariante por reparametrização do espaço de parâmetros [2], como também é a única métrica riemanniana (a menos de fator de escala) invariante por estatística suficiente [13]. Essas propriedades tornam-na a escolha natural para estudar a geometria de variedades estatísticas.

DOI: 10.5540/03.2023.010.01.0004

Considere uma curva $\xi\colon [0,1] \to \Xi$ no espaço de parâmetros e sua imagem $\gamma\colon [0,1] \to \mathcal{S}$ pela parametrização $\varphi\colon \Xi \to \mathcal{S}$, i.e., $\gamma = \varphi \circ \xi$. Na geometria de Fisher, o comprimento $l(\gamma)$ da curva γ pode ser calculado como

$$l(\gamma) \coloneqq \int_0^1 \sqrt{\langle \dot{\gamma}(t), \dot{\gamma}(t) \rangle_{G(\xi(t))}}\, \mathrm{d}t = \int_0^1 \sqrt{\dot{\xi}(t)^\mathsf{T} G(\xi(t)) \dot{\xi}(t)}\, \mathrm{d}t. \tag{2}$$

Dadas duas distribuições p_{ξ_1} e p_{ξ_2} em $\mathcal{S}$, o ínfimo do comprimento das curvas γ diferenciáveis por partes que as une define uma distância entre essas distribuições, chamada *distância de Fisher–Rao*:

$$d_{\mathrm{FR}}(\xi_1, \xi_2) \coloneqq d_{\mathrm{FR}}(p_{\xi_1}, p_{\xi_2}) \coloneqq \inf_{\gamma} \left\{ l(\gamma) \,:\, \gamma(0) = p_{\xi_1},\ \gamma(1) = p_{\xi_2} \right\}. \tag{3}$$

O teorema de Hopf–Rinow garante que se $(\mathcal{S}, d_{\mathrm{FR}})$ for conexo e completo como espaço métrico, então quaisquer dois pontos $p, q \in \mathcal{S}$ podem ser unidos por uma geodésica minimizante, i.e., uma curva cujo comprimento é igual à distância $d_{\mathrm{FR}}(p, q)$ e que é uma geodésica.

Infelizmente, calcular a distância de Fisher–Rao, em geral, não é uma tarefa trivial, uma vez que envolve encontrar as geodésicas na variedade estatística de interesse (e.g., resolvendo o sistema de equações diferenciais geodésicas) e avaliar a integral em (2). Por isso, formas fechadas são conhecidas apenas em algumas famílias de distribuições de probabilidade. Uma primeira contribuição [10] neste tema foi realizar uma curadoria de formas fechadas de distâncias de Fisher–Rao presentes na literatura, bem como aumentar o repertório com a introdução de novos exemplos. Alguns exemplos são apresentados na Tabela 1.

Tabela 1: Exemplos de distâncias de Fisher–Rao para distribuições contínuas.

	Distribuição	Parâmetros	Distância de Fisher–Rao
Rayleigh	$\frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right)$	$\sigma \in \mathbb{R}_+^*$	$2\left\lvert \log \sigma_1 - \log \sigma_2 \right\rvert$
Gaussiana	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+^*$	$2\sqrt{2}\,\operatorname{arctanh}\left(\sqrt{\frac{(\mu_1-\mu_2)^2+2(\sigma_1-\sigma_2)^2}{(\mu_1-\mu_2)^2+2(\sigma_1+\sigma_2)^2}}\right)$
Laplace	$\frac{1}{2b} \exp\left(-\frac{\lvert x-\mu \rvert}{b}\right)$	$(\mu, b) \in \mathbb{R} \times \mathbb{R}_+^*$	$2\operatorname{arctanh}\left(\sqrt{\frac{(\mu_1-\mu_2)^2+(b_1-b_2)^2}{(\mu_1-\mu_2)^2+(b_1+b_2)^2}}\right)$
Cauchy	$\frac{\gamma}{\pi[(x-x_0)^2+\gamma^2]}$	$(x_0, \gamma) \in \mathbb{R} \times \mathbb{R}_+^*$	$\sqrt{2}\operatorname{arctanh}\left(\sqrt{\frac{(x_{0,1}-x_{0,2})^2+(\gamma_1-\gamma_2)^2}{(x_{0,1}-x_{0,2})^2+(\gamma_1+\gamma_2)^2}}\right)$

No caso da distribuição categórica $p(x) = \sum_{i=1}^n p_i \mathbb{1}_{\{i\}}(x)$, definida para $x \in \{1, \ldots, n\}$, a variedade estatística respectiva é isomorfa ao simplexo $\Delta^{n-1} = \{(p_1, \ldots, p_n) \in [0,1]^n \,:\, \sum_{i=1}^n p_i = 1\}$, e ambos podem ser parametrizados pelas primeiras $(n-1)$ coordenadas $(p_1, \ldots, p_{n-1})$, tomando $p_n = 1 - \sum_{i=1}^{n-1} p_i$. Usando um argumento clássico [2], pode-se considerar a reparametrização $p_i \mapsto 2\sqrt{p_i}$ que leva o simplexo na esfera euclidiana de raio 2 e mostrar que a geometria dessa variedade estatística coincide com a geometria esférica. Nesse caso, a distância de Fisher–Rao é

$$d_{\mathrm{FR}}(p, q) = 2 \arccos\left(\sum_{i=1}^n \sqrt{p_i q_i} \right). \tag{4}$$

4.2 Aprendizado com a Função Perda de Fisher–Rao

O problema de classificação consiste em atribuir a cada entrada $\boldsymbol{x} \in \mathcal{X}$ a sua classe correspondente dentre $\mathcal{Y} = \{1, \ldots, K\}$. Um classificador é uma função $f\colon \mathcal{X} \to \mathbb{R}^K$ (e.g., rede neural),

que atribui à entrada $\boldsymbol{x}$ um vetor de probabilidades $\boldsymbol{p} = (p_1, \ldots, p_K) \coloneqq f(\boldsymbol{x})$, a partir do qual é possível escolher $\hat{y} = \arg\max_{1 \le i \le K} p_i$. O objetivo é encontrar (aprender, treinar) um classificador que cometa o menor erro possível, medido por uma *função perda* $L\colon \mathcal{Y} \times \mathbb{R}^K \to \mathbb{R}_+$. No caso de aprendizado supervisionado, temos acesso a um conjunto de treinamento $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, que pode ser usado para encontrar o classificador (paramétrico) $f \in \mathcal{F}$ por minimização do risco empírico, i.e., resolvendo $\arg\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^N L(y_i, f(\boldsymbol{x}_i))$, o que é feito por um algoritmo numérico de otimização, como método do gradiente.

Escolher uma função perda adequada é fundamental, pois afeta diretamente o desempenho do classificador resultante, assim como a dinâmica de aprendizado. Escolhas usuais incluem o erro quadrático médio (MSE) $L_{\mathrm{MSE}}(y, f(\boldsymbol{x})) = \|\boldsymbol{p}\|_2^2 - 2p_y + 1$, o erro médio absoluto (MAE) $L_{\mathrm{MAE}}(y, f(\boldsymbol{x})) = 1 - p_y$ e a entropia cruzada (CE) $L_{\mathrm{CE}}(y, f(\boldsymbol{x})) = -\log p_y$. Em [11], propomos e estudamos uma função perda baseada no quadrado da distância de Fisher–Rao (4) entre a saída do classificador $f(\boldsymbol{x})$ e o vetor de probabilidades ideal (que atribui 1 à classe correta y e 0 a todas as outras), dada por $L_{\mathrm{FR}}(y, f(\boldsymbol{x})) = \left(\arccos \sqrt{p_y}\right)^2$.

Consideramos em particular o problema de *ruído de rótulo* [6], i.e., quando o conjunto de treinamento *limpo* é substituído por uma versão *ruidosa* $\{(\boldsymbol{x}_i, \tilde{y}_i)\}_{i=1}^N$ em que alguns rótulos de classe $\tilde{y}_i$ estão incorretos ($\tilde{y}_i \neq y_i$). Uma maneira conveniente de lidar com esse tipo de problema é usar no treinamento uma função perda que seja inerentemente robusta a ruído de rótulo, i.e., tal que o desempenho do classificador resultante treinado com dados ruidosos seja tão bom quanto o de um classificador treinado com dados limpos. É possível mostrar que a perda MAE é robusta a ruído de rótulo uniforme, enquanto que a perda CE não o é de todo [6]. Por outro lado, a forma da perda MAE faz com que usá-la resulte em um treinamento lento, ao passo que o treinamento com a perda CE converge mais rapidamente.

Mostramos que a perda de Fisher–Rao oferece uma solução de compromisso entre robustez a ruído de rótulo e velocidade de aprendizado: provamos um limitante para o quanto o ruído de rótulo pode afetar o treinamento com a perda de Fisher–Rao [11, Prop. 3], enquanto que, analisando seu gradiente, observamos que ela fornece uma velocidade de treinamento intermediária entre as perdas MAE e CE. Em resultados numéricos com dados sintéticos e com o conjunto de dados MNIST, observamos que, sob ruído de rótulo, a perda de Fisher–Rao apresenta desempenho superior, sem prejuízo à velocidade de treinamento.

Agradecimentos

Agradeço à Profa. Sueli Costa pela orientação ao longo do mestrado. Este trabalho teve apoio do CNPq (processo 131387/2021-9) e da FAPESP (processo 2021/04516-8).

Referências

[1] S. Amari e H. Nagaoka. **Methods of Information Geometry**. Providence, RI: American Mathematical Society, 2000. ISBN: 978-0-8218-4302-4.

[2] O. Calin e C. Udrişte. **Geometric Modeling in Probability and Statistics**. Cham: Springer, 2014. ISBN: 978-3-319-07778-9.

[3] T. M. Cover e J. A. Thomas. **Elements of Information Theory**. 2nd. Hoboken, NJ: Wiley, 2006. ISBN: 978-0-4712-4195-9.

[4] T. Ericson e V. Zinoviev. **Codes on Euclidean Spheres**. Amsterdam: North-Holland, 2001. ISBN: 0-444-50329-3.

[5] E. Gassiat. **Universal Coding and Order Identification by Model Selection Mehtods**. Cham: Springer, 2018. ISBN: 978-3-0300-7167-7.

[6] A. Ghosh, H. Kumar e P. S. Sastry. "Robust loss functions under label noise for deep neural networks". Em: **Proceedings of the 31st AAAI Conference on Artificial Intelligence**. 2017, pp. 1919–1925.

[7] D. W. Lyons. "An elementary introduction to the Hopf fibration". Em: **Mathematics Magazine** 76.2 (2003), pp. 87–98. DOI: 10.2307/3219300.

[8] H. K. Miyamoto. "Geometria, estatística e aplicações a comunicações e aprendizado". Dissertação de mestrado. Universidade Estadual de Campinas, 2022. URL: https://hdl.handle.net/20.500.12733/6633.

[9] H. K. Miyamoto, S. I. R. Costa e H. N. Sá Earp. "Constructive spherical codes by Hopf foliations". Em: **IEEE Transactions on Information Theory** 67.12 (2021), pp. 7925–7939. DOI: 10.1109/TIT.2021.3114094.

[10] H. K. Miyamoto, F. C. C. Meneghetti e S. I. R. Costa. "On closed-form expressions for the Fisher–Rao distance". Em: **arXiv Preprints** (2023). DOI: 10.48550/arXiv.2304.14885.

[11] H. K. Miyamoto, F. C. C. Meneghetti e S. I. R. Costa. "The Fisher–Rao loss for learning under label noise". Em: **Information Geometry** 6 (2023), pp. 107–126. DOI: 10.1007/s41884-022-00076-8.

[12] H. K. Miyamoto e S. Yang. "Context-tree-based lossy compression and its application to CSI representation". Em: **IEEE Transactions on Communications** 70.7 (2022), pp. 4417–4428. DOI: 10.1109/TCOMM.2022.3173002.

[13] H. Vân Lê. "The uniqueness of the Fisher metric as information metric". Em: **Annals of the Institute of Statistical Mathematics** 69 (2017), pp. 879–896. DOI: 10.1007/s10463-016-0562-0.

[14] F. M. J. Willems, Y. M. Shtarkov e T. J. Tjalkens. "The context-tree weighting method: basic properties". Em: **IEEE Transactions on Information Theory** 41.3 (1995), pp. 653–664. DOI: 10.1109/18.382012.

[15] F. M. J. Willems, T. J. Tjalkens e Y. M. Shtarkov. "Context-tree maximizing". Em: **Proceedings of the 34th Annual Conference on Information Sciences and Systems**. 2000, TP6-7–TP6-12.

