Universidade Estadual de Campinas
Instituto de Matemática, Estatística
e Computação Científica

Henrique Koji Miyamoto

Geometria, Estatística e Aplicações a Comunicações e Aprendizado

Campinas

2022

Henrique Koji Miyamoto

Geometria, Estatística e Aplicações a Comunicações e Aprendizado

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Matemática Aplicada.

Orientadora: Sueli Irene Rodrigues Costa

Este trabalho corresponde à versão final da Dissertação defendida pelo aluno Henrique Koji Miyamoto e orientada pela Profa. Dra. Sueli Irene Rodrigues Costa.

Campinas

2022

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

M699g

Miyamoto, Henrique Koji, 1996-
Geometria, estatística e aplicações a comunicações e aprendizado / Henrique Koji Miyamoto. – Campinas, SP : [s.n.], 2022.

Orientador: Sueli Irene Rodrigues Costa.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Aprendizagem supervisionada (Aprendizado do computador). 2. Compressão de dados (Computação). 3. Empacotamento de esferas. 4. Geometria da informação. 5. Teoria da informação. I. Costa, Sueli Irene Rodrigues. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Informações Complementares

Título em outro idioma: Geometry, statistics and applications to communications and learning
Palavras-chave em inglês:
Supervised learning (Machine learning)
Data compression (Computer science)
Sphere packings
Information geometry
Information theory
Área de concentração: Matemática Aplicada
Titulação: Mestre em Matemática Aplicada
Banca examinadora:
Sueli Irene Rodrigues Costa [Orientador]
João Eloir Strapasson
Charles Casimiro Cavalcante
Data de defesa: 14-09-2022
Programa de Pós-Graduação: Matemática Aplicada

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0003-1131-2790
- Currículo Lattes do autor: http://lattes.cnpq.br/5273967542101123

Dissertação de Mestrado defendida em 14 de setembro de 2022 e aprovada pela banca examinadora composta pelos Profs. Drs.

Prof(a). Dr(a). SUELI IRENE RODRIGUES COSTA

Prof(a). Dr(a). JOÃO ELOIR STRAPASSON

Prof(a). Dr(a). CHARLES CASIMIRO CAVALCANTE

A Ata da Defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria de Pós-Graduação do Instituto de Matemática, Estatística e Computação Científica.

À memória do meu avô Takehiro Miyamoto.

Agradecimentos

Trilhar um mestrado não é um caminho que se faça sozinho. Por isso, agradeço, em primeiro lugar, à Profa. Sueli Costa por me guiar nesta trajetória. Sua orientação dedicada e olhar cuidadoso são exemplo para nós.

Agradeço aos colegas do LMDC pela companhia sempre espirituosa ao longo do percurso: Ana, Fábio, Franciele, Juliana, Makson, Marcelo e Maruan, estendendo ao Luiz, vizinho que se tornou membro honorário. Agradecimentos especiais ao Fábio e ao Luiz, pelas frutíferas discussões e colaborações.

E à minha família, pelo apoio incondicional para que eu continuasse nesta aventura.

* * *

Sou muito feliz de ter realizado este mestrado na Unicamp, que considero minha verdadeira *alma mater*. Viva a universidade pública brasileira!

* * *

Este trabalho foi realizado com o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), processo 131387/2021-9 (08/2021 a 11/2021), e da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), processo 2021/04516-8 (12/2021 a 09/2022), o que foi um privilégio em tristes tempos de desvalorização da ciência no Brasil.

"É preciso estar atento e forte
Não temos tempo de temer a morte"
(Caetano Veloso e Gilberto Gil)

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.

Abstract

This dissertation is composed of three contributions, which have in common the use of tools from geometry and/or statistics in applications to communications and learning. The first of them concerns the construction of spherical codes from a recursive procedure based on the sphere foliations given by the Hopf fibration. In the second one, we propose a method for lossy vector compression, formed by a data-adapted quantiser, followed by compression of the quantisation indices with a context-tree algorithm. The third consists in using a loss function based on the Fisher-Rao distance in the manifold of discrete distributions for training neural networks, particularly under label noise.

Keywords: data compression, information geometry, information theory, spherical packings, supervised learning.

Lista de Símbolos

$\wp(X)$	conjunto das partes do conjunto X
$\mathbb{C}$	conjunto dos números complexos
$\mathbb{N}$	conjunto dos números naturais, i.e., $\{0, 1, 2, \dots\}$
$\mathbb{N}^*$	conjunto dos números naturais positivos, i.e., $\{1, 2, 3, \dots\}$
$\mathbb{R}$	conjunto dos números reais
$\mathbb{R}_+$	conjunto dos números reais não-negativos
$\mathbb{R}_+^*$	conjunto dos números reais positivos
$\mathbb{O}$	conjunto dos octônios
$\mathbb{H}$	conjunto dos quatérnios
δ_{ij}	delta de Kronecker, i.e., a função que vale 1 se $i = j$, e 0 se $i \neq j$
$\mathbb{E}[\cdot]$	esperança matemática de uma variável aleatória
$\mathbb{1}_A(x)$	função indicadora, que vale 1 se $x \in A$, e 0 se $x \notin A$
i.i.d.	independentes e identicamente distribuídos
Δ^{n-1}	simplexo n-dimensional, i.e., $\left\{(x_1, \dots, x_n) \in \mathbb{R}^n \;\middle|\; x_i \geq 0,\ \sum_{i=1}^n x_i = 1\right\}$
$\bigsqcup$	união disjunta

Sumário

Capítulo 1

Introdução

Desde o trabalho inaugural de Shannon [47], publicado em 1948, a informação produzida por uma fonte tem sido definida em termos da sua distribuição de probabilidades, através de quantidades como entropia e informação mútua [17]. Também desde a publicação daquele trabalho, que estabeleceu os limites fundamentais da comunicação, iniciou-se uma corrida para encontrar códigos que pudessem deles se aproximar, incluindo abordagens geométricas [15, 21]. Nesta dissertação, 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 trata de um problema de geometria discreta com aplicações a comunicação, a saber, a construção de códigos esféricos, que pode ser visto como o problema clássico do empacotamento de esferas, na superfície de esferas em dimensão n. De um ponto de vista prático, o objetivo é não só alocar uma grande quantidade de pontos mutuamente distantes de uma certa distância, mas fazê-lo de maneira eficiente, i.e., com baixa complexidade computacional. Apresentamos um procedimento recursivo, baseado em folheações de esferas inspiradas pela fibração de Hopf, a partir de uma construção base em dimensão 4, e mostramos que essa solução constitui um compromisso entre boas taxas e construtibilidade efetiva com baixa complexidade computacional. Essa contribuição é apresentada no Capítulo 2.

A seguir, no Capítulo 3, tratamos do problema de compressão vetorial com perdas. Nesse caso, o problema passa a ser representar, da forma mais econômica possível, uma sequência de vetores, para uma dada restrição de distorção considerada aceitável. Apresentamos uma solução em dois passos, tendo em vista particularmente a aplicação de representação de vetores de informação do estado do canal (CSI) em comunicações sem-fio. O primeiro passo, com perdas, é formado por um quantizador baseado em *companders* paramétricos, adaptáveis à distribuição dos dados. O segundo passo é sem perdas e envolve a compressão dos índices de quantização usando um algoritmo baseado em árvores de contexto, que modelam cadeias de Markov de memória finita. Realizamos simulações em canais LTE para mostrar a efetividade

da abordagem proposta. Nesse caso, ferramentas de estatística são mobilizadas para resolver um problema de comunicações.

Finalmente, na terceira contribuição, usamos uma abordagem de geometria da informação, que é a subárea que estuda modelos estatísticos, i.e., famílias de distribuições de probabilidade, com ferramentas da geometria diferencial. Ao considerar tais famílias como variedades riemannianas, munidas da métrica de Fisher, é possível atribuir uma noção de distância geodésica entre duas distribuições da mesma família, a chamada distância de Fisher-Rao. No Capítulo 4, após introduzir os conceitos fundamentais dessa subárea, apresentamos exemplos explícitos de distâncias de Fisher-Rao. Além de colecionar exemplos apresentados na literatura, derivamos também novos exemplos, que não estão presentes nas referências clássicas. A seguir, interessamo-nos particularmente pela distância de Fisher-Rao das distribuições discretas, e estudamos uma função perda para aprendizado supervisionado baseada nela. Através da derivação de resultados teóricos e análise de resultados experimentais, comparamos o aprendizado de redes neurais usando essa função com outras funções perda usuais, e verificamos que nossa proposta constitui um compromisso natural entre velocidade de aprendizado e robustez a ruído de rótulo.

Esta dissertação está organizada em um formato misto. Visto que as duas primeiras contribuições se referem a trabalhos já publicados [41, 43], os Capítulos 2 e 3 apresentam uma visão geral de cada contribuição, e uma versão de cada artigo é reproduzida nos Anexos A e B. Já o Capítulo 4 apresenta em detalhes a terceira e mais recente contribuição. Alguns pré-requisitos associados são resumidos nos Apêndices A, B e C.

Capítulo 2

Construção de Códigos Esféricos por Folheações de Hopf

2.1 Visão geral

Códigos esféricos [21] são conjuntos discretos de pontos na superfície da esfera euclidiana n-dimensional $S^{n-1} \subset \mathbb{R}^n$, e problemas relacionados a eles consistem em encontrar boas distribuições de pontos com relação a algum parâmetro de interesse, como a distância mínima entre dois pontos. Por exemplo, o problema do *empacotamento esférico* em tais esferas pode ser formulado da seguinte forma: dada uma distância mínima $d > 0$, queremos encontrar o maior número de pontos que pode ser distribuído na esfera S^{n-1} de modo que a distância entre dois deles seja pelo menos 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, mas também permitam realizar facilmente codificação e decodificação. Aplicações clássicas de códigos esféricos em comunicações incluem a transmissão de sinais por canais gaussianos, como generalização da modulação PSK, e compressão de fonte, como quantizadores vetoriais. Por exemplo, códigos esféricos têm sido estudados para o projeto de constelações nos contextos de comunicações óticas e sem-fio.

Por outro lado, a fibração de Hopf [39] é a submersão $h : S^{2n-1} \to S^n$, definida por $(z_0, z_1) \mapsto \big(2z_0\overline{z_1}, |z_0|^2 - |z_1|^2\big)$, onde z_0 e z_1 são elementos de uma das álgebras de divisão normada: $\mathbb{R}$, $\mathbb{C}$, $\mathbb{H}$ ou $\mathbb{O}$, para $n \in \{1, 2, 4, 8\}$. Esse mapa mune a esfera S^{2n-1} de uma estrutura de fibrado $S^{n-1} \hookrightarrow S^{2n-1} \to S^n$ e permite descrevê-la como uma folheação por variedades $S^{n-1}_{\cos\eta} \times S^{n-1}_{\sin\eta}$, $\eta \in \big[0, \pi/2\big]$. De fato, é possível estender essa folheação de esferas S^{2n-1} para qualquer $n \in \mathbb{N}$, ao que chamamos *folheações de Hopf*.

Em [41], apresentamos uma construção de códigos esféricos baseada nas folheações de Hopf de esferas em dimensão 2^k, que consiste em um procedimento recursivo a partir de uma cons-

trução base em $S^3 \subset \mathbb{R}^4$. Os resultados numéricos mostram que nossa construção supera outros métodos conhecidos na literatura em termos de cardinalidade, em diferentes dimensões, e para vários regimes de distâncias mínimas. Calculamos limitantes assintóticos para a densidade dos códigos construídos, e argumentamos que nossa construção oferece um compromisso entre boas taxas e construtibilidade efetiva, para uma ampla faixa de valores de distância mínima e várias dimensões. Ainda, explicitamos um procedimento de codificação para nossos códigos com baixa complexidade temporal e de armazenamento. Por fim, propomos um algoritmo de decodificação sub-ótimo, porém eficiente, que não requer o armazenamento de todo o livro de códigos e tem baixa complexidade computacional. Verificamos experimentalmente que este método tem boa performance de decodificação para ruído gaussiano, ao passo que evita o alto custo computacional de um decodificador de máxima verossimilhança por força bruta.

2.2 Referência da contribuição

Esta contribuição [41] foi publicada em:

> H. K. Miyamoto, S. I. R. Costa e H. N. Sá Earp. Constructive Spherical Codes by Hopf Foliations. *IEEE Transactions on Information Theory*, vol. 67, n. 12, p. 7925–7939, dez. 2021, doi: 10.1109/TIT.2021.3114094.

Uma cópia da versão aceita do artigo é reproduzida no Anexo A.

Capítulo 3

Compressão com Perdas Baseada em Árvores de Contexto

3.1 Visão geral

Para diferentes tarefas em comunicações, como *feedback* e armazenamento, pode ser necessário representar vetores de informação de forma econômica, i.e., usando tão poucos bits quanto possível, para uma dada tolerância de distorção. Este é o problema de *compressão com perdas*, e o compromisso fundamental entre taxa da sequência codificada e distorção é conhecido para processos estacionários [17]. Uma forma de abordar o problema é quantizar os vetores de informação usando algum alfabeto finito e, em seguida, comprimir a sequência de índices de quantização. Se esta sequência for estacionária, pode ser comprimida sem perdas até uma taxa de bits tão baixa quanto a taxa de entropia do processo subjacente; para tanto, podem-se aplicar algoritmos de compressão universal, i.e., que não dependem do conhecimento da estatística da fonte, como os de Lempel-Ziv e *context-tree weighting* (CTW). No entanto, a aplicação direta desses métodos apresenta inconvenientes, como o fato de as sequências de saída serem de comprimento variável e poderem não ser imediatamente produzidas, o que não é adaptado para aplicações em tempo real, sensíveis a atrasos.

Para contornar tais dificuldades, propomos em [43] 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 com perdas: normalizamos os vetores (complexos) de CSI e quantizamos suas componentes de amplitude e fase separadamente, usando *companders* paramétricos adaptados à distribuição dos vetores de informação, seguidos de quantização uniforme. Além do conhecido μ-*compander*, propomos um novo β-*compander*, inspirado na distribuição beta. O segundo passo é sem perdas: as sequências de índices de quantização são comprimidas usando a distribuição condicional estimada pelo modelo de máximo *a posteriori* de cadeias de Markov de ordem limitada. Isso pode ser feito de modo eficiente a

partir de uma modificação do algoritmo CTW [53].

Realizamos simulações de canais LTE, em diferentes cenários de mobilidade e correlação de antenas, nos quais analisamos o desempenho do método proposto para compressão de vetores de CSI. Observamos que os quantizadores propostos têm melhor desempenho em termos de taxa *versus* distorção que outro método do estado da arte, e que, em conjunto com o método de compressão proposto, apresenta ganhos interessantes em relação à transmissão sem compressão. Finalmente, estudamos também a taxa de comunicação atingível por diferentes métodos e notamos que o limitante superior para esta taxa é atingido mais rapidamente quando os métodos propostos são utilizados.

3.2 Referência da contribuição

Esta contribuição [43] foi publicada em:

> H. K. Miyamoto e S. Yang. Context-Tree-Based Lossy Compression and Its Application to CSI Representation. *IEEE Transactions on Communications*, vol. 70, n. 7, p. 4417–4428, jul. 2022, doi: 10.1109/TCOMM.2022.3173002.

Uma cópia da versão aceita do artigo é reproduzida no Anexo B.

Capítulo 4

Geometria da Informação e Aprendizado

4.1 Introdução

A subárea de geometria da informação estuda a geometria intrínseca de famílias de distribuições de probabilidade, vistas como variedades riemannianas a que chamamos *variedades estatísticas* [1, 2, 5, 11]. Em famílias de distribuições paramétricas, a chamada *métrica de Fisher*, dada pela matriz de informação de Fisher, é essencialmente a única métrica riemanniana invariante por estatísticas suficientes, o que a torna a escolha natural de métrica em tais variedades. Em particular, isso torna possível calcular a *distância de Fisher-Rao* entre duas distribuições, que é a distância geodésica entre duas distribuições da mesma família, na variedade estatística correspondente. Resultados básicos dessa área são recapitulados na Seção 4.2.

Infelizmente, calcular a distância de Fisher-Rao entre duas distribuições pode ser uma tarefa não-trivial. De fato, expressões fechadas para essa distância são conhecidas apenas para algumas famílias de distribuições. Na Seção 4.3 colecionamos exemplos explícitos da literatura e também apresentamos exemplos que não são apresentados na literatura clássica. É interessante notar que alguns exemplos recuperam geometrias conhecidas, como hiperbólica e esférica.

É bem sabido que a geometria das variedades das distribuições discretas coincide com a geometria esférica, o que permite calcular facilmente a distância de Fisher-Rao entre duas distribuições dessa família. Na Seção 4.4, exploramos esse fato para propor uma função perda para treinamento de redes neurais, em problemas de classificação, que é proporcional ao quadrado da distância de Fisher-Rao das variedades de distribuições discretas. Como a distância de Fisher-Rao surge naturalmente, do ponto de vista da geometria diferencial, a partir da métrica de Fisher para essa variedade, é legítimo se indagar que vantagens o treinamento com uma função perda baseada nela pode oferecer. Estudamos este desempenho sob dois aspectos: robustez a ruído de rótulo e dinâmica de aprendizado, e mostramos, através de resultados teóricos e

experimentais, que essa proposta oferece um compromisso natural entre eles, em comparação com outras funções perda comumente usadas.

4.2 Preliminares de geometria da informação

Esta seção apresenta resultados básicos de geometria da informação e está baseada principalmente em [2, 11]. Pré-requisitos associados são apresentados nos apêndices A, B e C. Sejam $(\Omega, \mathcal{G}, P)$ um espaço de probabilidade, e $X\colon \Omega \to \mathcal{X}$ uma variável aleatória no espaço de medida $(\mathcal{X}, \mathcal{F}, \mu)$, com μ medida σ-finita. A derivada de Radon-Nikodym $p \coloneqq \frac{\mathrm{d}X_*P}{\mathrm{d}\mu}\colon \mathcal{X} \to \mathbb{R}$ pode ser vista como a função massa ou densidade de probabilidade (f.m.p. ou f.d.p.), respectivamente, nos casos de $\mathcal{X}$ ser discreto ou contínuo.

Um *modelo estatístico* ou *modelo paramétrico* $\mathcal{S}$ é uma família de distribuições de probabilidade parametrizadas por um vetor n-dimensional $\xi = \left(\xi^1, \ldots, \xi^n\right)$, i.e.,

$$\mathcal{S} \coloneqq \left\{ p_\xi = p(x;\xi) \;\middle|\; \xi = (\xi^1, \ldots, \xi^n) \in \Xi \subseteq \mathbb{R}^n \right\}, \tag{4.1}$$

de modo que a aplicação $\xi \mapsto p_\xi$ seja injetiva, e com Ξ conjunto aberto de $\mathbb{R}^n$. Em princípio, consideraremos modelos estatísticos nos quais o suporte de p_ξ não depende de ξ, e tomaremos $\mathcal{X} = \operatorname{supp}(p)$. Note que $\mathcal{S}$ está contido no espaço de dimensão infinita

$$\mathcal{P}(\mathcal{X}) \coloneqq \left\{ p \in L^1(\mu) \;\middle|\; p > 0,\ \int_{\mathcal{X}} p \,\mathrm{d}\mu = 1 \right\}, \tag{4.2}$$

onde $L^1(\mu)$ denota o espaço de Banach das funções μ-integráveis.

Para introduzir uma estrutura diferenciável em $\mathcal{S}$, consideraremos as seguintes hipóteses:

1. a parametrização $\varphi\colon \Xi \to \mathcal{P}(\mathcal{X}),\ \varphi(\xi) = p_\xi$ é um homeomorfismo sobre a imagem;
2. denotando $\partial_i \coloneqq \frac{\partial}{\partial \xi^i}$, as funções $\left\{\partial_1 p_\xi, \ldots, \partial_n p_\xi\right\}$ são linearmente independentes;
3. a aplicação $\xi \mapsto p_\xi(x)$ é de classe C^∞, para todo $x \in \mathcal{X}$;
4. as derivadas parciais $\partial_i p_\xi(x)$ comutam com as integrais.

Além disso, ao considerar parametrizações difeomorfas—i.e., tais que a mudança de parâmetros é um difeomorfismo—como equivalentes, $\mathcal{S} = \left\{ p_\xi \;\middle|\; \xi \in \Xi \right\}$ torna-se uma variedade diferenciável, a que chamamos *variedade estatística*. Note que qualquer parametrização $\xi \mapsto p_\xi$ é um sistema de coordenadas global para essa variedade.

Veremos a seguir que é possível munir a variedade estatística $\mathcal{S}$ de uma métrica riemanniana. Denotando $\ell(\xi) \coloneqq \log p_\xi$ a função log-verossimilhança, os elementos da *matriz de informação*

de Fisher, ou simplesmente *matriz de Fisher*, $G(\xi) = \left[g_{ij}(\xi)\right]_{i,j}$ são definidos como

$$g_{ij}(\xi) \coloneqq \mathbb{E}_{p_\xi}\left[\partial_i \ell(\xi) \partial_j \ell(\xi)\right], \tag{4.3}$$

para $1 \leq i, j \leq n$. Explicitamente,

$$g_{ij}(\xi) = \int_{\mathcal{X}} p_\xi \left(\frac{\partial}{\partial \xi^i} \log p_\xi\right) \left(\frac{\partial}{\partial \xi^j} \log p_\xi\right) \mathrm{d}\mu.$$

A seguir, apresentamos algumas representações alternativas para os elementos da matriz de Fisher.

Proposição 4.1 ([11, § 1.6])**.** *Os elementos da matriz de Fisher podem ser escritos como*

$$g_{ij}(\xi) = 4 \int_{\mathcal{X}} \partial_i \sqrt{p_\xi} \partial_j \sqrt{p_\xi} \, \mathrm{d}\mu.$$

Demonstração.

$$\begin{aligned} g_{ij}(\xi) &= \int_{\mathcal{X}} p_\xi \left(\partial_i \log p_\xi\right) \left(\partial_j \log p_\xi\right) \mathrm{d}\mu = \int_{\mathcal{X}} p_\xi \left(\frac{1}{p_\xi} \partial_i p_\xi\right) \left(\frac{1}{p_\xi} \partial_j p_\xi\right) \mathrm{d}\mu \\ &= 4 \int_{\mathcal{X}} \frac{\partial_i p_\xi}{2\sqrt{p_\xi}} \frac{\partial_i p_\xi}{2\sqrt{p_\xi}} \mathrm{d}\mu = 4 \int_{\mathcal{X}} \partial_i \sqrt{p_\xi} \partial_j \sqrt{p_\xi} \, \mathrm{d}\mu. \end{aligned}$$

■

Proposição 4.2 ([11, § 1.6])**.** *Os elementos da matriz de Fisher podem ser escritos como o negativo da esperança da hessiana da função log-verossimilhança:*

$$g_{ij}(\xi) = -\mathbb{E}_{p_\xi}\left[\partial_i \partial_j \ell(\xi)\right].$$

Demonstração. Como $\int_{\mathcal{X}} p_\xi \, \mathrm{d}\mu = 1$, derivando com respeito a ξ^i, obtemos $\int_{\mathcal{X}} \partial_i p_\xi \, \mathrm{d}\mu = 0$. Assim,

$$\mathbb{E}_{p_\xi}\left[\partial_i \log p_\xi\right] = \int_{\mathcal{X}} p_\xi \partial_i \log p_\xi \, \mathrm{d}\mu = \int_{\mathcal{X}} \partial_i p_\xi \, \mathrm{d}\mu = 0.$$

Derivando novamente, com respeito a ξ^j, temos então

$$\begin{aligned}
&\partial_j \int_{\mathcal{X}} p_\xi \partial_i \log p_\xi \, \mathrm{d}\mu = 0 \\
\iff &\int_{\mathcal{X}} \partial_j p_\xi \partial_i \log p_\xi + \int_{\mathcal{X}} p_\xi \partial_j \partial_i \log p_\xi \, \mathrm{d}\mu = 0 \\
\iff &\int_{\mathcal{X}} p_\xi \partial_j \log p_\xi \partial_i \log p_\xi \, \mathrm{d}\mu + \int_{\mathcal{X}} p_\xi \partial_j \partial_i \log p_\xi \, \mathrm{d}\mu = 0 \\
\iff &\mathbb{E}\left[\partial_i \log p_\xi \partial_j \log p_\xi\right] + \mathbb{E}\left[\partial_j \partial_i p_\xi\right] = 0 \\
\iff &g_{ij}(\xi) = -\mathbb{E}\left[\partial_j \partial_i p_\xi\right].
\end{aligned}$$

■

Proposição 4.3 ([11, § 4.4])**.** *A matriz de Fisher pode ser escrita como a hessiana da divergência de Kullback-Leibler, no seguinte sentido:*

$$g_{ij}(\xi_0) = \partial_i \partial_j D_{\mathrm{KL}}(p_{\xi_0} \| p_\xi)\big|_{\xi=\xi_0}.$$

Demonstração. A divergência de Kullback-Leibler é dada por $D_{\mathrm{KL}}(p_{\xi_0}\|p_\xi) = \int_{\mathcal{X}} p_{\xi_0} \log \frac{p_{\xi_0}}{p_\xi} \, \mathrm{d}\mu$. Com ξ_0 fixo, ao derivar com relação aos parâmetros ξ^i e ξ^j, obtemos

$$\partial_i \partial_j D_{\mathrm{KL}}(p_{\xi_0}\|p_\xi) = \partial_i \partial_j \int_{\mathcal{X}} p_{\xi_0} \log p_{\xi_0} \, \mathrm{d}\mu - \partial_i \partial_j \int_{\mathcal{X}} p_{\xi_0} \log p_\xi \, \mathrm{d}\mu = -\int_{\mathcal{X}} p_{\xi_0} \partial_i \partial_j \log p_\xi \, \mathrm{d}\mu.$$

Tomando $\xi = \xi_0$ e usando a Proposição 4.2, temos então

$$\partial_i \partial_j D_{\mathrm{KL}}(p_{\xi_0}\|p_\xi)\big|_{\xi=\xi_0} = -\int_{\mathcal{X}} p_\xi \partial_i \partial_j \log p_\xi \, \mathrm{d}\mu = -\mathbb{E}_{p_\xi}\left[\partial_i \partial_j \log p_\xi\right] = g_{ij}(\xi).$$

■

Observação 4.1. A matriz de Fisher aparece no limitante de Cramér-Rao. No contexto de estatística paramétrica, ao observar amostras $(x_1, \ldots x_N) \in \mathcal{X}^N$ geradas por uma distribuição p_ξ, gostaríamos de identificar o parâmetro ξ que gerou tais amostras. Esse é o problema de estimação, e um *estimador* é uma aplicação

$$\hat{\xi}_N : \mathcal{X}^N \to \Xi,$$

que associa, a cada conjunto de amostras, um parâmetro do modelo estatístico em questão. Considerando um vetor aleatório $(X_1, \ldots, X_N)$, $\hat{\xi}_N = \hat{\xi}_N(X_1, \ldots, X_N)$ também será uma variável aleatória. Uma sequência de estimadores será dita consistente, se[1] $\lim_{N\to\infty} \hat{\xi}_N = \xi$. Sob algu-

[1] O limite deve ser entendido em termos de convergência em probabilidade: $\forall\, \epsilon > 0,\ \Pr\left\{|\hat{\xi}_N - \xi| > \epsilon\right\} = 0$.

mas condições de regularidade, cf. [2, p. 83], temos então $\lim_{N\to\infty} \mathbb{E}[\hat{\xi}_N] = \xi$. Nesse contexto, o limitante (assintótico) de Cramér-Rao afirma que, para um estimador consistente $\hat{\xi}_N$, vale que

$$\lim_{N\to\infty} N\,\mathbb{E}\left[\left(\hat{\xi}_N - \xi\right)\left(\hat{\xi}_N - \xi\right)^{\mathsf{T}}\right] \succcurlyeq \left[G(\xi)\right]^{-1}, \tag{4.4}$$

onde $\succcurlyeq$ denota a ordem parcial de Lowener para matrizes simétricas semi-definidas positivas[2]. Dessa forma, a inversa da matriz de Fisher fornece um limitante inferior para a matriz de covariância do estimador em questão. Um estimador que atinge o limitante é dito ser *assintoticamente eficiente* ou *Fisher eficiente* [1, 2].

Proposição 4.4 ([11, § 1.6])**.** *A matriz de Fisher de qualquer modelo estatístico é simétrica definida positiva.*

Demonstração. A simetria da matriz de Fisher $G(\xi)$ é consequência direta da definição (4.3). Para verificar que é definida positiva, tome $v = (v^1, \ldots, v^n)^{\mathsf{T}} \neq 0$. Usando a Proposição 4.1, temos

$$\begin{aligned} v^{\mathsf{T}}[G(\xi)]v = \sum_{i,j} g_{ij}v^iv^j &= 4\int_{\mathcal{X}} \sum_{i,j} \partial_i\sqrt{p_\xi}\partial_j\sqrt{p_\xi}v^iv^j \,\mathrm{d}\mu \\ &= 4\int_{\mathcal{X}} \left(\textstyle\sum_i v^i\partial_i\sqrt{p_\xi}\right)\left(\sum_j v^j\partial_j\sqrt{p_\xi}\right)\mathrm{d}\mu \\ &= 4\int_{\mathcal{X}} \left(\textstyle\sum_i v^i\partial_i\sqrt{p_\xi}\right)^2\mathrm{d}\mu \geq 0. \end{aligned}$$

Temos ainda que

$$\begin{aligned} v^{\mathsf{T}}[G(\xi)]v = 0 &\iff \int_{\mathcal{X}} \left(\textstyle\sum_i v^i\partial_i\sqrt{p_\xi}\right)^2\mathrm{d}\mu = 0 \iff \textstyle\sum_i v^i\partial_i\sqrt{p_\xi} = 0,\ \mu\text{-q.t.p.} \\ &\iff \textstyle\sum_i v^i\partial_i p_\xi = 0,\ \mu\text{-q.t.p.} \iff v^i = 0,\ i \in \{1,\ldots,n\}, \end{aligned}$$

pois $\left\{\partial_1 p_\xi, \ldots, \partial_n p_\xi\right\}$ são linearmente independentes. Logo $G(\xi)$ é definida positiva. ■

Como consequência da Proposição 4.4, a matriz de Fisher define uma métrica riemanniana g_ξ na variedade estatística $\mathcal{S}$, chamada *métrica de Fisher*. Esta é a escolha natural para a geometria de $\mathcal{S}$, em vista dos resultados a seguir.

Proposição 4.5 ([11, § 1.6])**.** *Seja $X\colon \Omega \to \mathcal{X} \subseteq \mathbb{R}^n$ uma variável aleatória distribuída de acordo com p_ξ. A métrica de Fisher é invariante por reparametrização de $\mathcal{X}$.*

Demonstração. Considere a reparametrização pela bijeção $f\colon \mathcal{X} \to \mathcal{Y} \subseteq \mathbb{R}^n$ e denote $\tilde{p}_\xi$ a distribuição associada à variável aleatória $Y \coloneqq f(X)$. O determinante jacobiano da transformação f

[2]Sejam A e B duas matrizes simétricas semi-definidas positivas. Denotamos $A \succcurlyeq B$ se, e somente se, $A - B$ é simétrica semi-definida positiva.

medeia a relação entre as funções densidade de probabilidade:

$$p_\xi(x) = \tilde{p}_\xi(y) \left| \frac{\mathrm{d}f}{\mathrm{d}x}(x) \right|. \tag{4.5}$$

As funções log-verossimilhança são $\tilde{\ell}(\xi) = \log \tilde{p}_\xi(y) = \log \tilde{p}_\xi(f(x))$ e $\ell(\xi) = \log p_\xi(x) = \log \tilde{p}_\xi(y) + \log \left| \frac{\mathrm{d}f}{\mathrm{d}x}(x) \right|$. Como f não depende dos parâmetros ξ, temos $\partial_i \ell(\xi) = \partial_i \tilde{\ell}(\xi)$, logo

$$\begin{aligned} g_{ij}(\xi) &= \int_{\mathcal{X}} p_\xi \partial_i \ell(\xi) \partial_i \ell(\xi) \, \mathrm{d}\mu = \int_{\mathcal{X}} (\tilde{p}_\xi \circ f) \left| \frac{\mathrm{d}f}{\mathrm{d}x} \right| \partial_i \tilde{\ell}(\xi) \partial_i \tilde{\ell}(\xi) \, \mathrm{d}\mu \\ &= \int_{\mathcal{Y}} \tilde{p}_\xi \partial_i \tilde{\ell}(\xi) \partial_i \tilde{\ell}(\xi) \, \mathrm{d}(f_* \mu) = \tilde{g}_{ij}(\xi), \end{aligned}$$

onde $f_*\mu$ é a medida *pushfoward* de μ por f, definida em $\mathcal{Y}$. ■

Proposição 4.6 ([11, § 1.6])**.** *A métrica de Fisher é covariante por reparametrização do espaço de parâmetros Ξ, i.e., dados dois sistemas de coordenadas ξ e θ, relacionados pela bijeção $\xi = \xi(\theta)$, a matriz de Fisher muda de coordenadas como*

$$\tilde{G}(\theta) = \left[\frac{\mathrm{d}\xi}{\mathrm{d}\theta} \right]^{\mathsf{T}} G\left(\xi(\theta)\right) \left[\frac{\mathrm{d}\xi}{\mathrm{d}\theta} \right],$$

onde $\left[\frac{\mathrm{d}\xi}{\mathrm{d}\theta} \right]$ denota o jacobiano da transformação $\theta \mapsto \xi$.

Demonstração. Considere dois sistemas de coordenadas $\xi = (\xi^1, \ldots, \xi^n)$ e $\theta = (\theta^1, \ldots, \theta^n)$, relacionados pela bijeção $\xi = \xi(\theta)$. Denote $\tilde{p}_\theta \coloneqq p_{\xi(\theta)}$. Pela regra da cadeia, temos

$$\frac{\partial}{\partial \theta^i} \tilde{p}_\theta = \sum_{k=1}^{n} \frac{\partial \xi^k}{\partial \theta^i} \frac{\partial}{\partial \xi^k} p_\xi \qquad \text{e} \qquad \frac{\partial}{\partial \theta^j} \tilde{p}_\theta = \sum_{r=1}^{n} \frac{\partial \xi^r}{\partial \theta^j} \frac{\partial}{\partial \xi^r} p_\xi.$$

Então

$$\begin{aligned} \tilde{g}_{ij}(\theta) &= \int_{\mathcal{X}} \tilde{p}_\theta \left(\frac{\partial}{\partial \theta^i} \log \tilde{p}_\theta \right) \left(\frac{\partial}{\partial \theta^j} \log \tilde{p}_\theta \right) \mathrm{d}\mu = \int_{\mathcal{X}} \frac{1}{\tilde{p}_\theta} \left(\frac{\partial}{\partial \theta^i} \tilde{p}_\theta \right) \left(\frac{\partial}{\partial \theta^j} \tilde{p}_\theta \right) \mathrm{d}\mu \\ &= \sum_{k=1}^{n} \sum_{r=1}^{n} \left[\int_{\mathcal{X}} \frac{1}{p_{\xi(\theta)}} \left(\frac{\partial}{\partial \xi^k} p_\xi \right) \left(\frac{\partial}{\partial \xi^r} p_\theta \right) \mathrm{d}\mu \right] \frac{\partial \xi^k}{\partial \theta^i} \frac{\partial \xi^r}{\partial \theta^j} = \sum_{k=1}^{n} \sum_{r=1}^{n} g_{kr}(\xi(\theta)) \frac{\partial \xi^k}{\partial \theta^i} \frac{\partial \xi^r}{\partial \theta^j}. \end{aligned}$$

■

Teorema 4.7 ([5, Theorem 1.2])**.** *A métrica de Fisher é a única métrica invariante por estatística suficiente em variedades estatísticas, a menos de fator de escala.*

Versões deste teorema, em diferentes níveis de generalidade e abstração, podem foram enunciadas por Čencov [13, Theorem 11.1], Campbell [12, Theorem, p. 137], Ay *et al.* [4, Theorem 2.10], e Vân Lê [52, Theorem 4].

Com essa estrutura riemanniana especial, é possível introduzir uma noção de distância entre duas distribuições de probabilidade da mesma família, como proposto em [46]. Aplicar a métrica de Fisher a dois vetores $u = \mathrm{d}\varphi(\xi_u)$ e $v = \mathrm{d}\varphi(\xi_v)$ no espaço tangente $T_{p_\xi}\mathcal{S}$ é equivalente a calcular um produto interno, mediado pela matriz de Fisher $G(\xi)$, entre os respectivos vetores de coordenadas locais $\xi_u, \xi_v \in \mathbb{R}^n$:

$$g_\xi(u, v) = \langle u, v \rangle_{G(\xi)} \coloneqq \xi_u^\mathsf{T}[G(\xi)]\xi_v.$$

Considere uma curva $\gamma\colon [0,1] \to \mathcal{S}$, que é a imagem da curva $\xi\colon [0,1] \to \Xi$, pela parametrização $\varphi\colon \Xi \to \mathcal{S}$, i.e., $\gamma(t) = \varphi(\xi(t))$, $t \in [0,1]$. Na geometria de Fisher, o comprimento da curva γ pode ser calculado como

$$l(\gamma) \coloneqq \int_0^1 \sqrt{\left\langle \dot\gamma(t), \dot\gamma(t) \right\rangle_{G\left(\xi(t)\right)}}\, \mathrm{d}t = \int_0^1 \sqrt{[\dot\xi(t)]^\mathsf{T}[G(\xi(t))][\dot\xi(t)]}\, \mathrm{d}t. \tag{4.6}$$

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

$$d_{\mathrm{FR}}(\xi_1, \xi_2) \coloneqq d_{\mathrm{FR}}(p_{\xi_1}, p_{\xi_2}) \coloneqq \inf_\gamma \left\{ l(\gamma) \;\middle|\; \gamma(0) = p_{\xi_1},\ \gamma(1) = p_{\xi_2} \right\}. \tag{4.7}$$

Por abuso de linguagem, também nos referimos à distância de Fisher-Rao entre duas distribuições paramétricas como a distância de Fisher-Rao entre seus parâmetros.

Notamos que a distância de Fisher-Rao (4.7) define uma métrica[3] em $\mathcal{S}$ [34, Theorem 6.4.2]. Para verificar que $(\mathcal{S}, d_{\mathrm{FR}})$ é um espaço métrico, basta notar que (i) $d_{\mathrm{FR}}(\xi_1, \xi_2) \geq 0$, com igualdade se, e somente se, $\xi_1 = \xi_2$; (ii) $d_{\mathrm{FR}}(\xi_1, \xi_2) = d_{\mathrm{FR}}(\xi_2, \xi_1)$; (iii) $d_{\mathrm{FR}}(\xi_1, \xi_3) \leq d_{\mathrm{FR}}(\xi_1, \xi_2) + d_{\mathrm{FR}}(\xi_2, \xi_3)$.

O Teorema de Hopf-Rinow[4] fornece uma condição suficiente, mas não necessária, para que as geodésicas minimizantes entre quaisquer dois pontos da variedade sejam realizadas:

Teorema 4.8 (Hopf-Rinow [34, Theorem 6.4.6])**.** *Se $(\mathcal{S}, d_{\mathrm{FR}})$ é um espaço métrico completo e conexo, então quaisquer dois pontos p, q da variedade $\mathcal{S}$ podem ser ligados por uma geodésica minimizante, i.e., cujo comprimento é igual à distância de Fisher-Rao $d_{\mathrm{FR}}(p, q)$ entre eles.*

Infelizmente, encontrar a distância de Fisher-Rao entre duas distribuições de uma variedade

[3] I.e., uma *métrica* no sentido de espaços métricos; não confundir com métrica riemanniana.

[4] De fato, o Teorema de Hopf-Rinow elenca uma série de condições equivalentes, cada uma das quais implica a realização das geodésicas minimizantes; veja, por exemplo, [20, § 7.2].

estatística não é uma tarefa trivial, visto que envolve primeiro encontrar as geodésicas minimizantes, potencialmente através da resolução das equações diferenciais das geodésicas, e, em seguida, avaliar a integral em (4.6).

4.3 Zoológico de distâncias de Fisher-Rao

Nesta seção, colecionamos exemplos de variedades estatísticas conhecidas na literatura, para as quais é possível escrever explicitamente a distância de Fisher-Rao (binomial, Poisson, geométrica, binomial negativa, categórica, multinomial, exponencial, gaussiana, Pareto) [3, 4, 9, 11, 16, 38]. Apresentamos também alguns exemplos que não aparecem na literatura clássica (Rayleigh, Laplace, Cauchy), tendo utilizado o programa Wolfram Mathematica como suporte para a realização de alguns cálculos. Excepcionalmente, nesta parte do texto, denotaremos as distribuições de probabilidade como f, notação também usual para a função densidade de probabilidade, pois reservamos a letra p para indicar o parâmetro de algumas distribuições discretas.

No caso de variedades estatísticas de dimensão 1, i.e., cujas parametrizações são dadas por apenas um número real, é imediato calcular as distâncias de Fisher-Rao. Nesse caso, a matriz de Fisher $G(\xi) = [g_{11}(\xi)]$ contém apenas um elemento, também chamado *informação de Fisher*. Dados dois parâmetros ξ_1 e ξ_2, só há uma trajetória entre eles, cujo comprimento não depende da parametrização escolhida. Em particular, podemos tomar a parametrização pelo comprimento de arco $\xi(t) = t$, com $t \in [\xi_1, \xi_2]$, de modo que $\left|\dot{\xi}(t)\right| = 1$. Assim, a expressão para o comprimento da curva $\gamma = \varphi(\xi(t))$ em (4.6) torna-se

$$l(\gamma) = \int_{\xi_1}^{\xi_2} \sqrt{g_{11}(\xi(t))}\, \mathrm{d}t = \int_{\xi_1}^{\xi_2} \sqrt{g_{11}(\xi)}\, \mathrm{d}\xi,$$

e a distância de Fisher-Rao entre as distribuições parametrizadas por ξ_1 e ξ_2 é dada por

$$d_{\mathrm{FR}}(\xi_1, \xi_2) = \left| \int_{\xi_1}^{\xi_2} \sqrt{g_{11}(\xi)}\, \mathrm{d}\xi \right|. \tag{4.8}$$

Para variedades de dimensões maiores, as técnicas para encontrar geodésicas e a distância de Fisher-Rao se resumem à resolução direta das equações diferenciais das geodésicas, ou à analogia com outra geometria conhecida (e.g., esférica, hiperbólica), como veremos a seguir. As Tabelas 4.1 e 4.2 ao fim da seção resumem os resultados das distribuições discretas e contínuas, respectivamente.

4.3.1 Distribuições discretas

Exemplo 4.1 (Binomial [3, 9, 11])**.** Uma distribuição binomial representa a probabilidade de haver x sucessos ao realizar n experimentos i.i.d. com distribuição Bernoulli com parâmetro p. Sua f.m.p. é dada por $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$, definida para $x \in \{0, 1, \ldots, n\}$ e parametrizada por $p \in]0, 1[$. Nesse caso, $\partial_p \ell(p) = \frac{x}{p} - \frac{n-x}{1-p}$, de modo que

$$\begin{aligned} g_{11} &= \mathbb{E}\left[\left(\partial_p \ell(p)\right)^2\right] = \mathbb{E}\left[\left(\frac{X}{p} - \frac{n-X}{1-p}\right)^2\right] \\ &= \frac{\mathbb{E}[X^2]}{p^2} + \frac{2\mathbb{E}[X^2] - 2n\mathbb{E}[X]}{p(1-p)} + \frac{n - 2n\mathbb{E}[X] + \mathbb{E}[X^2]}{(1-p)^2} \\ &= \frac{n}{p(1-p)}, \end{aligned}$$

em que usamos que $\mathbb{E}[X] = np$ e $\mathbb{E}[X^2] = np - np^2 + n^2p^2$. A distância de Fisher-Rao é dada por

$$d_{\mathrm{FR}}(p_1, p_2) = \left| \int_{p_1}^{p_2} \sqrt{\frac{n}{t(1-t)}}\, \mathrm{d}t \right| = 2\sqrt{n} \left| \arcsin\sqrt{p_1} - \arcsin\sqrt{p_2} \right|.$$

Exemplo 4.2 (Poisson [3, 9, 11])**.** Uma distribuição de Poisson é dada por $f(x) = \lambda^x e^{-\lambda}/x!$, definida para $x \in \mathbb{N}$ e parametrizada por $\lambda \in \mathbb{R}_+^*$. Nesse caso, temos $\partial_\lambda \ell(\lambda) = \frac{x}{\lambda} - 1$, e a informação de Fisher é dada por

$$\begin{aligned} g_{11} &= \mathbb{E}\left[\left(\partial_\lambda \ell(\lambda)\right)^2\right] = \mathbb{E}\left[\left(\frac{X}{\lambda} - 1\right)^2\right] \\ &= \frac{1}{\lambda^2}\mathbb{E}\left[X^2\right] - \frac{2}{\lambda}\mathbb{E}\left[X\right] + 1 \\ &= \frac{1}{\lambda}, \end{aligned}$$

em que usamos que $\mathbb{E}[X] = \lambda$ e $\mathbb{E}[X^2] = \lambda(\lambda + 1)$. Dessa forma, a distância de Fisher-Rao entre dois parâmetros é dada por

$$d_{\mathrm{FR}}(\lambda_1, \lambda_2) = \left| \int_{\lambda_1}^{\lambda_2} \frac{1}{\sqrt{t}}\, \mathrm{d}t \right| = 2\left| \sqrt{\lambda_1} - \sqrt{\lambda_2} \right|.$$

Exemplo 4.3 (Geométrica [11])**.** Uma distribuição geométrica modela a quantidade de experimentos i.i.d. Bernoulli de parâmetro p necessários até obter um sucesso. Sua f.m.p. é dada por $f(x) = p(1-p)^{x-1}$, definida para $x \in \mathbb{N}^*$, e parametrizada por $p \in]0, 1[$. Temos

$\partial_p \ell(p) = \frac{1}{p} - \frac{x-1}{1-p}$, e a informação de Fisher é

$$\begin{aligned} g_{11} &= \mathbb{E}\left[\left(\partial_p \ell(p)\right)^2\right] = \mathbb{E}\left[\left(\frac{1}{p} - \frac{X-1}{1-p}\right)^2\right] \\ &= \frac{1}{p^2} + \frac{2 - 2\mathbb{E}[X]}{p(1-p)} + \frac{\mathbb{E}[X^2] - 2\mathbb{E}[X] + 1}{(1-p^2)} \\ &= \frac{1}{p^2(1-p)}, \end{aligned}$$

em que usamos que $\mathbb{E}[X] = \frac{1}{p}$ e que $\mathbb{E}[X^2] = \frac{2-p}{p^2}$. A distância de Fisher-Rao é dada por

$$d_{\mathrm{FR}}(p_1, p_2) = \left| \int_{p_1}^{p_2} \frac{1}{t\sqrt{1-t}} \, \mathrm{d}t \right| = 2 \left| \operatorname{arctanh} \sqrt{1-p_1} - \operatorname{arctanh} \sqrt{1-p_2} \right|.$$

Exemplo 4.4 (Binomial negativa [9])**.** Considere uma sequência de experimentos i.i.d. Bernoulli de parâmetro p. Uma distribuição binomial negativa modela o excesso da quantidade de experimentos necessários para que uma quantidade r de sucessos ocorra. Sua f.m.p. é dada por $f(x) = \binom{x+r-1}{r-1} p^r (1-p)^x$, está definida para $x \in \mathbb{N}$, e é parametrizada por $p \in]0, 1[$. Temos $\partial_p \ell(p) = \frac{r}{p} - \frac{x}{1-p}$, e a informação de Fisher é dada por

$$\begin{aligned} g_{11} &= \mathbb{E}\left[\left(\partial_p \ell(p)\right)^2\right] = \mathbb{E}\left[\left(\frac{r}{p} - \frac{X}{1-p}\right)^2\right] \\ &= \frac{r^2}{p^2} - \frac{2r\mathbb{E}[X]}{p(1-p)} + \frac{\mathbb{E}[X^2]}{(1-p^2)} \\ &= \frac{r}{p^2(1-p)}, \end{aligned}$$

em que usamos que $\mathbb{E}[X] = \frac{r(1-p)}{p}$ e $\mathbb{E}[X^2] = \frac{r(1-p)+r^2(1-p)^2}{(1-p)^2}$. A distância de Fisher-Rao é dada por

$$d_{\mathrm{FR}}(p_1, p_2) = \left| \int_{p_1}^{p_2} \frac{\sqrt{r}}{t\sqrt{1-t}} \, \mathrm{d}t \right| = 2\sqrt{r} \left| \operatorname{arctanh} \sqrt{1-p_1} - \operatorname{arctanh} \sqrt{1-p_2} \right|.$$

Exemplo 4.5 (Categórica [3, 4, 9, 11])**.** Considere uma variável aleatória que toma valores no espaço amostral $\mathcal{X} = \{1, 2, \ldots, n\}$ com probabilidades $p_1, \ldots, p_n$, de modo que $\sum_{i=1}^n p_i = 1$. A distribuição categórica, ou simplesmente *discreta*, tem como f.m.p. $f(x) = \sum_{i=1}^n p_i \mathbb{1}_{\{i\}}(x)$. A variedade estatística

$$\mathcal{S} = \left\{ f(x) = \textstyle\sum_{i=1}^n p_i \mathbb{1}_{\{i\}}(x) \;\middle|\; p_i \in]0,1[\,,\ \sum_{i=1}^n p_i = 1 \right\} \tag{4.9}$$

está em correspondência com o simplexo de probabilidades

$$\Delta^{n-1} = \left\{ \boldsymbol{p} = (p_1, \ldots, p_n) \;\middle|\; p_i \in]0,1[\,,\ \textstyle\sum_{i=1}^{n} p_i = 1 \right\} \tag{4.10}$$

através da bijeção $\iota \colon \Delta^{n-1} \to \mathcal{S}$, definida por $(p_1, \ldots, p_n) \mapsto \sum_{i=1}^{n} p_i \mathbb{1}_{\{i\}}(x)$. Cada uma dessas variedades podem ser parametrizadas pelos vetores do conjunto

$$\Xi = \left\{ \xi = (\xi^1, \ldots, \xi^{n-1}) \;\middle|\; \xi^i > 0,\ \textstyle\sum_{i=1}^{n-1} \xi^i < 1 \right\} \subset \mathbb{R}^{n-1}, \tag{4.11}$$

com $p_i = \xi^i$, para $1 \leq i \leq n-1$, e $p_n = 1 - \sum_{i=1}^{n-1} \xi^i$.

Para calcular a matriz de Fisher, escrevemos $f(x) = \sum_{i=1}^{n} p_i \mathbb{1}_{\{i\}}(x) = \sum_{i=1}^{n-1} \xi^i \mathbb{1}_{\{i\}}(x) + \left(1 - \sum_{i=1}^{n-1} \xi^i\right) \mathbb{1}_{\{n\}}(x)$. Temos assim

$$\partial_i \ell(\xi) = \frac{1}{\left(\sum_k p_k \mathbb{1}_{\{k\}}(x)\right)^2} \left(\mathbb{1}_{\{i\}}(x) - \mathbb{1}_{\{n\}}(x) \right), \tag{4.12}$$

de modo que os elementos da matriz de Fisher são dados por

$$\begin{aligned}
g_{ij} \coloneqq & g_{ij}(\xi) = \mathbb{E}\left[\left(\partial_i \ell(\xi)\right) \left(\partial_j \ell(\xi)\right) \right] \\
= & \mathbb{E}\left[\frac{\left(\mathbb{1}_{\{i\}}(X) - \mathbb{1}_{\{n\}}(X) \right) \left(\mathbb{1}_{\{j\}}(X) - \mathbb{1}_{\{n\}}(X) \right)}{\left(\sum_k p_k \mathbb{1}_{\{k\}}(X)\right)^2} \right] \\
= & \mathbb{E}\left[\frac{\mathbb{1}_{\{i\}}(X) \mathbb{1}_{\{j\}}(X)}{\left(\sum_k p_k \mathbb{1}_{\{k\}}(X)\right)^2} \right] - \mathbb{E}\left[\frac{\mathbb{1}_{\{i\}}(X) \mathbb{1}_{\{n\}}(X)}{\left(\sum_k p_k \mathbb{1}_{\{k\}}(X)\right)^2} \right] \\
& - \mathbb{E}\left[\frac{\mathbb{1}_{\{j\}}(X) \mathbb{1}_{\{n\}}(X)}{\left(\sum_k p_k \mathbb{1}_{\{k\}}(X)\right)^2} \right] + \mathbb{E}\left[\frac{\left(\mathbb{1}_{\{n\}}(X)\right)^2}{\left(\sum_k p_k \mathbb{1}_{\{k\}}(X)\right)^2} \right]
\end{aligned}$$

Notamos que

$$\mathbb{E}\left[\frac{\mathbb{1}_{\{i\}}(X) \mathbb{1}_{\{j\}}(X)}{\left(\sum_k p_k \mathbb{1}_{\{k\}}(X)\right)^2} \right] = \sum_{x \in \mathcal{X}} \frac{\mathbb{1}_{\{i\}}(x) \mathbb{1}_{\{j\}}(x)}{\sum_k p_k \mathbb{1}_{\{k\}}(x)} = \frac{\delta_{ij}}{p_i} = \frac{\delta_{ij}}{p_j}.$$

Dessa forma,

$$\begin{aligned}
g_{ij}(\xi) &= \frac{\delta_{ij}}{p_i} - \frac{\delta_{in}}{p_n} - \frac{\delta_{jn}}{p_n} + \frac{1}{p_n} \\
&= \frac{\delta_{ij}}{p_i} + \frac{1}{p_n} \\
&= \frac{\delta_{ij}}{\xi^i} + \frac{1}{1 - \sum_{k=1}^{n-1} \xi^k},
\end{aligned}$$

para $1 \leq i, j \leq n-1$.

Para obter as geodésicas e a distância de Fisher-Rao, no entanto, é conveniente considerar a reparametrização $p_i \mapsto 2\sqrt{p_i} =: z_i$, que leva o simplexo $\Delta^{n-1} \subset \mathbb{R}^n$ na parte positiva da esfera euclidiana de raio 2, denotada $S^{n-1}_{2,+} \subset \mathbb{R}^n$ [32, § 7.2]. Considere o difeomorfismo

$$\begin{aligned} \pi\colon \quad & \mathcal{S} \subset \mathcal{P}(\mathcal{X}) \to S^{n-1}_{2,+} \subset \mathbb{R}^n \\ & f = \sum_{i=1}^{n} p_i \mathbb{1}_{\{i\}} \mapsto (2\sqrt{p_1}, \ldots, 2\sqrt{p_n}) =: (z_1, \ldots, z_n) =: \boldsymbol{z}. \end{aligned}$$

Observamos que, de fato, $\|\boldsymbol{z}\|^2 = \sum_{i=1}^{n} (z_i)^2 = \sum_{i=1}^{n} (2\sqrt{p_i})^2 = 4$. Mostraremos que π é uma isometria entre $\mathcal{S}$ com a métrica de Fisher g, e $S^{n-1}_{2,+}$ com a restrição da métrica euclidiana ambiente [4, § 2.2], i.e., que $g_f(u,v) = \big\langle \mathrm{d}\pi_f(u), \mathrm{d}\pi_f(v) \big\rangle$, para todo $f \in \mathcal{S}$ e $u, v \in T_f\mathcal{S}$. Tomando a curva $\alpha_i(t) = \big(\xi^1, \ldots, \xi^i + t, \ldots, \xi^n\big)$, temos, cf. (C.1),

$$\mathrm{d}\pi_f \left(\frac{\partial}{\partial \xi^i} \right) = \mathrm{d}\pi_f \left(\alpha_i'(0) \right) = (\pi \circ \alpha_i)'(0).$$

Além disso,

$$(\pi \circ \alpha_i)(t) = \pi\left(\alpha_i(t)\right) = \left(2\sqrt{\xi^1}, \ldots, 2\sqrt{\xi^i + t}, \ldots, 2\sqrt{\xi^n}, 2\sqrt{1 - \sum_{k \neq i} \xi^k - (\xi^i + t)} \right),$$

de modo que

$$\begin{aligned} \mathrm{d}\pi_f \left(\frac{\partial}{\partial \xi^i} \right) &= \frac{\mathrm{d}}{\mathrm{d}t} (\pi \circ \alpha_i)(0) = \frac{\mathrm{d}}{\mathrm{d}t} (\pi \circ \alpha_i)(t) \Big|_{t=0} \\ &= \left(0, \ldots, 0, \frac{1}{\sqrt{\xi^i + t}}, 0, \ldots, 0, -\frac{1}{\sqrt{1 - \sum_{k \neq i} \xi^k - (\xi^i + t)}} \right) \Bigg|_{t=0} \\ &= \left(0, \ldots, \frac{1}{\sqrt{\xi^i}}, \ldots, 0, -\frac{1}{\sqrt{1 - \sum_k \xi^k}} \right). \end{aligned}$$

Portanto,

$$\begin{aligned} & \left\langle \mathrm{d}\pi_f \left(\frac{\partial}{\partial \xi^i} \right), \mathrm{d}\pi_f \left(\frac{\partial}{\partial \xi^j} \right) \right\rangle \\ &= \left\langle \left(0, \ldots, \frac{1}{\sqrt{\xi^i}}, \ldots, 0, -\frac{1}{\sqrt{1 - \sum_k \xi^k}} \right), \left(0, \ldots, \frac{1}{\sqrt{\xi^j}}, \ldots, 0, -\frac{1}{\sqrt{1 - \sum_k \xi^k}} \right) \right\rangle \\ &= \frac{\delta_{ij}}{\xi^i} + \frac{1}{1 - \sum_k \xi^k} = g_{ij} = g_f \left(\frac{\partial}{\partial \xi^i}, \frac{\partial}{\partial \xi^j} \right). \end{aligned}$$

Como $\left\{\frac{\partial}{\partial\xi^1},\dots,\frac{\partial}{\partial\xi^n}\right\}$ é base de $T_f\mathcal{S}$, isso é suficiente para mostrar que π é uma isometria. Portanto, a métrica de Fisher em $\mathcal{S}$ coincide com a métrica euclidiana restrita à esfera $S^{n-1}_{2,+}$. Essa identificação também permite estender a métrica de Fisher ao bordo da variedade estatística $\mathcal{S}$.

Na esfera, geodésicas são dadas por arcos de grandes círculos, então o comprimento da geodésica ligando as distribuições $f_p \coloneqq \iota(\boldsymbol{p})$ e $f_q \coloneqq \iota(\boldsymbol{q})$ é dado pelo dobro do ângulo α que separa $\boldsymbol{z}_p \coloneqq \pi(f_p)$ e $\boldsymbol{z}_q \coloneqq \pi(f_q)$ na esfera, i.e.,

$$2\alpha = 2\arccos\left\langle \frac{\boldsymbol{z}_p}{2}, \frac{\boldsymbol{z}_q}{2}\right\rangle = 2\arccos\left(\sum_{i=1}^{n}\sqrt{p_i q_i}\right).$$

Portanto, a distância de Fisher-Rao entre essas duas distribuições é[5]

$$d_{\mathrm{FR}}(\boldsymbol{p},\boldsymbol{q}) = 2\arccos\left(\sum_{i=1}^{n}\sqrt{p_i q_i}\right). \tag{4.13}$$

Observação 4.2. Podemos calcular a métrica de Fisher com respeito aos parâmetros $(\theta^1,\dots,\theta^{n-1})$ tais que $\theta^i \geq 0$ e $\sum_{i=1}^{n-1}(\theta^i)^2 \leq 4$, i.e., tais que $z_i = \theta^i$, para $1 \leq i \leq n-1$, e $z_n = \sqrt{4 - \sum_{i=1}^{n-1}(\theta^i)^2}$. Notando que $\frac{\partial\xi^k}{\partial\theta^i} = \frac{\theta^i}{2}\delta_{ki}$, podemos aplicar a Proposição 4.6 e deduzir que os elementos da matriz de Fisher nessa nova parametrização são

$$g_{ij}(\theta) = \delta_{ij} + \frac{z_i z_j}{z_n^2} = \delta_{ij} + \frac{\theta^i\theta^j}{4 - \sum_{i=1}^{n-1}(\theta^i)^2},$$

que de fato corresponde à métrica euclidiana de $\mathbb{R}^n$ restrita à esfera $S^{n-1}_{2,+}$.

Exemplo 4.6 (Multinomial [3, 9, 11])**.** A distribuição multinomial modela o resultado de m experimentos i.i.d. que seguem uma distribuição categórica com n possíveis resultados, de probabilidades $p_1,\dots,p_n$. Sua f.m.p. é dada por $f(x) = f(x_1,\dots,x_n) = \frac{n!}{x_1!\cdots x_n!}p_1^{x_1}p_2^{x_2}\cdots p_n^{x_n}$ e está definida no espaço amostral $\mathcal{X} = \left\{(x_1,\dots,x_n)\in\mathbb{N}^n \;\middle|\; \sum_{i=1}^{n} x_i = m\right\}$. É parametrizada pelos mesmos $\xi = (\xi^1,\dots,\xi^{n-1}) \in \Xi$ das categóricas, cf. (4.11), com $p_i = \xi^i$, para $1 \leq i \leq n-1$ e $p_n = 1 - \sum_{i=1}^{n-1}\xi^i$. Nesse caso, temos

$$\partial_i\ell(\xi) = \frac{x_i}{p_i} - \frac{x_n}{p_n} \qquad \text{e} \qquad \partial_j\partial_i\ell(\xi) = -\left(\frac{x_i}{p_i^2}\delta_{ij} + \frac{x_n}{p_n^2}\right).$$

[5] A rigor, os argumentos da distância de Fisher-Rao deveriam ser os vetores $\xi_p = (p_1,\dots,p_{n-1})$ e $\xi_q = (q_1,\dots,q_{n-1})$, que parametrizam cada distribuição. No entanto, por abuso de notação e para simplificar, usamos como argumentos os vetores $\boldsymbol{p} = (p_1,\dots,p_n)$ e $\boldsymbol{q} = (q_1,\dots,q_n)$ do simplexo.

Assim, os elementos da matriz de Fisher são dados por

$$\begin{aligned} g_{ij}(\xi) &= -\mathbb{E}\left[\partial_i \partial_j \ell(\xi)\right] = \mathbb{E}\left[\frac{X_i}{p_i^2}\delta_{ij} + \frac{X_n}{p_n^2}\right] \\ &= n\left(\frac{\delta_{ij}}{p_i} + \frac{1}{p_n}\right) \\ &= n\left(\frac{\delta_{ij}}{\xi^i} + \frac{1}{1-\sum_{k=1}^{n-1}\xi^k}\right), \end{aligned}$$

em que usamos que $\mathbb{E}[X_i] = np_i$. Observamos que essa matriz de Fisher é a mesma da distribuição categórica, a menos da multiplicação por n. Assim, a distância de Fisher-Rao entre duas distribuições $f_p = \iota(\boldsymbol{p})$ e $f_q = \iota(\boldsymbol{q})$ é dada por

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

4.3.2 Distribuições contínuas

Exemplo 4.7 (Exponencial [3])**.** Uma distribuição exponencial tem densidade $f(x) = \lambda e^{-\lambda x}$, definida para $x \in \mathbb{R}_+$, e parametrizada por $\lambda \in \mathbb{R}_+^*$. Nesse caso, temos $\partial_\lambda \ell(\lambda) = \frac{1}{\lambda} - x$, e a informação de Fisher é

$$\begin{aligned} g_{11} &= \mathbb{E}\left[\left(\partial_\lambda \ell(\lambda)\right)^2\right] = \mathbb{E}\left[\left(\frac{1}{\lambda} - X\right)^2\right] \\ &= \frac{1}{\lambda^2} - \frac{2}{\lambda}\mathbb{E}[X] + \mathbb{E}[X^2] \\ &= \frac{1}{\lambda^2}, \end{aligned}$$

em que usamos que $\mathbb{E}[X] = \frac{1}{\lambda}$ e $\mathbb{E}[X^2] = \frac{2}{\lambda^2}$. A distância de Fisher-Rao é dada por

$$d_{\mathrm{FR}}(\lambda_1, \lambda_2) = \left|\int_{\lambda_1}^{\lambda_2} \frac{1}{t}\,\mathrm{d}t\right| = \left|\log\lambda_1 - \log\lambda_2\right|.$$

Exemplo 4.8 (Rayleigh)**.** Uma distribuição de Rayleigh tem densidade $f(x) = \frac{x}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right)$, definida para $x \in \mathbb{R}_+$, e parametrizada por $\sigma \in \mathbb{R}_+^*$. Temos $\partial_\sigma \ell(\sigma) = \frac{x^2}{\sigma^3} - \frac{2}{\sigma}$, e a informação de

Fisher é

$$\begin{aligned} g_{11} &= \mathbb{E}\left[\left(\partial_\sigma \ell(\sigma)\right)^2\right] = \mathbb{E}\left[\left(\frac{X^2}{\sigma^3} - \frac{2}{\sigma}\right)^2\right] \\ &= \frac{\mathbb{E}[X^4]}{\sigma^6} - \frac{4\mathbb{E}[X^2]}{\sigma^4} + \frac{4}{\sigma^2} \\ &= \frac{4}{\sigma^2}, \end{aligned}$$

em que usamos que $\mathbb{E}[X^2] = 4\sigma^2$ e $\mathbb{E}[X^4] = 8\sigma^4$. Assim, a distância de Fisher-Rao é dada por

$$d_{\mathrm{FR}}(\sigma_1, \sigma_2) = \left| \int_{\sigma_1}^{\sigma_2} \frac{2}{t}\, \mathrm{d}t \right| = 2\left|\log \sigma_1 - \log \sigma_2\right|.$$

Exemplo 4.9 (Gaussiana [3, 9, 11, 16])**.** Uma distribuição gaussiana tem densidade $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$, definida para $x \in \mathbb{R}$, e parametrizada pelo par $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+^*$. Nesse caso, temos

$$\partial_\mu \ell \coloneqq \partial_\mu \ell(\mu, \sigma) = \frac{x-\mu}{\sigma^2} \qquad \text{e} \qquad \partial_\sigma \ell \coloneqq \partial_\sigma \ell(\mu, \sigma) = -\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3}.$$

Os elementos da matriz de Fisher são então

$$\begin{aligned} g_{11} &= \mathbb{E}\left[\left(\partial_\mu \ell\right)^2\right] = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma^2}\right)^2\right] \\ &= \frac{\mathbb{E}[(X-\mu)^2]}{\sigma^4} = \frac{1}{\sigma^2}, \\ g_{12} = g_{21} &= \mathbb{E}\left[\left(\partial_\mu \ell\right)\left(\partial_\sigma \ell\right)\right] = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma^2}\right)\left(-\frac{1}{\sigma} + \frac{(X-\mu)^2}{\sigma^3}\right)\right] \\ &= -\frac{\mathbb{E}[X-\mu]}{\sigma^3} + \frac{\mathbb{E}[(X-\mu)^3]}{\sigma^5} = 0, \\ g_{22} &= \mathbb{E}\left[\left(\partial_\sigma \ell\right)^2\right] = \mathbb{E}\left[\left(-\frac{1}{\sigma} + \frac{(X-\mu)^2}{\sigma^3}\right)^2\right] \\ &= \frac{1}{\sigma^2} - \frac{2\mathbb{E}[(X-\mu)^2]}{\sigma^4} + \frac{\mathbb{E}[(X-\mu)^4]}{\sigma^6} = \frac{2}{\sigma^2}, \end{aligned}$$

em que usamos que os momentos centralizados ímpares da variável aleatória gaussiana são nulos, e que $\mathbb{E}[(X-\mu)^2] = \sigma^2$ e $\mathbb{E}[(X-\mu)^4] = 3\sigma^4$. Assim, a matriz de Fisher é dada por

$$G = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}. \tag{4.14}$$

Em [11, § 2.1], as geodésicas dessa variedade são calculadas através da resolução das equações diferenciais das geodésicas. Em [3, 9, 16], a geometria dessa variedade estatística é estudada em associação com a métrica hiperbólica do semiplano de Poincaré. Em particular, [16] realiza uma cuidadosa leitura geométrica e estabelece conexões com outras medidas de divergência. Seguiremos esse trabalho na dedução a seguir.

O modelo do semiplano de Poincaré é $\mathcal{H} \coloneqq \left\{ z = x + iy \in \mathbb{C} \;\middle|\; \operatorname{Im}(z) > 0 \right\} \cong \mathbb{R} \times \mathbb{R}_+^*$ com a métrica

$$G_{\mathcal{H}} = \begin{pmatrix} \frac{1}{y^2} & 0 \\ 0 & \frac{1}{y^2} \end{pmatrix}. \tag{4.15}$$

A geometria dessa variedade é bem estudada e a distância geodésica entre dois pontos $z, w \in \mathcal{H}$ é dada pelas expressões equivalentes [6, § 7.2]

$$d_{\mathcal{H}}(z, w) = \log \frac{|z - \overline{w}| + |z - w|}{|z - \overline{w}| - |z - w|} = \operatorname{arccosh}\left(1 + \frac{|z - w|^2}{2 \operatorname{Im}(z) \operatorname{Im}(w)}\right) = \operatorname{arctanh}\left|\frac{z - w}{z - \overline{w}}\right|. \tag{4.16}$$

Comparando (4.14) e (4.15), podemos associar a distância de Fisher-Rao na variedade das gaussianas com a distância do semiplano de Poincaré, cf. [16]:

$$d_{\mathrm{FR}}\big((\mu_1, \sigma_1), (\mu_2, \sigma_2)\big) = \sqrt{2} d_{\mathcal{H}}\left(\frac{\mu_1}{\sqrt{2}} + i\sigma_1, \frac{\mu_2}{\sqrt{2}} + i\sigma_2\right).$$

Assim, a distância de Fisher-Rao é obtida como

$$d_{\mathrm{FR}}\big((\mu_1, \sigma_1), (\mu_2, \sigma_2)\big) = \sqrt{2} \operatorname{arctanh}\left(\sqrt{\frac{(\mu_1 - \mu_2)^2 + 2(\sigma_1 - \sigma_2)^2}{(\mu_1 - \mu_2)^2 + 2(\sigma_1 + \sigma_2)^2}}\right).$$

Exemplo 4.10 (Laplace)**.** Uma distribuição de Laplace tem densidade $f(x) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$, definida para $x \in \mathbb{R}$, com parâmetros $(\mu, b) \in \mathbb{R} \times \mathbb{R}_+^*$. Nesse caso, temos

$$\partial_\mu \ell \coloneqq \partial_\mu \ell(\mu, b) = \begin{cases} +\frac{1}{b}, & x > \mu \\ -\frac{1}{b}, & x < \mu \end{cases} \qquad \text{e} \qquad \partial_b \ell \coloneqq \partial_b \ell(\mu, b) = -\frac{1}{b} + \frac{|x - \mu|}{b^2}.$$

Os elementos da matriz de Fisher são dados por

$$\begin{aligned}
g_{11} &= \mathbb{E}\left[\left(\partial_\mu \ell\right)^2\right] = \mathbb{E}\left[\left(-\frac{1}{b}\mathbb{1}_{]-\infty,\mu[}(X) + \frac{1}{b}\mathbb{1}_{]\mu,\infty[}(X)\right)^2\right] \\
&= \frac{1}{b^2}, \\
g_{12} = g_{21} &= \mathbb{E}\left[\left(\partial_\mu \ell\right)\left(\partial_b \ell\right)\right] = \int_{-\infty}^{+\infty} f(x)\left(\partial_\mu \ell\right)\left(\partial_b \ell\right)\, \mathrm{d}x \\
&= \int_{-\infty}^{\mu} f(x)\left(-\frac{1}{b}\right)\left(-\frac{1}{b} + \frac{|x-\mu|}{b^2}\right)\mathrm{d}x + \int_{\mu}^{+\infty} f(x)\left(\frac{1}{b}\right)\left(-\frac{1}{b} + \frac{|x-\mu|}{b^2}\right)\mathrm{d}x \\
&= 0, \\
g_{22} &= \mathbb{E}\left[\left(\partial_\sigma b\right)^2\right] = \mathbb{E}\left[\left(-\frac{1}{b} + \frac{|X-\mu|}{b^2}\right)^2\right] \\
&= \frac{1}{b^2} - \frac{2\mathbb{E}[|X-\mu|]}{b^3} + \frac{\mathbb{E}[|X-\mu|^2]}{b^4} \\
&= \frac{1}{b^2},
\end{aligned}$$

em que usamos o fato de que a função $f(x)\left(-\frac{1}{b^2} + \frac{|x-\mu|}{b^3}\right)$ é simétrica em relação à média $x = \mu$, e que $\mathbb{E}[|X-\mu|] = b$ e $\mathbb{E}[|X-\mu|^2] = 2b^2$. Assim, a matriz de Fisher fica

$$G = \begin{pmatrix} \frac{1}{b^2} & 0 \\ 0 & \frac{1}{b^2} \end{pmatrix},$$

que coincide com a métrica hiperbólica do semiplano de Poincaré. Portanto, a distância de Fisher-Rao nessa variedade estatística é dada por

$$d_{\mathrm{FR}}\big((\mu_1, b_1), (\mu_2, b_2)\big) = d_{\mathcal{H}}\left(\mu_1 + ib_1, \mu_2 + ib_2\right) = \operatorname{arctanh}\left(\sqrt{\frac{(\mu_1-\mu_2)^2 + (b_1-b_2)^2}{(\mu_1-\mu_2)^2 + (b_1+b_2)^2}}\right).$$

Exemplo 4.11 (Pareto [38])**.** Uma distribuição de Pareto é tem densidade $f(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}}\mathbb{1}_{[x_m,\infty[}(x)$, parametrizada por $(\alpha, x_m) \in \mathbb{R}_+^* \times \mathbb{R}_+^*$. Notamos que seu suporte depende da parametrização, de modo que o modelo estatístico associado infringe uma das hipóteses acordadas na Seção 4.2. Ainda assim, é possível calcular uma métrica riemanniana a partir da definição da matriz de Fisher; porém, não é possível utilizar as fórmulas alternativas para obtenção dessa matriz. Temos

$$\partial_\alpha \ell \coloneqq \partial_\alpha \ell(\alpha, x_m) = \frac{1}{\alpha} + \log x_m - \log x \qquad \text{e} \qquad \partial_{x_m} \ell \coloneqq \partial_{x_m} \ell(\alpha, x_m) = \frac{\alpha}{x_m}.$$

Os elementos da matriz de Fisher são calculados como

$$\begin{aligned}
g_{11} &= \mathbb{E}\left[(\partial_\alpha \ell)^2\right] = \mathbb{E}\left[\left(\frac{1}{\alpha} + \log x_m - \log X\right)^2\right] \\
&= \frac{1}{\alpha^2} + (\log x_m)^2 + \mathbb{E}\left[(\log X)^2\right] + \frac{2}{\alpha}\left(\log x_m - \mathbb{E}[\log X]\right) - 2\log x_m \mathbb{E}[\log X] \\
&= \frac{1}{\alpha^2}, \\
g_{12} = g_{21} &= \mathbb{E}\left[(\partial_\alpha \ell)\,(\partial_{x_m} \ell)\right] = \mathbb{E}\left[\left(\frac{1}{\alpha} + \log x_m - \log X\right)\left(\frac{\alpha}{x_m}\right)\right] \\
&= \frac{1}{x_m} + \frac{\alpha}{x_m}\log x_m - \frac{\alpha}{x_m}\mathbb{E}[\log X] = 0, \\
g_{22} &= \mathbb{E}\left[(\partial_{x_m} \ell)^2\right] = \mathbb{E}\left[\left(\frac{\alpha}{x_m}\right)^2\right] \\
&= \frac{\alpha^2}{x_m^2},
\end{aligned}$$

em que usamos que $\mathbb{E}[\log X] = \frac{1}{\alpha} + \log x_m$ e $\mathbb{E}[(\log X)^2] = \frac{2}{\alpha^2} + \frac{2\log x_m}{\alpha} + (\log x_m)^2$. Assim, a matriz de Fisher é

$$G = \begin{pmatrix} \frac{1}{\alpha^2} & 0 \\ 0 & \frac{\alpha^2}{x_m^2} \end{pmatrix}.$$

É possível usar a matriz da métrica para explicitamente calcular os símbolos de Christoffel, encontrar as geodésicas e obter uma expressão para a distância de Fisher-Rao nessa variedade estatística. Um recente trabalho [38] adotou uma abordagem diferente, ao mostrar que $(\alpha, x_0) \mapsto (\log x_0, 1/\alpha)$ é uma isometria entre a variedade estatística em questão e o semiplano de Poincaré. Assim, foi possível escrever a expressão para a distância de Fisher-Rao como

$$\begin{aligned}
d_{\mathrm{FR}}\big((\alpha_1, x_{m,1}), (\alpha_2, x_{m,2})\big) &= d_{\mathcal{H}}\left(\log x_{m,1} + i\frac{1}{\alpha_1}, \log x_{m,2} + i\frac{1}{\alpha_2}\right) \\
&= \operatorname{arctanh}\left(\sqrt{\frac{(\alpha_1\alpha_2 \log(x_{m,1}/x_{m,2}))^2 + (\alpha_1 - \alpha_2)^2}{(\alpha_1\alpha_2 \log(x_{m,1}/x_{m,2}))^2 + (\alpha_1 + \alpha_2)^2}}\right).
\end{aligned}$$

Exemplo 4.12 (Cauchy)**.** Uma distribuição de Cauchy tem densidade $f(x) = \frac{\gamma}{\pi\left[(x-x_0)^2+\gamma^2\right]}$, definida para $x \in \mathbb{R}$, e parametrizada por $(x_0, \gamma) \in \mathbb{R} \times \mathbb{R}_+^*$. Temos

$$\partial_{x_0}\ell := \partial_{x_0}\ell(x_0, \gamma) = \frac{2(x - x_0)}{(x - x_0)^2 + \gamma^2} \qquad \text{e} \qquad \partial_{x_0}\ell := \partial_{x_0}\ell(x_0, \gamma) = \frac{1}{\gamma} - \frac{2\gamma}{(x - x_0)^2 + \gamma^2}.$$

Os elementos da matriz de Fisher são dados por

$$\begin{aligned}
g_{11} &= \mathbb{E}\left[(\partial_{x_0}\ell)^2\right] = \mathbb{E}\left[\left(\frac{2(X-x_0)}{(X-x_0)^2+\gamma^2}\right)^2\right] \\
&= 4\mathbb{E}\left[\frac{(X-x_0)^2}{[(X-x_0)^2+\gamma^2]^2}\right] = \frac{1}{\gamma^2}, \\
g_{12} = g_{21} &= \mathbb{E}\left[(\partial_{x_0}\ell)\left(\partial_\gamma\ell\right)\right] = \mathbb{E}\left[\left(\frac{2(X-x_0)}{(X-x_0)^2+\gamma^2}\right)\left(\frac{1}{\gamma}-\frac{2\gamma}{(X-x_0)^2+\gamma^2}\right)\right] \\
&= \frac{2}{\gamma}\mathbb{E}\left[\frac{X-x_0}{(X-x_0)^2+\gamma^2}\right] - 4\gamma\mathbb{E}\left[\frac{X-x_0}{[(X-x_0)^2+\gamma^2]^2}\right] = 0, \\
g_{22} &= \mathbb{E}\left[\left(\partial_\gamma\ell\right)^2\right] = \mathbb{E}\left[\left(\frac{1}{\gamma}-\frac{2\gamma}{(X-x_0)^2+\gamma^2}\right)^2\right] \\
&= \frac{1}{\gamma^2} - 4\mathbb{E}\left[\frac{1}{(X-x_0)^2+\gamma^2}\right] + 4\gamma^2\mathbb{E}\left[\frac{1}{[(X-x_0)^2+\gamma^2]^2}\right] \\
&= \frac{1}{2\gamma^2},
\end{aligned}$$

em que os resultados foram obtidos calculando as esperanças: $\mathbb{E}\left[\frac{(X-x_0)^2}{[(X-x_0)^2+\gamma^2]^2}\right] = \frac{1}{8\gamma^2}$, $\mathbb{E}\left[\frac{X-x_0}{(X-x_0)^2+\gamma^2}\right] = \mathbb{E}\left[\frac{X-x_0}{[(X-x_0)^2+\gamma^2]^2}\right] = 0$, $\mathbb{E}\left[\frac{1}{(X-x_0)^2+\gamma^2}\right] = \frac{1}{2\gamma^2}$, $\mathbb{E}\left[\frac{1}{[(X-x_0)^2+\gamma^2]^2}\right] = \frac{3}{8\gamma^4}$. Assim, a matriz de Fisher é

$$G = \begin{pmatrix} \frac{1}{2\gamma^2} & 0 \\ 0 & \frac{1}{2\gamma^2} \end{pmatrix} = \frac{1}{2}\begin{pmatrix} \frac{1}{\gamma^2} & 0 \\ 0 & \frac{1}{\gamma^2} \end{pmatrix},$$

que é a métrica hiperbólica do semiplano de Poincaré, a menos de fator multiplicativo. Isso permite calcular a distância de Fisher-Rao como

$$\begin{aligned}
d_{\mathrm{FR}}\big((x_{0,1},\gamma_1),(x_{0,2},\gamma_2)\big) &= \frac{1}{\sqrt{2}} d_{\mathcal{H}}\left(x_{0,1}+i\gamma_1, x_{0,2}+i\gamma_2\right) \\
&= \frac{1}{\sqrt{2}}\operatorname{arctanh}\left(\sqrt{\frac{(x_{0,1}-x_{0,2})^2+(\gamma_1-\gamma_2)^2}{(x_{0,1}-x_{0,2})^2+(\gamma_1+\gamma_2)^2}}\right).
\end{aligned}$$

4.4 Aprendizado com a função perda de Fisher-Rao

O problema de classificação supervisionada é um problema importante em aprendizado de máquina [7, 10, 27]. O treinamento de um classificador (e.g., uma rede neural) pode ser feito por minimização do risco empírico: um algoritmo numérico de otimização é aplicado para encontrar os parâmetros do modelo que minimizam o valor médio da função perda no conjunto de dados de treinamento. Nesse contexto, escolher uma função perda adequada é fundamental,

Tabela 4.1: Resumo das distâncias de Fisher-Rao para distribuições discretas.

	Distribuição	Suporte	Parâmetros	Matriz de Fisher	Distância de Fisher-Rao	Referências
Binomial	$\binom{n}{x}p^x(1-p)^{n-x}$	$x \in \mathbb{N}$	$p \in]0,1[$	$\left(\frac{n}{p(1-p)}\right)$	$2\sqrt{n}\left\lvert\arcsin\sqrt{p_1} - \arcsin\sqrt{p_2}\right\rvert$	[3, 9, 11]
Poisson	$\lambda^x e^{-\lambda}/x!$	$x \in \mathbb{N}$	$\lambda \in \mathbb{R}_+^*$	$\left(\frac{1}{\lambda}\right)$	$2\left\lvert\sqrt{\lambda_1} - \sqrt{\lambda_2}\right\rvert$	[3, 9, 11]
Geométrica	$p(1-p)^{x-1}$	$x \in \mathbb{N}^*$	$p \in]0,1[$	$\left(\frac{1}{p^2(1-p)}\right)$	$2\left\lvert\operatorname{arctanh}\sqrt{1-p_1} - \operatorname{arctanh}\sqrt{1-p_2}\right\rvert$	[11]
Binomial negativa	$\binom{x+r-1}{r-1}p^r(1-p)^x$	$x \in \mathbb{N}$	$p \in]0,1[$	$\left(\frac{r}{p^2(1-p)}\right)$	$2\sqrt{r}\left\lvert\operatorname{arctanh}\sqrt{1-p_1} - \operatorname{arctanh}\sqrt{1-p_2}\right\rvert$	[9]
Categórica	$\sum_{i=1}^n p_i \mathbb{1}_{\{i\}}(x)$	$x \in \{1,\dots,n\}$	$(p_1,\dots,p_{n-1}) \in [0,1]^{n-1}$, $\sum_i p_i \leq 1$, $p_n \coloneqq 1 - \sum_i p_i$	$\left(\frac{1}{p_i} + \frac{1}{p_n}\right)_{i,j}$	$2\arccos\left(\sum_{i=1}^n \sqrt{p_i q_i}\right)$	[3, 4, 9, 11]
Multinomial	$\frac{n!}{x_1!\cdots x_n!}p_1^{x_1}\cdots p_n^{x_n}$	$(x_1,\dots,x_n) \in \mathbb{N}^n$, $\sum_i x_i = m$	$(p_1,\dots,p_{n-1}) \in [0,1]^{n-1}$, $\sum_i p_i \leq 1$, $p_n \coloneqq 1 - \sum_i p_i$	$\left(\frac{n}{p_i} + \frac{n}{p_n}\right)_{i,j}$	$2\sqrt{n}\arccos\left(\sum_{i=1}^n \sqrt{p_i q_i}\right)$	[3, 9, 11]

Tabela 4.2: Resumo das distâncias de Fisher-Rao para distribuições contínuas.

	Distribuição	Suporte	Parâmetros	Matriz de Fisher	Distância de Fisher-Rao	Referências
Exponencial	$\lambda e^{-\lambda x}$	$x \in \mathbb{R}_+$	$\lambda \in \mathbb{R}_+^*$	$\left(\frac{1}{\lambda^2}\right)$	$\lvert\log\lambda_1 - \log\lambda_2\rvert$	[3]
Rayleigh	$\frac{x}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right)$	$x \in \mathbb{R}_+$	$\sigma \in \mathbb{R}_+^*$	$\left(\frac{4}{\sigma^2}\right)$	$2\lvert\log\sigma_1 - \log\sigma_2\rvert$	$*$
Gaussiana	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$x \in \mathbb{R}$	$(\mu,\sigma) \in \mathbb{R}\times\mathbb{R}_+^*$	$\begin{pmatrix}\frac{1}{\sigma^2} & 0\\ 0 & \frac{2}{\sigma^2}\end{pmatrix}$	$\sqrt{2}\operatorname{arctanh}\left(\sqrt{\frac{(\mu_1-\mu_2)^2+2(\sigma_1-\sigma_2)^2}{(\mu_1-\mu_2)^2+2(\sigma_1+\sigma_2)^2}}\right)$	[3, 9, 11, 16]
Laplace	$\frac{1}{2b}\exp\left(-\frac{\lvert x-\mu\rvert}{b}\right)$	$x \in \mathbb{R}$	$(\mu,b) \in \mathbb{R}\times\mathbb{R}_+^*$	$\begin{pmatrix}\frac{1}{b^2} & 0\\ 0 & \frac{1}{b^2}\end{pmatrix}$	$\operatorname{arctanh}\left(\sqrt{\frac{(\mu_1-\mu_2)^2+(b_1-b_2)^2}{(\mu_1-\mu_2)^2+(b_1+b_2)^2}}\right).$	$*$
Pareto	$\frac{\alpha x_m^\alpha}{x^{\alpha+1}}\mathbb{1}_{[x_m,\infty[}(x)$	$x \in [x_m,\infty[$	$(\alpha,x_m) \in \mathbb{R}_+^*\times\mathbb{R}_+^*$	$\begin{pmatrix}\frac{1}{\alpha^2} & 0\\ 0 & \frac{\alpha^2}{x_m^2}\end{pmatrix}$	$\operatorname{arctanh}\left(\sqrt{\frac{(\alpha_1\alpha_2\log(x_{m,1}/x_{m,2}))^2+(\alpha_1-\alpha_2)^2}{(\alpha_1\alpha_2\log(x_{m,1}/x_{m,2}))^2+(\alpha_1+\alpha_2)^2}}\right)$	[38]
Cauchy	$\frac{\gamma}{\pi\left[(x-x_0)^2+\gamma^2\right]}$	$x \in \mathbb{R}$	$(x_0,\gamma) \in \mathbb{R}\times\mathbb{R}_+^*$	$\begin{pmatrix}\frac{1}{2\gamma^2} & 0\\ 0 & \frac{1}{2\gamma^2}\end{pmatrix}$	$\frac{1}{\sqrt{2}}\operatorname{arctanh}\left(\sqrt{\frac{(x_{0,1}-x_{0,2})^2+(\gamma_1-\gamma_2)^2}{(x_{0,1}-x_{0,2})^2+(\gamma_1+\gamma_2)^2}}\right)$	$*$

Os resultados marcados com $*$ foram derivados neste trabalho.

pois diferentes escolhas podem afetar o desempenho do classificador resultante, assim como a dinâmica de treinamento.

A saída da rede neural treinada para classificação é frequentemente interpretada como representando uma distribuição de probabilidade condicional $P(y|\boldsymbol{x})$ da classe y, dado o vetor de entrada $\boldsymbol{x}$, o que motiva o uso da entropia cruzada como função perda. Apesar de originalmente usado para problemas de regressão, o erro quadrático médio também é usado como função perda, e diversos trabalhos têm comparado essas duas perdas [18, 26, 30, 31, 33]. Além disso, a busca por outras funções perda tem sido um tema de pesquisa ativo, e perdas com diversas inspirações têm sido propostas, não raro para problemas ou contextos específicos [14, 23, 29, 48].

A proposta aqui é usar a distância de Fisher-Rao na variedade de distribuições discretas como função e estudar seu desempenho em comparação com outras perdas comumente usadas. Em particular, estamos interessados nesse estudo no caso de haver ruído de rótulo, i.e., quando alguns dos rótulos das classes podem estar incorretos. Este é um problema clássico em aprendizado de máquina e várias soluções têm sido propostas [22]. Uma solução robusta, facilmente implementável, e que permite derivar garantias teóricas é adotar funções perda que sejam inerentemente tolerantes a esse tipo de ruído. Mostraremos que a função perda baseada na distância de Fisher-Rao oferece um compromisso natural entre robustez a ruído de rótulo e velocidade de aprendizado. Esta contribuição pode ser encontrada em [42].

4.4.1 Formulação do problema

Consideramos o problema de classificação supervisionada: $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^n$ denota o *vetor de entrada*, que pertence a exatamente uma *classe*, rotulada por $\mathcal{Y} \coloneqq \{1, \dots, K\}$, e consideramos que o conjunto de dados segue uma distribuição conjunta $(\boldsymbol{x}, y) \sim \mathcal{D}$. O objetivo é realizar o aprendizado de um classificador (e.g., uma rede neural) $f\colon \mathcal{X} \to \mathbb{R}^K$ que associa a cada vetor de entrada $\boldsymbol{x}$ um vetor de *pontuações* $\boldsymbol{s} = (s_1, \dots, s_K) \coloneqq f(\boldsymbol{x})$, que pode ser usado para induzir uma decisão $\hat{y} = \arg\max_{1 \le i \le K} s_i$. Ao aplicar a função *softmax* $\sigma\colon \mathbb{R}^K \to \Delta^{K-1}$ às pontuações $\boldsymbol{s}$, obtém-se um vetor interpretado como uma distribuição de probabilidade condicional $P(y|\boldsymbol{x})$ em $\mathcal{Y}$, dado por $\boldsymbol{p} = (p_1, \dots, p_K) \coloneqq (\sigma \circ f)(\boldsymbol{x}) = \sigma\big((s_1, \dots, s_K)\big)$, com $p_i = e^{s_i} / \sum_{j=1}^{K} e^{s_j}$, para $1 \le i \le K$.

O treinamento do classificador pode ser feito por minimização do risco empírico, que é a versão populacional do problema $\min_f \mathbb{E}_{\mathcal{D}}\big[L(y, f(\boldsymbol{x})\big]$, sob uma escolha adequada de *função perda* $L\colon \mathcal{Y} \times \mathbb{R}^K \to \mathbb{R}_+$. Definimos o *risco* de uma função perda L como seu valor esperado:

$$R_L \coloneqq R_L(f) \coloneqq \mathbb{E}_{\mathcal{D}}\Big[L\big(y, f(\boldsymbol{x})\big)\Big]. \tag{4.17}$$

Em um problema de classificação supervisionada, o classificador tem acesso a um conjunto de

treinamento $\left\{(\boldsymbol{x}_i, y_i)\right\}_{i=1}^N$, de modo que o *risco empírico* é dado por

$$\bar{R}_L \coloneqq \bar{R}_L(f) \coloneqq \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, f(\boldsymbol{x}_i)\right),$$

e o problema torna-se encontrar $\min_f \bar{R}_L(f)$.

4.4.2 Funções perda

A seguir, apresentaremos algumas perdas conhecidas na literatura. Denotamos $\boldsymbol{e}^{(y)} \in \mathbb{R}^K$ o vetor que contém 1 na y-ésima posição e 0 nas posições restantes, i.e., sua i-ésima componente $e_i^{(y)}$ é igual ao delta de Kroenecker δ_{iy}, para $1 \leq i \leq K$.

Erro quadrático médio (MSE). A perda do *erro quadrático médio* (MSE) é definida como

$$L_{\text{MSE}}\left(y, f(\boldsymbol{x})\right) \coloneqq \|\boldsymbol{e}^{(y)} - (\sigma \circ f)(\boldsymbol{x})\|_2^2 = \|\boldsymbol{p}\|_2^2 - 2p_y + 1, \tag{4.18}$$

que é o quadrado da distância euclidiana no espaço ambiente $\mathbb{R}^K$, restrita ao simplexo Δ^{K-1}.

Erro absoluto médio (MAE). A perda do *erro médio absoluto* (MAE) é definida como[6]

$$L_{\text{MAE}}\left(y, f(\boldsymbol{x})\right) \coloneqq \frac{1}{2}\|\boldsymbol{e}^{(y)} - (\sigma \circ f)(\boldsymbol{x})\|_1 = 1 - p_y, \tag{4.19}$$

que é proporcional à norma ℓ_1 no espaço ambiente $\mathbb{R}^K$, restrita ao simplexo Δ^{K-1}.

Entropia cruzada (CE). A perda da *entropia cruzada* (CE) é definida como

$$L_{\text{CE}}\left(y, f(\boldsymbol{x})\right) \coloneqq -\sum_{i=1}^{K} e_i^{(y)} \log[(\sigma \circ f)(\boldsymbol{x})]_i = -\log p_y, \tag{4.20}$$

onde $[\boldsymbol{v}]_i$ denota a i-ésima coordenada do vetor $\boldsymbol{v}$. Note que

$$L_{\text{CE}}\left(y, f(\boldsymbol{x})\right) = D_{KL}\left(\boldsymbol{e}^{(y)} \| (\sigma \circ f)(\boldsymbol{x})\right) + \underbrace{H\left(\boldsymbol{e}^{(y)}\right)}_{=0},$$

i.e., neste caso, a entropia cruzada é igual à divergência de Kullback-Leibler entre a distribuição verdadeira $\boldsymbol{e}^{(y)}$, que é determinística, e a distribuição predita $\boldsymbol{p} = (\sigma \circ f)(\boldsymbol{x})$.

[6]Normalmente, essa perda é definida como $\|\boldsymbol{e}^{(y)} - (\sigma \circ f)(\boldsymbol{x})\|_1$, mas adotamos o fator 1/2 por uma questão de normalização na comparação com outras funções perda. Neste caso, ela coincide com a perda *unhinged* [51] e com a distância da variação total [50, § 2.4].

q-entropia cruzada (q-CE). Uma generalização da função perda da entropia cruzada pode ser definida ao se considerar o q-logaritmo de Tsallis[7] [49]:

$$\log_q(x) \coloneqq \begin{cases} \frac{x^{1-q}-1}{1-q}, & q \neq 1, \\ \log(x), & q = 1, \end{cases}$$

definido para $x > 0$ e $q \in \mathbb{R}$. Em particular, ao substituir o logaritmo usual em (4.20) pelo q-logaritmo de Tsallis, com $q \in [0, 1)$, obtém-se a perda definida[8] em [54]:

$$L_{q\text{-CE}}(y, f(\boldsymbol{x})) \coloneqq -\sum_{i=1}^{K} e_i^{(y)} \log_q[(\sigma \circ f)(\boldsymbol{x})]_i = -\log_q p_y, \tag{4.21}$$

a que chamaremos *q-entropia cruzada* (q-CE). Note que, ao tomar $q = 0$, esta perda coincide com a perda MAE, enquanto que, para $q = 1$, coincide com a perda CE usual.

Distância de Fisher-Rao. Propomos uma função perda que é proporcional ao quadrado da distância de Fisher-Rao na variedade das distribuições discretas. Tomar o quadrado, além de não modificar o problema de otimização, é razoável, pois as perdas MSE e CE também se comportam, ao menos localmente, como distâncias ao quadrado.

De fato, usando a expressão da distância de Fisher-Rao para distribuições discretas (4.13), temos

$$d_{\text{FR}}^2\left(\boldsymbol{e}^{(y)}, (\sigma \circ f)(\boldsymbol{x})\right) = 4\left(\arccos\sqrt{[(\sigma \circ f)(\boldsymbol{x})]_y}\right)^2 = 4\left(\arccos\sqrt{p_y}\right)^2.$$

Desprezando a constante multiplicativa (o que corresponde a alterar a taxa de aprendizado pelo mesmo fator), definimos então a perda de *Fisher-Rao* como

$$L_{\text{FR}}(y, f(\boldsymbol{x})) \coloneqq \left(\arccos\sqrt{[(\sigma \circ f)(\boldsymbol{x})]_y}\right)^2 = \left(\arccos\sqrt{p_y}\right)^2. \tag{4.22}$$

Distância de Hellinger. A reparametrização usada no Exemplo 4.5 para obter a distância de Fisher-Rao da variedade das distribuições discretas, que leva o simplexo Δ^{K-1} na esfera $S_{2,+}^{K-1}$, imediatamente sugere uma aproximação para essa distância. Se a distância de Fisher-Rao é dada pelo comprimento de um arco na esfera, podemos aproximá-la pela distância cordal da

[7] Aqui, $\log_q$ denota o q-logaritmo de Tsallis, que não deve ser confundido com o logaritmo na base q.

[8] Essa perda foi originalmente definida em [54] em termos do negativo da transformação de Box-Cox, que está relacionada ao q-logaritmo de Tsallis. Além disso, o parâmetro q que usamos aqui corresponde ao parâmetro $1 - q$ daquele trabalho.

esfera, i.e., a distância euclidiana no espaço ambiente $\mathbb{R}^K$. Usando a notação daquele exemplo:

$$\|\boldsymbol{z}_p - \boldsymbol{z}_q\|^2 = 2\left(\sum_{i=1}^{K}\left(\sqrt{p_i} - \sqrt{q_i}\right)\right)^{1/2}.$$

Eis que esta quantidade é o dobro de uma conhecida distância entre distribuições de probabilidade, a chamada distância de Hellinger [50, § 2.4], que é dada por $d_{\mathrm{H}}(\boldsymbol{p}, \boldsymbol{q}) = \left(\sum_{i=1}^{K}\left(\sqrt{p_i} - \sqrt{q_i}\right)\right)^{1/2}$. Assim, podemos também considerar uma função perda que é igual ao quadrado dessa distância:

$$d_{\mathrm{H}}^2\left(\boldsymbol{e}^{(y)}, (\sigma \circ f)(\boldsymbol{x})\right) = \sum_{i=1}^{K}\left(\sqrt{\boldsymbol{e}_i^{(y)}} - \sqrt{\boldsymbol{p}_i}\right)^2 = 2(1 - \sqrt{p_y}).$$

Portanto, podemos definir a perda de *Hellinger* como

$$L_{\mathrm{H}}(y, f(\boldsymbol{x})) \coloneqq 2\left(1 - \sqrt{[(\sigma \circ f)(\boldsymbol{x})]_y}\right) = 2\,(1 - \sqrt{p_y}), \tag{4.23}$$

que também é mencionada em [10, § 3.9]. É interessante notar que este também é um caso particular da perda q-CE, para $q = 1/2$.

Relação entre as diferentes perdas

A proposição a seguir estabelece relações entre diferentes funções perda.

Proposição 4.9. *Sejam L_{CE}, L_{FR} e L_{H} as funções perda da entropia cruzada, de Fisher-Rao e de Hellinger em* (4.20), (4.22) *e* (4.23), *respectivamente. Então temos:*

1. $L_{\mathrm{FR}}(y, f(\boldsymbol{x})) = L_{\mathrm{H}}(y, f(\boldsymbol{x})) + O\big(L_{\mathrm{H}}^2(y, f(\boldsymbol{x}))\big)$;
2. $L_{\mathrm{FR}}(y, f(\boldsymbol{x})) = L_{\mathrm{CE}}(y, f(\boldsymbol{x})) + O\big(L_{\mathrm{CE}}^2(y, f(\boldsymbol{x}))\big)$.

Ademais:

3. $L_{\mathrm{H}}(y, f(\boldsymbol{x})) \leq L_{\mathrm{FR}}(y, f(\boldsymbol{x})) \leq L_{\mathrm{CE}}(y, f(\boldsymbol{x}))$.

Demonstração. O item 1 é consequência direta da aproximação arco-corda que permitiu obter o dobro da distância de Hellinger como aproximação para a distância de Fisher das distribuições discretas. Especificamente, essas distâncias se relacionam por [32, § 7.2.2]

$$d_{\mathrm{FR}}(p, q) = 4 \arcsin\left(\frac{d_{\mathrm{H}}(p, q)}{2}\right). \tag{4.24}$$

Então, tomando a expansão em série de Taylor da função arco seno, obtemos

$$d_{\mathrm{FR}}(p,q) = 2d_{\mathrm{H}}(p,q) + O\left(d_{\mathrm{H}}^3(p,q)\right).$$

Para o item 2, isolando p_y em (4.22) e substituindo-o em (4.20), obtendo

$$L_{\mathrm{CE}}(y, f(\boldsymbol{x})) = -2\log\cos\sqrt{L_{\mathrm{FR}}(y, f(\boldsymbol{x}))}. \tag{4.25}$$

A seguir, expandimos a série de Taylor de $g(x) = -2\log\cos\sqrt{x}$ em torno de zero, até a primeira ordem, o que resulta em $g(x) = x + O(x^2)$ e fornece o resultado desejado.

A primeira desigualdade do item 3 é imediata de (4.24), i.e., do fato que a distância cordal é menor ou igual que a distância geodésica na esfera. Para a segunda desigualdade, note que a expansão de Taylor de primeira ordem de $g(x)$ em torno do zero é $g(x) = x + R_1$, com

$$R_1 = \frac{x^2}{2}g''(x^*) = \frac{x^2}{4}\left(\frac{\sec^2\sqrt{x^*}}{x^*} - \frac{\tan\sqrt{x^*}}{(x^*)^{3/2}}\right) \geq 0,$$

para $x^* \in (0, x) \subseteq (0, \pi^2/4)$. Logo $g(x) \geq x$, donde $L_{\mathrm{CE}}(y, f(\boldsymbol{x})) \geq L_{\mathrm{FR}}(y, f(\boldsymbol{x}))$. ■

4.4.3 Robustez a ruído de rótulo *versus* velocidade de aprendizado

Consideraremos agora o problema de aprendizado na presença de ruído de rótulo, cf. [24, 25]. Isso ocorre quando o classificador não tem acesso a um conjunto de treinamento $\left\{(\boldsymbol{x}_i, y_i)\right\}_{i=1}^N$ dito *limpo*, mas sim a uma versão *ruidosa* $\left\{(\boldsymbol{x}_i, \tilde{y}_i)\right\}_{i=1}^N$, na qual $\tilde{y}_i$ denota os rótulos ruidosos. Supomos que esses dados seguem uma distribuição conjunta de probabilidade $(\boldsymbol{x}, \tilde{y}) \sim \mathcal{D}_\eta$.

Vamos nos concentrar no caso de ruído de rótulo *uniforme* ou *simétrico*, i.e., quando os rótulos ruidosos $\tilde{y}_i$ não dependem dos vetores de entrada $\boldsymbol{x}_i$ nem dos rótulos verdadeiros y_i correspondentes. Neste caso, o ruído de rótulo é modelado por

$$\Pr(\tilde{y}_i = j | y_i = k) = \begin{cases} 1-\eta, & j = k, \\ \frac{\eta}{K-1}, & j \neq k, \end{cases} \tag{4.26}$$

para uma constante $\eta \in [0, 1]$, chamada *taxa de ruído*.

Analogamente ao risco $R_L(f)$ em (4.17), denotamos $R_L^\eta(f) \coloneqq \mathbb{E}_{\mathcal{D}_\eta}\left[L(\tilde{y}, f(\boldsymbol{x}))\right]$ o risco com relação aos dados ruidosos. Sejam f^* e $\hat{f}$ minimizadores globais de $R_L(f)$ e $R_L^\eta(f)$, respectivamente. Dizemos que a minimização do risco com respeito à função perda L é *tolerante a ruído* se o classificador $\hat{f}$ tem a mesma probabilidade de erro na classificação que f^* [24, 40]. Em outras palavras, o desempenho do classificador treinado com dados ruidosos é tão bom quanto o de um classificador treinado com dados limpos.

Robustez a ruído de rótulo

Uma contribuição importante de [24] foi fornecer uma condição suficiente para uma função perda ser robusta a ruído de rótulo uniforme:

Teorema 4.10 ([24, Theorem 1])**.** *Uma função perda L é tolerante a ruído de rótulo uniforme com taxa de ruído $\eta < \frac{K-1}{K}$, se satisfaz*

$$\sum_{i=1}^{K} L(i, f(\boldsymbol{x})) = C, \quad \forall x \in \mathcal{X},\ \forall f, \tag{4.27}$$

para alguma constante C.

Além disso, os autores observam que a perda MAE satisfaz essa condição, enquanto as perdas MSE e CE, não. De fato:

$$\begin{aligned}
\sum_{i=1}^{K} L_{\mathrm{MAE}}(i, f(\boldsymbol{x})) &= \sum_{i=1}^{K} (1 - p_i) = K - 1, \\
\sum_{i=1}^{K} L_{\mathrm{MSE}}(i, f(\boldsymbol{x})) &= \sum_{i=1}^{K} \left(\|\boldsymbol{p}\|_2^2 - 2p_i + 1 \right) = K \left(\|\boldsymbol{p}\|_2^2 + 1 \right) - 2, \\
\sum_{i=1}^{K} L_{\mathrm{CE}}(i, f(\boldsymbol{x})) &= \sum_{i=1}^{K} (-\log p_i) = \sum_{i=1}^{K} \log \frac{1}{p_i}.
\end{aligned}$$

No entanto, mesmo que uma função perda L não satisfaça à condição (4.27), se a quantidade $\sum_{i=1}^{K} L(i, f(\boldsymbol{x}))$ for limitada, ainda é possível obter alguma garantia teórica quanto à robustez a ruído de rótulo uniforme. Isso é feito para a perda q-CE em [54, Theorem 1]. Inspirados por esse trabalho, derivamos um resultado similar.

Lema 4.11. *A perda de Fisher-Rao L_{FR} satisfaz*

$$K \left(\arccos \frac{1}{\sqrt{K}} \right)^2 \leq \sum_{i=1}^{K} L_{\mathrm{FR}}(i, f(\boldsymbol{x})) \leq \frac{\pi^2}{4}(K - 1). \tag{4.28}$$

Demonstração. Usaremos multiplicadores de Lagrange para encontrar os pontos críticos de $F(\boldsymbol{p}) \coloneqq \sum_{j=1}^{K} (\arccos \sqrt{p_j})^2$, sujeito a $G(\boldsymbol{p}) \coloneqq \left(\sum_{j=1}^{K} p_j \right) - 1 = 0$. Como $\frac{\partial F}{\partial p_i}(\boldsymbol{p}) = -\frac{\arccos \sqrt{p_i}}{\sqrt{p_i(1-p_i)}}$ e $\frac{\partial G}{\partial p_i}(\boldsymbol{p}) = 1$, pontos críticos no interior de Δ^{K-1} devem ser da forma $\frac{\partial F}{\partial p_i}(\boldsymbol{p}) = \lambda$, para $1 \leq i \leq K$. Isso será satisfeito se, e somente se, $p_1 = p_2 = \cdots = p_K = \frac{1}{K}$, resultando em $F(\boldsymbol{p}) = K \left(\arccos \frac{1}{\sqrt{K}} \right)^2$. Por outro lado, se houver pontos críticos na fronteira de Δ^{K-1}, eles terão zero em alguma coordenada, e.g., $\boldsymbol{p} = (0, p_2, \ldots, p_K)$. Repetindo o mesmo argumento, encontramos que os pontos críticos são da forma $p_1 = 0$, $p_2 = \cdots = p_K = \frac{1}{K-1}$, de modo

que $F(\boldsymbol{p}) = \frac{\pi^2}{4} + (K-1)\left(\arccos\frac{1}{\sqrt{K-1}}\right)^2$. Procedendo por indução, temos que qualquer ponto crítico de F é da forma

$$F(\boldsymbol{p}) = (K-i)\frac{\pi^2}{4} + i\left(\arccos\frac{1}{\sqrt{i}}\right)^2, \quad 1 \leq i \leq K.$$

Comparando esses valores, encontramos o máximo em $i = K$ e o mínimo em $i = 1$, o que resulta nas desigualdades desejadas. ■

Teorema 4.12. *Sejam f^* e $\hat{f}$ minimizadores globais de $R_{L_{\mathrm{FR}}}(f)$ e $R^{\eta}_{L_{\mathrm{FR}}}(f)$, respectivamente, para a perda de Fisher-Rao. Sob ruído de rótulo uniforme com $\eta < \frac{K-1}{K}$, temos*

$$0 \leq R^{\eta}_{L_{\mathrm{FR}}}(f^*) - R^{\eta}_{L_{\mathrm{FR}}}(\hat{f}) \leq A_{\mathrm{FR}} \tag{4.29}$$

e

$$B_{\mathrm{FR}} \leq R_{L_{\mathrm{FR}}}(f^*) - R_{L_{\mathrm{FR}}}(\hat{f}) \leq 0, \tag{4.30}$$

com

$$A_{\mathrm{FR}} \coloneqq A_{\mathrm{FR}}(K,\eta) \coloneqq \eta\left(\frac{\pi^2}{4} - \frac{K}{K-1}\left(\arccos\frac{1}{\sqrt{K}}\right)^2\right)$$

e

$$B_{\mathrm{FR}} \coloneqq B_{\mathrm{FR}}(K,\eta) \coloneqq \eta\frac{K\left(\arccos\frac{1}{\sqrt{K}}\right)^2 - \frac{\pi^2}{4}(K-1)}{K-1-\eta K}.$$

Demonstração. As desigualdades que envolvem 0 em (4.29) e (4.30) decorrem diretamente das definições de f^* e $\hat{f}$, respectivamente. Vamos então provar as desigualdades com A_{FR} e B_{FR}. Como em [54, Theorem 1], temos

$$\begin{aligned}
R^{\eta}_{L_{\mathrm{FR}}}(f) &= \mathbb{E}_{(\boldsymbol{x},\tilde{y})}\left[L_{\mathrm{FR}}(\tilde{y}, f(\boldsymbol{x}))\right] = \mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{y|\boldsymbol{x}}\mathbb{E}_{\tilde{y}|\boldsymbol{x},y}\left[L_{\mathrm{FR}}(\tilde{y}, f(\boldsymbol{x}))\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{y|\boldsymbol{x}}\left[(1-\eta)L_{\mathrm{FR}}(y, f(\boldsymbol{x})) + \frac{\eta}{K-1}\sum_{i\neq y} L_{\mathrm{FR}}(i, f(\boldsymbol{x}))\right] \\
&= (1-\eta)R_{L_{\mathrm{FR}}}(f) + \frac{\eta}{K-1}\mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{y|\boldsymbol{x}}\left[\sum_{i\neq y} L_{\mathrm{FR}}(i, f(\boldsymbol{x}))\right] \\
&= \left(1 - \frac{\eta K}{K-1}\right)R_{L_{\mathrm{FR}}}(f) + \frac{\eta}{K-1}\mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{y|\boldsymbol{x}}\left[\sum_{i=1}^{K} L_{\mathrm{FR}}(i, f(\boldsymbol{x}))\right].
\end{aligned}$$

Usando o Lema 4.11, obtemos

$$\left(1-\frac{\eta K}{K-1}\right)R_{L_{\mathrm{FR}}}(f)+\frac{\eta K}{K-1}\left(\arccos\frac{1}{\sqrt{K}}\right)^2\leq R^{\eta}_{L_{\mathrm{FR}}}(f)\leq\left(1-\frac{\eta K}{K-1}\right)R_{L_{\mathrm{FR}}}(f)+\frac{\eta\pi^2}{4}. \tag{4.31}$$

Escrevendo as desigualdades em termos de $R_{L_{\mathrm{FR}}}(f)$, obtemos

$$\frac{(K-1)\left(R^{\eta}_{L_{\mathrm{FR}}}(f)-\frac{\eta\pi^2}{4}\right)}{K-1-\eta K}\leq R_{L_{\mathrm{FR}}}(f)\leq\frac{(K-1)\,R^{\eta}_{L_{\mathrm{FR}}}(f)-\eta K\left(\arccos\frac{1}{\sqrt{K}}\right)^2}{K-1-\eta K}. \tag{4.32}$$

Usando (4.31) para f^* e $\hat{f}$, obtemos

$$\begin{aligned}R^{\eta}_{L_{\mathrm{FR}}}(f^*)-R^{\eta}_{L_{\mathrm{FR}}}(\hat{f})&\leq\frac{\eta\pi^2}{4}+\left(1-\frac{\eta K}{K-1}\right)\left(R_{L_{\mathrm{FR}}}(f^*)-R_{L_{\mathrm{FR}}}(\hat{f})\right)-\frac{\eta K}{K-1}\left(\arccos\frac{1}{\sqrt{K}}\right)^2\\&\leq\eta\left(\frac{\pi^2}{4}-\frac{K}{K-1}\left(\arccos\frac{1}{\sqrt{K}}\right)^2\right)=A_{\mathrm{FR}},\end{aligned}$$

tendo usado que $R_{L_{\mathrm{FR}}}(f^*)-R_{L_{\mathrm{FR}}}(\hat{f})\leq 0$ e $\eta<\frac{K-1}{K}$.

Analogamente, usando (4.32) para f^* e $\hat{f}$, obtemos

$$\begin{aligned}R_{L_{\mathrm{FR}}}(f^*)-R_{L_{\mathrm{FR}}}(\hat{f})&\geq\frac{(K-1)\left(R^{\eta}_{L_{\mathrm{FR}}}(f^*)-R^{\eta}_{L_{\mathrm{FR}}}(\hat{f})\right)}{K-1-\eta K}+\frac{\eta K\left(\arccos\frac{1}{\sqrt{K}}\right)^2-\frac{\eta\pi^2}{4}(K-1)}{K-1-\eta K}\\&\geq\eta\frac{K\left(\arccos\frac{1}{\sqrt{K}}\right)^2-\frac{\pi^2}{4}(K-1)}{K-1-\eta K}=B_{\mathrm{FR}},\end{aligned}$$

tendo usado que $R^{\eta}_{L_{\mathrm{FR}}}(f^*)-R^{\eta}_{L_{\mathrm{FR}}}(\hat{f})\geq 0$ e $\eta<\frac{K-1}{K}$. ■

Esse resultado fornece limitantes para a degradação de desempenho de um classificador treinado com a perda de Fisher-Rao sob ruído de rótulo uniforme, em termos da taxa de ruído η e do número de classes K. Nota-se que não só o impacto desse tipo de ruído é limitado, como torna-se arbitrariamente pequeno quando o número de classes K aumenta, visto que

$$\lim_{K\to\infty}A_{\mathrm{FR}}(K,\eta)=\lim_{K\to\infty}\eta\left(\frac{\pi^2}{4}-\frac{K}{K-1}\left(\arccos\frac{1}{\sqrt{K}}\right)^2\right)=0,$$

e

$$\lim_{K\to\infty}B_{\mathrm{FR}}(K,\eta)=\lim_{K\to\infty}\eta\frac{K\left(\arccos\frac{1}{\sqrt{K}}\right)^2-\frac{\pi^2}{4}(K-1)}{K-1-\eta K}=0.$$

Para a perda CE, não é possível obter um tal limitante superior para a degradação de desempenho sob ruído de rótulo, pois a expressão de (4.20) não é limitada superiormente. Para

a perda MAE, por outro lado, a diferença de desempenho é nula, uma vez que esta perda é robusta a ruído de rótulo uniforme. O resultado a seguir, de [54], permite comparar com perdas q-CE em geral:

Teorema 4.13 ([54, Theorem 1])**.** *Sejam f^* e $\hat{f}$ minimizadores globais de $R_{L_{q\text{-CE}}}(f)$ e $R^{\eta}_{L_{q\text{-CE}}}(f)$, respectivamente, para a perda q-entropia cruzada. Sob ruído de rótulo uniforme com $\eta < \frac{K-1}{K}$, temos*

$$0 \leq R^{\eta}_{L_{q\text{-CE}}}(f^*) - R^{\eta}_{L_{q\text{-CE}}}(\hat{f}) \leq A_{q\text{-CE}} \tag{4.33}$$

e

$$B_{q\text{-CE}} \leq R_{L_{q\text{-CE}}}(f^*) - R_{L_{q\text{-CE}}}(\hat{f}) \leq 0, \tag{4.34}$$

com $A_{q\text{-CE}} \coloneqq A_{q\text{-CE}}(K,\eta) \coloneqq \eta\frac{K^q-1}{(1-q)(K-1)}$ e $B_{q\text{-CE}} \coloneqq B_{q\text{-CE}}(K,\eta) \coloneqq \eta\frac{1-K^q}{(1-q)(K-1-\eta K)}$.

Note que, para essas perdas, também temos, para η fixo,

$$\lim_{K\to\infty} A_{q\text{-CE}}(K,\eta) = \lim_{K\to\infty} \eta\frac{K^q-1}{(q-1)(K-1)} = 0$$

e

$$\lim_{K\to\infty} B_{q\text{-CE}}(K,\eta) = \lim_{K\to\infty} \eta\frac{1-K^q}{(q-1)(K-1-\eta K)} = 0.$$

Em particular, esse resultado se aplica à perda de Hellinger, tomando $q = 1/2$.

Para a perda MSE, derivamos a seguir resultados análogos.

Lema 4.14. *A perda do erro quadrático médio L_{MSE} satisfaz*

$$K - 1 \leq \sum_{i=1}^{K} L_{\text{MSE}}(i, f(\boldsymbol{x})) \leq 2(K-1). \tag{4.35}$$

Demonstração. De (4.18), temos $L_{\text{MSE}}(y, f(\boldsymbol{x})) = 1 + \|\boldsymbol{p}\|_2^2 - 2p_y$, de modo que

$$\sum_{i=1}^{K} L_{\text{MSE}}(i, f(\boldsymbol{x})) = \sum_{i=1}^{K} \left(1 + \|\boldsymbol{p}\|_2^2 - 2p_i\right) = K\left(1 + \|\boldsymbol{p}\|_2^2\right) + 2$$

Além disso, $F(\boldsymbol{p}) \coloneqq \|\boldsymbol{p}\|_2^2 = \sum_{i=1}^{K} p_i^2$ é uma função convexa em $\boldsymbol{p} = (p_1, \ldots, p_K)$, e, com $\boldsymbol{p}$ restrito ao simplexo Δ^{K-1}, temos $\frac{1}{K} \leq \|\boldsymbol{p}\|_2^2 \leq 1$, donde

$$K - 1 = K\left(1 + \frac{1}{K}\right) - 2 \leq \sum_{i=1}^{K} L_{\text{MSE}}(i, f(\boldsymbol{x})) \leq K\,(1+1) - 2 = 2(K-1).$$

■

Teorema 4.15. *Sejam f^* e $\hat{f}$ minimizadores globais de $R_{L_{\text{MSE}}}(f)$ e $R^{\eta}_{L_{\text{MSE}}}(f)$, respectivamente, para a perda do erro quadrático médio. Sob ruído de rótulo uniforme com $\eta < \frac{K-1}{K}$, temos*

$$0 \leq R^{\eta}_{L_{\text{MSE}}}(f^*) - R^{\eta}_{L_{\text{MSE}}}(\hat{f}) \leq A_{\text{MSE}} \tag{4.36}$$

e

$$B_{\text{MSE}} \leq R_{L_{\text{MSE}}}(f^*) - R_{L_{\text{MSE}}}(\hat{f}) \leq 0, \tag{4.37}$$

com $A_{\text{MSE}} \coloneqq A_{\text{MSE}}(\eta) \coloneqq \eta$ e $B_{\text{MSE}} \coloneqq B_{\text{MSE}}(K, \eta) \coloneqq -\eta \frac{K-1}{K-1-\eta K}$.

Demonstração. A demonstração é análoga à do Teorema 4.12. Apresentamos uma versão resumida. Temos

$$R^{\eta}_{L_{\text{MSE}}}(f) = \left(1 - \frac{\eta K}{K-1}\right) R_{L_{\text{MSE}}}(f) + \frac{\eta}{K-1} \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{y|\boldsymbol{x}} \left[\sum_{i=1}^{K} L_{\text{MSE}}(i, f(\boldsymbol{x}))\right].$$

Usando o Lema 4.14, obtemos

$$\left(1 - \frac{\eta K}{K-1}\right) R_{L_{\text{MSE}}}(f) + \eta \leq R^{\eta}_{L_{\text{MSE}}}(f) \leq \left(1 - \frac{\eta K}{K-1}\right) R_{L_{\text{MSE}}}(f) + 2\eta$$

e

$$\frac{(K-1) R^{\eta}_{L_{\text{MSE}}}(f)}{K-1-\eta K} - \frac{2\eta(K-1)}{K-1-\eta K} \leq R^{\eta}_{L_{\text{MSE}}}(f) \leq \frac{(K-1) R^{\eta}_{L_{\text{MSE}}}(f)}{K-1-\eta K} - \frac{\eta(K-1)}{K-1-\eta K}.$$

Desse modo,

$$R^{\eta}_{L_{\text{MSE}}}(f^*) - R^{\eta}_{L_{\text{MSE}}}(\hat{f}) \leq \left(1 - \frac{\eta K}{K-1}\right) \left(R_{L_{\text{MSE}}}(f^*) - R_{L_{\text{MSE}}}(\hat{f})\right) + \eta \leq \eta$$

e

$$R_{L_{\text{MSE}}}(f^*) - R_{L_{\text{MSE}}}(\hat{f}) \geq \frac{(K-1)\left(R^{\eta}_{L_{\text{MSE}}}(f^*) - R^{\eta}_{L_{\text{MSE}}}(\hat{f})\right)}{K-1-\eta K} - \frac{\eta(K-1)}{K-1-\eta K} \geq -\frac{\eta(K-1)}{K-1-\eta K},$$

usando que f^* e $\hat{f}$ são minimizam $R_{L_{\text{MSE}}}(f)$ e $R^{\eta}_{L_{\text{MSE}}}(f)$, respectivamente, e que $\eta < \frac{K-1}{K}$. ■

Note que $A_{\text{MSE}}(\eta) = \eta > 0$ não depende do número de classes K, enquanto que

$$\lim_{K \to \infty} B_{\text{MSE}}(K, \eta) = \lim_{K \to \infty} \eta \frac{K-1}{K-1-\eta K} = \frac{\eta}{1-\eta} > 0.$$

A Tabela 4.3 resume os limitantes de desempenho sob ruído de rótulo uniforme discutidos até aqui. Na Figura 4.1, traçamos os limitantes de diferentes funções perda, para valores de $2 \leq K \leq 300$, com $\eta = 0{,}8 - 1/K$. Para a perda q-CE, escolhemos o valor $q = 0{,}7$, que é o

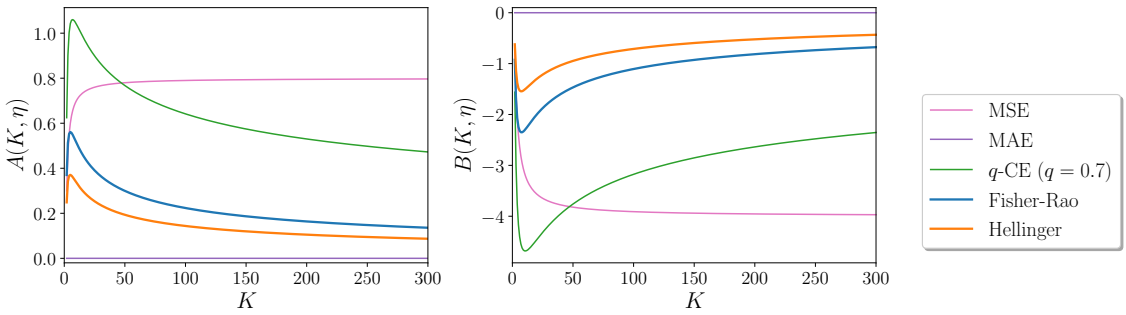


Figura 4.1: Limitantes $A(K,\eta)$ e $B(K,\eta)$ para diferentes funções perda, com $\eta = 0{,}8 - 1/K$.

usado nos experimentos em [54] por apresentar "bom compromisso entre rápida convergência e robustez a ruído". Observamos que a melhor robustez a ruído de rótulo uniforme é obtida com a perda MAE, seguida da perda de Hellinger, de Fisher-Rao e da q-CE com $q = 0{,}7$. Para valores altos de K, a perda MSE se torna a menos robusta dentre as comparadas na figura. Observamos que a perda CE não aparece nos gráficos, pois ela não permite obter um tal limitante superior.

Tabela 4.3: Resumo dos limitantes $A(K,\eta)$ e $B(K,\eta)$ para diferentes funções perda.

Função perda	$A(K,\eta)$	$B(K,\eta)$
Erro médio quadrado (MSE)	η	$-\eta\frac{K-1}{K-1-\eta K}$
Erro médio absoluto (MAE)	0	0
Entropia cruzada (CE)	$+\infty$	$-\infty$
q-entropia cruzada	$\eta\frac{K^q-1}{(1-q)(K-1)}$	$\eta\frac{1-K^q}{(1-q)(K-1-\eta K)}$
Fisher-Rao	$\eta\left(\frac{\pi^2}{4} - \frac{K}{K-1}\left(\arccos\frac{1}{\sqrt{K}}\right)^2\right)$	$\eta\frac{K\left(\arccos\frac{1}{\sqrt{K}}\right)^2 - \frac{\pi^2}{4}(K-1)}{K-1-\eta K}$
Hellinger ($q = 1/2$)	$\eta\frac{2(\sqrt{K}-1)}{K-1}$	$\eta\frac{2(1-\sqrt{K})}{(K-1-\eta K)}$

Velocidade de aprendizado

Apesar de a perda MAE ser a única, dentre as estudadas, rigorosamente robusta a ruído de rótulo uniforme, sabe-se que ela resulta em aprendizado lento quando usada para o treinamento de redes neurais [31, 36, 54]. A perda CE, por outro lado, apresenta boa dinâmica de treinamento, mas não é de todo robusta a ruído de rótulo. Idealmente, não gostaríamos de trocar robustez por velocidade, nem o contrário. Para investigar o compromisso que há por trás dessas duas características, estudaremos a seguir a velocidade de aprendizado das funções perda apresentadas.

O aprendizado em redes neurais envolve a minimização do risco empírico por meio de um algoritmo numérico, como o método do gradiente. Nesse caso, os parâmetros $\boldsymbol{w}$ da rede neural $f_{\boldsymbol{w}} \coloneqq f$ são atualizados iterativamente de forma proporcional ao gradiente do risco empírico $\bar{R}_L$ (4.4.1), com relação aos parâmetros $\boldsymbol{w}$:

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \gamma \nabla_{\boldsymbol{w}} \bar{R}_L,$$

onde o parâmetro $\gamma > 0$ é chamado *taxa de aprendizado*. Assim, o vetor gradiente

$$\nabla_{\boldsymbol{w}} \bar{R}_L = \frac{1}{N} \sum_{i=1}^{N} \nabla_{\boldsymbol{w}} L(y_i, f_{\boldsymbol{w}}(\boldsymbol{x}_i)) \tag{4.38}$$

indica a direção e a magnitude de cada passo em direção a um mínimo local de R_L.

É interessante notar que as perdas MAE, CE, q-CE, Fisher-Rao e Hellinger podem todas ser escritas como função da probabilidade $p_y = \sigma(s_y)$ associada à classe correta, enquanto que a perda MSE, por outro lado, depende de todas as coordenadas do vetor $\boldsymbol{p} = (\sigma \circ f)(\boldsymbol{x})$. Isso significa que a minimização desta tenta, simultaneamente, tornar a probabilidade predita p_y da classe correta próxima de 1, e minimizar a probabilidade p_i de todas as outras classes $i \neq y$.

Especificamente, as expressões das perdas para o primeiro grupo podem ser escritas na forma

$$L(y, f(\boldsymbol{x})) = h\big([(\sigma \circ f)(\boldsymbol{x})]_y\big) = h(p_y), \tag{4.39}$$

para alguma função $h\colon [0,1] \to \mathbb{R}$ diferenciável e monotonicamente decrescente, com $h(1) = 0$. Nesse caso, a expressão do gradiente (4.38) ganha a forma

$$\nabla_{\boldsymbol{w}} \bar{R}_L = \frac{1}{N} \sum_{i=1}^{N} \nabla_{\boldsymbol{w}} L(y_i, f_{\boldsymbol{w}}(\boldsymbol{x}_i)) = \frac{1}{N} \sum_{i=1}^{N} h'(p_{y_i}) \nabla_{\boldsymbol{w}} [(\sigma \circ f_{\boldsymbol{w}})(\boldsymbol{x}_i)]_{y_i}. \tag{4.40}$$

Como o fator $\nabla_{\boldsymbol{w}} [(\sigma \circ f_{\boldsymbol{w}})(\boldsymbol{x}_i)]_{y_i}$ não depende da escolha da função perda, basta estudar $|h'(p_{y_i})|$ para comparar como varia a dinâmica de aprendizado sob diferentes escolhas de perda L. A Tabela 4.4 apresenta as expressões das funções $h(p_y)$ e $|h'(p_y)|$ para diferentes perdas, e a Figura 4.2 mostra seu gráfico, com $q = 0{,}7$ para a q-CE, como usado nos experimentos em [54].

Como anteriormente observado [36, 54], o lento aprendizado com uso da perda MAE deve-se ao fato de o gradiente de seu risco ser constante. Isso significa que, independentemente do quão grande for o erro cometido na predição atual, o tamanho do passo no espaço de parâmetros da rede neural será o mesmo. Com a perda CE, ao contrário, o módulo da derivada $|h'(p_y)|$ aumenta com a diminuição de p_y, significando que, quanto mais longe do mínimo local, maiores serão os passos dados em sua direção. A forma dessa derivada também implica que a ênfase do treinamento é colocada sobre os exemplos cuja predição está mais distante do rótulo prescrito,

TABELA 4.4: FUNÇÕES $h(p_y)$ E SUAS DERIVADAS $|h'(p_y)|$ PARA DIFERENTES FUNÇÕES PERDA.

Função perda	$h(p_y)$	$\lvert h'(p_y)\rvert$
Erro médio absoluto (MAE)	$1 - p_y$	1
Entropia cruzada (CE)	$-\log p_y$	$\frac{1}{p_y}$
q-entropia cruzada	$-\log_q p_y$	$\frac{1}{(p_y)^q}$
Fisher-Rao	$\left(\arccos\sqrt{p_y}\right)^2$	$\frac{\arccos\sqrt{p_y}}{\sqrt{p_y(1-p_y)}}$
Hellinger ($q = 1/2$)	$2\left(1-\sqrt{p_y}\right)$	$\frac{1}{\sqrt{p_y}}$

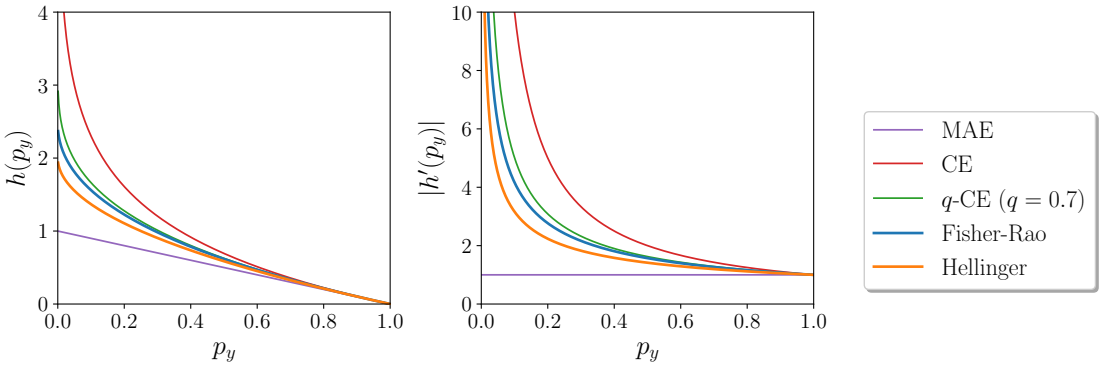


Figura 4.2: Funções $h(p_y)$ e suas derivadas $h'(p_y)$ para diferentes funções perda.

o que explica tanto a rápida convergência, quanto a baixa tolerância a ruído de rótulo.

As derivadas $|h'(p_y)|$ das perdas q-CE, Fisher-Rao e Hellinger também aumentam com a diminuição de p_y, indicando uma melhor dinâmica de aprendizado que com a perda MAE. Esses valores são, no entanto, menores que para a perda CE, o que sugere que os efeitos discutidos anteriormente para esta perda ocorrem naquelas em menor intensidade. Assim, usando dados limpos, poderíamos esperar, por exemplo, que a ordenação das perdas com respeito à velocidade de convergência seria CE, q-CE com $q = 0{,}7$, Fisher-Rao, Hellinger e MAE, que é precisamente a ordem inversa daquela observada para robustez a ruído de rótulo.

Dessa forma, as funções perda q-CE. Fisher-Rao e Hellinger podem ser vistas como meios de obtenção de uma *solução de compromisso* entre robustez a ruído de rótulo e velocidade de aprendizado. Em particular, os resultados teóricos sugerem que a perda de Fisher-Rao é capaz de fornecer uma modesta melhoria na dinâmica de aprendizado ao custo de pequena redução na robustez a ruído de rótulo, quando comparada com a perda de Hellinger. Na próxima subseção, compararemos experimentalmente o desempenho de algumas dessas funções perda com relação a esses dois aspectos.

4.4.4 Resultados experimentais

Realizamos experimentos numéricos para estudar o desempenho de diferentes funções perda, em um conjunto de dados sintético e em outro real. Os experimentos usam redes neurais simples e objetivam ilustrar os resultados teóricos desenvolvidos na seção anterior, particularmente o compromisso fornecido pela perda de Fisher-Rao—e não reproduzir resultados do estado da arte.

Avaliamos o desempenho dos classificadores treinados com dados limpos e com ruído de rótulo uniforme de taxa de ruído η. Apenas os conjuntos de treinamento são contaminados com ruído, enquanto os conjuntos de teste permanecem limpos. Em todos os experimentos, a função de ativação é ReLU e o treinamento é feito com o método do gradiente estocástico. As taxas de aprendizado para cada função perda foram manualmente ajustadas de modo a fornecer o melhor resultado. Reportamos as médias e desvios padrões dos resultados obtidos nos experimentos.

Conjunto de dados sintético

Inicialmente, consideramos um conjunto de dados sintético formado por vetores de dimensão 100 divididos em 10 classes. Os dados foram gerados por distribuições gaussianas centradas nos vértices de um hipercubo, usando o método `make_classification` da biblioteca scikit-learn [44], e são formados por 8.000 exemplos de treinamento e 2.000 exemplos de teste. A Figura 4.3 mostra um exemplo desses dados sintéticos. Configuramos uma rede *multilayer perceptron* (MLP) com três camadas internas de 80, 40 e 20 neurônios, respectivamente. O tamanho de *batch* é 20 e o modelo é treinado por 20 épocas.

Realizamos experimentos sob ruído de rótulo uniforme de taxas $\eta = 0$ (sem ruído), $\eta = 0{,}3$ e $\eta = 0{,}5$. A evolução das acurácias de treinamento e de teste são mostradas na Figura 4.4 e as acurácias de teste finais são resumidas na Tabela 4.5, em que os melhores resultados estão destacados em negrito. Observamos que, na ausência de ruído, a acurácia final atingida por todas as perdas são similares, com a perda CE fornecendo o treinamento mais rápido, como observado anteriormente. Entretanto, sob ruído de rótulo uniforme, o treinamento coma as perdas de Fisher-Rao e Hellinger consistentemente resulta em melhores acurácias.

Tabela 4.5: Acurácia de teste (%) para conjunto de dados sintético.

Função perda	$\eta = 0$	$\eta = 0{,}3$	$\eta = 0{,}5$
Erro quadrático médio (MSE)	88,39 (±0,70)	74,43 (±0,41)	64,08 (±0,70)
Entropia cruzada (CE)	**90,21 (±1,27)**	73,68 (±0,99)	60,78 (±1,15)
Fisher-Rao	**89,64 (±0,80)**	**77,83 (±0,71)**	**67,38 (±0,46)**
Hellinger	89,36 (±1,18)	**78.43 (±0,66)**	**68,49 (±1,07)**

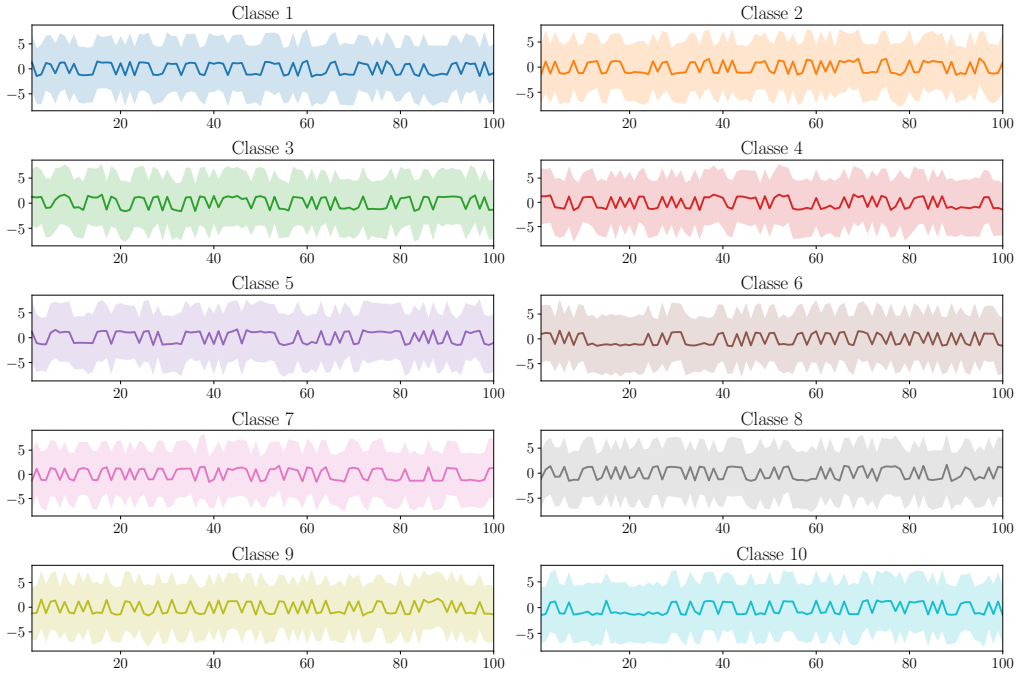
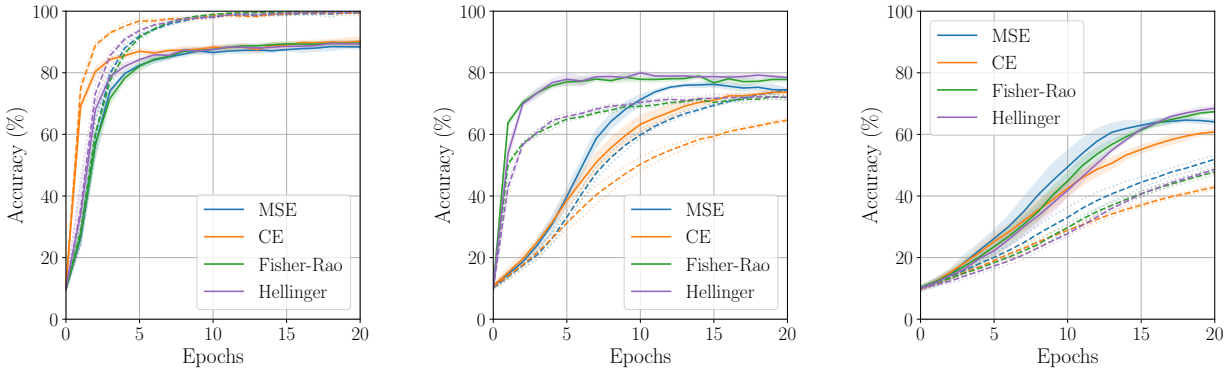


Figura 4.3: Exemplo de dados sintéticos. Média (linha contínua) e desvio padrão (área sombreada) dos vetores de dimensão 100 de cada uma das 10 classes.



(a) $\eta = 0$ **(b)** $\eta = 0.3$ **(c)** $\eta = 0.5$

Figura 4.4: Acurácia de treinamento (linhas tracejadas) e de teste (linhas sólidas) para o conjunto de dados sintético.

Conjunto de dados MNIST

O conjunto de dados MNIST [37] contém 60.000 exemplos de treinamento e 10.000 exemplos de teste. É formado por imagens 28×28 em escala de cinza de dígitos manuscritos de 0 a 9, cf. Figura 4.5. Para esse experimento, configuramos uma rede MLP composta por duas camadas internas, de 300 e 100 neurônios, respectivamente. O tamanho de *batch* é 64 e a rede é treinada por 40 épocas.

Os resultados para taxas de ruído uniforme $\eta = 0$ (sem ruído), $\eta = 0{,}3$ e $\eta = 0{,}5$ são

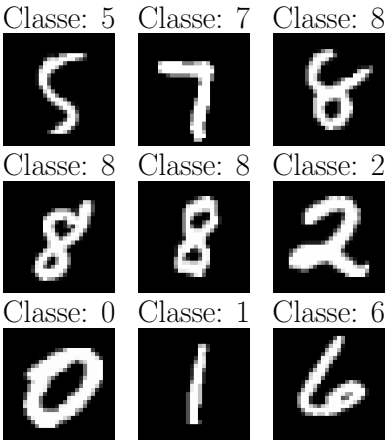
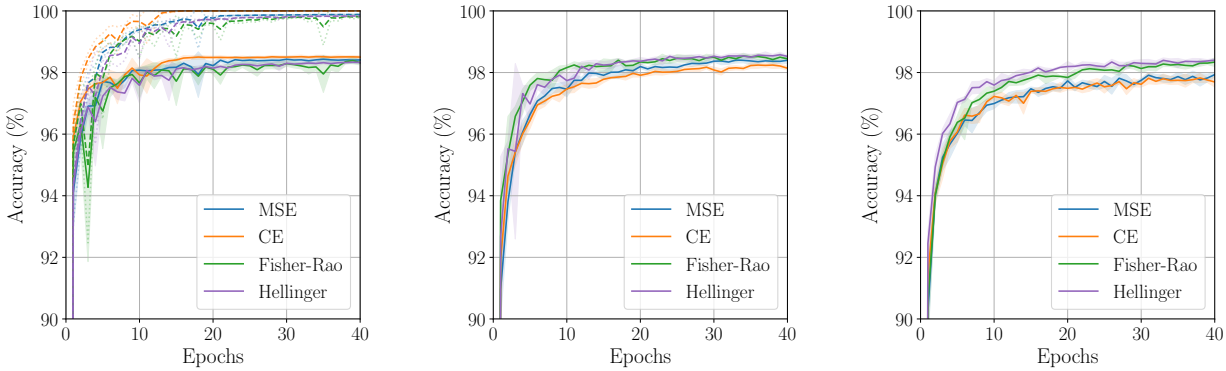


Figura 4.5: Exemplo de dados do conjunto MNIST.



(a) $\eta = 0$ **(b)** $\eta = 0.3$ **(c)** $\eta = 0.5$

Figura 4.6: Acurácia de treinamento (linhas tracejadas) e de teste (linhas sólidas) para o conjunto de dados MNIST.

similarmente reportados na Figura 4.6 e na Tabela 4.6. Notamos que, nesses experimentos, as perdas de Fisher-Rao e Hellinger atingem acurácias competitivas com dados limpos. Na presença de ruído, apresentam acurácias ligeiramente maiores, i.e., maior robustez ao ruído, em comparação com as outras perdas.

Tabela 4.6: Acurácia de teste (%) para conjunto de dados MNIST.

Função perda	$\eta = 0$	$\eta = 0{,}3$	$\eta = 0{,}5$
Erro quadrático médio (MSE)	**98,41 (±0,09)**	98,40 (±0,10)	97,93 (±0,07)
Entropia cruzada (CE)	**98,50 (±0,04)**	98,14 (±0,06)	97,69 (±0,16)
Fisher-Rao	98,32 (±0,07)	**98,44 (±0,05)**	**98,34 (±0,14)**
Hellinger	98,33 (±0,05)	**98,53 (±0,03)**	**98,40 (±0,06)**

Capítulo 5

Conclusão

Nesta dissertação, apresentamos três contribuições que mostram como as ferramentas de geometria e/ou estatística podem ser utilizadas para tratar problemas em comunicações e aprendizado. A seguir, elencamos perspectivas futuras de cada tema.

Construção de códigos esféricos por folheações de Hopf

- Focamos particularmente na construção recursiva a partir de um caso base em dimensão 4 para construir códigos esféricos em dimensões 2^k. Uma perspectiva é estender a construção recursiva ao considerar construções base em outras dimensões.

- Uma aplicação clássica de códigos esféricos é como quantizadores vetoriais. Pode-se, assim, estudar o desempenho dos códigos SCHF para quantização em comparação com outros métodos da literatura, particularmente para fontes gaussianas.

- Outra aplicação relacionada é a construção de códigos para fontes que têm alfabeto contínuo. Nesse caso, o problema se traduz no empacotamento de curvas em esferas. Seria possível investigar a adaptação da estrutura dos códigos SCHF para esse problema.

Compressão com perdas baseada em árvores de contexto

- Apresentamos uma solução para compressão vetorial com perdas, construindo um quantizador e um compressor universal dos índices de quantização. No entanto, não está claro se essa solução em dois passos é ótima; seria, portanto, importante caracterizar o compromisso fundamental de taxa-distorção para fontes com distribuição desconhecida.

- A aplicação direta de compressores universais existentes a dados de dimensão elevada poderia falhar na prática devido ao grande tamanho do alfabeto de índices de quantização. Uma possibilidade é separar os vetores em sub-vetores de dimensão menor (como feito na abordagem apresentada); outra opção seria realizar redução de dimensionalidade antes da

compressão. Gostaríamos de comparar o desempenho dessas duas estratégias sub-ótimas, em termos de sua redundância assintótica.

Aprendizado com uma função perda de Fisher-Rao

- Realizamos experimentos simples para o treinamento de redes neurais com a função perda proposta, com o objetivo de demonstrar seu potencial de desempenho. Perspectivas futuras incluem a realização de experimentos com problemas mais complexos e usando arquiteturas do estado da arte, de modo a verificar o desempenho nesses casos.

- Outra perspectiva consiste em explorar uma possível extensão do método proposto para problemas de regressão. Nesse caso, o modelo deve estimar uma distribuição condicional da saída, dada a entrada apresentada, dentre uma família previamente acordada. Para isso, é necessário ter uma expressão fechada para a distância de Fisher-Rao na variedade estatística em questão, bem como definir uma distribuição objetivo, a ser comparada com a predição a cada instante.

Referências

[1] S. Amari. *Information Geometry and Its Applications*. Springer, Tokyo, 2016.

[2] S. Amari e H. Nagaoka. *Methods of Information Geometry*. American Mathematical Society, Providence, RI, 2000.

[3] C. Atkinson e A. F. S. Mitchell. Rao's distance measure. *Sankhyā: Indian J. Stat., Ser. A*, v. 43, n. 3, p. 345–365, 1981.

[4] N. Ay, J. Jost, H. Vân Lê, e L. Schwachhöfer. Information geometry and sufficient statistics. *Probab. Theory Relat. Fields*, v. 162, p. 327–364, 2015.

[5] N. Ay, J. Jost, H. Vân Lê, e L. Schwachhöfer. *Information Geometry*. Springer, Cham, 2017.

[6] A. F. Beardon. *The Geometry of Discrete Groups*. Springer, New York, NY, 1983.

[7] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, NY, 2006.

[8] V. I. Bogachev. *Measure Theory*. Springer, Berlin, 2ª edição, 2007.

[9] J. Burbea. Informative geometry of probability spaces. *Expo. Math.*, n. 4, p. 347–378, 1986.

[10] O. Calin. *Deep Learning Architectures: A Mathematical Approach*. Springer, Cham, 2020.

[11] O. Calin and C. Udrişte. *Geometric Modeling in Probability and Statistics*. Springer, Cham, 2014.

[12] L. L. Campbell. An extended Čencov characterization of the information metric. *Proc. Am. Math. Soc.*, v. 98, p. 135–141, 1986.

[13] N. N. Chentsov (Čencov). *Statistical Decision Rules and Optimal Inference*. American Mathematical Society, Providence, RI, 1982.

[14] J. Clough, N. Byrne, I. Oksuz, V. A. Zimmer, J. A. Schnabel, e A. King. A topological loss function for deep-learning based image segmentation using persistent homology. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020. Early access.

[15] S. I. R. Costa, F. Oggier, A. Campello, J.-C. Belfiore, e E. Viterbo. *Lattices Applied to Coding for Reliable and Secure Communications*. Springer, 2017.

[16] S. I. R. Costa, S. A. Santos, e J. E. Strapasson. Fisher information distance: A geometrical reading. *Discrete Appl. Math.*, v. 197, p. 59–69, 2015.

[17] T. M. Cover e J. A. Thomas. *Elements of Information Theory*. Wiley, Hoboken, NJ, 2ª edição, 2006.

[18] A. Demirkaya, J. Chen, e S. Oymak. Exploring the role of loss functions in multiclass classification. Em *54th Annu. Conf. Inf. Sci. Syst. (CISS)*, p. 1–5, 2020.

[19] M. P. do Carmo. *Geometria Diferencial de Curvas e Superfícies*. SBM, Rio de Janeiro, 6ª edição, 2014.

[20] M. P. do Carmo. *Geometria Riemanniana*. IMPA, Rio de Janeiro, 6ª edição, 2019.

[21] T. Ericson e V. Zinoviev. *Codes on Euclidean Spheres*. North-Holland, Amsterdam, 2001.

[22] B. Frenay e M. Verleysen. Classification in the presence of label noise: A survey. *IEEE Trans. Neural Netw. Learn. Syst.*, v. 25, n. 5, p. 845–869, 2014.

[23] C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, e T. Poggio. Learning with a Wasserstein loss. Em *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, p. 2053–2061, 2015.

[24] A. Ghosh, H. Kumar, e P. S. Sastry. Robust loss functions under label noise for deep neural networks. Em *Proc. 31st AAAI Conf. Artif. Intell.*, p. 1919–1925, 2017.

[25] A. Ghosh, N. Manwani, e P. Sastry. Making risk minimization tolerant to label noise. *Neurocomputing*, v. 160, p. 93–107, 2015.

[26] P. Golik, P. Doetsch, e H. Ney. Cross-entropy vs. squared error training: a theoretical and experimental comparison. Em *Proc. Interspeech 2013*, p. 1756–1760, 2013.

[27] I. Goodfellow, Y. Bengio, e A. Courville. *Deep Learning*. MIT Press, Cambridge, MA, 2016.

[28] R. M. Gray. *Entropy and Information Theory*. Springer, New York, NY, 2ª edição, 2011.

[29] L. Hou, C.-P. Yu, e D. Samaras. Squared earth movers distance loss for training deep neural networks on ordered-classes. Em *31st Conf. Neural Inf. Process. Syst. (NIPS)*, 2017.

[30] L. Hui e M. Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. Em *9th Int. Conf. Learn. Representations (ICLR)*, 2021.

[31] K. Janocha e W. M. Czarnecki. On loss functions for deep neural networks in classification. *Schedae Informaticae*, v. 25, 2017.

[32] R. E. Kass e P. W. Vos. *Geometrical Foundations of Asymptotic Inference*. Wiley, New York, NY, 1997.

[33] D. M. Kline e V. L. Berardi. Revisiting squared-error and cross-entropy functions for training neural network classifiers. *Neural Comp. Appl.*, v. 14, n. 4, p. 310–318, 2005.

[34] W. Klingenberg. *A Course in Differential Geometry*. Springer, New York, NY, 1978.

[35] A. Kovalev. Notas de aula de geometria diferencial. Disponível em: `https://www.dpmms.cam.ac.uk/~agk22/teaching.html`, 2019. Acesso em: 5 ago. 2022.

[36] H. Kumar e P. S. Sastry. Robust loss functions for learning multi-class classifiers. Em *IEEE Int. Conf. Syst. Man Cybern. (SMC)*, p. 687–692, 2018.

[37] Y. LeCun, C. Cortes, e C. J. C. Burges. The MNIST database of handwritten digits. Disponível em: `http://yann.lecun.com/exdb/mnist/`. Acesso em: 5 ago. 2022.

[38] M. Li, H. Sun, e L. Peng. Fisher–Rao geometry and Jeffreys prior for Pareto distribution. *Commun. Stat. – Theory Methods*, v. 51, n. 6, p. 1895–1910, 2022.

[39] D. W. Lyons. An elementary introduction to the Hopf fibration. *Math. Mag.*, v. 76, n. 2, p. 87–98, 2003.

[40] N. Manwani e P. S. Sastry. Noise tolerance under risk minimization. *IEEE Trans. Cybern.*, v. 43, n. 3, p. 1146–1151, 2013.

[41] H. K. Miyamoto, S. I. R. Costa, e H. N. Sá Earp. Constructive spherical codes by Hopf foliations. *IEEE Trans. Inf. Theory*, v. 67, n. 12, p. 7925–7939, 2021.

[42] H. K. Miyamoto, F. C. C. Meneghetti, e S. I. R. Costa. The Fisher-Rao loss for learning under label noise. *arXiv:2210.16401*, 2022. Aceito para publicação em *Inf. Geom.*

[43] H. K. Miyamoto e S. Yang. Context-tree-based lossy compression and its application to CSI representation. *IEEE Trans. Commun.*, v. 70, n. 7, p. 4417–4428, 2022.

[44] F. Pedregosa *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, v. 12, p. 2825–2830, 2011.

[45] Y. Polyanskiy e Y. Wu. Lecture notes on information theory. Disponível em: `http://www.stat.yale.edu/~yw562/teaching/itlectures.pdf`, 2019. Acesso em: 5 ago. 2022.

[46] C. R. Rao. *Information and the Accuracy Attainable in the Estimation of Statistical Parameters*, p. 235–247. Springer, New York, NY, 1992.

[47] C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, v. 27, n. 3, p. 379–423, 1948.

[48] A. Singh e J. C. Príncipe. A loss function for classification based on a robust similarity metric. Em *Int. Joint Conf. Neural Netw. (IJCNN)*, p. 1–6, 2010.

[49] C. Tsallis. What are the numbers that experiments provide? *Química Nova*, v. 17, p. 468–471, 1994.

[50] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, NY, 2009.

[51] B. van Rooyen, A. Menon, e R. C. Williamson. Learning with symmetric label noise: The importance of being unhinged. Em *Adv. Neural Inf. Process. Syst. (NIPS)*, 2015.

[52] H. Vân Lê. The uniqueness of the Fisher metric as information metric. *Ann. Inst. Stat. Math.*, v. 69, p. 879–896, 2017.

[53] F. Willems, Y. Shtarkov, e T. Tjalkens. The context-tree weighting method: basic properties. *IEEE Trans. Inf. Theory*, v. 41, n. 3, p. 653–664, 1995.

[54] Z. Zhang e M. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. Em *32nd Conf. Neural Inf. Process. Syst. (NIPS)*, 2018.

Apêndices

Apêndice A

Pré-Requisitos de Probabilidade

Este apêndice reúne, de forma resumida, definições e resultados básicos de probabilidade, usando a linguagem de teoria da medida. Está baseado principalmente em [11, 28]. Demonstrações são omitidas e podem ser encontradas nessas referências.

Definição A.1 (σ-álgebra)**.** Uma *σ-álgebra* sobre o conjunto $\mathcal{X}$ é uma coleção $\mathcal{F} \subseteq \wp(\mathcal{X})$, tal que (i) $\mathcal{X} \in \mathcal{F}$, (ii) se $A \in \mathcal{F}$, então $\mathcal{X} \setminus A \in \mathcal{F}$, e (iii) se $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathcal{F}$, então $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{F}$. O par $(\mathcal{X}, \mathcal{F})$ é chamado *espaço mensurável* e os elementos de $\mathcal{X}$ são ditos $\mathcal{F}$-mensuráveis.

Definição A.2 (Medida)**.** Uma *medida* μ no espaço mensurável $(\mathcal{X}, \mathcal{F})$ é uma função $\mu\colon \mathcal{F} \to [0, \infty]$, que satisfaz: (i) $\mu(\emptyset) = 0$, e (ii) se $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathcal{F}$ são mutuamente disjuntos, então $\mu\left(\bigsqcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} \mu(A_n)$. A tripla $(\mathcal{X}, \mathcal{F}, \mu)$ é chamada *espaço de medida*.

Uma medida μ em $(\mathcal{X}, \mathcal{F})$ é dita *σ-finita* se existe $\{A_n\}_{n\in\mathbb{N}}$ tal que $\mathcal{X} = \bigcup_{n\in\mathbb{N}} A_n$ e $\mu(A_n) < \infty$ para todo $n \in \mathbb{N}$. Uma medida μ é dita *absolutamente contínua* com relação a outra medida λ, o que é denotado $\mu \ll \lambda$, se $\lambda(A) = 0$ implica $\mu(A) = 0$, para todo $A \in \mathcal{F}$.

São exemplos de espaços de medida:

1. Espaço de medida discreto: $\mathcal{X} = \{1, 2, \ldots, n\}$, $\mathcal{F} = \wp(\mathcal{X})$ o conjunto das partes, e $\mu(A) = |A|$, $A \in \mathcal{F}$ a medida de contagem.

2. Espaço de medida contínuo: $\mathcal{X} = \mathbb{R}^n$, $\mathcal{F} = \mathcal{L}(\mathbb{R}^n)$ a σ-álgebra de Lebesgue[1], e μ a medida de Lebesgue, que satisfaz $\mu\left([a, b]^n\right) = (b - a)^n$.

Dizemos que uma propriedade vale μ em quase toda parte, abreviado μ-q.t.p., se o conjunto no qual ela não é válida tem medida μ nula.

[1] A σ-ágebra de Lebesgue $\mathcal{L}(\mathbb{R}^n)$ é o completamento da σ-álgebra de Borel $\mathcal{B}(\mathbb{R}^n)$. A σ-álgebra de Borel $\mathcal{B}(\mathbb{R}^n)$ é a menor σ-álgebra que contém os abertos de $\mathbb{R}^n$. Um espaço de medida $(\mathcal{X}, \mathcal{F}, \mu)$ é dito completo se todo subconjunto de um conjunto de medida nula é $\mathcal{F}$-mensurável (e tem medida nula).

Uma medida P em $(\Omega, \mathcal{F})$ tal que $P(\Omega) = 1$ é dita *medida de probabilidade*. Nesse caso, $(\Omega, \mathcal{F}, P)$ é chamado *espaço de probabilidade* e os elementos de $\mathcal{F}$ são chamados *eventos*. Chamamos Ω *espaço amostral*, que é o conjunto abstrato de todos os estados que podem ocorrer como resultado de um experimento aleatório.

Definição A.3 (Função mensurável)**.** Uma função $f : \mathcal{Y} \to \mathcal{X}$ entre dois espaços mensuráveis $(\mathcal{Y}, \mathcal{G})$ e $(\mathcal{X}, \mathcal{F})$ é dita *mensurável* se $f^{-1}(A) \in \mathcal{G}$, para todo $A \in \mathcal{F}$.

Definição A.4 (Variável aleatória)**.** Uma *variável aleatória* do espaço de probabilidade $(\Omega, \mathcal{F}, P)$ em outro espaço mensurável $(\mathcal{X}, \mathcal{F})$ é uma função mensurável $X : \Omega \to \mathcal{X}$.

Uma tal variável aleatória induz uma medida de probabilidade em em $(\mathcal{X}, \mathcal{F})$, chamada *medida pushforward*, que é denotada X_*P e definida por $(X_*P)(A) \coloneqq P\left(X^{-1}(A)\right)$, $A \in \mathcal{F}$. Pelo teorema de mudança de variáveis [8, Theorem 3.6.1], uma função mensurável g em $\mathcal{X}$ será integrável com respeito a X_*P se, e somente se, $g \circ X$ for integrável com respeito a P. Nesse caso,

$$\int_{\mathcal{X}} g \, \mathrm{d}(X_*P) = \int_{\Omega} g \circ X \, \mathrm{d}P.$$

Teorema A.1 (Radon-Nikodym)**.** *Sejam λ e μ medidas σ-finitas definidas em $(\mathcal{X}, \mathcal{F})$, com $\lambda \ll \mu$. Então existe uma função $f \coloneqq \frac{\mathrm{d}\lambda}{\mathrm{d}\mu} \geq 0$, $\mathcal{F}$-mensurável tal que*

$$\mu(A) = \int_A f \, \mathrm{d}\mu, \quad \forall A \in \mathcal{F}.$$

Ademais, f é única μ-q.t.p., e é chamada derivada de Radon-Nikodym *de λ com respeito a μ.*

Sejam $X \colon \Omega \to \mathcal{X}$ uma variável aleatória do espaço de probabilidade $(\Omega, \mathcal{G}, P)$ no espaço mensurável $(\mathcal{X}, \mathcal{F})$, e μ medida σ-finita em $(\mathcal{X}, \mathcal{F})$. Se X_*P é absolutamente contínua com respeito a μ, então a derivada de Radon-Nikodym $f \coloneqq \frac{\mathrm{d}X_*P}{\mathrm{d}\mu}$ existe e

$$P(X \in A) = \int_{X^{-1}(A)} \mathrm{d}P = \int_A \mathrm{d}(X_*P) = \int_A f \, \mathrm{d}\mu, \quad A \in \mathcal{X}.$$

Nesse caso, f faz o papel da função massa ou densidade de probabilidade (f.m.p. ou f.d.p.), respectivamente nos casos de $\mathcal{X}$ ser discreto ou contínuo.

Definição A.5 (Esperança)**.** Seja $(\Omega, \mathcal{G}, P)$ espaço de probabilidade e $(\mathcal{X}, \mathcal{F})$ outro espaço mensurável. A *esperança* ou *valor esperado* de uma variável aleatória $X \colon \Omega \to \mathcal{X}$ é

$$\mathbb{E}[X] \coloneqq \int_{\Omega} X \, \mathrm{d}P.$$

Se μ é medida em $(\mathcal{X}, \mathcal{F})$ tal que a derivada de Radon-Nikodym $f \coloneqq \frac{\mathrm{d}X_*P}{\mathrm{d}\mu}$ existe, então

$$\mathbb{E}[X] = \int_\Omega \mathrm{id} \circ X \ \mathrm{d}P = \int_\mathcal{X} xf \ \mathrm{d}\mu,$$

onde id denota a função identidade.

Proposição A.2 (Lei do estatístico inconsciente)**.** *Seja g uma função mensurável em $\mathcal{X}$ e $X\colon \Omega \to \mathcal{X}$ uma variável aleatória do espaço de probabilidade $(\Omega, \mathcal{G}, P)$ no espaço de medida $(\mathcal{X}, \mathcal{F}, \mu)$. Se a derivada de Radon-Nikodym $f \coloneqq \frac{\mathrm{d}X_*P}{\mathrm{d}\mu}$ existe, então*

$$\mathbb{E}\left[g(X)\right] = \int_\Omega g \circ X \ \mathrm{d}P = \int_\mathcal{X} gf \ \mathrm{d}\mu.$$

Apêndice B

Pré-Requisitos de Teoria da Informação

Este apêndice apresenta conceitos e resultados básicos de teoria da informação. Está baseado em [17, 45], onde deduções omitidas podem ser encontradas. Para uma variável aleatória discreta $X\colon \Omega \to \mathcal{X}$, denotaremos sua f.m.p. $P_X(x) \coloneqq P(X = x)$.

Definição B.1 (Entropia de Shannon)**.** A *entropia* de uma variável aleatória discreta X é

$$H(X) \coloneqq -\mathbb{E}\left[\log P(X)\right] = -\sum_{x\in\mathcal{X}} P_X(x)\log P_X(x).$$

Usaremos a convenção que $0\log 0 = 0$ na expressão acima. Alternativamente, é possível dar uma caracterização axiomática para a entropia—ver, e.g., [17]. Se X é uma variável aleatória contínua que admite densidade p, também é possível definir sua *entropia diferencial* como $h(X) \coloneqq -\mathbb{E}\left[\log p(X)\right] = -\int_{\mathcal{X}} p\log p \,\mathrm{d}\mu$; no entanto, essa quantidade tem a propriedade um tanto inconveniente de admitir valores negativos.

Definição B.2 (Entropia condicional)**.** Sejam $X\colon \Omega \to \mathcal{X}$ e $Y\colon \Omega \to \mathcal{Y}$ duas variáveis aleatórias discretas. A *entropia condicional* de X, dado Y, é

$$\begin{aligned} H(X|Y) &\coloneqq -\mathbb{E}\left[\log P(X|Y)\right] \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{X,Y}(x,y)\log P_{X|Y}(x|y) \\ &= -\sum_{y\in\mathcal{Y}} P_Y(y)\sum_{x\in\mathcal{X}} P_{X|Y}(x|y)\log P_{X|Y}(x|y). \end{aligned}$$

Teorema B.1 (Algumas propriedades da entropia)**.** *Seja $X\colon \Omega \to \mathcal{X}$ uma variável aleatória discreta; sua entropia $H(X)$ satisfaz às seguintes propriedades.*

1. *$H(X) \geq 0$, com igualdade se, e somente se, X é constante.*
2. *Se $\mathcal{X}$ é finito, então $H(X) \leq \log|\mathcal{X}|$, com igualdade se, e somente se, X tem distribuição uniforme em $\mathcal{X}$.*

3. $H(X|Y) \leq H(X)$, *com igualdade se, e somente se,* X *e* Y *são independentes.*

4. *(Regra da cadeia)* $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$.

Definição B.3 (Divergência de Kullback-Leibler)**.** Sejam P e Q duas medidas de probabilidade. A *divergência de Kullback-Leibler* ou *entropia relativa* entre elas é

$$D_{\mathrm{KL}}(P\|Q) \coloneqq \begin{cases} \mathbb{E}_P\left[\log \frac{\mathrm{d}P}{\mathrm{d}Q}\right], & \text{se } P \ll Q, \\ +\infty, & \text{caso contrário.} \end{cases}$$

Para distribuições discretas, podemos tomar $\frac{\mathrm{d}P}{\mathrm{d}Q}$ como a razão das f.m.p.; para distribuições contínuas, podemos podemos tomar $\frac{\mathrm{d}P}{\mathrm{d}Q}$ como a razão das f.d.p. A divergência de Kullback-Leibler $D_{\mathrm{KL}}(P\|Q)$ mede a dissimilaridade entre as distribuições P e Q; apesar de não ser simétrica nem satisfazer a desigualdade triangular, vale o

Teorema B.2. $D_{\mathrm{KL}}(P\|Q) \geq 0$, *com* $D_{\mathrm{KL}}(P\|Q) = 0 \iff P = Q$.

Definição B.4 (Informação mútua)**.** Sejam X e Y variáveis aleatórias discretas de distribuição conjunta $P_{X,Y}$ e distribuições marinais, respectivamente, P_X e P_Y. A *informação mútua* entre elas é

$$I(X;Y) \coloneqq D_{\mathrm{KL}}(P_{X,Y}\|P_X \otimes P_Y),$$

onde $P_X \otimes P_Y$ denota a distribuição produto entre as marginais de X e Y.

Teorema B.3 (Algumas propriedades da informação mútua)**.** *Sejam* X *e* Y *duas variáveis aleatórias discretas; a informação mútua* $I(X;Y)$ *entre elas satisfaz às seguintes propriedades.*

1. $I(X;Y) = I(Y;X)$.

2. $I(X;Y) \geq 0$, *com igualdade se, e somente se,* X *e* Y *são independentes.*

3. $I(X;X) = H(X)$.

4. $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$.

Intuitivamente, uma estatística suficiente é uma transformação de variável aleatória que não envolve "perda de informação". A definição de informação mútua permite dar a seguinte caracterização, que tomaremos como a definição de estatística suficiente.

Definição B.5 (Estatística suficiente)**.** Seja $\mathcal{S} = \left\{p_\xi \;\middle|\; \xi \in \Xi\right\}$ uma família paramétrica de distribuições de probabilidade e $X\colon \Omega \to \mathcal{X}$ uma variável aleatória distribuída de acordo com uma distribuição dessa família. Uma função $T\colon \mathcal{X} \to \mathcal{Y}$ é uma *estatística suficiente* de X para ξ se, e somente se,

$$I(\xi;X) = I(\xi;T(X)).$$

Apêndice C

Pré-Requisitos de Geometria Diferencial

Este apêndice trata, resumidamente, de conceitos básicos de geometria diferencial, com base em [19, 20, 35], onde demonstrações omitidas podem ser encontradas.

Definição C.1 (Espaço topológico)**.**

1. Um *espaço topológico* (X, τ) é formado por um conjunto X munido de uma *topologia*, i.e., uma família $\tau \subset \wp(X)$, cujos elementos são chamados *abertos*, de forma que (i) $\emptyset$ e X são abertos, (ii) a interseção de finitos abertos é um aberto, e (iii) a união arbitrária de abertos é um aberto.

2. Um espaço topológico (X, τ) é dito *Hausdorff* se quaisquer dois pontos possuem vizinhanças que não se intersectam, i.e., dados $x_1, x_2 \in X$, existem abertos $U_1 \ni x_1$ e $U_2 \ni x_2$, tais que $U_1 \cap U_2 = \emptyset$.

3. Um espaço topológico (X, τ) é dito *segundo contável* se possui base enumerável, i.e., existe uma coleção enumerável $\mathcal{B} \subset \tau$ de forma que qualquer aberto $U \in \tau$ pode ser escrito como união de elementos de $\mathcal{B}$.

Definição C.2 (Variedade diferenciável)**.** Uma *variedade C^k-diferenciável* M de dimensão n (também denotada M^n) é um espaço topológico (M, τ) Haussdorff, segundo contável e *localmente euclidiano*, i.e., existe um *atlas* $\mathcal{A} = \left\{(U_\alpha, \varphi_\alpha^{-1}) \;\middle|\; \alpha \in A\right\}$ formado por homeomorfismos locais

$$\varphi_\alpha^{-1} \colon U_\alpha \subset M \to \varphi_\alpha^{-1}(U_\alpha) \subset \mathbb{R}^n$$

chamados *cartas*, tal que (i) $M = \bigcup_{\alpha \in A} U_\alpha$; (ii) quaisquer duas cartas são compatíveis, i.e., a mudança de coordenadas locais

$$\varphi_\alpha^{-1} \circ \varphi_\beta \colon \varphi_\beta^{-1}(U_{\alpha\beta}) \to \varphi_\alpha^{-1}(U_{\alpha\beta}), \quad U_{\alpha\beta} \coloneqq U_\alpha \cap U_\beta$$

é de classe C^k; (iii) o atlas é maximal, i.e., se (U, φ^{-1}) é uma carta compatível com todas as cartas $(U_\alpha, \varphi_\alpha^{-1})$, então $(U, \varphi^{-1}) \in \mathcal{A}$.

Observamos que sempre que as condições (i) e (ii) forem satisfeitas é possível completar o atlas, de modo a satisfazer (iii). Uma coleção $\{(U_\alpha, \varphi_\alpha^{-1})\colon \alpha \in A\}$ que satisfaz (i)–(iii) é chamada *estrutura diferenciável*. Estaremos particularmente interessados em variedades C^∞-diferenciáveis, também chamadas *variedades diferenciáveis* ou simplesmente *variedades*. Os pares $(U_\alpha, \varphi_\alpha)$, com $\varphi_\alpha\colon \varphi_\alpha^{-1}(U_\alpha) \subset \mathbb{R}^n \to M$ são chamados *parametrizações locais* de M; muitas vezes optaremos por trabalhar com parametrizações locais em detrimento de cartas. Dada uma parametrização local $(U_\alpha, \varphi_\alpha)$ e $p \in U_\alpha$, é usual denotar $\left(x^1, \ldots, x^n\right) \coloneqq \left(x^1(p), \ldots, x^n(p)\right) \coloneqq \varphi_\alpha^{-1}(p)$.

Definição C.3 (Função diferenciável)**.** Sejam M^m e N^n duas variedades diferenciáveis. Uma função $f\colon M \to N$ é *diferenciável* (em $p \in M$), se existem parametrizações (U, φ) em p e (V, ψ) em $f(p)$ tais que

$$\psi^{-1} \circ f \circ \varphi\colon \mathbb{R}^m \to \mathbb{R}^n$$

é uma função diferenciável. Se, além disso, f é inversível e sua inversa é diferenciável, então é dita um *difeomorfismo*.

Definição C.4 (Vetor tangente)**.** Seja M^n uma variedade diferenciável e $\alpha\colon (-\epsilon, \epsilon) \to M$ uma curva diferenciável em M. O *vetor tangente à curva α em $t = 0$* é a função $\alpha'(0)\colon f \mapsto \frac{\mathrm{d}}{\mathrm{d}t}(f \circ \alpha)\Big|_{t=0}$, com f uma função em M diferenciável em $\alpha(0)$. Um *vetor tangente em p* é o vetor tangente em $t = 0$ de alguma curva $\alpha\colon (-\epsilon, \epsilon)$ tal que $\alpha(0) = p$. O conjunto de todos os vetores tangentes a M em um ponto $p \in M$ forma o *espaço tangente* a M em p e é denotado T_pM.

Dada uma escolha de parametrização local (U, φ), podemos escrever localmente a curva α como $\alpha(t) = \varphi\left(x^1(t), \ldots, x^n(t)\right)$, de modo que o vetor tangente à curva α em $p = \alpha(0) = \varphi\left(x_0^1, \ldots, x_0^n\right) \in U$ expressa-se como

$$\begin{aligned}\alpha'(0)(f) &= \frac{\mathrm{d}}{\mathrm{d}t}(f \circ \alpha)\bigg|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t}(f \circ \varphi)\left(x^1(t), \ldots, x^n(t)\right)\bigg|_{t=0} \\ &= \sum_{i=1}^n \left(\frac{\mathrm{d}}{\mathrm{d}t}x^i(0)\right)\left(\frac{\partial f}{\partial x^i}(p)\right) = \left(\sum_{i=1}^n (x^i)'(0)\frac{\partial}{\partial x^i}(p)\right) f.\end{aligned}$$

Assim, temos

$$\alpha'(0) = \sum_{i=1}^n (x^i)'(0)\frac{\partial}{\partial x^i}(p),$$

mostrando também que a escolha de uma parametrização local (U, φ) determina uma base para T_pM dada por $\left\{\frac{\partial}{\partial x^1}(p), \ldots, \frac{\partial}{\partial x^n}(p)\right\}$. Além disso, observamos, em particular, que o vetor

tangente no ponto $p = \varphi\left(x_0^1, \ldots, x_0^n\right)$ a curvas do tipo $\alpha_i(t) = \varphi(x_0^1, \ldots, x_0^i + t, \ldots x_0^n)$ é da forma

$$\alpha_i'(0) = \frac{\partial}{\partial x^i}(p). \tag{C.1}$$

Definição C.5 (Diferencial de uma função)**.** Sejam M^m e N^n duas variedades diferenciáveis e $f\colon M \to N$ uma função diferenciável. A cada $p \in M$ associamos a aplicação linear $\mathrm{d}f_p\colon T_pM \to T_{f(p)}N$ chamada *diferencial de f em p* e definida da seguinte forma. Para $v \in T_pM$, tome uma curva diferenciável $\alpha\colon (-\epsilon, \epsilon) \to M$ tal que $\alpha(0) = p$ e $\alpha'(0) = v$. Então $\mathrm{d}f_p(v) \coloneqq (f \circ \alpha)'(0)$.

A diferencial de uma função está bem definida, pois, como é possível mostrar [20, Proposição 2.7], a definição não depende da escolha da curva α.

Definição C.6 (Métrica riemanniana)**.** Seja M^n uma variedade diferenciável. Uma *métrica riemanniana* em M é um produto interno, i.e., uma forma bilinear simétrica definida positiva

$$g_p\colon T_pM \times T_pM \to \mathbb{R},$$

definida em todo $p \in M$, de sorte que, dado um sistema de coordenadas locais $\left(x^1, \ldots, x^n\right)$, as funções $p \mapsto g_p\left(\frac{\partial}{\partial x^i}(p), \frac{\partial}{\partial x^j}(p)\right) =: g_{ij}$ são diferenciáveis com relação a p, para $1 \leq i, j \leq n$. O par (M, g) é chamado *variedade riemanniana*.

Definição C.7 (Isometria)**.** Sejam (M, g) e (N, h) duas variedades riemannianas. Um difeomorfismo $f\colon M \to N$ é chamado *isometria* se

$$g_p\left(u, v\right) = h_{f(p)}\left(\mathrm{d}f_p(u), \mathrm{d}f_p(v)\right),$$

para todo $p \in M$ e todo $u, v \in T_pM$.

Definição C.8 (Geodésica)**.** Seja (M^n, g) variedade rimeanniana. Uma curva $\gamma\colon I \to M$ é uma *geodésica* se sua derivada covariante é nula em todo ponto da curva. Em coordenadas locais, isso pode ser expresso através das equações diferenciais geodésicas:

$$\ddot{x}_k + \sum_{i,j=1}^{n} \Gamma_{ij}^k \dot{x}_i \dot{x}_j = 0, \quad k = 1, \ldots, n.$$

onde Γ_{ij}^k são os *símbolos de Christoffel*, que podem ser obtidos através da equação

$$g_{\ell k}\Gamma_{ij}^k = \frac{1}{2}\sum_{\ell=1}^{n}\left(\frac{\partial}{\partial x_i}g_{j\ell} + \frac{\partial}{\partial x_j}g_{\ell i} - \frac{\partial}{\partial x_\ell}g_{ij}\right).$$

O resultado a seguir, que usa alguns detalhes técnicos que omitiremos, mostra que as geodésicas minimizam localmente o comprimento de arco. Para mais detalhes, veja [20, § 3.3].

Proposição C.1. *Sejam M uma variedade riemanniana, $p \in M$, U uma vizinhança normal de p e $B \subset U$ uma bola normal de centro p. Seja $\gamma\colon [0,1] \to B$ um segmento de geodésica com $\gamma(0) = p$. Se $\alpha\colon [0,1] \to M$ é outra curva diferenciável por partes ligando $\gamma(0)$ a $\gamma(1)$, então $l(\gamma) \leq l(\alpha)$. E, se a igualdade vale, então $\gamma\big([0,1]\big) = \alpha\big([0,1]\big)$.*

Anexos

Anexo A

Cópia do Artigo [41]

A seguir, anexamos uma cópia da versão aceita do artigo apresentado no Capítulo 2; a versão final está disponível em [41].

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2021.3114094

Constructive Spherical Codes by Hopf Foliations

Henrique K. Miyamoto, *Student Member, IEEE,* Sueli I. R. Costa, *Member, IEEE,* and Henrique N. Sá Earp

***Abstract*—We present a new systematic approach to constructing spherical codes in dimensions 2^k, based on Hopf foliations. Using the fact that a sphere S^{2n-1} is foliated by manifolds $S^{n-1}_{\cos\eta} \times S^{n-1}_{\sin\eta}$, $\eta \in [0, \pi/2]$, we distribute points in dimension 2^k via a recursive algorithm from a basic construction in $\mathbb{R}^4$. Our procedure outperforms some current constructive methods in several small-distance regimes and constitutes a compromise between achieving a large number of codewords for a minimum given distance and effective constructiveness with low encoding computational cost. Bounds for the asymptotic density are derived and compared with other constructions. The encoding process has storage complexity $O(n)$ and time complexity $O(n \log n)$. We also propose a sub-optimal decoding procedure, which does not require storing the codebook and has time complexity $O(n \log n)$.**

***Index Terms*—Asymptotic density, encoding and decoding complexity, Hopf foliation, spherical codes.**

I. Introduction

A SPHERICAL code $\mathcal{C}(M, n, d) \coloneqq \{x_1, x_2, \ldots, x_M\} \subset S^{n-1}$ is a set of M points on the unit Euclidean sphere in $\mathbb{R}^n$ with minimum Euclidean distance at least d, cf. [2]. Problems with spherical codes involve finding optimal distributions of points relative to some parameter of interest, and they lend themselves to several applications. From a practical point of view, it is also desirable that a code exhibits algebraic constructions or geometric regularities, which can provide lower complexity in the encoding and decoding processes. The spherical packing problem in spherical code design can be considered in the following presentation: given a minimum Euclidean distance $d > 0$, to find the largest possible number M of points on S^{n-1} with all mutual distances at least d. The solution is trivial for $n = 2$, namely a regular polygon, but few optimal solutions are known for higher dimensions. Special codes and some best known codes for a given distance in selected dimensions are presented in [2] and [3].

Among the most well-known constructive spherical codes, we highlight the so-called *apple-peeling* [4], *wrapped* [5] and *laminated* [6] methods, the last two being asymptotically dense. The *torus layers spherical codes* (TLSC) [7], while not asymptotically dense, have a more homogeneous structure, in the sense that points on the same leaf are indistinguishable with respect to distance profile, and have been shown to compare favorably with other codes for non-asymptotic minimum distances. This method foliates the sphere S^{2n-1} by flat manifolds $S^1 \times \cdots \times S^1$ and distributes points using good packing density lattices in the half-dimension [8], [9]. Other recent contributions to this topic include codes obtained by partitioning the sphere into regions of equal area [10], bounds for constructible codes near the Shannon bound [11], commutative group codes [12], [13] and cyclic group codes [14]. One main challenge for the application of spherical codes is the effective constructiveness for a large range of distances at a reasonable computational cost, which we propose to address in this work.

Classical applications of spherical codes in communications include channel coding, as a generalization of PSK modulation, and source coding, using shape-gain vector quantizers [15], [16]. The problem of optimal constellation design for signalling in non-coherent communications can be formulated as a sphere packing on the Grassmannian manifold of lines [17], which, in turn, is associated to an antipodal spherical code [18]. A recent example of such approach can be found in [19]. Furthermore, spherical codes have been used in schemes to improve power efficiency of communication systems in the context of MIMO communications [20], [21].

In the context of coherent optical communications, four-dimensional modulations have been considered in order to exploit the physical nature of the electromagnetic field. In [22]–[24], the performance of four-dimensional modulations is studied and spherical codes are also considered. In [25], the authors observe that, at low spectral efficiencies, in dimensions two and four, spherical codes have optimal or close to optimal performance. The performance of modulations in dimensions 8 and 16 has also been addressed in [26], [27].

We propose a construction of spherical codes inspired by the TLSC method and the Hopf fibration, which gives a somewhat 'natural' foliation of S^3, S^7 and S^{15}, and which also appears in problems in physics and communications [24], [28], [29]. Our procedure exploits Hopf foliations in dimensions 2^k to construct a family of *spherical codes by Hopf foliations* (SCHF), by means of a recursive algorithm for any given minimum distance $d \in]0, 2]$. The initial step is a flat

The work of H. K. Miyamoto was supported by São Paulo Research Foundation (FAPESP) under grant 16/05126-0. The work of S. I. R. Costa was supported by Brazilian National Council for Scientific and Technological Development (CNPq) under grant 313326/2017-7 and by FAPESP under grant 13/25977-7. The work of H. N. Sá Earp was supported by CNPq under grant 307217/2017-5 and by FAPESP under grants 17/20007-0 and 18/21391-1. This paper was presented in part at the 2019 IEEE International Symposium on Information Theory [1].

H.K. Miyamoto was with the School of Electrical and Computer Engineering (FEEC), University of Campinas (Unicamp), Campinas, SP 13083-852 Brazil. He is now with the Institute of Mathematics, Statistics and Scientific Computing (IMECC), University of Campinas (Unicamp), Campinas, SP 13083-859 Brazil (e-mail: miyamotohk@gmail.com).

S. I. R. Costa and H. N. Sá Earp are with the Institute of Mathematics, Statistics and Scientific Computing (IMECC), University of Campinas (Unicamp), Campinas, SP 13083-859 Brazil (e-mail: sueli@unicamp.br; henrique.saearp@ime.unicamp.br).

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2021.3114094

model in $\mathbb{R}^4$, for which this construction is equivalent to TLSC via special lattices in dimension two. For higher dimensions, the construction is qualitatively different and, besides defining a much simpler algorithm, for certain minimum distances, it outperforms known TLSC implementations in terms of code cardinality. Although we focus on codes in dimensions 2^k with basic dimension 4, the procedure presented here can be applied to any even dimension $2n$, if provided with a family of spherical codes in dimension n. The performance analysis of the proposed codes includes the comparison with other known constructions, determining their asymptotic density, and computing the complexity of the encoding and decoding processes.

This paper is organized as follows: Section II is an introduction to Hopf foliations. Section III introduces the SCHF, and derives some of their properties and the recursive construction procedure. In Section IV, we present numerical results for constructions in dimensions 4, 8, 16, 32 and 64. In Section V, we derive asymptotic density bounds for our family of codes, which can be closely approached in the simulations in Section IV. Section VI discusses the encoding complexity, showing that this construction has storage complexity $O(n)$ and time complexity $O(n \log n)$. In Section VII, we provide a suboptimal decoding algorithm with time complexity $O(n \log n)$ and storage complexity $O(1)$, which avoids the high-complexity of the ML decoder, while keeping reasonable decoding performance in terms of error rate. Finally, in Section VIII, we draw some conclusions and perspectives for subsequent work.

II. Hopf Fibration and Sphere Foliations

We denote the Euclidean sphere at the origin of $\mathbb{R}^n$, with radius r, by

$$S_r^{n-1} \coloneqq \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| = r\},$$

and the unit sphere simply by $S^{n-1} \coloneqq S_1^{n-1}$. In real dimensions $n \in \{1, 2, 4, 8\}$, let $\mathbb{A} \cong \mathbb{R}^n$ be the corresponding normed division algebra: respectively, the real numbers $\mathbb{R}$, the complex numbers $\mathbb{C}$, the quaternions $\mathbb{H}$ or the octonions $\mathbb{O}$, cf. [30]. Identifying $\mathbb{R}^{2n} \cong \mathbb{A}^2$ by $(x_1, \ldots, x_n; x_{n+1}, \ldots, x_{2n}) \leftrightarrow (z_0, z_1)$ and $\mathbb{R}^{n+1} \cong \mathbb{A} \times \mathbb{R}$ by $(x_1, \ldots, x_n; x_{n+1}) \leftrightarrow (z; x_{n+1})$, the unit $(2n-1)$- and n-spheres can be described respectively by

$$S^{2n-1} = \{(z_0, z_1) \subset \mathbb{A}^2 : |z_0|^2 + |z_1|^2 = 1\} \tag{1}$$

and

$$S^n = \{(z; x_{n+1}) \subset \mathbb{A} \times \mathbb{R} : |z|^2 + x_{n+1}^2 = 1\}.$$

In this description, for $n \in \{1, 2, 4, 8\}$, the *Hopf fibration* [31], [32] is the (submersion) map

$$\begin{aligned} h : S^{2n-1} &\to S^n \\ (z_0, z_1) &\mapsto \left(2 z_0 \overline{z_1}, |z_0|^2 - |z_1|^2\right) \end{aligned} \tag{2}$$

in which $z_0,\ z_1 \in \mathbb{A}$ (see Fig. 1).

Since $|z_0|^2 + |z_1|^2 = 1$, there is a unique $\eta \in [0, \pi/2]$ such that $|z_0| = \cos\eta$ and $|z_1| = \sin\eta$. Each value of η determines a height $x_{n+1} = |z_0|^2 - |z_1|^2 = \cos 2\eta$ in the image S^n, cutting out a $(n-1)$-sphere $S_{\sin 2\eta}^{n-1} \subset S^n$ to which we refer as a *parallel slice*. Furthermore, the preimage $h^{-1}(P)$ of each point $P \in S^n$ under the Hopf fibration (2) is a great sphere $S^{n-1} \subset S^{2n-1}$, called the *fiber of h over P*.

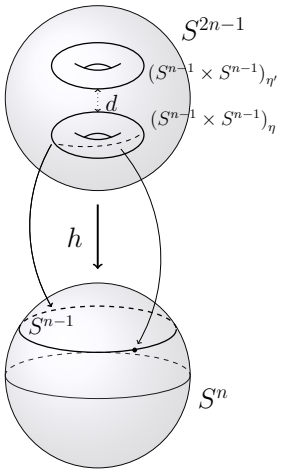


Fig. 1. The generalized Hopf map.

Varying P over such a parallel slice spans a preimage in the total space S^{2n-1} comprising the union of the corresponding fibers, and it can thus be described as the product $S_{\cos\eta}^{n-1} \times S_{\sin\eta}^{n-1} \subset S^{2n-1}$. Hence, by considering all parallel slices in the base sphere S^n, we characterize S^{2n-1} as a disjoint union of product manifolds

$$T_\eta^{2n-2} \coloneqq \left(S^{n-1} \times S^{n-1}\right)_\eta \coloneqq S_{\cos\eta}^{n-1} \times S_{\sin\eta}^{n-1}.$$

This decomposition is an instance of a *foliation*, the *leaves* of which are the generalized tori T_η^{2n-2}; in other words, the sphere S^{2n-1} is said to be *foliated* by tori T_η^{2n-2}. So we will incorporate this vocabulary from differential topology, but we will not invoke any substantial results from that theory in this paper.

As it turns out, this structure, which we call *Hopf foliation*, is not restricted to the cases $n \in \{1, 2, 4, 8\}$, indeed it extends to any $n \in \mathbb{N}^*$, regardless of the existence of an associated normed division algebra in that dimension:

Assertion 1. *For every $n \in \mathbb{N}^*$, the sphere $S^{2n-1} \subset \mathbb{R}^{2n}$ is foliated by manifolds $T_\eta^{2n-2} = \left(S^{n-1} \times S^{n-1}\right)_\eta$.*

Explicitly, write $\mathbf{x} = (x_1, \ldots, x_{2n}) \in S^{2n-1} \subset \mathbb{R}^{2n}$ as

$$\mathbf{x} = \left(\alpha \frac{(x_1, \ldots, x_n)}{\alpha}; \beta \frac{(x_{n+1}, \ldots, x_{2n})}{\beta}\right) \tag{3}$$

for $\alpha \coloneqq \|(x_1, \ldots, x_n)\|$, $\beta \coloneqq \|(x_{n+1}, \ldots, x_{2n})\|$ and $\alpha,\ \beta \neq 0$. For $\alpha = 0$ or $\beta = 0$, we have degenerate manifolds $\mathbf{0} \times S_{\sin\eta}^{n-1}$ or $S_{\cos\eta}^{n-1} \times \mathbf{0}$. Since for any $\mathbf{x} \in S^{2n-1}$ we have $\alpha^2 + \beta^2 = 1$, there is a unique $\eta \in [0, \pi/2]$ such that $\alpha = \cos\eta$ and $\beta = \sin\eta$, so

$$\mathbf{x} = (\cos\eta\, \mathbf{v}_1; \sin\eta\, \mathbf{v}_2), \quad \mathbf{v_1}, \mathbf{v_2} \in S^{n-1}.$$

This describes the foliation of the unit sphere $S^{2n-1} \subset \mathbb{R}^{2n}$ by products of spheres $S_{\cos\eta}^{n-1} \times S_{\sin\eta}^{n-1}$ of radii $\cos\eta$ and $\sin\eta$.

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2021.3114094

In particular, the Hopf fibration in dimension 4 ($n = 2$) gives a foliation of S^3 by two-dimensional flat tori $T^2_\eta = S^1_{\cos\eta} \times S^1_{\sin\eta}$:

$$\begin{aligned}\iota : \left[0, \frac{\pi}{2}\right] \times [0, 2\pi[^2 \;&\to\; S^3 \\ (\eta; \xi_1, \xi_2) \;&\mapsto\; \underbrace{\left(e^{\boldsymbol{i}\xi_1} \cos\eta,\, e^{\boldsymbol{i}\xi_2} \sin\eta\right)}_{\cong(\cos\eta\,(\cos\xi_1, \sin\xi_1);\ \sin\eta\,(\cos\xi_2, \sin\xi_2))}.\end{aligned} \tag{4}$$

For each angle $\eta \in [0, \pi/2]$, the induced map

$$\iota_\eta : [0, 2\pi[^2 \to S^3 \tag{5}$$

spans the 2-torus

$$T_\eta \coloneqq T^2_\eta = S^1_{\cos\eta} \times S^1_{\sin\eta} = \operatorname{im} \iota_\eta \subset S^3 \tag{6}$$

of Euclidean radii $\cos\eta$ and $\sin\eta$. When $\eta \in \{0, \pi/2\}$, the parametrization describes circles, which are degenerate tori. Taking $\eta \notin \{0, \pi/2\}$ and $\mathbf{c} = (c_1, c_2) \coloneqq (\cos\eta, \sin\eta)$, the image $\iota_\eta(u/\cos\eta, v/\sin\eta) = \Phi_{\mathbf{c}}(u, v)$ coincides with the flat tori map defined in [7]. Moreover, $\iota_\eta([0, 2\pi[^2) = \Phi_{\mathbf{c}}([0, 2\pi c_1[\times [0, 2\pi c_2[)$ and $\Phi_{\mathbf{c}}$ is a local isometry, which maps the rectangle into the flat torus in $\mathbb{R}^4$ by gluing its parallel boundary segments.

III. Construction of Spherical Codes

Our construction of spherical codes, inspired by the Hopf fibration, uses the foliations of Assertion 1 to algorithmically distribute points on spheres $S^{2n-1} \subset \mathbb{R}^{2n}$, given a minimum mutual Euclidean distance $d \in]0, 2]$. Each part of the code constructed on a Cartesian product $\left(S^{n-1} \times S^{n-1}\right)_\eta$ corresponds to a *direct sum* [2, Section 1.7] of codes on each copy of S^{n-1}. We construct each code $\mathcal{C}(M, n, d)$ as the union of several such products.

A. Choosing the Leaves

The next result, obtained by straightforward calculation, is used to choose the layers of leaves $\left(S^{n-1} \times S^{n-1}\right)_\eta$, all along our recursive procedure. We remark that, restricted to $n = 2$, this proposition is the same as [7, Proposition 1]. We denote henceforth the integer points of an interval $[a, b]$ by $[\![a, b]\!] \coloneqq [a, b] \cap \mathbb{Z}$.

Proposition 1. *The minimum distance between two leaves $T^{2n-2}_\eta = \left(S^{n-1} \times S^{n-1}\right)_\eta$ and $T^{2n-2}_{\eta'} = \left(S^{n-1} \times S^{n-1}\right)_{\eta'}$ is*

$$d\left(T^{2n-2}_\eta, T^{2n-2}_{\eta'}\right) = 2\sin\left(\frac{\eta - \eta'}{2}\right), \tag{7}$$

which coincides with the Euclidean distance between two points of angles η and η' on the first quadrant of S^1.

Proof: Adopting the notation $\mathbf{v}_1 \coloneqq (v_1, \ldots, v_n)$ and $\mathbf{v}_2 \coloneqq (v_{n+1}, \ldots, v_{2n})$, take two points

$$\mathbf{x} = (\cos\eta\, \mathbf{v}_1;\, \sin\eta\, \mathbf{v}_2) \in \left(S^{n-1} \times S^{n-1}\right)_\eta$$

and

$$\mathbf{x}' = (\cos\eta'\, \mathbf{v}'_1;\, \sin\eta'\, \mathbf{v}'_2) \in \left(S^{n-1} \times S^{n-1}\right)_{\eta'},$$

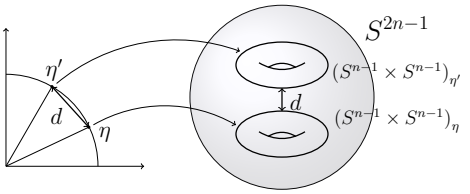


Fig. 2. The distance between leaves $T^{2n-2}_\eta = \left(S^{n-1} \times S^{n-1}\right)_\eta$ and $T^{2n-2}_{\eta'} = \left(S^{n-1} \times S^{n-1}\right)_{\eta'}$ in $\mathbb{R}^{2n}$, viewed as a chordal distance between points determined by the angles η and η' in S^1.

with $\|\mathbf{v}_i\| = \|\mathbf{v}'_i\| = 1$, for $i \in \{1, 2\}$. For the squared Euclidean distance $d^2(\mathbf{x}, \mathbf{x}')$ we have

$$\begin{aligned} d^2(\mathbf{x}, \mathbf{x}') &= \|\mathbf{x} - \mathbf{x}'\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{x}'\|^2 - 2\langle \mathbf{x}, \mathbf{x}'\rangle \\ &= 2 - 2\left(\cos\eta\cos\eta'\langle \mathbf{v}_1, \mathbf{v}'_1\rangle + \sin\eta\sin\eta'\langle \mathbf{v}_2, \mathbf{v}'_2\rangle\right) \\ &\geq 2 - 2\left(\cos\eta\cos\eta' + \sin\eta\sin\eta'\right) \\ &= 2\left[1 - \cos(\eta - \eta')\right] \\ &= 2\left[1 - \left(1 - 2\sin^2\left(\frac{\eta - \eta'}{2}\right)\right)\right] \\ &= 4\sin^2\left(\frac{\eta - \eta'}{2}\right), \end{aligned}$$

and equality holds if, and only if, $\mathbf{v}_1 = \mathbf{v}'_1$ and $\mathbf{v}_2 = \mathbf{v}'_2$. Therefore the minimum distance between the sets $\left(S^{n-1} \times S^{n-1}\right)_\eta$ and $\left(S^{n-1} \times S^{n-1}\right)_{\eta'}$ is $d\left(T^{2n-2}_\eta, T^{2n-2}_{\eta'}\right) = 2\sin\left(\frac{\eta - \eta'}{2}\right)$, which is the chordal distance between points determined by angles η and η' on the first quadrant of the circle $S^1 \subset \mathbb{R}^2$ (see Fig. 2). ∎

Corollary 1. *In the context of Proposition 1:*

a) The minimum angular interval between η and η' respecting the minimum distance d is

$$\Delta\eta \coloneqq |\eta - \eta'| = 2\arcsin(d/2). \tag{8}$$

b) The maximum number of leaves separated by distance d is $t(d) + 1$, with

$$t(d) = \left\lfloor \frac{\pi}{4\arcsin(d/2)} \right\rfloor. \tag{9}$$

c) We may choose the leaves $S^{n-1}_{\cos\eta} \times S^{n-1}_{\sin\eta}$ separated by at least d, considering

i) $\eta = \eta_0 + k\Delta\eta$, for $k \in [\![0, t(d)]\!]$ and $0 \leq \eta_0 \leq (\pi/2 - t(d)\Delta\eta)/2$, or

ii) $\eta = \pi/4 \pm k\Delta\eta$, for $k \in [\![0, \lfloor t(d)/2 \rfloor]\!]$.

In the latter case, leaves are symmetrically chosen around $\eta = \pi/4$, the leaf of greatest 'area'.

Once the leaves have been chosen with a guaranteed minimum mutual distance d, we proceed to construct a spherical code in S^{2n-1}, by considering codes on each leaf T^{2n-2}_η with the desired minimum distance. We illustrate this idea with an example, to show that there are several ways of choosing these leaves.

Example 1. For minimum distance $d = 1$, we have $\Delta\eta = \pi/3$ and $t(1) = 1$. We can choose different sets of leaves, for

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2021.3114094

instance, $\eta = \frac{\pi}{4}$, $\eta \in \{0, \frac{\pi}{3}\}$ or $\eta \in \{\frac{\pi}{12}, \frac{5\pi}{12}\}$. In dimension $n = 4$, we can construct one of the following codes:

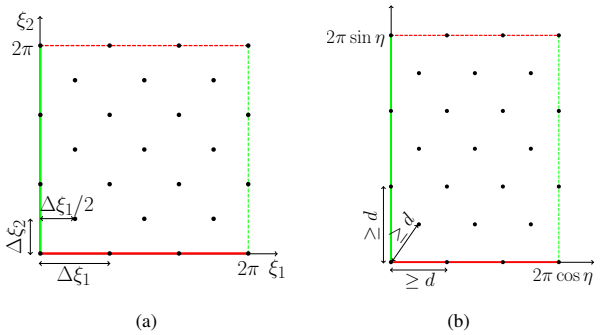
1) Case $\eta = \frac{\pi}{4}$ (only one leaf). Consider the code in $S^1_{1/\sqrt{2}} \times S^1_{1/\sqrt{2}}$ as the product code $\mathcal{C} = \mathcal{C}_{\text{bi}} \times \mathcal{C}_{\text{bi}}$, where $\mathcal{C}_{\text{bi}}$ is the biorthogonal code in $S^1_{1/\sqrt{2}}$, given as the set of all permutations of $(\pm\sqrt{1/2}, 0)$, which has minimum distance 1 and 16 codewords.
2) Case $\eta \in \{0, \frac{\pi}{3}\}$ (two leaves). For $\eta = 0$, consider the code $\mathcal{C}_1 = \mathcal{C}_{\text{hex}} \times \{(0,0)\}$, where $\mathcal{C}_{\text{hex}}$ is the hexagonal code in S^1_1. For $\eta = \frac{\pi}{3}$, consider $\mathcal{C}_2 = \mathcal{C}_{\text{anti}} \times \mathcal{C}_{\text{pen}}$, where $\mathcal{C}_{\text{pen}}$ is the pentagon in $S^1_{\sqrt{3}/2}$ and $\mathcal{C}_{\text{anti}}$ is the set of two antipodal points in $S^1_{1/2}$. The final code $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2$ is a spherical code with 16 codewords.
3) For $\eta \in \{\frac{\pi}{12}, \frac{5\pi}{12}\}$ (two symmetrical leaves), consider the codes $\mathcal{C}_1 = \mathcal{C}_{\text{pen}} \times \mathcal{C}_{\text{one}}$ and $\mathcal{C}_2 = \mathcal{C}_{\text{one}} \times \mathcal{C}_{\text{pen}}$, where $\mathcal{C}_{\text{one}}$ is a single-point code in $S^1_{\sin(\pi/12)} = S^1_{\cos(5\pi/12)}$ and $\mathcal{C}_{\text{pen}}$ is the pentagonal code in $S^1_{\cos(\pi/12)} = S^1_{\sin(5\pi/12)}$. Each $\mathcal{C}_1$, $\mathcal{C}_2$ has 5 points, and the final code $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2$ has 10 points, which is less than the 12 points obtained by taking $\eta \in \{0, \frac{\pi}{2}\}$ (degenerate tori).
4) Note that Proposition 1 provides a sufficient condition for the minimum distance. But in this case we can see that, besides the codewords of item 1), we can also consider points in the degenerate tori defined by $\eta \in \{0, \frac{\pi}{2}\}$, i.e., the codewords $(0, 0, \pm\sqrt{1/2}, \pm\sqrt{1/2})$ and $(\pm\sqrt{1/2}, \pm\sqrt{1/2}, 0, 0)$, and still have minimum distance 1 between the 24 codewords. This spherical code is the best known for $d = 1$ in $\mathbb{R}^4$ [2] and a similar code for this distance with $4\binom{n}{2}$ codewords can be obtained in $\mathbb{R}^n$.

In the algorithm we will formulate shortly, we set a procedure based on Proposition 1, considering different choices for the leaves, such as the ones in items 1), 2) and 3), and a few special codes such as the one in item 4). These examples illustrate the fact that the general construction by leaves is complex, as it requires choosing good codes in the half-dimension. This motivates us to propose a recursive procedure for dimensions 2^k, using dimension $n = 4$ as the basic case.

B. Basic Case: Spherical Codes in $\mathbb{R}^4$

Given a minimum distance $d \in]0, 2]$, our procedure is based on a two-step process:

1) Choose a set of parameters $H = \{\eta_1, \ldots, \eta_p\} \subset [0, \pi/2]$, generating a family of tori $\{T_\eta = S^1_{\cos\eta} \times S^1_{\sin\eta} : \eta \in H\}$ mutually distant by at least d, as sets in $\mathbb{R}^4$, cf. Corollary 1–c).
2) On each torus T_η, distribute points with minimal mutual distance d, following three steps:
   a) choose n *internal circles*, i.e., the images $\iota_\eta(\xi_1, \xi_2)$, for $\xi_1 \in [0, 2\pi[$ and fixed ξ_2, mutually distant by at least d and separated by $\Delta\xi_2$;
   b) on each such circle, distribute m equidistant points, separated by $\Delta\xi_1$;
   c) shift the distributions of consecutive internal circles by $\Delta\xi_1/2$, so as to bring those circles closer and improve the point density.



(a) (b)

Fig. 3. Distribution of points on a torus T_η, seen as the preimage by ι_η (a) and $\Phi_{\mathbf{c}}$ (b), where $\mathbf{c} = (\cos\eta, \sin\eta)$. In this case, $d = 0.9$ and $\eta = 2\arcsin(d/2)$, so that $m = 3$ and $n = 6$. Moreover, $\Delta\xi_1 = 2\pi/m = 2\pi/3$ and $\Delta\xi_2 = 2\pi/n = \pi/3$. Note the displacement of $\Delta\xi_1/2$ between consecutive internal circles.

An illustration of these parameters is given in Fig. 3. The next result provides a way to determine the number n of internal circles and the number m of points on each such circle, within each T_η.

Proposition 2. *On each torus T_η, defined as in* (6)*, for $\eta \in [0, \pi/2]$, we have:*

a) The maximum number $m = m(d, \eta)$ of points that can be distributed on a internal circle, respecting the minimum mutual distance d, is

$$m(d,\eta) = \begin{cases} \left\lfloor \dfrac{\pi}{\arcsin\left(d/(2\cos\eta)\right)} \right\rfloor, & \text{if} \quad d \leq 2\cos\eta, \\ 1, & \text{otherwise.} \end{cases} \tag{10}$$

b) The maximum number $n = n(d, \eta)$ of internal circles that can be distributed on T_η, with shifting angle π/m, $m = m(d, \eta)$, such that the mutual distance among their points is at least d, is

$$n(d,\eta) = \max\{\tilde{n}, 1\}, \tag{11}$$

with $\tilde{n} = 2\lfloor \min\{n_1, n_2\}/2 \rfloor$ and

$$n_1 = \left\lfloor \frac{\pi}{\arcsin\left[\left((d^2/4)\csc^2\eta - \cot^2\eta \sin^2(\pi/2m)\right)^{\frac{1}{2}}\right]} \right\rfloor \tag{12}$$

$$n_2 = \begin{cases} \left\lfloor \dfrac{2\pi}{\arcsin\left(d/(2\sin\eta)\right)} \right\rfloor, & \text{if} \quad d \leq 2\sin\eta, \\ 1, & \text{otherwise.} \end{cases} \tag{13}$$

Proof: The calculations are straightforward using that the squared distance between the image by ι_η as in (5) of two points determined by angles (ξ_1, ξ_2) and (ξ'_1, ξ'_2) is

$$\begin{aligned} &d^2\big(\iota_\eta(\xi_1,\xi_2), \iota_\eta(\xi'_1,\xi'_2)\big) \\ &\quad = |e^{\boldsymbol{i}\xi_1}\cos\eta - e^{\boldsymbol{i}\xi'_1}\cos\eta|^2 + |e^{\boldsymbol{i}\xi_2}\sin\eta - e^{\boldsymbol{i}\xi'_2}\sin\eta|^2. \end{aligned} \tag{14}$$

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2021.3114094

In particular, when the points are in the same internal circle determined by ξ_2 and are displaced by $\Delta\xi_1$, i.e., $\xi_1' = \xi_1 + \Delta\xi_1$ and $\xi_2' = \xi_2$, we have

$$d\big(\iota_\eta(\xi_1,\xi_2), \iota_\eta(\xi_1+\Delta\xi_1,\xi_2)\big) = 2\cos\eta\sin(\Delta\xi_1/2). \quad (15)$$

Note that $m = m(d,\eta)$ is obtained by ensuring minimum distance between points in the same internal circle. Using that $m = \lfloor 2\pi/\Delta\xi_1 \rfloor$, where $\Delta\xi_1 = 2\arcsin\big(d/(2\cos\eta)\big)$ is obtained by inverting (15) and setting the distance to d, yields (10).

Now, for $n = n(d,\eta)$, we have to ensure minimum distance between points both in shifted and aligned internal circles (see Fig. 3). First, we compute the distance between two points in the image by ι_η, one in each of two consecutive internal circles, with point distributions shifted by $\Delta\xi_1/2 = \pi/m$, with $m = m(d,\eta)$. Hence, setting $\xi_1' = \xi_1 + \Delta\xi_1/2$ and $\xi_2' = \xi_2 + \Delta\xi_2$ in (14) gives

$$d\big(\iota_\eta(\xi_1,\xi_2),\, \iota_\eta(\xi_1+\Delta\xi_1/2, \xi_2+\Delta\xi_2)\big) = \left(4\cos^2\eta\sin^2\frac{\pi}{2m} + 4\sin^2\eta\sin^2\frac{\Delta\xi_2}{2}\right)^{\frac{1}{2}}. \quad (16)$$

Note that $n_1 = \lfloor 2\pi/\Delta\xi_2 \rfloor$, with

$$\Delta\xi_2 = 2\arcsin\left[\big((d^2/4)\csc^2\eta - \cot^2\eta\sin^2(\pi/2m)\big)^{\frac{1}{2}}\right]$$

obtained by inverting (16) and setting the distance to d.

Finally, the distance between the image by ι_η of two points aligned with respect to ξ_1 in two (alternate) internal circles parametrized by ξ_2 and $\xi_2 + 2\Delta\xi_2$ is obtained by setting $\xi_1' = \xi_1$ and $\xi_2' = \xi_2 + 2\Delta\xi_2$ in (14):

$$d\big(\iota_\eta(\xi_1,\xi_2), \iota_\eta(\xi_1,\xi_2+2\Delta\xi_2)\big) = 2\sin\eta\sin\Delta\xi_2. \quad (17)$$

Similarly, we have $n_2 = \lfloor 2\pi/\Delta\xi_2 \rfloor$, with $\Delta\xi_2 = \arcsin\big(d/(2\sin\eta)\big)$ obtained from (17) with distance d.

As we have to ensure minimum distances both in (16) and in (17), we choose the minimum between n_1 and n_2. Notice moreover that, if we put more than one internal circle, the number of circles $\tilde{n}$ effectively has to be even, so that first and last circles (which are neighboring circles in the torus T_η) have different displacements, thus ensuring minimum mutual distance between their points (see Fig. 3). ∎

It is possible to describe the generation of points as described above in complex variables, once again referring to the Hopf foliation and noticing that a rotation in $\mathbb{R}^2 \cong \mathbb{C}$ corresponds to multiplication by a unit complex number. Thus, on each torus T_η, points take the form

$$(z_0, z_1) = \big(e^{\boldsymbol{i}(j\Delta\xi_1 + k\Delta\xi_1/2)}\cos\eta,\, e^{\boldsymbol{i}(k\Delta\xi_2)}\sin\eta\big),$$

with $j \in [\![0, m-1]\!]$ and $k \in [\![0, n-1]\!]$. This description can compare favorably, for instance, to the use of rotation matrices in [7], because it reduces several matrix products to scalar and complex products. In this work, we have considered the complex description for its simplicity in implementation.

C. Recursive Generalization: Spherical Codes in $\mathbb{R}^{2^k}$

With the generalized foliation of Assertion 1, the following natural two-step algorithm for $S^{2n-1} \subset \mathbb{R}^{2n}$ emerges:

1) Vary the parameter $\eta \in [0, \pi/2]$, generating a family of leaves $S^{n-1}_{\cos\eta} \times S^{n-1}_{\sin\eta}$ separated by at least d.
2) On each leaf $S^{n-1}_{\cos\eta} \times S^{n-1}_{\sin\eta}$, distribute points recursively on each of the spheres $S^{n-1}_{\cos\eta}$ and $S^{n-1}_{\sin\eta}$, at scaled minimum distances $d/\cos\eta$ and $d/\sin\eta$, respectively.

We shall focus on dimensions $2^k,\ k \geq 2$, starting from $\mathbb{R}^4$. For instance, to construct a spherical code in $S^{15} \subset \mathbb{R}^{16}$, we foliate it by manifolds $\big(S^7 \times S^7\big)_\eta$, and each copy of S^7 is itself foliated by $\big(S^3 \times S^3\big)_{\eta'}$. The distribution on each copy of $S^3 \subset \mathbb{R}^4$ is known from the basic case.

In our implementation, the standard algorithm exploits in particular the symmetry of the leaves $\big(S^{n-1} \times S^{n-1}\big)_\eta$ around $\eta = \pi/4$. The first chosen leaf is $\eta_0 = \pi/4$, the distribution is done for $\eta \in\,]\pi/4, \pi/2]$ and the points for $\eta \in [0, \pi/4[$ are obtained by coordinate permutations.

As a direct result of the proposed construction, we immediately derive the following Proposition 3:

Proposition 3. *The cardinality $M(2n,d)$ of a SCHF in dimension $2n$, with minimum distance d, constructed by our standard procedure, is given by the recursive expressions* (18) *and* (19) *at the bottom of the page, with $\eta_i \coloneqq \pi/4 + i\Delta\eta$ as in* (8)*, $t \coloneqq t(d)$ as in* (9)*, $m_i \coloneqq m(d,\eta_i)$ as in* (10) *and $n_i \coloneqq n(d,\eta_i)$ as in* (11)*.*

Example 2. To construct a code in $\mathbb{R}^8$ with minimum distance $d = 0.5$ by our standard procedure, we consider the foliation of S^7 by $\big(S^3 \times S^3\big)_\eta$ and use Proposition 1 to choose the set of parameters $\eta \in \{0.2800, \pi/4, 1.2908\}$. Next, for each leaf $S^3_{\cos\eta} \times S^3_{\sin\eta}$, we take the Cartesian product of codes in the 3-spheres of radii $\cos\eta$ and $\sin\eta$ in $\mathbb{R}^4$. On each of these 3-spheres, we apply the basic-case algorithm for minimum distances $d/\cos\eta$ and $d/\sin\eta$, namely: choose a family of tori T_η and distribute points on each. For instance, $\big(S^3 \times S^3\big)_{1.2908} = S^3_{0.2764} \times S^3_{0.9610}$. On the first-component sphere, it is only possible to choose one torus, with $\eta = \pi/4$. On the second-component sphere, we choose the tori with $\eta \in \{0.2591, \pi/4, 1.3117\}$. Due to the symmetry about $\eta = \pi/4$, in both cases, it suffices to calculate half of the points

$$M(2n,d) = \begin{cases} \Big(M(n,\sqrt{2}d)\Big)^2 + 2\displaystyle\sum_{i=1}^{\lfloor t/2 \rfloor} M(n, d/\cos\eta_i) M(n, d/\sin\eta_i), & \text{for } n > 2, \quad (18) \\ m_0 n_0 + 2\displaystyle\sum_{i=1}^{\lfloor t/2 \rfloor} m_i n_i, & \text{for } n = 2\ . \quad (19) \end{cases}$$

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2021.3114094

TABLE I
CARDINALITY OF FOUR-DIMENSIONAL SPHERICAL CODES FOR DIFFERENT MINIMUM DISTANCES d

d	SCHF	TLSC [7]	Apple-peeling [4]	Wrapped [5]	Laminated [6]
0.5	168	172	170	*	*
0.4	321	308	342	*	*
0.3	774	798	826	*	*
0.2	$2,683$	$2,718$	$2,822$	*	*
0.1	$22,164$	$22,406$	$22,740$	$17,198$	$16,976$
0.01	2.27×10^7	2.27×10^7	1.97×10^7	$2.31 \times 10^{7\dagger}$	2.31×10^7
0.001	2.27×10^{10}	2.27×10^{10}	2.27×10^{10}	$2.59 \times 10^{10\dagger}$	2.59×10^{10}

* unknown values, † estimated values

TABLE II
CARDINALITY OF n-DIMENSIONAL SPHERICAL CODES FOR DIFFERENT MINIMUM DISTANCES d

n	d	SCHF	TLSC (k elements) [33]	TLSC (polygon layers) [33]
8	0.5	$4,206$	$2,748$	$2,312$
	0.3	$150,200$	$45,252$	$89,945$
	0.1	3.89×10^8	6.47×10^6	4.09×10^8
	0.01	4.28×10^{15}	7.66×10^{10}	5.19×10^{15}
16	0.5	$471,912$	$69,984$	$195,312$
	0.3	2.77×10^8	1.17×10^8	7.17×10^7
	0.1	4.90×10^{15}	2.41×10^{12}	2.39×10^{15}
	0.01	6.48×10^{30}	3.66×10^{20}	*
32	0.5	2.47×10^7	32	$32,768$
	0.3	4.95×10^{12}	2.68×10^{12}	1.41×10^{12}
	0.1	1.87×10^{27}	6.81×10^{21}	7.02×10^{24}
	0.01	3.96×10^{58}	2.48×10^{38}	*
64	0.5	4.98×10^9	64	2.14×10^9
	0.3	4.61×10^{17}	2.40×10^{11}	9.22×10^{18}
	0.1	9.35×10^{44}	1.08×10^{38}	2.90×10^{37}

* unknown values

and obtain the symmetric ones by permuting their coordinates. Summing across all the leaves, there are 2,608 points in total.

D. Modifications

One can consider some small modifications of the previous standard procedure, in order to improve the cardinality of the code.

1) When choosing leaves, in the context of Corollary 1, we may choose not only symmetrically distributed leaves around $\eta = \pi/4$, but consider the following choices – and even a combination of them – in different dimensions:
   a) $\eta = \pi/4 \pm k\Delta\eta$, for $k \in [\![0, \lfloor t(d)/2 \rfloor]\!]$;
   b) $\eta = k\Delta\eta$, for $k \in [\![0, t(d)]\!]$;
   c) $\eta = \pi/2 - k\Delta\eta$, for $k \in [\![0, t(d)]\!]$.
2) When distributing points on a torus T_η, in the context of Proposition 2, we may consider only the 'diagonal' internal circles, i.e., the images $\iota_\eta(\xi_1, \xi_2)$ with $\xi_1 = \xi_2$, whenever that is more advantageous than the standard distribution. As those circles have unit radius, the number of points that can be placed on them, respecting minimum mutual distance d, is $\lfloor \pi / \arcsin(d/2) \rfloor$.
3) Whenever possible and more advantageous, we can consider explicit *ad hoc* constructions [2]: optimal codes in $\mathbb{R}^4$ for cardinalities 2, 3, 4, 5, 8, 10 and 24 (minimum distances 2, $\sqrt{3}$, $\sqrt{8/3}$, $\sqrt{5/2}$, $\sqrt{2}$, $\sqrt{5/3}$, 1, respectively), as well as the biorthogonal codes which place $2n$ points with minimum distance $\sqrt{2}$ in any dimension n.

Remark 1. In the proposed standard procedure, any dimension n can be considered as a basic case for codes in $\mathbb{R}^{2n}$, so long as good constructive codes are available in $\mathbb{R}^n$ for a wide range of minimum distances. However much this may provide greater density, as discussed in Section V, in this paper we focus on the construction in $\mathbb{R}^{2^k}$ with basic case $\mathbb{R}^4$ due to its effective constructiveness and low complexity.

IV. NON-ASYMPTOTIC PERFORMANCE ANALYSIS

We compare the cardinality of our codes with other constructive spherical codes, in different dimensions, for many non-asymptotic minimum distance regimes. In dimension 4, we compare our results with apple-peeling[1] [4], wrapped [5], laminated [6] and the torus layers spherical codes (TLSC) [7] (Table I). In higher dimensions, we compare with two TLSC implementations by Naves [33] that differ in the choice of the subcode: either with k elements or on polygon layers (Table II). In these dimensions, we have also considered codes

[1]We follow the description in [5], using the implementation generously shared by its authors.

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2021.3114094

TABLE III
CARDINALITY OF n-DIMENSIONAL SPHERICAL CODES FOR DIFFERENT MINIMUM DISTANCES d

n	d	SCHF	EQ codes [10]
4	0.27944	918	500
	0.23707	1,540	1,000
	0.10374	19,768	10,000
8	0.51282	3,228	500
	0.47025	5,889	1,000
	0.31379	100,074	10,000
16	0.56498	23,882	500
	0.51483	182,424	1,000
	0.40868	2.06×10^6	10,000
32	0.45847	4.07×10^7	500
	0.44805	1.68×10^8	1,000
	0.41207	6.59×10^8	10,000

TABLE IV
CARDINALITY OF FOUR-DIMENSIONAL SPHERICAL CODES FOR DIFFERENT MINIMUM DISTANCES d

d	SHCF	CGC [12]
0.330158	556	200
0.237033	1,586	400
0.193059	2,988	600
0.16806	4,535	800
0.149405	6,450	1,000

TABLE V
CARDINALITY OF n-DIMENSIONAL SPHERICAL CODES FOR DIFFERENT MINIMUM DISTANCES d

n	d	SCHF	CGC [13]
4	0.012706	11,067,004	141,180
	0.00733585	57,610,534	423,540
	0.00465076	226,265,570	1,053,780
	0.00423537	299,595,092	1,270,620
8	0.707107	416	648
	0.541196	2,342	2,048
	0.437016	9,700	5,000
	0.366025	38,244	10,368

generated by the equal area sphere partitioning algorithm (EQ codes) [10] (Table III), concatenated MPSK [34, p. 36], commutative group codes (CGC) [12], [13] (Tables IV and V) and some codes from Sloane *et al.* [3]. The SCHF considered in these tables use the modifications introduced in Section III-D. The proposed SCHF construction was implemented in *Wolfram Mathematica* and *Python*.

We see that the performance of SCHF in $\mathbb{R}^4$ is, as expected, similar to TLSC and not far from some of the best known spherical codes. In higher dimensions, SCHF can achieve a higher cardinality than TLSC (k elements) in most regimes and, in many of them, higher than TLSC (polygon layers) too. Commutative group codes, which have a powerful algebraic structure, are outperformed by SCHF in nearly all considered minimum distance regimes.

A more complete picture is given on Fig. 4, 5, 6, 7, showing the binary rate per dimension $R = (\log_2 M)/n$ for codes in dimensions 4, 8, 16 and 32, respectively. These computations for the proposed recursive SCHF, both with and without the modifications of Section III-D, show in each dimension and for small values of d a good approximation of the asymptotic bounds derived in Section V. Indeed, SCHF generally outperforms the other plotted constructions.

TABLE VI
CARDINALITY OF FOUR-DIMENSIONAL SPHERICAL CODES FOR DIFFERENT MINIMUM DISTANCES d

d	SCHF	Rodrigues *et al.* [24]
0.488876	174	112
0.389872	344	128

We also acknowledge the recent appearance, on a somewhat different vein, of some Hopf fibration formalism in the context of optical communications [24]. A design for higher order modulations is presented, based on an interesting use of the Hopf preimage under the so-called *sampled discrete Hopf fibration* and in close relation to physical properties of light. That construction relies on the choice of a polytope on the base space (such as the tetrakis hexahedron) and apparently does not address the spherical packing problem for any given minimum distance. Particularities aside, SCHF outperform the two four-dimensional modulations presented in [24] at the same minimum distance, cf. Table VI.

V. ASYMPTOTIC DENSITY

We now analyze the density of our spherical codes. Consider the gamma function $\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} \, \mathrm{d}t$. The Euclidean $(n-1)$-dimensional volume (hypersurface area) of the sphere $S^{n-1} \subset \mathbb{R}^n$ is given [8] by

$$\mathbb{S}_n \coloneqq \frac{n\pi^{n/2}}{\Gamma(1+n/2)},$$

and the corresponding n-dimensional volume of the ball bounded by S^{n-1} is

$$\mathbb{V}_n \coloneqq \frac{\pi^{n/2}}{\Gamma(1+n/2)}.$$

Note that the spherical code with minimum distance d has minimum angular separation $\theta(d) = 2\arcsin(d/2)$. The $(n-1)$-dimensional volume of a spherical cap on the sphere S^{n-1} with angular radius $\theta(d)/2$ is

$$\mathbb{S}_C(n,d) \coloneqq \mathbb{S}_{n-1} \int_0^{\theta(d)/2} \sin^{n-2} x \, \mathrm{d}x.$$

Hence the density $\Delta(\mathcal{C})$ of a n-dimensional spherical code $\mathcal{C}(M,n,d)$ is the ratio of the total area covered by the $M(n,d)$ spherical caps, with angular radius $\theta(d)/2$ centered at the codewords, by the total surface area:

$$\Delta(\mathcal{C}(n,d)) \coloneqq \frac{M(n,d)\,\mathbb{S}_C(n,d)}{\mathbb{S}_n}. \tag{20}$$

In what follows, we will write $f(d) \simeq g(d)$ when

$$\lim_{d\to 0} \frac{f(d)}{g(d)} = 1.$$

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2021.3114094

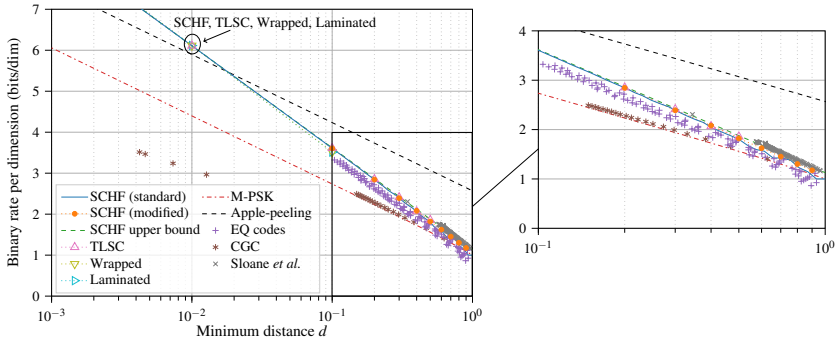


Fig. 4. Binary rate per dimension for different codes in dimension 4 with detail.

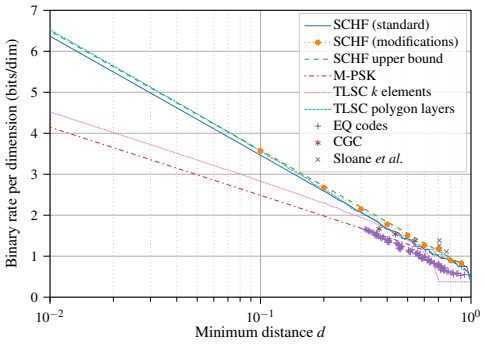


Fig. 5. Binary rate per dimension for different codes in dimension 8.

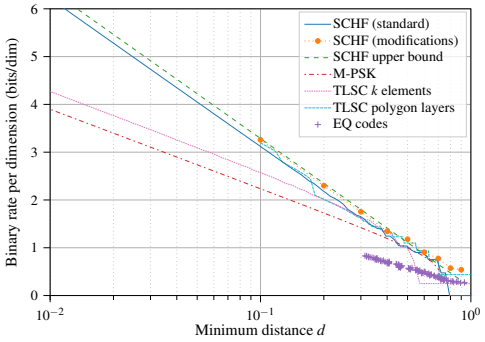


Fig. 6. Binary rate per dimension for different codes in dimension 16.

For small values of d, $\mathbb{S}_C(n,d)$ can be approximated [5] by

$$\mathbb{S}_C(n,d) = \mathbb{V}_{n-1}\left(\frac{d}{2}\right)^{n-1} + O(d^{n+1}),$$

which implies $\mathbb{S}_C(n,d) \simeq \mathbb{V}_{n-1}\left(\frac{d}{2}\right)^{n-1}$ and

$$\Delta(\mathcal{C}(n,d)) \simeq \frac{M(n,d)\mathbb{V}_{n-1}}{\mathbb{S}_n}\left(\frac{d}{2}\right)^{n-1}. \tag{21}$$

The *center density* of a spherical code $\mathcal{C}(M,n,d)$ is defined as $\Delta_c(\mathcal{C}(n,d)) := \Delta(\mathcal{C}(n,d))/\mathbb{V}_{n-1}$ and its asymptotic value

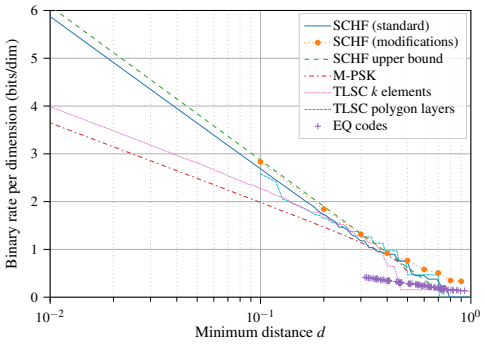


Fig. 7. Binary rate per dimension for different codes in dimension 32.

for a family of spherical codes, constructed for different minimum distances d, is

$$\overline{\Delta}_c(\mathcal{C}(n)) := \lim_{d\to 0} \frac{M(n,d)}{\mathbb{S}_n}\left(\frac{d}{2}\right)^{n-1}. \tag{22}$$

It provides a means of comparing packings of different constructions in a given dimension, for small d.

Lemma 1. *The asymptotic center density of the SCHF in dimension* 4 *is*

$$\overline{\Delta}_c(\mathrm{SCHF}[4]) = \frac{1}{4\sqrt{3}}. \tag{23}$$

Proof: In $\mathbb{R}^4$, the asymptotic SCHF and TLSC coincide, and their densities can be approached by the density of the lattice product $A_2\times\mathbb{Z}$ [7, Proposition 6]. Considering the well-known center densities of lattices A_2 and $\mathbb{Z}$ [8], we then have:

$$\begin{aligned}\overline{\Delta}_c(\mathrm{SCHF}[4]) = \Delta_c(A_2\times\mathbb{Z}) &= \Delta_c(A_2)\cdot\Delta_c(\mathbb{Z})\\ &= \frac{1}{2\sqrt{3}}\cdot\frac{1}{2} = \frac{1}{4\sqrt{3}}.\end{aligned}$$

■

Proposition 4. *The asymptotic center density of the SCHF in dimension* $2n$*, constructed from a family of codes* $\mathcal{C}(n)$ *in dimension* n*, which achieves asymptotic density* $\overline{\Delta}_c(\mathcal{C}(n))$*, is*

$$\overline{\Delta}_c(\mathrm{SCHF}[2n;\mathcal{C}(n)]) = \frac{1}{2}\left(\overline{\Delta}_c(\mathcal{C}(n))\right)^2.$$

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2021.3114094

Proof: For small d, we have

$$\overline{\Delta}_c(\mathrm{SCHF}[2n;\mathcal{C}(n)]) \simeq \frac{M(2n,d)}{\mathbb{S}_{2n}}\left(\frac{d}{2}\right)^{2n-1} \tag{24}$$

and

$$\overline{\Delta}_c(\mathcal{C}(n)) \simeq \frac{M(n,d)}{\mathbb{S}_{n}}\left(\frac{d}{2}\right)^{n-1}. \tag{25}$$

From (25), we have

$$M(n,d) \simeq \mathbb{S}_n\left(\frac{2}{d}\right)^{n-1}\overline{\Delta}_c(\mathcal{C}(n)). \tag{26}$$

In the asymptotic behavior, the particular choice of leaves as introduced in Section III-D is irrelevant. Therefore, we consider, for simplicity, the leaves $\eta_i = i\Delta\eta$. From the construction of SCHF, we have

$$\begin{aligned} M(2n,d) &= \sum_{i=0}^{t(d)} M(n,d/\cos\eta_i)M(n,d/\sin\eta_i) \\ &\simeq \sum_{i=0}^{t(d)}\Bigg[\left(\mathbb{S}_n\left(\frac{2\cos\eta_i}{d}\right)^{n-1}\overline{\Delta}_c(\mathcal{C}(n))\right) \\ &\qquad \times\left(\mathbb{S}_n\left(\frac{2\sin\eta_i}{d}\right)^{n-1}\overline{\Delta}_c(\mathcal{C}(n))\right)\Bigg] \\ &= [\overline{\Delta}_c(\mathcal{C}(n)]^2(\mathbb{S}_n)^2\frac{2^{n-1}}{d^{2n-2}}\sum_{i=0}^{t(d)}[\sin(2\eta_i)]^{n-1}. \end{aligned}$$

Therefore, from (24),

$$\begin{aligned} &\overline{\Delta}_c(\mathrm{SCHF}\,[2n;\mathcal{C}(n)]) \\ &\qquad \simeq \frac{[\overline{\Delta}_c(\mathcal{C}(n)]^2}{2^n}\frac{(\mathbb{S}_n)^2}{\mathbb{S}_{2n}}\sum_{i=0}^{t(d)}[\sin(2\eta_i)]^{n-1}\,d. \end{aligned}$$

For small d, we have $\Delta\eta \simeq d$, $\eta_i \simeq id$ and $t(d) \simeq \pi/2d$, hence the last summation approaches the corresponding integral, which implies

$$\begin{aligned} &\overline{\Delta}_c\left(\mathrm{SCHF}[2n;\mathcal{C}(n)]\right) \\ &\qquad = \frac{[\overline{\Delta}_c(\mathcal{C}(n)]^2}{2^n}\frac{(\mathbb{S}_n)^2}{\mathbb{S}_{2n}}\int_0^{\pi/2}(\sin 2\eta)^{n-1}\,\mathrm{d}\eta \\ &\qquad = \frac{[\overline{\Delta}_c(\mathcal{C}(n)]^2}{2^{n+1}}\frac{(\mathbb{S}_n)^2}{\mathbb{S}_{2n}}\int_0^{\pi}(\sin x)^{n-1}\,\mathrm{d}x. \end{aligned} \tag{27}$$

The sphere S^{2n-1} is foliated by the leaves $S^{n-1}_{\cos\eta_i}\times S^{n-1}_{\sin\eta_i}$, with $\eta_i \in [0,\pi/2]$, and the distance between the leaves is the chordal distance in S^1 (Proposition 1). Hence, for small d, the arc-chord approximation $d \simeq \Delta\eta$ yields

$$\begin{aligned} \mathbb{S}_{2n} &\simeq \sum_{i=0}^{t(d)}\mathsf{Vol}(S^{n-1}_{\cos\eta_i}\times S^{n-1}_{\sin\eta_i})\Delta\eta \\ &= \sum_{i=0}^{t(d)}\mathsf{Vol}(S^{n-1}_{\cos\eta_i})\mathsf{Vol}(S^{n-1}_{\sin\eta_i})\Delta\eta \\ &= \sum_{i=0}^{t(d)}\mathbb{S}_n(\cos\eta_i)^{n-1}\,\mathbb{S}_n(\sin\eta_i)^{n-1}\,\Delta\eta \\ &= \frac{(\mathbb{S}_n)^2}{2^{n-1}}\sum_{i=0}^{t(d)}[\sin(2\eta_i)]^{n-1}\,\Delta\eta, \end{aligned}$$

where $\mathsf{Vol}(\cdot)$ denotes the volume of the object, and, when $d\to 0$, as in (27),

$$\mathbb{S}_{2n} = \frac{(\mathbb{S}_n)^2}{2^n}\int_0^{\pi}(\sin x)^{n-1}\,\mathrm{d}x. \tag{28}$$

Substituting (28) in (27) yields the claim:

$$\overline{\Delta}_c\left(\mathrm{SCHF}[2n;\mathcal{C}(n)]\right) = \frac{1}{2}\big(\overline{\Delta}_c(\mathcal{C}(n))\big)^2. \tag{29}$$

■

Remark 2. Since the maximum asymptotic center density for spherical codes in $S^{n-1}\subset\mathbb{R}^n$ is the highest center packing density of $\mathbb{R}^{n-1}$, denoted by $\Delta_c(\mathbb{R}^{n-1})$ (cf. Proposition 4), the asymptotic density of a SCHF in $S^{2n-1}\subset\mathbb{R}^{2n}$ is bounded above and asymptotic to

$$\overline{\Delta}_c\left(\mathrm{SCHF}[2n;\mathcal{C}(n)]\right) \le \frac{1}{2}\big(\Delta_c(\mathbb{R}^{n-1})\big)^2. \tag{30}$$

By recursively applying Proposition 4 in dimensions 2^k, for $k\ge 2$ (cf. Lemma 1), we get

Corollary 2. *The asymptotic center density of the recursive SCHF in dimension $n=2^k$, $k\ge 2$ is*

$$\begin{aligned} \overline{\Delta}_c(\mathrm{SCHF}[2^k]) &\coloneqq \overline{\Delta}_c\big(\mathrm{SCHF}[2^k;\mathrm{SCHF}[2^{k-1};\cdots\mathrm{SCHF}[4]]]\big) \\ &= (2)^{1-(3)2^{k-2}}(3)^{-2^{k-3}}. \end{aligned} \tag{31}$$

Moreover, using (26) and Corollary 2, we get:

Corollary 3. *The cardinality $M(n,d)$ of the recursive SCHF in dimension $n=2^k$ is bounded above and, as $d\to 0$, asymptotic to*

$$\overline{M}(2^k,d) = \frac{2^{k+2^{k-2}}\;3^{-2^{k-3}}\;\pi^{2^{k-1}}}{(2^{k-1})!\;d^{2^k-1}}. \tag{32}$$

Proof: For small values of d, we have the approximation

$$\overline{\Delta}_c \simeq \frac{M(n,d)}{\mathbb{S}_n}\left(\frac{d}{2}\right)^{n-1}.$$

Using (31), we get

$$\begin{aligned} \overline{M}(2^k,d) &= \overline{\Delta}_c(\mathrm{SCHF}[2^k])\;\mathbb{S}_{2^k}\left(\frac{2}{d}\right)^{2^k-1} \\ &= \left((2)^{1-(3)2^{k-2}}(3)^{-2^{k-3}}\right)\frac{2^k\pi^{2^{k-1}}}{(2^{k-1})!}\left(\frac{2}{d}\right)^{2^k-1} \\ &= \frac{2^{k+2^{k-2}}\;3^{-2^{k-3}}\;\pi^{2^{k-1}}}{(2^{k-1})!\;d^{2^k-1}}. \end{aligned}$$

■

For a fixed dimension n, the asymptotic center density of different spherical code constructions allows one to compare their respective numbers of codewords for the same small minimum distance d, cf. (26).

Table VII compares some asymptotic center densities for spherical codes in dimensions 4, 8, 16 and 32. We consider both the SCHF recursive procedure with basic case $\mathbb{R}^4$ (Corollary 2) and the SCHF procedure that uses asymptotic dense codes in the half dimension (Remark 2). For each dimension, we also include the TLSC upper bound as in [7, Proposition 6], the apple-peeling bound as in [5, Lemma 3] and highest

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2021.3114094

TABLE VII
ASYMPTOTIC CENTER DENSITY FOR DIFFERENT n-DIMENSIONAL SPHERICAL CODES. Δ_n DENOTES THE HIGHEST CENTER DENSITY OF A SPHERE IN $\mathbb{R}^n$: $\Delta_n \coloneqq \Delta_c(\mathbb{R}^{n-1})$. FOR DIMENSIONS 8, 16 AND 32 THE CALCULATIONS ASSUME THE BEST KNOWN CENTER DENSITIES.

	SCHF (recursive)	SCHF (half-dimension)	TLSC	Apple-peeling	$n-1$ packing
n	$\left(2^{\frac{3n}{4}-1}3^{\frac{n}{8}}\right)^{-1}$	$\frac{1}{2}\left(\Delta_{\frac{n}{2}-1}\right)^2$	$\Delta_{\frac{n}{2}}\Delta_{\frac{n}{2}-1}$	$\frac{\mathbb{V}_{n-2}}{\mathbb{V}_{n-1}}\frac{\Delta_{n-2}}{2}\beta\left(\frac{n}{2},\frac{1}{2}\right)$	Δ_{n-1}
4	$\frac{1}{4\sqrt{3}} \approx 0.1443$	–	$\frac{1}{4\sqrt{3}} \approx 0.1443$	$\frac{1}{4\sqrt{3}} \approx 0.1443$	$\frac{1}{4\sqrt{2}} \approx 0.1768$
8	$\frac{1}{96} \approx 0.0104$	$\frac{1}{64} \approx 0.0156$	$\frac{1}{32\sqrt{2}} \approx 0.0221$	$\frac{1}{16\sqrt{3}} \approx 0.0361$	$\frac{1}{16} = 0.0625$
16	$\frac{1}{18,432} \approx 5.43 \times 10^{-5}$	$\frac{1}{512} \approx 0.0020$	$\frac{1}{256} \approx 0.0039$	$\frac{1}{32\sqrt{3}} \approx 0.0180$	$\frac{1}{16\sqrt{2}} \approx 0.0442$
32	$\frac{1}{2^{23}\cdot 3^4} \approx 1.47 \times 10^{-9}$	$\frac{1}{1,024} \approx 0.0010$	$\frac{1}{256\sqrt{2}} \approx 0.0028$	$\frac{3^{13.5}}{2^{23}} \approx 0.3292$	$\frac{3^{15}}{2^{23.5}} \approx 1.2095$

asymptotic center density for spherical codes (from the best known packing in the previous dimension) [8], [35].

We can see that the ratios between the center densities of these different constructions show how much smaller the number of codewords achieved by recursive SCHF construction is, when compared SCHF using half-dimension codes, TLSC, apple-peeling and, of course, the highest possible asymptotic density in each dimension, which can be theoretically achieved by wrapped or laminated codes. The trade-off to emphasize here is the high constructibility of recursive SCHF for any given minimum distance and its low complexity in the encoding and decoding processes, which will be discussed in Sections VI and VII.

One should also point out that there may be a difference between asymptotic density bounds and the density effectively achieved, especially for higher dimensions. This is due to the characteristics of each of the analyzed constructions. Half-dimension SCHF depend on the existence of good codes in the half dimension; TLSC require the use of the best codes and lattices in the half dimension; wrapped codes rely on the choice of a lattice in the previous dimension; laminated codes have been approached in dimensions from 2 to 49 and may have slower convergence than wrapped codes; apple-peeling construction is based on spherical codes in the previous dimension. We note that, in the case of TLSC, the implementations carried out so far do not seek to construct the densest theoretically possible codes, but rather good, feasible ones – as in [33], which proposes different approaches for the construction of the subcodes. On the other hand, in dimensions 2^k, the construction of recursive SCHF does not depend on any choice, can be done for any given minimum distance d and the asymptotic bound is indeed approached in the results shown in Fig. 4, 5, 6, 7.

The construction of SCHF using better available constructions in the half dimension offers an easy way of obtaining good codes. One could consider, for instance, using a family of wrapped codes in dimension 25 (as in [15], based on the Leech lattice) to construct codes in dimension 50 by Hopf foliations, with low addition in encoding complexity, cf. Section VI (a wrapped spherical code in this dimension would require the use of a good lattice in dimension 49).

Finally, it is also interesting to compare the asymptotic behavior of recursive SCHF with the more structured spherical commutative group codes (CGC). From [12, Proposition 7], the number $M(n,d)$ of codewords of a CGC in dimension n is bounded above by

$$M(n,d) < \Delta_c(\Lambda_{n/2})\left(\frac{4\pi}{d\sqrt{n/2}}\right)^{n/2}, \tag{33}$$

where $\Delta_c(\Lambda_{n/2})$ is the maximum center density of a lattice packing in $\mathbb{R}^{n/2}$. This implies that the asymptotic center density is equal to zero, which is expected, since those codes must be contained in a n-dimensional flat torus. Note that, for a fixed dimension $n = 2^k$, the cardinality of a CGC grows with $O(1/d^{2^{k-1}})$, while for a recursive SCHF (cf. Corollary 3), it grows with $O(1/d^{2^k-1})$, i.e., there exists a value d beyond which the cardinality of SCHF outperforms CGC (see Tables IV and V).

VI. ENCODING COMPLEXITY ANALYSIS

We now present a complexity analysis of the encoding algorithm for the standard SCHF construction. Lachaud and Stern [36] propose the following definition for polynomial complexity of a spherical code. Let Σ be a finite alphabet and consider spherical codes $\mathcal{C} = \mathcal{C}(M,n,d) \subset S^{n-1}$ that are images of maps $F : \Sigma^k \to S^{n-1}$. We say that a family $(\mathcal{C}_i)$ of spherical codes is *polynomially constructible* if there is a sequence (F_i) of maps $F_i : \Sigma^{k_i} \to S^{n_i-1}$ such that: (i) F_i is one-to-one from Σ^{k_i} to $\mathcal{C}_i$, and (ii) for every $a \in \Sigma^{k_i}$, the point $F_i(a)$ is computable from i and a in polynomial time, with respect to the dimension n_i of $\mathcal{C}_i$.

A. Basic Case: Spherical Codes in $\mathbb{R}^4$

We first analyze the encoding complexity in dimension $n = 4$. The injection F can be decomposed as $F = \iota \circ \chi$, where ι is as in (4) and $\chi(a) = (\eta; \xi_1, \xi_2)$ as in Algorithm 1. We assume that, in the construction of the code $\mathcal{C}(M,4,d)$, we store a table that contains information on each leaf T_η: each

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2021.3114094

Algorithm 1 Encoding algorithm for $\mathbb{R}^4$ (map $F = \iota \circ \chi$).

Input: a, d
Output: $\mathbf{x} = (x_1, x_2, x_3, x_4)$
1: $t \leftarrow \lfloor \pi/4 \arcsin(d/2) \rfloor$
2: **for** $i \in [\![-\lfloor t/2 \rfloor, \lfloor t/2 \rfloor]\!]$ **do**
3: $M_i \leftarrow$ consult i-th line of table
4: **if** $a \geq M + M_i$ **then**
5: $M \leftarrow M + M_i$
6: $i \leftarrow i + 1$
7: **else**
8: $\eta \leftarrow \frac{\pi}{4} + 2i \arcsin \frac{d}{2}$
9: $m \leftarrow$ as in Proposition 2, item a)
10: $n \leftarrow$ as in Proposition 2, item b)
11: $j \leftarrow (a - M) \mod m$
12: $k \leftarrow \lfloor (a - M)/m \rfloor$
13: $\xi_1 \leftarrow j\frac{2\pi}{m} + k\frac{\pi}{m}$
14: $\xi_2 \leftarrow k\frac{2\pi}{n}$
15: **return** $\mathbf{x} \leftarrow$
$(\cos\eta\cos\xi_1, \cos\eta\sin\xi_1, \sin\eta\cos\xi_2, \sin\eta\sin\xi_2)$
16: **end if**
17: **end for**

line contains the index i of the leaf, the parameter η_i and the number of points in that leaf M_i. The length of this table is $t(d) = \lfloor \pi/4 \arcsin(d/2) \rfloor$ (Corollary 1), hence the storage complexity is $O(t)$. We assume that accessing data in this table has constant complexity.

Each individual line of Algorithm 1 has constant complexity (constant number of additions, multiplications, trigonometric functions etc.). In the worst-case scenario, the main loop (line 2) will be executed $t = t(d)$ times. Note that

$$t(d) = \left\lfloor \frac{\pi}{4\arcsin(d/2)} \right\rfloor \leq \frac{\pi}{4\arcsin(d/2)} \leq \frac{\pi}{2d},$$

so the computational complexity of the algorithm is $O(t) = O(d^{-1})$.

B. General Case: Spherical Codes in $\mathbb{R}^{2n}$

We consider now the algorithm that implements the map $F(a) = (x_1, \ldots, x_{2n}) \in \mathcal{C}(M, 2n, d)$. This injection can be decomposed with the help of the following maps:

$$\begin{aligned} \sigma : \Sigma &\to \left[0, \frac{\pi}{2}\right] \times \Sigma_1 \times \Sigma_2 \\ a &\mapsto (\eta; a_1, a_2), \end{aligned} \tag{34}$$

$$\begin{aligned} \iota \circ \chi : \Sigma &\to \mathcal{C}(M, 4, d) \\ a &\mapsto (\cos\eta\cos\xi_1, \cos\eta\sin\xi_1, \sin\eta\cos\xi_2, \sin\eta\sin\xi_2), \end{aligned} \tag{35}$$

and

$$\begin{aligned} \Phi : \left[0, \frac{\pi}{2}\right] &\times \mathcal{C}(|\Sigma_1|, n, d/\cos\eta) \times \mathcal{C}(|\Sigma_2|, n, d/\sin\eta) \\ &\to \mathcal{C}(|\Sigma_1||\Sigma_2|, 2n, d) \\ (\eta;\, &(x_1, \ldots, x_n),\, (y_1, \ldots, y_n)) \\ &\mapsto (\cos\eta\,(x_1, \ldots, x_n);\, \sin\eta\,(y_1, \ldots, y_n)), \end{aligned} \tag{36}$$

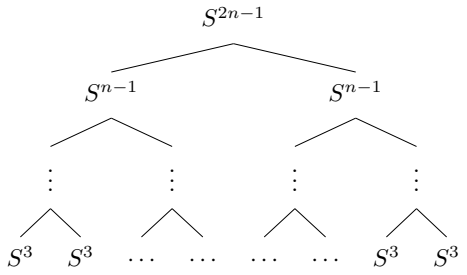


Fig. 8. Decomposition tree for S^{2n-1}. Note that the number of nodes is $\sum_{i=0}^{k-2} 2^i = 2^{k-1} - 1 = n - 1$, with $k = \log_2(2n)$.

where Σ is the alphabet of the total code ($|\Sigma| = M$) and Σ_1, Σ_2 are the alphabets of each half-dimension code.

We also assume that we have stored tables with information on each leaf $\left(S^{\tilde{n}-1} \times S^{\tilde{n}-1}\right)_\eta$, for dimensions $\tilde{n} \in \{n, n/2, \ldots, 2\}$, during the construction of the code. Each line of such a table contains the index i of the leaf, its parameter η_i, and the number of points $M_{i,1}$ and $M_{i,2}$ on each of the half-dimension spheres $S^{\tilde{n}-1}_{\cos\eta_i}$ and $S^{\tilde{n}-1}_{\sin\eta_i}$, respectively (see Table VIII).

TABLE VIII
EXAMPLE OF STORAGE TABLE FOR $\mathcal{C}(M, 2n, d)$

i	η_i	$M_{i,1}$	$M_{i,2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
-1	$\pi/4 - \Delta\eta$	$M_{-1,1}$	$M_{-1,2}$
0	$\pi/4$	$M_{0,1}$	$M_{0,2}$
1	$\pi/4 + \Delta\eta$	$M_{1,1}$	$M_{1,2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

The length of each table is equal to the number $\tilde{t}$ of leaves in the corresponding dimension. For S^{2n-1}, this number is $t(d) = \lfloor \pi/4 \arcsin(d/2) \rfloor$ (Corollary 1). There will be $n - 1$ tables, one for each sphere (node) of the decomposition tree (Fig. 8). Note that, when halving the dimension of the sphere ($S^{2n-1} \to S^{n-1}$), the size of the table in the new dimension cannot increase:

$$d \leq \min\left\{ \frac{d}{\cos\eta}, \frac{d}{\sin\eta} \right\},$$

therefore

$$\left\lfloor \frac{\pi}{4\arcsin(d/2)} \right\rfloor \geq \max\left\{ \left\lfloor \frac{\pi}{4\arcsin(d/2\cos\eta)} \right\rfloor, \left\lfloor \frac{\pi}{4\arcsin(d/2\sin\eta)} \right\rfloor \right\}.$$

Thus the storage space needed is no greater than $(n-1)t(d)$, and the storage complexity is $O(nt) = O(nd^{-1})$, which is linear in the dimension n.

The algorithm that implements the map $\iota \circ \chi$ in (35) is the encoding algorithm for the basic case $\mathbb{R}^4$ (Algorithm 1), and it has complexity $O(t)$. The map σ in (34) is implemented by Algorithm 2. Each individual line has constant complexity and,

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2021.3114094

Algorithm 2 Algorithm implementing map σ.

Input: a, d
Output: η, a_1, a_2
1: $M \leftarrow 0$
2: $\tilde{t} \leftarrow \lfloor \pi/4 \arcsin(d/2) \rfloor$
3: **for** $i \in [\![-\lfloor \tilde{t}/2 \rfloor, \lfloor \tilde{t}/2 \rfloor]\!]$ **do**
4: $(M_{i,1}, M_{i,2}) \leftarrow$ consult i-th line of table
5: **if** $a > M + M_{i,1}M_{i,2}$ **then**
6: $M \leftarrow M + M_{i,1}M_{i,2}$
7: $i \leftarrow i + 1$
8: **else**
9: $\eta \leftarrow \frac{\pi}{4} + 2i \arcsin \frac{d}{2}$
10: $a_1 \leftarrow (a - M) \mod M_{i,1}$
11: $a_2 \leftarrow \lfloor (a - M)/M_{i,1} \rfloor$
12: **return** $(\eta; a_1, a_2)$
13: **end if**
14: **end for**

Algorithm 3 Encoding algorithm for $\mathbb{R}^n$ (map F).

Input: n, a, d
Output: $\mathbf{x} = (x_1, \ldots, x_n)$
1: **if** $n = 4$ **then**
2: **return** $\mathbf{x} \leftarrow \iota \circ \chi(a, d)$
3: **else**
4: $(\eta; a_1, a_2) \leftarrow \sigma(a, d)$
5: $(w_1, \ldots, w_{n/2}) \leftarrow F(n/2, a_1, d/\cos\eta)$
6: $(z_1, \ldots, z_{n/2}) \leftarrow F(n/2, a_2, d/\sin\eta)$
7: **return** $\mathbf{x} \leftarrow \big(\cos\eta\,(w_1, \ldots, w_{n/2}),\, \sin\eta\,(z_1, \ldots, z_{n/2})\big)$
8: **end if**

in the worst-case scenario, the main loop (line 3) is repeated $\tilde{t} \leq t$ times, hence it has complexity $O(t) = O(d^{-1})$.

Finally, the implementation of map F is represented in Algorithm 3. The general step for dimension n computes $\sigma(a, d)$ with $O(d^{-1})$ (line 4), calls itself twice with parameter $n/2$ (lines 5 and 6), and performs n multiplications (line 7). If we have a good family of codes in dimension $n/2$, with encoding complexity $O(f(n))$, and we apply one iteration of Algorithm 3 to double the dimension with SCHF construction, the encoding complexity of the new code with respect to the dimension n will be $O(\max\{2f(n), n\})$. In the recursive case, the number of steps of the recurrence is characterized by

$$T(n) = 2T\left(\frac{n}{2}\right) + O(n) + O(d^{-1}).$$

Using the master theorem [37, p. 73], we find that, for fixed d, this algorithm has complexity $O(n \log n)$.

We can compare this complexity with known TLSC implementations [33]. Codes obtained via a subcode with k elements have linear time complexity; in spite of the low complexity, they have the weakest performance among TLSC implementations and are outperformed by recursive SCHF in most scenarios (see Section IV). Codes on polygon layers have the best performance among TLSC implementations and the closest to SCHF; nonetheless, their exact complexity has not been established and, based on the code structure and computing time required for tested examples, seems to be higher than recursive SCHF. For instance, the results in Table II for this construction, with $d = 0.01$ in dimensions 16 and 32, could not be computed using the implementation provided in [33] under the same time and storage resources as the other two codes.

VII. Decoding

Given a vector $\mathbf{y} \in \mathbb{R}^n$ and a spherical code $\mathcal{C}(M, n, d)$, the maximum likelihood (ML) decoding consists in finding the vector $\mathbf{x} \in \mathcal{C}(M, n, d)$ such that

$$\mathbf{x} = \arg\min_{\mathbf{x}_i \in \mathcal{C}} \|\mathbf{y} - \mathbf{x}_i\|. \tag{37}$$

As shown in [7], to decode a received vector $\mathbf{y}$ in a spherical code, we can consider $\mathbf{y}$ to be a unit vector and the problem is equivalent to

$$\mathbf{x} = \arg\max_{\mathbf{x}_i \in \mathcal{C}} \langle \mathbf{x}_i, \mathbf{y} \rangle. \tag{38}$$

For small codes, it is feasible to obtain (38) by computing all inner products and choosing the maximizing codeword. But, to avoid high-complexity of ML decoding on larger codes, we introduce a sub-optimal decoding algorithm for standard SCHF construction, which is inspired by [7] and does not require storage of the whole codebook. As previously, we start with a procedure for the basic case $\mathbb{R}^4$ and then generalize it recursively to $\mathbb{R}^{2n}$.

A. Basic Case: Spherical Codes in $\mathbb{R}^4$

Using the Hopf foliation, a unit vector $\mathbf{y} = (y_1, y_2, y_3, y_4)$ may be written as

$$(y_1 + \boldsymbol{i} y_2,\, y_3 + \boldsymbol{i} y_4) = (e^{\boldsymbol{i}\xi_1} \cos\eta,\, e^{\boldsymbol{i}\xi_2} \sin\eta),$$

where

$$\eta = \arctan\left(\sqrt{\frac{y_3^2 + y_4^2}{y_1^2 + y_2^2}}\right), \tag{39}$$

$$\xi_1 = \arctan(y_2/y_1), \tag{40}$$

$$\xi_2 = \arctan(y_4/y_3). \tag{41}$$

In general, however, the triplet $(\eta; \xi_1, \xi_2)$ does not parametrize a point of the codebook. So our objective is to find the triplet $(\hat{\eta}; \hat{\xi_1}, \hat{\xi_2})$ which parametrizes the codeword closest to $\mathbf{y}$. Let us denote our guess by $\hat{\mathbf{x}} = (e^{\boldsymbol{i}\hat{\xi_1}} \cos\hat{\eta},\, e^{\boldsymbol{i}\hat{\xi_2}} \sin\hat{\eta})$. We propose a two-step decoding method, as follows.

1) The first step is to search for the torus $T_{\hat{\eta}}$ closest to received point $\mathbf{y}$. Thanks to Proposition 1, this is equivalent to finding

$$\hat{\eta} = \arg\min_{\eta' \in H} |\eta' - \eta|, \tag{42}$$

where $H = \{\frac{\pi}{4} + 2i \arcsin\frac{d}{2},\ -\lfloor t(d)/2 \rfloor \leq i \leq \lfloor t(d)/2 \rfloor\}$ is the set of η-parameters used in the code.
2) Once $\hat{\eta}$ is determined, we project $\mathbf{y}$ on $T_{\hat{\eta}}$, obtaining $(e^{\boldsymbol{i}\xi_1} \cos\hat{\eta},\, e^{\boldsymbol{i}\xi_2} \sin\hat{\eta})$. This is the point on $T_{\hat{\eta}}$ which is closest to the received vector $\mathbf{y}$. To obtain $\hat{\xi_1}$ and $\hat{\xi_2}$, we compute

$$\hat{k} = \lfloor \xi_2/\Delta\xi_2 \rceil \mod n \tag{43}$$

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2021.3114094

Algorithm 4 Decoding algorithm in $\mathbb{R}^4$.

Input: $\mathbf{y}$, d
Output: $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4)$
1: $\mathbf{y} \leftarrow \mathbf{y}/\|\mathbf{y}\|$
2: $(\eta, \xi_1, \xi_2) \leftarrow$ as in (39), (40), (41)
3: $\hat{\eta} \leftarrow \left\lfloor \frac{\eta - \pi/4}{\Delta\eta} \right\rceil \Delta\eta + \frac{\pi}{4}$, with $\Delta\eta = 2\arcsin(d/2)$
4: $m \leftarrow$ as in Proposition 2, item a)
5: $n \leftarrow$ as in Proposition 2, item b)
6: $(\hat{\xi}_1, \hat{\xi}_2) \leftarrow$ as in (43), (44), (45), (46)
7: **return** $\hat{\mathbf{x}} \leftarrow$ $(\cos\hat{\eta}\cos\hat{\xi}_1, \cos\hat{\eta}\sin\hat{\xi}_1, \sin\hat{\eta}\cos\hat{\xi}_2, \sin\hat{\eta}\sin\hat{\xi}_2)$

and

$$\hat{j} = \left\lfloor \frac{\xi_1 - \hat{k}\Delta\xi_1/2}{\Delta\xi_1} \right\rceil \mod m, \tag{44}$$

where $\lfloor\cdot\rceil$ denotes the rounding function and $\Delta\xi_1 = 2\pi/m$, $\Delta\xi_2 = 2\pi/n$. Then,

$$\hat{\xi}_1 = \hat{j}\Delta\xi_1 + \hat{k}\Delta\xi_2, \tag{45}$$

$$\hat{\xi}_2 = \hat{k}\Delta\xi_2. \tag{46}$$

These steps are detailed in Algorithm 4. To further approach the minimum distance solution, additional steps can be considered. If $\hat{d} \coloneqq \|\hat{\mathbf{x}} - \mathbf{y}\| < d/2$, the decoding is finished. Otherwise, the closest point $\mathbf{x}^*$ may be on another torus $T_{\eta*}$. We can look for the set of tori with parameters $\{\eta_1, \ldots, \eta_\nu\}$ for which $d_i = \|\hat{\mathbf{x}}_i - \mathbf{y}\| < \hat{d}$, where

$$\hat{\mathbf{x}}_i = \left(e^{\boldsymbol{i}\xi_{1,i}}\cos\eta_i,\, e^{\boldsymbol{i}\xi_{2,i}}\sin\eta_i\right)$$

is obtained from the projection on torus T_{η_i}, $i \in \{1, \ldots, \nu\}$ and we choose $\mathbf{x}^* = \hat{\mathbf{x}}_i$ that minimizes d_i. As for the coding design, in dimension 4, this procedure approaches the one proposed for TLSC [7].

B. General Case: Spherical Codes in $\mathbb{R}^{2n}$

Let us generalize the decoding procedure to codes on S^{2n-1}. As before, write the arbitrary received vector $\mathbf{y} \in S^{2n-1}$ as

$$\mathbf{y} = \left(\cos\eta\,(y_1, \ldots, y_n),\, \sin\eta\,(y_{n+1}, \ldots, y_{2n})\right),$$

with

$$\eta = \arctan\left(\sqrt{\frac{\sum_{i=n+1}^{2n} y_i^2}{\sum_{j=1}^{n} y_j^2}}\right). \tag{47}$$

A similar two-step procedure can be deduced, as follows.

1) Find the closest leaf $\left(S^{n-1} \times S^{n-1}\right)_{\hat{\eta}}$ to point $\mathbf{y}$, i.e., the value $\hat{\eta}$ that parametrizes a leaf used in the code closest to η, as in (42).
2) We can then split S^{2n-1} into $S^{n-1}_{\cos\hat{\eta}}$ and $S^{n-1}_{\sin\hat{\eta}}$ and recursively apply the procedure all the way down to the basic case $S^3 \subset \mathbb{R}^4$. The additional steps can also be applied for each leaf $\left(S^{n-1} \times S^{n-1}\right)_{\eta}$.

A pseudocode for the recursive method is presented in Algorithm 5.

Algorithm 5 Decoding algorithm in $\mathbb{R}^n$.

Input: $\mathbf{y}$, d, n
Output: $\hat{\mathbf{x}} = (\hat{x}_1, \ldots, \hat{x}_n)$
1: **if** $n = 4$ **then**
2: Apply Algorithm 4 to $\mathbf{y}$ and d
3: **else**
4: $\mathbf{y} \leftarrow \mathbf{y}/\|\mathbf{y}\|$
5: Compute η from $\mathbf{y}$ using (47)
6: $\hat{\eta} \leftarrow \left\lfloor \frac{\eta - \pi/4}{\Delta\eta} \right\rceil \Delta\eta + \frac{\pi}{4}$, with $\Delta\eta = 2\arcsin(d/2)$
7: $(w_1, \ldots, w_{n/2}) \leftarrow$ apply decoding in $\mathbb{R}^{n/2}$ to $(y_1, \ldots, y_{n/2})$ with $d/\cos\hat{\eta}$
8: $(z_1, \ldots, z_{n/2}) \leftarrow$ apply decoding in $\mathbb{R}^{n/2}$ to $(y_{(n/2)+1}, \ldots, y_n)$ with $d/\sin\hat{\eta}$
9: **return** $\hat{\mathbf{x}} \leftarrow \left(\cos\hat{\eta}\,(w_1, \ldots, w_{n/2}),\, \sin\hat{\eta}\,(z_1, \ldots, z_{n/2})\right)$
10: **end if**

C. Decoding Performance

We analyze the performance of decoding using the previously presented standard SCHF procedure. The only information required to store are minimum distance and dimension, so this algorithm has storage complexity $O(1)$.

The number of operations $T(n)$ in the general loop of Algorithm 5 follows the recursive expression

$$T(n) = 2T\left(\frac{n}{2}\right) + O(n),$$

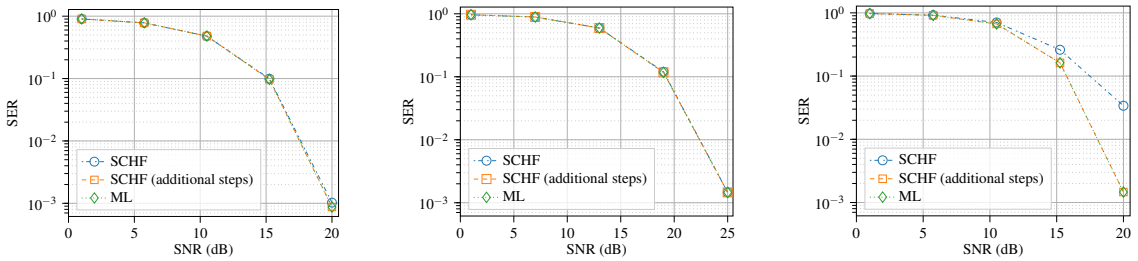
where $O(n)$ accounts for computing $\mathbf{y}$ (line 4), η from (47) (line 5) and the products in line 9. Using the master theorem [37, p. 73], it follows that this algorithm has complexity $O(n\log n)$. Compare this with the complexity of a brute-force ML decoder, which has time complexity $O(Mn)$ and storage complexity $O(M)$. In [34], two decoding algorithms are proposed for laminated spherical codes: one uses $O(\sqrt{M})$ space and $O(\log M)$ time and the other uses $O(1)$ space and $O(\sqrt{M})$ time.

To analyze the performance of this sub-optimal decoder, we have performed the following test: for a given code $\mathcal{C}(M, n, d)$, we add i.i.d. centered Gaussian noise $\mathbf{z}_i \sim \mathcal{N}(0, \sigma^2)$ to each point $\mathbf{x}_i$ in the code and decode each $\mathbf{y_i} = \mathbf{x}_i + \mathbf{z}_i$. We compute the symbol error rate (SER) for different signal-to-noise ratios (SNR), as well as the average CPU time[2] required for decoding one codeword using the proposed algorithm without additional steps, with some additional steps (if $\hat{d} \geq d/2$, we consider the two adjacent leaves $\eta \pm \Delta\eta$), and the brute-force ML decoder. Results are presented in Fig. 9 and Table IX.

While the SER of the proposed decoder without additional steps is higher than brute-force ML for higher SNR, the average time required to decode one codeword by the latter method can be up to ten times higher. On the other hand, when allowing simple additional steps, the decoding performance practically matches brute-force ML, while keeping low time complexity. This justifies the use of the proposed sub-optimal decoder.

[2]Using *Python 3.7.6* on a 8GB RAM, Intel Core i5-7200U @ 2.50GHz machine.

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2021.3114094



(a) $\mathcal{C}(52, 4, 0.7)$ (b) $\mathcal{C}(152, 4, 0.5)$ (c) $\mathcal{C}(360, 8, 0.7)$

Fig. 9. Symbol error rate (SER) for decoding different SCHF $\mathcal{C}(M, n, d)$ with sub-optimal SCHF methods and brute-force ML decoder.

TABLE IX
AVERAGE CPU TIME (ms) FOR DECODING ONE CODEWORD

	SCHF without additional steps	SCHF with additional steps	Brute-force ML
$\mathcal{C}(52, 4, 0.7)$	0.109	0.139	0.409
$\mathcal{C}(152, 4, 0.5)$	0.114	0.139	1.169
$\mathcal{C}(360, 8, 0.7)$	0.287	0.582	2.837

VIII. CONCLUSION

We propose a construction for spherical codes in dimensions 2^k by a recursive procedure that is based on the Hopf foliations of S^{2n-1} by $\left(S^{n-1} \times S^{n-1}\right)_\eta$ and uses $\mathbb{R}^4$ as basic case. In fact, this construction can be applied to any even dimension $2n$ as long as a family of spherical codes is provided in dimension n.

Given a minimum distance $d > 0$, the standard method chooses leaves $S^{n-1}_{\cos\eta} \times S^{n-1}_{\sin\eta}$, parametrized by $\eta \in [0, \pi/2]$, that foliate S^{2n-1} while mutually distant by at least d. On each leaf, we recursively distribute points on each of the spheres $S^{n-1}_{\cos\eta}$ and $S^{n-1}_{\sin\eta}$, with scaled minimum distances and combine the results as a Cartesian product. In the basic case $\mathbb{R}^4$, the sphere S^3 is foliated by tori T_η, each of which is divided in internal circles mutually distant by d, where points are equidistantly distributed.

In non-asymptotic regime, SCHF compare favorably to other constructive methods. Asymptotic upper bounds for the recursive and half-dimension SCHF are derived and compared with other constructions. An encoding algorithm is presented, the time and storage complexities of which are respectively $O(n \log n)$ and $O(n)$. A sub-optimal decoder with time complexity $O(n \log n)$ and storage complexity $O(1)$ is also proposed. We verify in some examples that, by allowing additional steps, its SER is close to that of ML decoder, while keeping the time required significantly lower.

Perspectives for the extension of this work include investigating, in several $2n$ dimensions, SCHF constructed from good available codes in dimension n; considering the structure of quaternions and octonions in the construction of codes; and analyzing the proposed SCHF for vector quantitation of Gaussian sources.

ACKNOWLEDGMENT

The authors are grateful for some generous contributions by C. Torezzan, V. Vaishampayan and L. Naves, as well as to J. Hamkins and K. Zeger, for sharing their apple-peeling implementation. We also thank the referees for their important suggestions, which have meaningfully improved the original manuscript.

REFERENCES

[1] H. K. Miyamoto, H. N. Sá Earp, and S. I. R. Costa, "Constructive spherical codes in 2^k dimensions," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2019, pp. 1612–1616.
[2] T. Ericson and V. Zinoviev, *Codes on Euclidean spheres*. Amsterdam, The Netherlands: North-Holland, 2001.
[3] N. J. A. Sloane *et al.* Tables of spherical codes. [Online]. Available: http://neilsloane.com/packings/
[4] A. E. Gamal, L. Hemachandra, I. Shperling, and V. Wei, "Using simulated annealing to design good codes," *IEEE Trans. Inf. Theory*, vol. 33, no. 1, pp. 116–123, 1987.
[5] J. Hamkins and K. Zeger, "Asymptotically dense spherical codes. I. Wrapped spherical codes," *IEEE Trans. Inf. Theory*, vol. 43, no. 6, pp. 1774–1785, 1997.
[6] ——, "Asymptotically dense spherical codes. II. Laminated spherical codes," *IEEE Trans. Inf. Theory*, vol. 43, no. 6, pp. 1786–1798, 1997.
[7] C. Torezzan, S. I. R. Costa, and V. A. Vaishampayan, "Constructive spherical codes on layers of flat tori," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6655–6663, 2013.
[8] J. H. Conway and N. J. A. Sloane, *Sphere packings, lattices and groups*. New York, USA: Springer, 1999.
[9] S. Costa, F. Oggier, A. Campello, J.-C. Belfiore, and E. Viterbo, *Lattices applied to coding for reliable and secure communications*. Cham, Switzerland: Springer, 2017.
[10] P. Leopardi, "A partition of the unit sphere into regions of equal area and small diameter," *Electron. Trans. Numer. Anal.*, vol. 25, pp. 309–327, 2006.
[11] P. Solé and J.-C. Belfiore, "Constructive spherical codes near the Shannon bound," *Des. Codes Cryptogr.*, vol. 66, pp. 17–26, 2013.
[12] R. M. Siqueira and S. I. R. Costa, "Flat tori, lattices and bounds for commutative group codes," *Des. Codes Cryptogr.*, vol. 49, pp. 307–321, 2008.
[13] C. Alves and S. I. R. Costa, "Commutative group codes in $\mathbb{R}^4$, $\mathbb{R}^6$, $\mathbb{R}^8$ and $\mathbb{R}^{16}$—approaching the bound," *Discrete Math.*, vol. 313, no. 16, pp. 1677–1687, 2013.
[14] R. M. Taylor, L. Mili, and A. Zaghloul, "Structured spherical codes with asymptotically optimal distance distributions," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2017, pp. 2188–2192.
[15] J. Hamkins and K. Zeger, "Gaussian source coding with spherical codes," *IEEE Trans. Inf. Theory*, vol. 48, no. 11, pp. 2980–2989, 2002.
[16] F. B. Miranda and C. Torezzan, "A shape-gain approach for vector quantization based on flat tori," *Adv. in Math. Commun.*, vol. 14, no. 3, pp. 467–476, 2020.
[17] L. Zheng and D. N. C. Tse, "Communication on the Grassmann manifold: a geometric approach to the noncoherent multiple-antenna channel," *IEEE Trans. Inf. Theory*, vol. 48, no. 2, pp. 359–383, 2002.

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2021.3114094

[18] J. H. Conway, R. H. Hardin, and N. J. A. Sloane, "Packing lines, planes, etc.: packings in Grassmannian spaces," *Exp. Math.*, vol. 5, no. 2, pp. 139–159, 1996.

[19] K. Ngo, A. Decurninge, M. Guillaud, and S. Yang, "Cube-split: A structured Grassmannian constellation for non-coherent SIMO communications," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1948–1964, 2020.

[20] M. A. Sedaghat, R. R. Müller, and C. Rachinger, "(Continuous) phase modulation on the hypersphere," *IEEE Trans. Wireless Commun.*, vol. 15, no. 8, pp. 5763–5774, 2016.

[21] C. Rachinger, R. R. Müller, and J. B. Huber, "Phase shift keying on the hypersphere: Peak power-efficient MIMO communications," 2016, arXiv:1611.01009v3.

[22] E. Agrell and M. Karlsson, "Power-efficient modulation formats in coherent transmission systems," *J. Lightw. Technol.*, vol. 27, no. 22, pp. 5115–5126, 2009.

[23] M. Karlsson and E. Agrell, "Four-dimensional optimized constellations for coherent optical transmission systems," in *Proc. 36th Eur. Conf. Exhib. Opt. Commun.*, 2010, pp. 1–6.

[24] F. Rodrigues, G. Temporão, and J. P. von der Weid, "Constructive methods for the design and labeling of four-dimensional modulations," *J. Commun. Inf. Syst.*, vol. 33, no. 1, 2018.

[25] M. Karlsson and E. Agrell, "Multidimensional modulation and coding in optical transport," *J. Lightw. Technol.*, vol. 35, no. 4, pp. 876–884, 2017.

[26] M. Reimer, S. O. Gharan, A. D. Shiner, and M. O'Sullivan, "Optimized 4 and 8 dimensional modulation formats for variable capacity in optical networks," in *Proc. Opt. Fiber Commun. Conf. Exhib. (OFC)*, 2016, pp. 1–3.

[27] G. Rademacher, B. J. Puttnam, R. S. Luís, Y. Awaji, N. Wada, E. Agrell, and K. Petermann, "Experimental investigation of a 16-dimensional modulation format for long-haul multi-core fiber transmission," in *Eur. Conf. on Opt. Commun. (ECOC)*, 2015, pp. 1–3.

[28] H. K. Urbantke, "The Hopf fibration—seven times in physics," *J. Geom. Phys.*, vol. 46, no. 2, pp. 125–150, 2003.

[29] R. Mosseri, "Two-qubit and three-qubit geometry and Hopf fibrations," in *Topology in Condensed Matter*, M. I. Monastyrsky, Ed. Berlin, Heidelberg, Germany: Springer, 2006, pp. 187–203.

[30] J. C. Baez, "The octonions," *Bull. Amer. Math. Soc.*, vol. 39, no. 2, pp. 145–206, 2001.

[31] D. W. Lyons, "An elementary introduction to the Hopf fibration," *Math. Mag.*, vol. 76, no. 2, pp. 87–98, 2003.

[32] M. Nakahara, *Geometry, topology and physics*, 2nd ed. Bristol, UK: Institute of Physics Publishing, 2003.

[33] L. R. B. Naves, "Códigos esféricos em canais grampeados," Ph.D. dissertation, Univ. of Campinas, Campinas, Brazil, 2016.

[34] J. Hamkins, "Design and analysis of spherical codes," Ph.D. dissertation, Univ. of Illinois, Urbana-Champaign, USA, 1996.

[35] G. Nebe and N. J. A. Sloane. Table of densest packings presently known. [Online]. Available: http://www.math.rwth-aachen.de/~Gabriele.Nebe/LATTICES/density.html

[36] G. Lachaud and J. Stern, "Polynomial-time construction of codes. II. Spherical codes and the kissing number of spheres," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1140–1146, 1994.

[37] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*. Cambridge, USA: MIT Press, 2009.

Henrique K. Miyamoto (Student Member, IEEE) was born in Salvador, BA, Brazil. He received the B.S. degree in electrical engineering from the University of Campinas (Unicamp), Campinas, SP, Brazil, in 2021, with a two-year period at CentraleSupéc, Gif-sur-Yvette, France.

He held short-term internships at the Laboratory of Signals and Systems (L2S), CentraleSupélec, in 2019, and at the Mathematical and Algorithmic Sciences Lab, Huawei France, in 2020. He is currently pursuing the M.Sc. degree in applied mathematics at Unicamp. His research interests include information theory and coding.

Sueli I. R. Costa (Member, IEEE) received the Ph.D. degree from the University of Campinas (Unicamp), Campinas, SP, Brazil and did her postdoctoral studies at the Institute for Advanced Study, Princeton, NJ, USA.

She is currently a Professor at the Institute of Mathematics, Statistics and Scientific Computing, Unicamp. Her activities related to research development include the coordination of the thematic project Information Theory and Coding – FAPESP and invited short-term visits to the Imperial College London, London, to the Bernoulli Intertfacultaire Centre, EPFL, Lausanne, and to the AT&T Research Lab, NJ. She has co-authored the book *Lattices applied to coding for reliable and secure communications* (Springer, 2017) and more than 60 papers and book chapters. Her research topics of interest include lattice codes and applications, discrete and continuous spherical codes, coding for storage and information geometry.

Prof. Costa has served as co-chair of the 2011 IEEE Information Theory Workshop and of the 2018 Latin American Week on Coding and Information, as chair of the IEEE Information Society Brazil Chapter, and is currently a member of the IEEE Richard W. Hamming Medal committee.

Henrique N. Sá Earp was born in Rio de Janeiro, RJ, Brazil in 1981. He received the B.S. degree in mathematics from the Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, in 2001, the M.S. degree in applied mathematics from UFRJ in 2002, the CAS (Part III of the mathematical tripos) in theoretical physics from the University of Cambridge, Cambridge, UK, in 2004, and the Ph.D. degree in mathematics from Imperial College London, London, UK, in 2008. He received the *Livre-docente* associateship title from the University of Campinas (Unicamp), Campinas, SP, Brazil, in 2019.

From 2009 to 2010, he was a Postdoctoral Fellow at Unicamp. He is currently an Associate Professor of mathematics at Unicamp. His research has been concerned with gauge theory in higher dimensions, manifolds with special holonomy, nonlinear partial differential equations, as well as string/M-theory and the geometry of information. He is the coordinator of a CAPES-Cofecub Brazil-France collaboration, and a MathAmSud Argentina-Brazil-Chile-France collaboration, and he is a UK Royal Society Newton Mobility Fellow.

Prof. Sá Earp is Executive Manager of BI0S – Brazilian Institute of Data Science.

Anexo B

Cópia do Artigo [43]

A seguir, anexamos uma cópia da versão aceita do artigo apresentado no Capítulo 3; a versão final está disponível em [43].

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2022.3173002

Context-Tree-Based Lossy Compression and Its Application to CSI Representation

Henrique K. Miyamoto, *Graduate Student Member, IEEE,* and Sheng Yang, *Member, IEEE*

***Abstract*—We propose novel compression algorithms for time-varying channel state information (CSI) in wireless communications. The proposed scheme combines (lossy) vector quantisation and (lossless) compression. First, the new vector quantisation technique is based on a class of parametrised companders applied on each component of the normalised CSI vector. Our algorithm chooses a suitable compander in an intuitively simple way whenever empirical data are available. Then, the sequences of quantisation indices are compressed using a context-tree-based approach. Essentially, we update the estimate of the conditional distribution of the source at each instant and encode the current symbol with the estimated distribution. The algorithms have low complexity, are linear-time in both the spatial dimension and time duration, and can be implemented in an online fashion. We run simulations to demonstrate the effectiveness of the proposed algorithms in such scenarios.**

***Index Terms*—Data compression, MIMO systems, vector quantisation.**

I. Introduction

WIRELESS communication systems feature an ever growing dimension due to larger antenna arrays, denser network deployments, and an increasing number of terminals and devices. To maintain connectivity in such systems, a colossal amount of channel measurements, often referred to as channel state information (CSI), are necessary. Efficiently representing the CSI is crucial for storage and dissemination. A typical example is the downlink transmission from a base station (BS) with multiple antennas (potentially a large number, i.e., massive MIMO) to multiple users simultaneously. The BS should steer the signal for user j in such a way that the interference with any other user $k \neq j$ is low enough. Such beamforming techniques rely on precise CSI at the transmitter side, e.g., [2]. For the BS to acquire the CSI, however, it usually requires that each user feeds back the CSI measurements in a timely and accurate fashion. How to reduce the bandwidth cost of such feedback traffic, which is highly non-negligible, is becoming a crucial problem. This is essentially a lossy data compression problem.

CSI measurements are typically correlated in space and time according to the propagation environment, and the mobility of users and obstacles. The spatial correlation for a single antenna array is inherent to the antenna structure and can be used to reduce CSI dimension so that only a few coefficients are needed to describe the channel state. For large antenna arrays, recent works apply deep learning and compressed sensing techniques to further exploit the correlation and sparsity of channel, e.g., [3]–[5] and references therein. Independent channel coefficients (e.g., when correlation is ignored or after decorrelation) are then quantised with a vector quantiser into symbols from a finite set (codebook), e.g., [6], [7] and references therein. Further spatial compression can be achieved with entropic encoding (e.g., arithmetic coding) on the bit representation of the quantisation indices [5].

The temporal correlation of CSI measurements, on the other hand, is less exploited for feedback. Indeed, the sequence of quantised symbols, considered as a random process, can be losslessly compressed to a bit-stream. If the sequence is stationary, then the bit rate can theoretically be as low as the entropy rate of the underlying process. A possible approach towards CSI compression is therefore to directly apply any universal compression algorithm [8]–[10], such as Lempel-Ziv [11], [12] (known as LZ77 and LZ78) to the quantisation indices.

Another universal compressor is the *context-tree weighting* (CTW) algorithm [13], which learns the distribution of a given sequence in an efficient way. The learned distribution can then be used as the coding distribution to compress the sequence in combination with arithmetic coding. It has been shown that, in this case, Rissanen lower bound [13] is achieved, in the sense of having optimal rate of convergence to the entropy for tree sources with unknown parameters. Extensions of the algorithm can be found in [14]–[18]. A modification of CTW based on the minimum-description principle yields the *context-tree maximising* (CTM) algorithm [19], which can produce maximum *a posteriori* (MAP) probability tree models [20]. Connections between CTW/CTM algorithms and Bayesian inference have been explored in [8], [21], [22]. In particular, in [22], the authors extended the CTM algorithm to find the k *a posteriori* most likely models, under the name of *Bayesian context-tree*, and generalised some results.

Directly applying these algorithms to compress quantisation indices presents, nonetheless, some difficulties. First, the output bit-stream is of variable length, making the feedback difficult to implement. Second, in Lempel-Ziv methods, the input symbol block is also of variable length, since it depends on parsing the original sequence. This means that the encoder may need to wait for an indefinite number of time slots to output an indefinite number of bits for feedback. Finally,

An earlier version of this paper was presented in part at the International Zurich Seminar on Information and Communication (IZS 2022) [1].

H. K. Miyamoto was with with the Laboratory of Signals and Systems (L2S), CentraleSupélec, Paris-Saclay University, 91190 Gif-sur-Yvette, France. He is now with the Institute of Mathematics, Statistics and Scientific Computing (IMECC), University of Campinas (Unicamp), Campinas, 13083-859, Brazil (email: hmiyamoto@ime.unicamp.br).

S. Yang is with the Laboratory of Signals and Systems (L2S), CentraleSupélec, Paris-Saclay University, 91190 Gif-sur-Yvette, France (email: sheng.yang@centralesupelec.fr).

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2022.3173002

arithmetic coding assumes that computations are carried out with infinite precision, while, in practice, it has to be carefully implemented so as to deal with finite precision constraints, e.g., [23], [24]. Trying to avoid such difficulties motivates us to propose new compression algorithms adapted to communication scenarios.

In this work, we focus on the problem of online lossy compression of a sequence of CSI vectors, for which we propose a two-step solution. The first step is lossy: we normalise the CSI vector and quantise the amplitude and phase components separately using a data-adapted compander, followed by uniform quantiser. In particular, we consider the widely used μ-compander and a new one called β-compander, inspired by the beta distribution. The second step is lossless: we compress the sequences of quantisation indices with the coding distribution estimated via a context-tree method. Two solutions can be considered: 1) to directly use CTW with arithmetic coding, or 2) to apply CTM to estimate the conditional distribution of the upcoming symbol at each time instant, and use this probability to compress the symbol. In the second case, we encode each symbol with a fixed number of levels to limit the fluctuation of the encoded bits flow, which is a desirable property in communication systems.

An important difference from previous works using deep learning techniques [3]–[5] is that they consider almost static channels, whereas our work investigates low, medium and high mobility scenarios. In addition, our scheme requires much less training samples: in fact, it can even be initialised without any training data, and training can be done online, as the sequence of coefficients is observed.

Our algorithms are linear-time in both the spatial dimension and time duration, and can be implemented in an online fashion. Although we propose the two steps as an ensemble, they are actually modular. This means that the new quantiser design and the new compression algorithms can be used independently, and combined with other existing quantisation or compression methods, if desired. Implementation codes are available in [25].

The remainder of the paper is organised as follows. In Section II we introduce the system model and review basic concepts of vector quantisation and context-tree representation. Our quantiser design is described in Section III, while the compression algorithm is presented in Section IV. Numerical simulations of CSI acquisition are analysed in Section V. Finally, we draw some conclusions in Section VI.

Notation: Throughout this paper, we use the following notational conventions. Vectors are denoted by bold italic lowercase (e.g., $\boldsymbol{v}$), and their L_2-norm is denoted by $\|\boldsymbol{v}\|$. Random variables are denoted by non-italic upper case letters (e.g., X). A binary string is denoted by bold non-italic lower case (e.g, $\mathbf{c}$). Logarithms are to the base 2. We denote $[n] \coloneqq \{1, \dots, n\}$. The indicator function $\mathbb{1}_{\{P\}}$ takes value 1 if the argument P is a true statement, and 0 otherwise.

II. Problem Formulation and Preliminaries

A. Main Problem

Let us consider a network composed of a transmitter (e.g., base station) and N_r receivers (e.g., mobile users). Assume that the *channel state information* (CSI) between the transmitter and receiver k at time t can be described by a complex vector $\boldsymbol{h}_k[t] \in \mathbb{C}^{N_\mathrm{t}\times 1}$, for $k \in [N_\mathrm{r}]$. The dimension N_t of the vector can depend on the number of antennas and subcarriers used by the transmitter. For different purposes, such as feedback and storage, each receiver is required to represent its state sequence in an 'economical' way, i.e., to use as few bits as possible to describe the sequence, for a given distortion constraint. This is known as the lossy source coding problem, and the fundamental trade-off between the rate of the encoded sequence and the distortion for a stationary process is known for a given distribution [9].

In this work, we are interested in compression algorithms for a source with unknown distribution that can be implemented in an online fashion, i.e., each $\boldsymbol{h}_k[t]$ can be successively compressed, while exploiting the time correlation of the sequence. In most practical scenarios, the norm of the vectors $\boldsymbol{h}_k$ (i.e., the strength) is less important than the relative strength of the components (i.e., the direction). Therefore, our goal is to compress the normalised vector $\boldsymbol{h}_k[t]/\|\boldsymbol{h}_k[t]\|$. Before presenting our scheme, we recall some basic notions of vector quantisation, lossless compression and context-tree representation.

B. Vector Quantisation

A *vector quantiser* [26] of dimension p and size M, is a mapping $q : \mathbb{R}^p \to \mathcal{C} \coloneqq \{\boldsymbol{y}_0, \boldsymbol{y}_1, \dots, \boldsymbol{y}_{M-1}\} \subset \mathbb{R}^p$ that assigns each vector $\boldsymbol{x} \in \mathbb{R}^p$ to a codeword $\hat{\boldsymbol{x}} \coloneqq q(\boldsymbol{x}) = \boldsymbol{y}_k$, for some $k \in \{0, 1, \dots, M-1\}$. To a sequence of vector symbols $\boldsymbol{x}_1^n \coloneqq \boldsymbol{x}_1\boldsymbol{x}_2\cdots\boldsymbol{x}_n$ we can apply vector-by-vector quantisation. In this case, the vector quantiser outputs a sequence of quantised vectors $\hat{\boldsymbol{x}}_1^n \coloneqq \hat{\boldsymbol{x}}_1\hat{\boldsymbol{x}}_2\cdots\hat{\boldsymbol{x}}_n$ and a sequence of quantisation indices $k_1^n \coloneqq k_1k_2\cdots k_n$, where $\hat{\boldsymbol{x}}_i = \boldsymbol{y}_{k_i}$, for each $i \in [n]$.

Two important parameters to assess the performance of a vector quantiser are the quantisation rate and the mean distortion. The *quantisation rate*, defined as $R \coloneqq (\log_2 M)/p$, is an indicator of the cost to describe the vector, while the *mean distortion* measures the error induced by the quantisation. A commonly used distortion measure is the mean squared chordal distance (MSCD). Specifically, the MSCD between the original vector $\boldsymbol{x}$ and the quantised vector $\hat{\boldsymbol{x}}$ is defined as

$$\mathrm{MSCD}(\boldsymbol{x}, \hat{\boldsymbol{x}}) \coloneqq 1 - \mathbb{E}\left[\frac{|\langle \boldsymbol{x}, \hat{\boldsymbol{x}}\rangle|^2}{\|\boldsymbol{x}\|^2\|\hat{\boldsymbol{x}}\|^2}\right]. \tag{1}$$

Note that the MSCD is invariant to scalar rotations, what is adapted to our application of CSI representation.

C. Lossless Compression and Universality

Consider a sequence of symbols (e.g., quantisation indices) from an m-ary discrete alphabet $\mathcal{A} = \{0, \dots, m-1\}$. It is well known that, in the context of source coding, if the distribution P of a source is known, Shannon's code can be used to generate a codeword with length $\lceil -\log P(x_1^n)\rceil$ for any source sequence x_1^n, where $P(x_1^n)$ is the probability of the realisation x_1^n. The expected length of such a code is within 1 bit of the entropy lower bound $H(\mathrm{X}_1^n) \coloneqq \mathbb{E}[-\log P(\mathrm{X}_1^n)]$.

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2022.3173002

Therefore, the coding rate, as the number of encoded bits per input symbol, can be arbitrarily close to $\frac{1}{n}H(\mathrm{X}_1^n)$.

If, however, P is not known and another probability distribution (also called coding distribution) Q_n is used instead, the codeword length becomes $\lceil -\log Q_n(x_1^n) \rceil$, incurring a *redundancy*

$$\begin{aligned} R(P, Q_n) &\coloneqq \mathbb{E}_P[-\log Q_n(\mathrm{X}_1^n)] - \mathbb{E}_P[-\log P(\mathrm{X}_1^n)] \\ &= D(P \,\|\, Q_n), \end{aligned} \tag{2}$$

which coincides with the Kullback-Leibler divergence $D(P \,\|\, Q_n)$ between P and Q_n. Without the knowledge of P, it is desirable to have low redundancy for every distribution in a given class of distributions. A coding distribution Q_n (and the corresponding code) is said to be *(weakly) universal* [10] for a class $\mathcal{P}$ of processes if $\frac{1}{n}R(P, Q_n) \to 0,\ \forall P \in \mathcal{P}$. For instance, both the Lempel-Ziv codes and the CTW algorithm with arithmetic coding are universal for the class of stationary ergodic sources [8], [9], [13].

D. Variable-Order Markov Chain and Context-Tree Representation

Let us denote $x_i^j \coloneqq x_i x_{i+1} \cdots x_j$ a scalar sequence over an m-ary alphabet $\mathcal{A} = \{0, 1, \ldots, m-1\}$, generated by a source with probability distribution P. We denote $l(x_i^j) \coloneqq j - i + 1$ the length of sequence x_i^j. A *variable-order Markov chain* with order or memory D (also called *bounded memory tree source*) is a random process for which the probability of a new symbol, given the whole past, only depends on the last D symbols, i.e., $P(x_i|x_{-\infty}^{i-1}) = P(x_i|x_{i-D}^{i-1})$. The main reason for our interest in Markov chains here is that any stationary ergodic source can be approximated by a Markov chain with sufficiently large order. Specifically, the entropy rate of a D-th order Markov chain approximation of a stationary ergodic process becomes arbitrarily close to that of the original process when $D \to \infty$ [8], [9]. In many practical cases, a small D is enough to describe a given process.

The statistical behaviour of a variable-order Markov chain can be described by a *context set* $\mathcal{S}$ (also known as *suffix set* or *model*), which is a subset of $\bigcup_{i=0}^{D} \mathcal{A}^i$ that is proper (no element in $\mathcal{S}$ is a proper suffix of any other) and complete (each $x_{-\infty}^n$ has a suffix in $\mathcal{S}$, which is unique by properness). The *context function* $c : \mathcal{A}^D \to \mathcal{S}$ maps each context x_{i-D}^{i-1} with length D to a suffix $c(x_{-\infty}^{i-1}) = c(x_{i-D}^{i-1}) = x_{i-j}^{i-1},\ j \le D$. Furthermore, each suffix $s \in \mathcal{S}$ is associated with a parameter $\boldsymbol{\theta}_s \coloneqq \big(\theta_s(0), \theta_s(1), \ldots, \theta_s(m-1)\big)$, where $\theta_s(j) \coloneqq P(j|s)$. The *parameter vector* $\Theta \coloneqq \Theta_{\mathcal{S}} \coloneqq \{\boldsymbol{\theta}_s : s \in \mathcal{S}\}$ groups all parameters in context set $\mathcal{S}$. Therefore, the Markov chain is completely characterised by the couple $(\mathcal{S}, \Theta_{\mathcal{S}})$. We use $\mathcal{C}_D$ to denote the class of all context sets of order up to D, and we define $L_D(\mathcal{S}) \coloneqq |\{s \in \mathcal{S} :\ l(s) = D\}|$ the number of contexts with length D.

Since the context set $\mathcal{S}$ is proper, its elements can be represented as leaf nodes of a tree $\mathcal{T}_D \supseteq \mathcal{S}$, called *context-tree*. For a given sequence x_1^n, each leaf node $s \in \mathcal{S}$ is associated with a *counter* $\boldsymbol{a}_s \coloneqq \boldsymbol{a}_s(x_1^n) \coloneqq \big(a_s(0), a_s(1), \ldots, a_s(m-1)\big)$, where $a_s(j)$ stores the number of times that symbol $j \in \mathcal{A}$ follows context s in x_1^n. The counter of each inner node of the tree is recursively defined as the sum of the counters of its children nodes, i.e., $\boldsymbol{a}_s \coloneqq \sum_{j \in \mathcal{A}} \boldsymbol{a}_{js},\ \forall\, s \in \mathcal{T}_D \setminus \mathcal{S}$. We use the empty string λ to denote the root of the tree. An illustrative example of these concepts is given in Fig. 1.

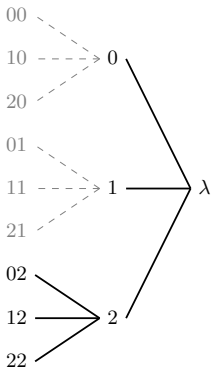


Fig. 1: Example of model $\mathcal{S} = \{0, 1, 02, 12, 22\} \subseteq \mathcal{T}_D$ with $m = 3$ and $D = 2$. In this case, we have $L_D(\mathcal{S}) = |\{02, 12, 22\}| = 3$. The suffix of the sequence $x_{-\infty}^0 = \cdots 01$ is $c(x_{-\infty}^0) = c(01) = 1$. The (conditional) probability of the string $x_1^5 = 02212$ given the past symbols $x_{-1}^0 = 01$ is $P(x_1^5|x_{-1}^0) = \theta_1(0) \cdot \theta_0(2) \cdot \theta_{02}(2) \cdot \theta_{22}(1) \cdot \theta_1(2)$. After processing this sequence, the counter for context $s = 1$ is $\boldsymbol{a}_1(x_1^5) = (a_1(0), a_1(1), a_1(2)) = (1, 0, 1)$.

With the above definitions and the Markov property for a D-th order Markov chain, if both $\mathcal{S}$ and $\Theta_{\mathcal{S}}$ are known, the probability of a sequence can be written, as in [22],

$$\begin{aligned} P(x_1^n|x_{D-1}^0, \mathcal{S}, \Theta_{\mathcal{S}}) &= \prod_{i=1}^{n} P(x_i|x_{i-D}^{i-1}, \mathcal{S}, \Theta_{\mathcal{S}}) \\ &= \prod_{i=1}^{n} P(x_i|c(x_{i-D}^{i-1}), \mathcal{S}, \Theta_{\mathcal{S}}) \\ &= \prod_{i=1}^{n} \theta_{c(x_{i-D}^{i-1})}(x_i) = \prod_{s \in \mathcal{S}} \prod_{j \in \mathcal{A}} \theta_s(j)^{a_s(j)}, \end{aligned}$$

as a simple function of the parameters $\mathcal{S}$, $\Theta_{\mathcal{S}}$, as well as the counters of the leaf nodes.

If only the model $\mathcal{S}$ is known, but not its parameters $\Theta_{\mathcal{S}}$, the *marginal distribution* of a sequence x_1^n, given its past x_{1-D}^0 and model $\mathcal{S}$, is

$$P(x_1^n|x_{1-D}^0, \mathcal{S}) = \int P(x_1^n|x_{1-D}^0, \mathcal{S}, \Theta)\pi(\Theta|\mathcal{S})\ \mathrm{d}\Theta, \tag{3}$$

assuming the distribution of the parameters $\pi(\Theta|\mathcal{S})$ is known. While this distribution is unknown in general, using the so-called *Jeffrey's prior* is asymptotically optimal in the minimax sense [8]. This choice corresponds to setting $\pi(\Theta|\mathcal{S})$ to be the Dirichlet distribution with parameters $\big(\frac{1}{2}, \cdots, \frac{1}{2}\big)$. It turns out that, under this assumption, the marginal distribution (3) can be simplified to the so-called *Krichevsky–Trofimov (KT) distribution*, which can be easily computed as

$$P(x_1^n|x_{1-D}^0, \mathcal{S}) = \prod_{s \in \mathcal{S}} P_e(\boldsymbol{a}_s), \tag{4}$$

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2022.3173002

where

$$P_e(\boldsymbol{a}_s) = \frac{\prod_{j\in\mathcal{A}} \left(\frac{1}{2}\right)\left(\frac{3}{2}\right)\cdots\left(\boldsymbol{a}_s(j)-\frac{1}{2}\right)}{\left(\frac{m}{2}\right)\left(\frac{m}{2}+1\right)\cdots\left(\frac{m}{2}+M_s-1\right)}, \quad s\in\mathcal{T}_D, \quad (5)$$

with $M_s \coloneqq \sum_{j=0}^{m-1} \boldsymbol{a}_s(j)$. We note that other prior distributions $\pi(\Theta|\mathcal{S})$ have been considered in the literature and lead to different marginal distributions [27], [28].

Finally, if the model $\mathcal{S}$ is also unknown, then we shall marginalise over $\mathcal{S}$ with a given prior distribution π_D on all models $\mathcal{S}$ of maximal depth D. Fixing $\gamma \in]0,1[$ and

$$\pi_D(\mathcal{S}) \coloneqq (1-\gamma)^{\frac{|\mathcal{S}|-1}{m-1}} \gamma^{|\mathcal{S}|-L_D(\mathcal{S})}, \quad (6)$$

we obtain a mixture of different distributions (4), corresponding to the coding distribution of CTW [8], [22]:

$$Q_n(x_1^n|x_{1-D}^0) \coloneqq \sum_{\mathcal{S}\in\mathcal{C}_D} \pi_D(\mathcal{S}) \prod_{s\in\mathcal{S}} P_e(\boldsymbol{a}_s). \quad (7)$$

Not only is this coding distribution universal for the class of stationary ergodic sources, but also it can be recursively computed so that complexity is linear in n. The essence of the context-tree weighting (CTW) algorithm is based on the following definitions and results.

Definition 1. For $\gamma \in]0,1[$, to each node $s \in \mathcal{T}_D$, with $l(s) = d$, we assign a *weighted probability* P_w^s, defined as

$$P_w^s \coloneqq \begin{cases} \gamma P_e(\boldsymbol{a}_s) + (1-\gamma)\prod_{j=0}^{m-1} P_w^{js}, & 0 \le d < D, \\ P_e(\boldsymbol{a}_s), & l(s) = D. \end{cases} \quad (9)$$

The context-tree together with the weighted probabilities of the nodes is called *weighted context-tree*.

Lemma 1 (See [13], [22])**.** *The weighted probability P_w^λ of the root node $\lambda \in \mathcal{T}_D$ satisfies*

$$P_w^\lambda = \sum_{\mathcal{S}\in\mathcal{C}_D} \pi_D(\mathcal{S}) \prod_{s\in\mathcal{S}} P_e(\boldsymbol{a}_s). \quad (10)$$

This lemma shows that the CTW probability $Q_n(x_1^n|x_{1-D}^0)$ is indeed the weighted probability P_w^λ of the root node λ. Therefore, to compute the CTW probability of x_1^n, the CTW algorithm updates P_w^s sequentially on $x_1, \ldots, x_1^i, \ldots, x_1^n$. Details are omitted and can be found in [13].

A modification of the CTW algorithm yields the context-tree maximising (CTM) algorithm [19], which computes the maximum *a posteriori* model for a given sequence.

Definition 2. For $\gamma \in]0,1[$, to each node $s \in \mathcal{T}_D$, with $l(s) = d$, we assign a *maximised probability* P_m^s, defined as

$$P_m^s \coloneqq \begin{cases} \max\{\gamma P_e(\boldsymbol{a}_s), (1-\gamma)\prod_{j=0}^{m-1} P_m^{js}\}, & 0 \le l(s) < D \\ P_e(\boldsymbol{a}_s), & l(s) = D. \end{cases} \quad (11)$$

The *maximising set* $\mathcal{S}_m^s$ is computed as shown in (12). The context-tree, together with the maximised probability distribution and the maximising sets, is called *maximised context-tree*.

Lemma 2 (See [19], [22])**.** *The maximised coding distribution P_m^s of the root node $\lambda \in \mathcal{T}_D$ satisfies*

$$P_m^\lambda = \pi_D(\mathcal{S}_m^\lambda) \prod_{s\in\mathcal{S}_m^\lambda} P_e(\boldsymbol{a}_s) = \max_{\mathcal{S}\in\mathcal{C}_D} \pi_D(\mathcal{S}) \prod_{s\in\mathcal{S}} P_e(\boldsymbol{a}_s). \quad (13)$$

It follows that the maximising set $\mathcal{S}_m^\lambda$, which is associated to the maximising probability P_m^λ, corresponds to the maximum *a posteriori* model:

$$\begin{aligned} \mathcal{S}_m^\lambda &= \arg\max_{\mathcal{S}\in\mathcal{C}_D} P(\mathcal{S}|x) = \arg\max_{\mathcal{S}\in\mathcal{C}_D} \frac{\pi_D(\mathcal{S})P(x|\mathcal{S})}{P(x)} \\ &= \arg\max_{\mathcal{S}\in\mathcal{C}_D} \pi_D(\mathcal{S}) \prod_{s\in\mathcal{S}} P_e(\boldsymbol{a}_s). \end{aligned}$$

Proof of Lemmas 1 and 2 were given, for the special case $m = 2, \gamma = 1/2$, in [13] and [19], respectively, while the general case is addressed in [22].

III. Quantiser Design

The proposed vector quantisation consists in vector normalisation, decomposition into real components, and individual scalar quantisation based on parametric companders, as indicated in Fig. 2. In the following, we elaborate each step.

A. Vector Normalisation

In this step, the input vector $\boldsymbol{x} \coloneqq (x(1), \ldots, x(N_\mathrm{t}))$ is normalised by the component with the largest absolute value, i.e., $\bar{\boldsymbol{x}} = \boldsymbol{x}/x(i^*)$ where $i^* \coloneqq \arg\max_{i\in[N_\mathrm{t}]} |x(i)|$. Note that $\bar{x}(i^*) = 1$, while the other normalised components are complex in general with absolute value in $[0,1]$. The i^*-th component can skip the following steps and be directly assigned a special index indicating it as the strongest component. Because of this special symbol, our compression alphabet sizes m have one more element than the number of quantisation levels M, i.e., $m = M + 1$.

B. Decomposition

Before scalar quantisation, each complex component has to be decomposed into real values. Two straightforward options are: 1) Cartesian decomposition into real real and imaginary parts, and 2) polar decomposition into amplitude and phase. We consider the polar decomposition, since the amplitude and phase are usually less correlated in wireless applications, therefore providing a less 'redundant' representation of the bounded complex number. Indeed, the real and imaginary parts of the normalised complex components tend to be correlated, e.g., strong real part implies weak imaginary part since the amplitude is bounded by 1.

$$\mathcal{S}_m^s \coloneqq \begin{cases} \bigcup_{j=0}^{m-1} \mathcal{S}_m^{js} \times \{j\}, & \text{if } (1-\gamma)\prod_{j=0}^{m-1} P_m^{js} > P_e(\boldsymbol{a}_s) \text{ and } 0 \le d < D, \\ \{\lambda\}, & \text{if } (1-\gamma)\prod_{j=0}^{m-1} P_m^{js} \le P_e(\boldsymbol{a}_s) \text{ and } 0 \le d < D, \\ \{\lambda\}, & \text{if } d = D. \end{cases} \quad (12)$$

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2022.3173002

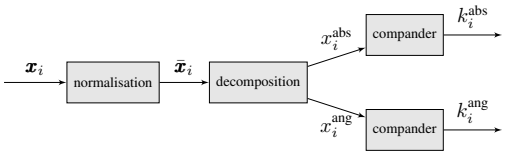


Fig. 2: Block diagram for vector quantisation.

C. Quantisation with Parametric Companders

The amplitude and phase are separately quantised with different scalar quantisers of M_{abs} and M_{ang} quantisation levels, respectively.

If the input symbols of the quantiser are uniformly distributed, then a uniform quantiser is optimal. In general, however, uniform quantisation can be far from optimal in the rate-distortion sense [9]. Let X be a random variable representing the input, following some distribution P with support $[0, 1]$. The idea of using a compander [29] is to apply a non-linear and non-decreasing mapping $g : [0, 1] \to [0, 1]$ to the signal (*compression*) before quantising it, so that the signal is more 'uniform' in the image space. To recover the signal, the inverse mapping $g^{-1} : [0, 1] \to [0, 1]$ is used (*expansion*); hence the name *compander*. It is practical to use parametric companders, i.e., differentiable mappings g that can be described by a few number of parameters, such as the μ-law compander, characterised by a single parameter $\mu > 0$. Note that, as compared to the Lloyd quantiser [9], compander-based quantisers have much lower complexity of quantisation and representation.

We propose a data-driven design for companders parametrised by some ξ (which can contain multiple scalar parameters). Assume that we have a set of training data $\{x_1, \ldots, x_n\}$. Our design follows a two-step procedure: 1) uniformisation of the data, and 2) adjustment of the compander parameters, as follows.

1) Uniformisation of the Data: We assume that the training data are independent samples from some distribution P. If we knew the cumulative distribution function (cdf) F_P of P, we could apply the mapping F_P such that $\{F_P(x_1), \ldots, F_P(x_n)\}$ are samples from a uniform distribution. If, however, we are restricted to a class of companders $\{g_\xi \colon \xi \in \Xi\}$, for some set Ξ, then we have to approximate F_P by g_ξ. Since a compander, as defined above, is non-decreasing from 0 to 1, it is equivalent to a cdf. Thus, a sensible criterion for the approximation is through the Kullback-Leibler divergence:

$$\begin{aligned}\xi^* &= \arg\min_{\xi\in\Xi} D(P \,\|\, g_\xi) \\ &= \arg\min_{\xi\in\Xi} \left\{-H(\mathrm{X}) - \mathbb{E}_P[\log(g'_\xi(\mathrm{X}))]\right\} \\ &= \arg\max_{\xi\in\Xi} \mathbb{E}_P[\log(g'_\xi(\mathrm{X}))]. \end{aligned} \tag{14}$$

Interestingly, this is equivalent to maximising the differential entropy of $g_\xi(\mathrm{X})$. As the uniform distribution maximises differential entropy among all bounded support distributions [9], the criterion (14) indeed returns the best 'uniformiser'. Note that since g_ξ is a cdf, g'_ξ is the corresponding probability density function (pdf).

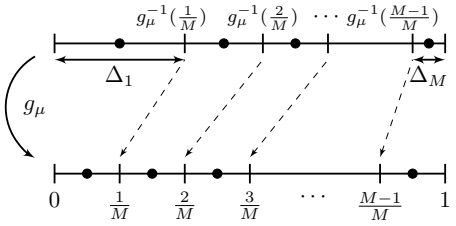


Fig. 3: μ-law compander quantiser of size M.

The true distribution of the data is, however, unknown in most practical scenarios. But we can adapt the probabilistic criterion (14) into a data-driven one by replacing the expectation with the sample mean:

$$\arg\max_{\xi\in\Xi} \frac{1}{n}\sum_{i=1}^{n} \log(g'_\xi(x_i)). \tag{15}$$

In this paper, we consider the μ-law compander and another one that we call β-law compander, as shown in Table I. The β-law compander is equivalent to the beta cdf, parametrised by $\alpha > 0$ and $\beta > 0$. An attractive feature of the β-law compander is that the corresponding pdf is log-concave in (α, β) [30, Theorem 6], so that the maximisation (15) is concave and can thus be easily solved.

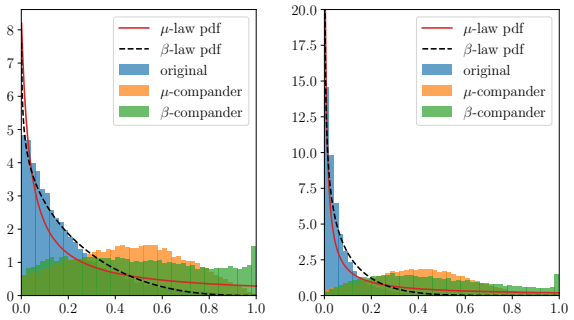
2) Adjustment of the Compander Parameters: Note that uniformising the input is not enough in the sense of rate-distortion. If we perform uniform quantisation right after this step, then the uniformisation only makes the number of samples in each quantisation interval as uniform as possible. If the number of samples are exactly the same in each interval, the average distortion is dominated by the one generated in the largest interval—we illustrate this argument in Fig. 3 with a μ-law compander. Therefore, we need to adjust the parameter to balance the distortion generated in different intervals, which is the role of the second step. While the exact solution is hard to find, we provide a heuristic, yet efficient way to make the adjustment.

Consider a quantiser with M levels. If we assume that the distortion generated in the i-th interval is proportional to the squared length Δ_i^2 of that interval, then the average distortion is proportional to $\sum_{i=0}^{M-1} N_i\Delta_i^2$, where N_i is the number of samples inside the i-th interval. Here, each interval i contributes with $N_i\Delta_i^2$. Starting with the solution given by step 1, all N_i's are similar, and the largest interval contributes the most in the average distortion; similarly, the smallest interval contributes the least. The idea is therefore to reduce the largest interval until $N_S\Delta_S^2 \geq N_L\Delta_L^2$, where 'S' and 'L' stand for the 'smallest' and 'largest' intervals, respectively. For instance, with the μ-law compander, the smallest interval is always the first one and the largest, the last one. Reducing μ towards 0 would reduce the gap between the extreme interval sizes. With the β-law compander, the smallest and largest intervals depend on the parameters (α, β), but letting (α, β) go towards $(1, 1)$ also reduces the gap between the extreme interval sizes. In Fig. 4, we plot the histogram of some data from CSI vectors and the output of the two companders. We

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2022.3173002

TABLE I: Some Compander Functions.

Compander type	Parameters ξ	cdf $g_\xi(x)$	pdf $g'_\xi(x)$
μ-law	$\mu > 0$	$\frac{\ln(1+\mu x)}{\ln(1+\mu)}$	$\frac{\mu}{(1+\mu x)\ln(1+\mu)}$
β-law	$\alpha > 0,\ \beta > 0$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^x t^{\alpha-1}(1-t)^{\beta-1}\,\mathrm{d}t$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$



(a) Amplitude. (b) Phase.

Fig. 4: Normalised histograms from CSI vector sequences (EVA70, high correlation, cf. Section V) before and after applying μ-law and β-law companders.

see that the histogram is indeed flattened after applying the companders.

Although the presented compander design is based on training data, we could also start with a uniform compander and update it regularly when more data is available. A great advantage of the parametric compander design is the negligible communication overhead of the (few) quantisation parameters.

D. Quantisation Resolutions for Amplitude and Phase

Since we quantise the amplitude and phase of a complex symbol separately, we would like to find out the 'optimal' resolutions for both quantisers. Let M_{abs} and M_{ang} be the respective number of quantisation levels, then $M = M_{\text{abs}} M_{\text{ang}}$ is the total number of quantisation levels for a complex number. The question is thus to find out the optimal values, M^*_{abs} and M^*_{ang}, for a given M. While the exact solution would depend on the distribution of the complex number, we are interested here in finding a rule of thumb based on sensible assumptions. We can show that the optimal solution is such that the respective quantisation errors satisfy $\epsilon_{\text{abs}} \approx \mathbb{E}\left[|\mathrm{X}|^2\right] \epsilon_{\text{ang}}$, where X is the complex input.

The problem is formulated as an optimisation of the MSE of the complex variable $\mathrm{X} = Ae^{\mathrm{j}\Phi}$, i.e., to minimise $\epsilon := \mathbb{E}\big[|Ae^{\mathrm{j}\Phi} - \hat{A}e^{\mathrm{j}\hat{\Phi}}|^2\big]$, where $A \in [0, \infty[$ and $\Phi \in [0, 2\pi[$. Letting $\tilde{A} := A - \hat{A}$ and $\tilde{\Phi} := \Phi - \hat{\Phi}$, we have $\epsilon = \mathbb{E}\big[\tilde{A}^2\big] + 2\mathbb{E}\big[A\hat{A}(1 - \cos(\tilde{\Phi}))\big]$. Let us reasonably assume that 1) $\mathbb{E}\big[\tilde{A}\big] = 0$; 2) $\hat{A}$ and $\tilde{A}$ are uncorrelated; and 3) $A\hat{A}$ and $1 - \cos(\tilde{\Phi})$ are uncorrelated. Then, using the approximation $1 - \cos(x) \approx x^2/2$, we have

$$\epsilon \approx \mathbb{E}\big[\tilde{A}^2\big] + \mathbb{E}\big[\hat{A}^2\big]\mathbb{E}\big[\tilde{\Phi}^2\big] = \epsilon_{\text{abs}} + (E_{\text{abs}} - \epsilon_{\text{abs}})\epsilon_{\text{ang}}, \quad (16)$$

where $E_{\text{abs}} := \mathbb{E}\big[A^2\big] = \mathbb{E}\big[|\mathrm{X}|^2\big]$, $\epsilon_{\text{abs}} := \mathbb{E}\big[\tilde{A}^2\big]$, and $\epsilon_{\text{ang}} := \mathbb{E}\big[\tilde{\Phi}^2\big]$. Since the regime of interest is such that $\epsilon_{\text{ang}} \leq 1$, the approximation (16) is an increasing function of ϵ_{abs} and of ϵ_{ang}. Intuitively, the overall quantisation error is increasing with the individual quantisation errors.

To finally find the optimal M^*_{abs} and M^*_{ang} that minimise (16), we make the following 'mild' assumption: both ϵ_{abs} and ϵ_{ang} decrease with M_{abs} and M_{ang} as

$$\epsilon_{\text{abs}}/E_{\text{abs}} \approx c_{\text{abs}} M_{\text{abs}}^{-2}, \quad \epsilon_{\text{ang}}/E_{\text{ang}} \approx c_{\text{ang}} M_{\text{ang}}^{-2}, \quad (17)$$

where $E_{\text{ang}} := \mathbb{E}\left[\Phi^2\right]$, and $c_{\text{abs}}, c_{\text{ang}}$ are constants depending on the respective marginal law of the normalised amplitude and phase. This assumption is supported by the rate-distortion theorem [9] and is in general observed for medium to high rate quantisers. Plugging in the constraint $M_{\text{abs}} M_{\text{ang}} = M$, relaxed from integers to all positive numbers, we obtain the constraint on ϵ_{abs} and ϵ_{ang}

$$\epsilon_{\text{abs}} \epsilon_{\text{ang}} \approx c_{\text{abs}} c_{\text{ang}} M^{-2}, \quad (18)$$

i.e., the product of ϵ_{abs} and ϵ_{ang} is a constant. Under this constraint, minimising (16) is equivalent to minimising $\epsilon_{\text{abs}} + E_{\text{abs}}\epsilon_{\text{ang}}$ subject to $\epsilon_{\text{abs}}\epsilon_{\text{ang}} = \text{const}$. It can be readily shown that the optimal solution is such that $\epsilon^*_{\text{abs}} = E_{\text{abs}}\epsilon^*_{\text{ang}}$ for any constant. From (17), we see that

$$M^*_{\text{ang}} \approx M^*_{\text{abs}} \sqrt{E_{\text{ang}} c_{\text{ang}} / c_{\text{abs}}}. \quad (19)$$

For example, if the normalised amplitude $A/\sqrt{E_{\text{abs}}}$ and phase $\Phi/\sqrt{E_{\text{ang}}}$ have the same distribution (with bounded support), i.e., $c_{\text{abs}} = c_{\text{ang}}$, then $M^*_{\text{ang}} \approx M^*_{\text{abs}} \sqrt{E_{\text{ang}}}$. If, in addition, they are both uniformly distributed, then $E_{\text{ang}} = \frac{4}{3}\pi^2$, and $\sqrt{E_{\text{ang}}} \approx 3.6$. Therefore, a rule of thumb for the number of quantisation bits is to use two more bits for the phase than for the amplitude.

Remark 1. It is well known that, followed by entropic encoding, a uniform quantiser is asymptotically optimal in the high-rate regime. We emphasise, however, that we do not operate here in the high-rate regime, unlike many other applications. More importantly, a large alphabet size would make the following context-tree-based compression highly inefficient. Hence, a carefully designed quantiser is crucial for the overall performance.

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2022.3173002

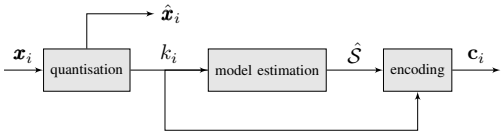


Fig. 5: Block diagram for vector processing.

IV. Compression Algorithm

In order to compress the sequence of quantisation indices, one option is to directly apply CTW algorithm with arithmetic coding. In this section, we describe an alternative context-tree-based solution that limits the fluctuation of the output bit-stream. It consists in first estimating a tree model $\hat{\mathcal{S}}$, and then using the probabilities derived from that model to encode each symbol, as depicted in Fig. 5.

A. Tree Model Estimation

Given a scalar sequence k_1^n of quantisation indices from an alphabet $\mathcal{A} = \{0, 1, \dots, m-1\}$, we use the CTM algorithm (cf. Section II-D) to find the maximum *a posteriori* tree model $\hat{\mathcal{S}}$ that describes that sequence. This algorithm consists in building the same tree $\mathcal{T}_D$ as in CTW algorithm, followed by a pruning procedure, as described by (12). Both the computational and storage complexity of CTM algorithm are known to be $O(nmD)$, i.e., linear with sequence length n, alphabet size m and maximum tree depth D, cf. [22].

When training data are available, we can apply the CTM algorithm on the training data to estimate the MAP model $\hat{\mathcal{S}}$, and use it to estimate symbol probabilities and encode the incoming sequence. This, however, is not mandatory: we can also initialise the full tree $\mathcal{T}_D$ with empty counts, keep updating them with incoming data, and regularly prune a copy of this tree to have an updated estimate of the MAP model $\hat{\mathcal{S}}$. Similarly, if the sequence is not stationary, we can also make the model forget symbols it has seen in a distant past by decreasing the counts of the tree model $\mathcal{T}_D$. With this, at each instant, the model is built upon the observation of a sequence of only the most recent symbols. This can be done without increasing the complexity of the algorithm.

B. Prediction and Encoding

Once a tree model $\hat{\mathcal{S}}$ is estimated, we can encode a sequence k_1^n according to the probabilities issued from that model. Note that, given a model $\hat{\mathcal{S}}$ and past symbols k_{1-D}^0, the estimated probability of k_1^n can be computed via the KT estimator, using (4) and (5). In particular, denoting $s \coloneqq c(k_{i-D}^{i-1})$, we can compute the probabilities $\hat{P}(\cdot) \coloneqq P(\cdot|\hat{\mathcal{S}})$ that the next symbol is $k_i = j$, for all $j \in \mathcal{A}$, as

$$\begin{aligned}\hat{P}(j|k_{i-D}^{i-1}) &= \frac{\hat{P}(k_{i-D}^{i})}{\hat{P}(k_{i-D}^{i-1})} = \frac{\prod_{s' \in \hat{\mathcal{S}}} P_e(\boldsymbol{a}_{s'}(k_1^i))}{\prod_{s' \in \hat{\mathcal{S}}} P_e(\boldsymbol{a}_{s'}(k_1^{i-1}))} \\ &= \frac{P_e(\boldsymbol{a}_s(k_1^i))}{P_e(\boldsymbol{a}_s(k_1^{i-1}))} = \frac{a_s(j) + \frac{1}{2}}{\frac{m}{2} + \sum_{j' \in \mathcal{A}} a_s(j')}. \end{aligned} \tag{20}$$

With $\hat{P}$, one could apply arithmetic coding to encode k_i. But the encoded bit-stream would have a variable length depending on both $\hat{P}$ and k_i, and reducing the fluctuation of the coded bit length is important in practical communication systems. On the other extreme, a fixed length coding does not exploit the knowledge of the coding distribution $\hat{P}$ and does not compress at all. Here, we propose an encoding scheme with three possible codeword lengths.

We assume that both encoder and decoder keep a synchronised version of the tree. In addition, we use an auxiliary lower resolution quantiser to apply on least probable symbols. Fix two integers $q_1, q_2 \leq \log m$ such that $m_1 \coloneqq 2^{q_1}$, $m_2 \coloneqq 2^{q_2}$, and $m_1 + m_2 \leq m$. Each incoming symbol $k_i \in \mathcal{A}$ at instant i is encoded as follows.

- If k_i is among the m_1 most probable symbols (tie could be broken with a fixed rule) according to $\hat{P}$, then the encoded bit string is $\mathbf{c}_i = 0$ followed by q_1 bits indicating the position of k_i in the list of the m_1 most probable symbols.
- Otherwise, if k_i is among the next m_2 most probable symbols, the encoded bit string $\mathbf{c}_i$ is 10 followed by q_2 bits indicating the position of k_i in the second list.
- Finally, if k_i is not among the $m_1 + m_2$ most probable symbols, the encoded bit string is $\mathbf{c}_i$ is 11 followed by $\lceil \log m_3 \rceil$ bits corresponding to the index $\tilde{k}_i$ from a lower resolution quantiser of $m_3 - 1$ levels.

Note that, with this scheme, the codeword length is either $1 + q_1$, $2 + q_2$, or $2 + \lceil \log m_3 \rceil$. The proposed scheme can be extended to more levels if desired. In the following, we fix $q_1 = 0$ so that $m_1 = 1$.

It is worth pointing out that the decoder does not have access to the original high-resolution index k_i when the third case happens at time i. This prevents the decoder from normally updating its tree. But encoder and decoder must update the tree in the same way so that the codebooks remain synchronised at both sides. A workaround is to let both the encoder and the decoder update the tree with a projection of the low-resolution index $\tilde{k}_i$ back to the high-resolution codebook. Specifically, we reconstruct the vector using the low-resolution quantiser, re-quantise it with the high-resolution quantiser, and use the corresponding index. Another way could be to simply ignore this symbol for tree update.

We have two main reasons to consider this strategy. First, encoding a symbol and decoding a binary string can be immediately done. Furthermore, the length of the encoded bit-stream is within a fixed number of levels. This aspect is a difference from arithmetic encoding, in which the output is of variable length, which may lead to difficulties when implementing CSI feedback in real systems. In exchange, we cannot expect the asymptotically optimal two-bits redundancy enjoyed by that method.

C. Multiple Trees

In practice, we may want to compress many processes simultaneously, as in the application to CSI representation. Multiple trees come both from the decomposition of complex components into amplitude and phase, and from the fact that the BS has multiple antennas. While each tree provides the marginal distribution of the given sequence, all the marginal

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2022.3173002

distributions can be jointly used to encode the parallel streams together, in order to improve the coding rate. The intuitive idea is that if the indices of all the processes at a given time instant agree with the prediction of the respective models, they need not be individually encoded. If, however, this is not the case, only information about the indices that differ from the model prediction need be transmitted.

Consider that, at each time instant, we have N_t (complex) processes, thus $2N_\mathrm{t}$ scalar symbols to compress (amplitude and phase indices). For the sake of explanation, let us introduce an auxiliary variable (*flag*) Δ_l, defined as follows. For each symbol k_l to be compressed, where, here, the index $l \in [2N_\mathrm{t}]$ denotes the process (and not time):

- if k_l is the most probable symbol (tie could be broken with a fixed rule) according to $\hat{P}$, then $\Delta_l = 0$;
- otherwise, if k_l is among the next 2^{q_l} most probable symbols, $\Delta_l = 1$;
- finally, if k_l is not among the $1 + 2^{q_l}$ most probable symbols, then it is encoded with a lower resolution codebook of size $m_{L,l}$, and $\Delta_l = 2$.

Note that, with this notation, the individual compression rate, described in the previous subsection, is given by (21).

$$R_\text{individual} = \sum_{l=1}^{2N_\mathrm{t}} \Big(\mathbb{1}_{\{\Delta_l=0\}} + \mathbb{1}_{\{\Delta_l=1\}}(2+q_l) + \mathbb{1}_{\{\Delta_l=2\}}(2+\lceil \log m_{L,l} \rceil) \Big). \quad (21)$$

Now, the joint description is composed of two parts: the *state indicator* that contains information about which of the N_t processes have 'varied', i.e., did not follow the correspondent model prediction, and the *change*, which represents the symbols that varied. Therefore the joint rate is written as the sum of the rates of the state indicator and the change parts:

$$R_\text{joint} = R_\text{indicator} + R_\text{change}.$$

Two joint strategies are described in the following.

1) Simple Strategy: A simple way to encode the state indicator is as follows: if all sequences follow the model, encode that with a 0. Otherwise, use N_t bits to indicate which processes (antennas) had some sequence (either amplitude or phase) that varied with respect to the model prediction. This requires

$$R_\text{indicator} = \left(1 + N_\mathrm{t}\mathbb{1}_{\{\Delta_1+\Delta_2+\cdots+\Delta_{2N_\mathrm{t}}>0\}}\right)/N_\mathrm{t} \quad (22)$$

bits per process.

To describe the variation with respect to the model prediction, we need

$$R_\text{change} = C_\text{abs} + C_\text{ang},$$

with C_abs and C_ang given by

$$C_i = \sum_{l=1}^{N_\mathrm{t}} \Big(\mathbb{1}_{\{\tilde{\Delta}_l=0\}}(1+q_i) + \mathbb{1}_{\{\tilde{\Delta}_l=1\}}(1+q_i) + \mathbb{1}_{\{\tilde{\Delta}_l=2\}}(1+\lceil \log m_{L,i} \rceil) \Big), \quad (23)$$

where i is either 'abs' or 'ang', each sum is over the N_t processes of the respective type (amplitude or phase), and $\tilde{\Delta}_l$ is the Δ_l of the corresponding type (amplitude or phase). In addition, q_i is the number of bits needed to describe the list of most probable symbols of an amplitude or phase component, and $m_{L,i}$ is the lower resolution alphabet sizes for amplitude or phase.

Note that we have to encode both the variation of amplitude and phase, as soon as at least one of them varied, for a given antenna, since the indicator part only informs that a sequence has varied, but does not indicate which of them (amplitude or phase). Moreover, in addition to the number of bits needed to describe the symbol—either as the index in the list of most probable symbols or in the lower resolution alphabet—, an additional bit is needed to inform which of these cases has happened.

2) Context-Tree Compression of State Indicator: Alternatively, a more sophisticated strategy is to consider the sequence of state indicators. We consider the state indicator, an array of N_t bits, as an integer number between 0 and $2^{N_\mathrm{t}} - 1$. Then, the sequence of state indicators can be itself compressed using the proposed context-tree-based method (without the lower resolution codebook). In this case, the indicator is described with

$$R_\text{indicator} = R_\text{encoded}/N_\mathrm{t} \quad (24)$$

bits per process, where R_encoded is the length of the binary output generated by the compression scheme.

The change part is as before, i.e.,

$$R_\text{change} = C_\text{abs} + C_\text{ang}, \quad (25)$$

with C_abs and C_ang given by (23).

V. Simulation Results

A. Simulation Setup

We use the MATLAB LTE Toolbox [31] to simulate LTE MIMO downlink channels. The model is configured according to the parameters in Table II. In particular, we consider low (EPA5, Doppler 5 Hz), moderate (EVA30, Doppler 30 Hz) and high mobility (EVA70, Doppler 70 Hz) scenarios, with low or high correlation between antennas at the base station. Other relevant parameters used for quantisation and compression are presented in Table III.

B. Quantiser Design

First, we are interested in evaluating the performance of the quantiser design. We consider three quantisation schemes: the μ-law compander, the β-law compander, and the cube-split quantiser [6]. Interestingly, the cube-split quantiser can be regarded as a complex compander adapted to the distribution of normalised complex Gaussian vectors.

In Fig. 6 we plot the MSCD versus the feedback bit rate per antenna for the three quantisers, with no compression, for low and high antenna correlation, in the EPA5 scenario, with $N_\mathrm{t} = N_\mathrm{r} = 4$. The plotted points correspond to the envelope formed by the best quantisation parameters (different codebook sizes) among those that were tested.

We see that, for low antenna correlation, the cube-split and the proposed quantisers achieve almost the same results. On

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2022.3173002

TABLE II: Simulation Parameters for MATLAB LTE Toolbox.

Field	Parameter	Value
Cell-wide settings	RC	R.12
	DuplexMode	FDD
	TotSubframes	10
Propagation channel configurations	NRxAnts	1
	MIMOCorrelation	High, Low
	NormalizeTxAnts	On
	DelayProfile	EPA, EVA
	DopplerFreq	5, 30, 70
	InitTime	0 to 9.99
	NTerms	16
	ModelType	GMEDS
	NormalizePathGains	On
	InitPhase	Random
Timing and frequency offset	toffset	7
	foffset	0

TABLE III: Simulation Parameters for Quantisation and Compression.

Parameter	Symbol	Value
Total sequence length	n	10^4
Channel signal-to-noise ratio	SNR	30 dB
Tree maximum depth	D	2
Context-tree weighting coefficient	γ	0.5
Training sequence size (% of total length)	–	20%
Interval between tree updates (in symbols)	–	100

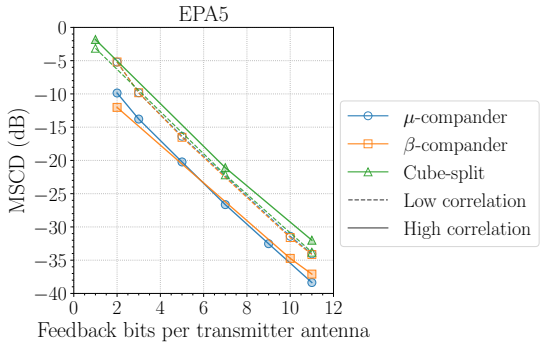


Fig. 6: MSCD distortion for different quantisers.

the other hand, when antenna correlation is high, both proposed quantisers have similar performances and are noticeably better than the cube-split, which assumes uniformity of the distribution by design. The behaviour for EVA30 and EVA70 is similar and has been omitted.

Remark 2. Although in this application we find that there is not much difference in using μ-law or β-law compander, this may not be the case in other applications. The latter, having more adjustable parameters, could provide more flexibility in fitting experimental distributions.

C. Compression Algorithm

Now, we fix the β-law compander as quantiser, and evaluate the performance of different compression schemes. In all cases, we assess the MSCD versus the average number of CSI bits per antenna, and plot the envelope formed by the best quantisation parameters (different codebook sizes) over those that have been tested.

We compare different variations of the proposed context-tree two-level scheme—individual compression, joint compression with the simple strategy, and joint compression with context-tree (CT) compression of the indicator sequence, cf. Section IV-C—with uncompressed and ideal CTW combined with arithmetic coding. The ideal CTW case [13] is simply evaluated with $\frac{1}{n}\left(\left\lceil -\log Q_n(x_1^n|x_{1-D}^0)\right\rceil + 1\right)$. The performance of the different methods is studied for different mobilities and antenna correlations.

Fig. 7 shows the case $N_\mathrm{t} = N_\mathrm{r} = 4$, for both low and high antenna correlation. Regarding the different CTM variations, we note that individual encoding of each sequence generally uses more bits, which is expected, as it does not exploit the spatial correlation between the antennas. Jointly encoding the sequences with the simple strategy can reduce the CSI bit rate, especially in low bit rate regime, and the CT compression of the indicator sequence can further compress the sequence.

Now, comparing the CTM with CT compression of the indicator sequence with uncompressed and CTW schemes, we see that the compression gains are significant and can reduce the CSI bit rate by up to a quarter in low correlation, and a half in high correlation, both in low rate regime. For EPA5, in the higher rate regime, the proposed CTM scheme can reduce the feedback in at least 4.5 bits and is 1.5 bits away from the CTW performance, approximately. For higher mobilities, the gains are more modest, due to the lower time correlation. Nevertheless, for EVA70, in the higher rate regime, the proposed CTM can save approximately 2.5 bits, and CTW, 4 bits, at least. Furthermore, for EVA30 and EVA70 in the extreme low rate regime, the proposed CTM slightly outperforms CTW, thanks to the auxiliary lower resolution quantiser.

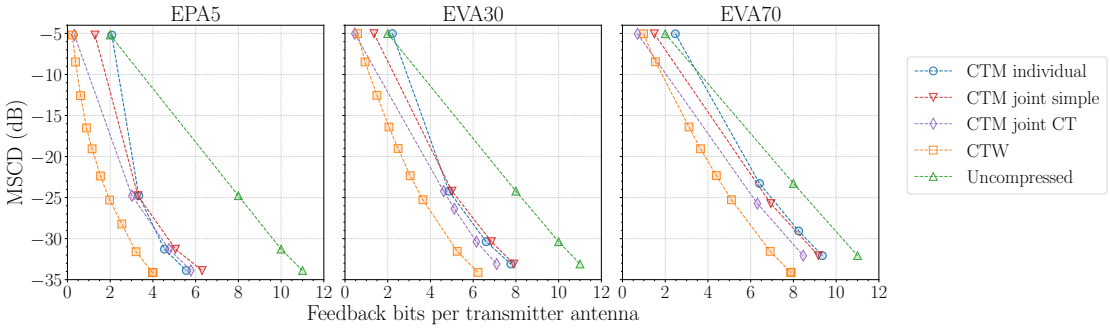
Similarly, Fig. 8 shows the performances for the case $N_\mathrm{t} = 16$, $N_\mathrm{r} = 8$. In the lower rate regime, there is approximately a gain of 1 bit when using the CTM scheme with simple joint strategy, and an additional gain of up to 1 bit if CTW is used instead. In the high rate regime, the gains are more pronounced in the low mobility scenario. For instance, in EPA5, the CTM schemes can provide a saving of approximately 5 bits, which is less than 2 bits away from the CTW performance, in the low correlation case.

D. Communication Rate

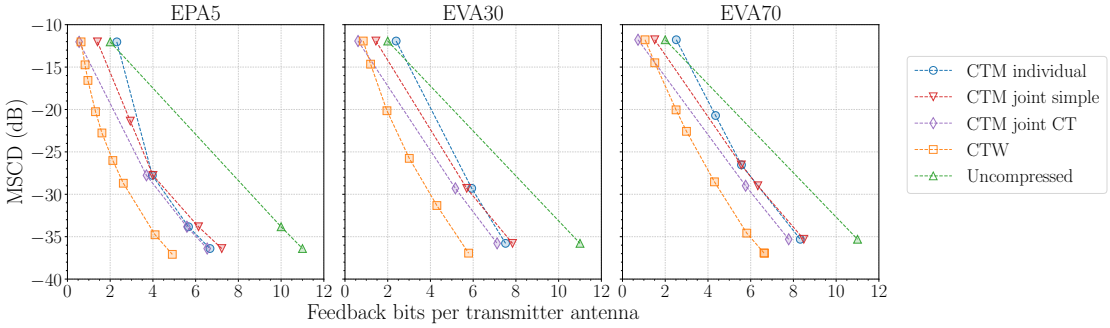
We also illustrate the gains in terms of the downlink communication sum rate with zero-forcing beamforming, evaluated approximately using the formula provided in [2, Eq. (20)], for low antenna correlation. The results are normalised by the achievable rate when perfect (i.e., noiseless) CSI is available to the BS, and are presented in Fig. 9, for different mobilities and number of antennas.

We emphasise that the communication rates converge much faster to the rate achieved by analogue CSI (i.e., with no compression) when some of the proposed compression schemes

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2022.3173002

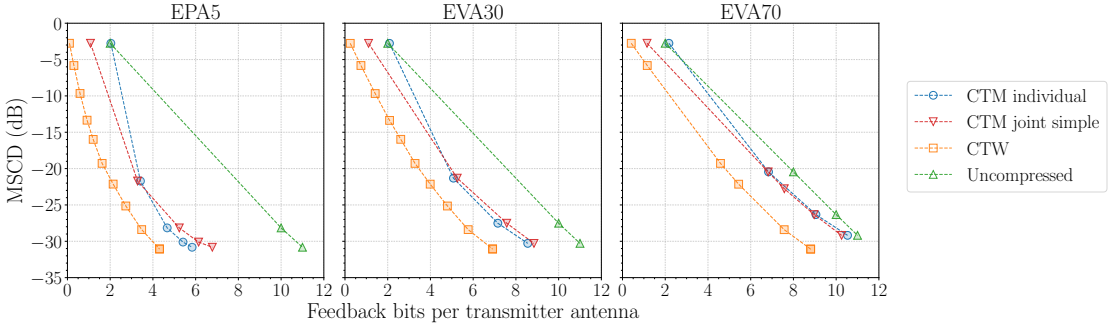


(a) Low antenna correlation.

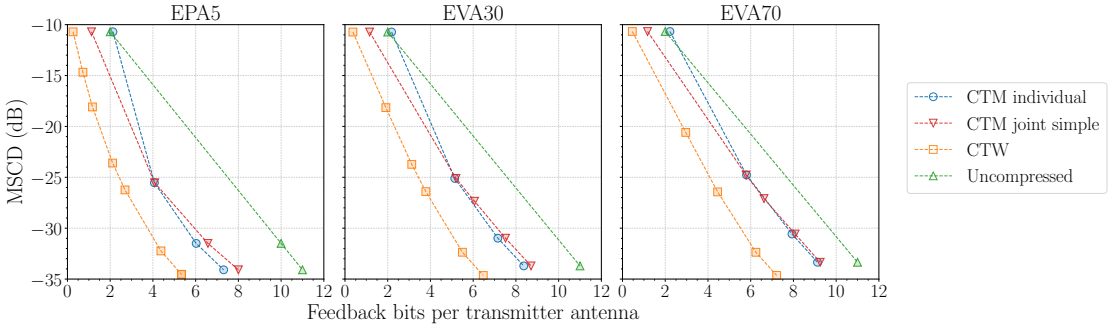


(b) High antenna correlation.

Fig. 7: MSCD distortion using β-law compander, for $N_{\mathrm{t}} = 4$, $N_{\mathrm{r}} = 4$.



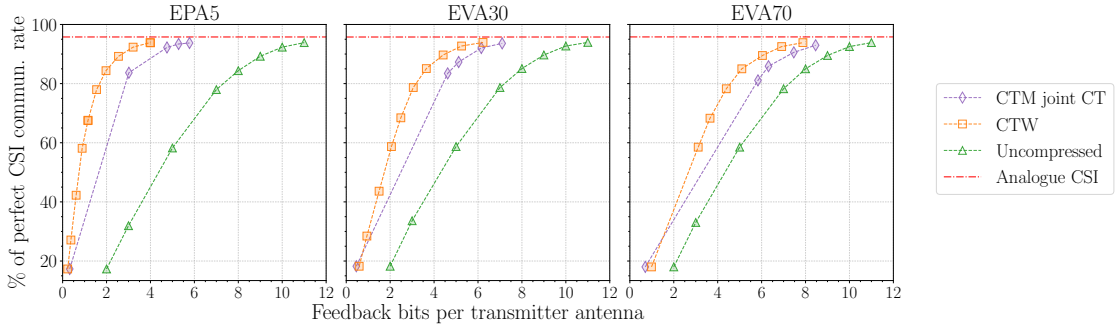
(a) Low antenna correlation.



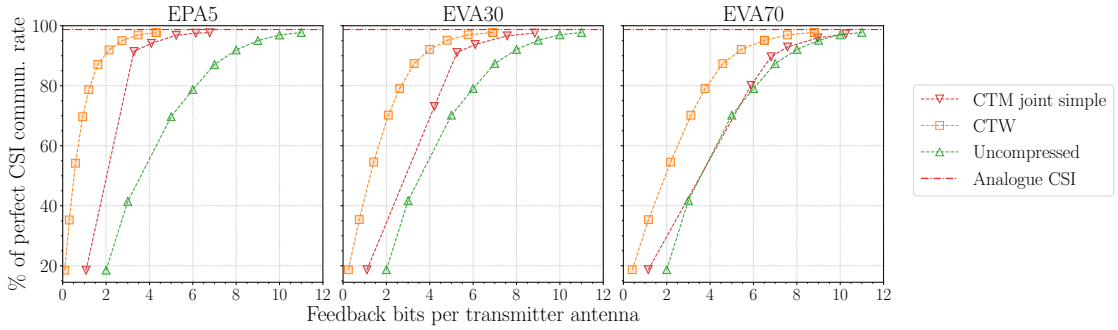
(b) High antenna correlation.

Fig. 8: MSCD distortion using β-law compander, for $N_{\mathrm{t}} = 16$, $N_{\mathrm{r}} = 8$.

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2022.3173002



(a) $N_t = 4$ and $N_r = 4$.



(b) $N_t = 16$ and $N_r = 8$.

Fig. 9: Communication sum rate, using β-law compander.

is employed. For instance, for EPA5 and $N_t = N_r = 4$, it takes the uncompressed scheme 11 feedback bits per antenna to achieve a communication rate close to the analogue upper-bound. Using the proposed CTM scheme, the same communication rate can be achieved with less than 6 bits, and, with ideal CTW, with 4 bits. In higher mobility, the gap between the two proposed scheme is smaller, while still presenting an advantage over not compressing at all. Similar gains are observed for $N_t = 16$, $N_r = 8$, even with the simple joint strategy.

VI. Conclusion

We have proposed a novel method for compressing CSI, combining lossy vector quantisation and lossless compression. The proposed vector quantiser is based on applying a data-adapted compander to the components of normalised vectors. The compression algorithm uses the estimated probability provided by the CTM model to encode a symbol, following a simple rule within a fixed number of levels. Simulations of LTE channels show the effectiveness of our approach in different scenarios.

More importantly, the proposed schemes have low complexity, can be implemented in an online fashion, and are modular. The context-tree-based compression scheme can be applied on any other quantisers, including those recently designed with neural networks, e.g., [5]. Similarly, the proposed quantiser can be combined with any other lossless compression schemes.

References

[1] H. K. Miyamoto and S. Yang, "A CSI compression scheme using context trees," in *Int. Zurich Seminar Inf. and Commun. (IZS 2022)*, 2022, pp. 24–28.

[2] G. Caire *et al.*, "Multiuser MIMO achievable rates with downlink training and channel state feedback," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2845–2866, 2010.

[3] C.-K. Wen, W.-T. Shih, and S. Jin, "Deep learning for massive MIMO CSI feedback," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 748–751, 2018.

[4] J. Guo, C.-K. Wen, S. Jin, and G. Y. Li, "Convolutional neural network-based multiple-rate compressive sensing for massive MIMO CSI feedback: Design, simulation, and analysis," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2827–2840, 2020.

[5] M. B. Mashhadi, Q. Yang, and D. Gündüz, "Distributed deep convolutional compression for massive MIMO CSI feedback," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2621–2633, 2021.

[6] A. Decurninge and M. Guillaud, "Cube-split: Structured quantizers on the Grassmannian of lines," in *2017 IEEE Wireless Commun. and Netw. Conf. (WCNC)*, 2017, pp. 1–6.

[7] N. Shlezinger and Y. C. Eldar, "Deep task-based quantization," *Entropy*, vol. 23, no. 1, 2021.

[8] E. Gassiat, *Universal Coding and Order Identification by Model Selection Methods*. Cham, Switzerland: Springer, 2018.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[10] I. Csiszár and P. Shields, "Information theory and statistics: A tutorial," *Found. and Trends in Commun. and Inf. Theory*, vol. 1, no. 4, pp. 417–528, 2004.

[11] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inf. Theory*, vol. 23, no. 3, pp. 337–343, 1977.

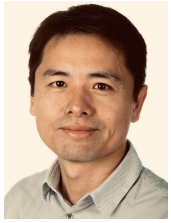
[12] ——, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inf. Theory*, vol. 24, no. 5, pp. 530–536, 1978.

[13] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: basic properties," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 653–664, 1995.

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2022.3173002

[14] T. J. Tjalkens, Y. M. Shtarkov, and F. M. J. Willems, "Context tree weighting: Multi-alphabet sources," in *14th Symp. Inf. Theory in the Benelux*, Veldhoven, The Netherlands, 1993, pp. 128–135.
[15] ——, "Sequential weighting algorithms for multi-alphabet sources," in *6th Joint Swedish-Russian Int. Workshop Inf. Theory*, Molle, Sweden, 1993, pp. 230–234.
[16] T. J. Tjalkens, F. Willems, and Y. Shtarkov, "Multi-alphabet universal coding using a binary decomposition context tree weighting algorithm," in *15th Symp. Inf. Theory in the Benelux*, Louvain-la-Neuve, Belgium, 1994, pp. 259–265.
[17] R. Begleiter, R. El-Yaniv, and G. Yona, "On prediction using variable order Markov models," *J. Artif. Int. Res.*, vol. 22, no. 1, pp. 385–421, 2004.
[18] R. Begleiter and R. El-Yaniv, "Superior guarantees for sequential prediction and lossless compression via alphabet decomposition," *J. Mach. Learn. Res.*, vol. 7, no. 13, pp. 379–411, 2006.
[19] F. Willems, T. Tjalkens, and Y. Shtarkov, "Context-tree maximizing," in *Proc. 34th Annu. Conf. Inf. Sciences and Syst.*, Princeton, New Jersey, 2000, pp. TP6–7–TP6–12.
[20] F. M. J. Willems, A. Nowbakht, and P. Volf, "Maximum a posteriori probability tree models," in *4th Int. ITG Conf. Source and Channel Coding*, Berlin, Germany, 2002, pp. 335–340.
[21] L. Mertzanis *et al.*, "Deep tree models for 'big' biological data," in *2018 IEEE 19th Int. Workshop Signal Process. Advances in Wireless Commun. (SPAWC)*, 2018, pp. 1–5.
[22] I. Kontoyiannis *et al.*, "Bayesian context trees: Modelling and exact inference for discrete time series," *arXiv*, 2020. [Online]. Available: https://arxiv.org/pdf/2007.14900v1.pdf
[23] I. H. Witten, R. M. Neal, and J. G. Cleary, "Arithmetic coding for data compression," *Commun. ACM*, vol. 30, no. 6, pp. 520–540, 1987.
[24] A. Moffat, R. M. Neal, and I. H. Witten, "Arithmetic coding revisited," *ACM Trans. Inf. Syst.*, vol. 16, no. 3, p. 256–294, Jul. 1998.
[25] "Context-tree based CSI compression." [Online]. Available: https://miyamotohk.github.io/context-tree-compression
[26] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston, MA, USA: Kluwer, 1992.
[27] T. J. Tjalkens, P. A. J. Volf, and F. M. J. Willems, "A context-tree weighting method for text generating sources," in *Proc. DCC '97. Data Compression Conf.*, 1997, p. 472.
[28] P. A. J. Volf, "Weighting techniques in data compression: Theory and algorithms," Ph.D. dissertation, Technische Universiteit Eindhoven, The Netherlands, 2002.
[29] W. R. Bennett, "Spectra of quantized signals," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 446–472, 1948.
[30] S. S. Dragomir, R. P. Agarwal, and N. S. Barnett, "Inequalities for beta and gamma functions via some classical and new integral inequalities," *RGMIA Res. Rep. Coll.*, vol. 2, no. 3, 1999.
[31] MathWorks. MATLAB LTE toolbox (R2020b). [Online]. Available: https://www.mathworks.com/help/lte/

Henrique K. Miyamoto (Graduate Student Member, IEEE) was born in Salvador, Brazil. He received the B.S. degree in electrical engineering from the University of Campinas (Unicamp), Campinas, Brazil, and the *Diplôme d'Ingénieur* degree from CentraleSupélec, Gif-sur-Yvette, France, both in 2021. He is currently pursuing the M.Sc. degree in applied mathematics with Unicamp. He held short-term internships at the Laboratory of Signals and Systems (L2S), CentraleSupélec, in 2019, and at the Mathematical and Algorithmic Sciences Lab, Huawei France, in 2020. His research interests include information theory and coding.



Sheng Yang (Member, IEEE) received the B.E. degree in electrical engineering from Jiaotong University, Shanghai, China, in 2001, and both the engineer degree and the M.Sc. degree in electrical engineering from Telecom ParisTech, Paris, France, in 2004. In 2007, he obtained the Ph.D. degree from Université Pierre et Marie Curie (Paris VI). From October 2007 to November 2008, he was with Motorola Research Center in Gif-sur-Yvette, France, as a Senior Staff Research Engineer. Since December 2008, he has joined CentraleSupélec, Paris-Saclay University, where he is currently a Full Professor. From April 2015, he also holds an Honorary Associate Professorship in the Department of Electrical and Electronic Engineering of the University of Hong Kong (HKU). He received the 2015 IEEE ComSoc Young Researcher Award for the Europe, Middle East, and Africa Region (EMEA). He was an Associate Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2015 to 2020. He is currently an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY.

Anexo C

Aviso Legal sobre Uso dos Artigos

"Em referência ao material do IEEE protegido por direitos autorais que é usado com permissão nesta dissertação, o IEEE não endossa qualquer produto ou serviço da Universidade Estadual de Campinas. Uso interno ou pessoal desse material é permitido. Se houver interesse em reimprimir/republicar material do IEEE protegido por direitos autorais para fins publicitários ou promocionais, ou para a criação de novos trabalhos coletivos para revenda ou redistribuição, por favor visite `http://www.ieee.org/publications_standards/publications/rights/rights_link.html` para saber como obter uma Licença RightsLink. Se aplicável, University Microfilms e/ou ProQuest Library, ou Archives of Canada podem fornecer cópias avulsas da dissertação."

"In reference to IEEE copyrighted material which is used with permission in this dissertation, the IEEE does not endorse any of Universidade Estadual de Campinas's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to `http://www.ieee.org/publications_standards/publications/rights/rights_link.html` to learn how to obtain a License from RightsLink. If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation."