
UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO

Fernando Zarpellon

# Análise do Problema de Compressão da Informação em Redes Neurais

CAMPINAS
2022

Fernando Zarpellon

# Análise do Problema de Compressão da Informação em Redes Neurais

Dissertação apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica, na Área de Engenharia de Computação.

Orientador: Prof. Dr. Romis Ribeiro de Faissol Attux

Este trabalho corresponde à versão final da dissertação defendida pelo aluno Fernando Zarpellon, e orientada pelo Prof. Dr. Romis Ribeiro de Faissol Attux.

CAMPINAS

2022

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

Z19a

Zarpellon, Fernando, 1990-
Análise do problema de compressão da informação em redes neurais / Fernando Zarpellon. – Campinas, SP : [s.n.], 2022.

Orientador: Romis Ribeiro de Faissol Attux.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Teoria da informação. 2. Aprendizagem profunda. I. Attux, Romis Ribeiro de Faissol, 1978-. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Informações Complementares

**Título em outro idioma:** Analysis of the information compression problem in neural networks
**Palavras-chave em inglês:**
Information theory
Deep learning
**Área de concentração:** Engenharia de Computação
**Titulação:** Mestre em Engenharia Elétrica
**Banca examinadora:**
Romis Ribeiro de Faissol Attux [Orientador]
Wanessa Carla Gazzoni
Levy Bocatto
**Data de defesa:** 11-10-2022
**Programa de Pós-Graduação:** Engenharia Elétrica

**Identificação e informações acadêmicas do(a) aluno(a)**
- ORCID do autor: https://orcid.org/0000-0002-7811-9277
- Currículo Lattes do autor: https://lattes.cnpq.br/8202469619522020

COMISSÃO JULGADORA – TESE DE MESTRADO

Candidato: Fernando Zarpellon RA: 226410

Data da defesa: 11 de outubro de 2022

Título da Tese: “Análise do Problema de Compressão da Informação em Redes Neurais”

Prof. Dr. Romis Ribeiro de Faissol Attux (Presidente)

Profa. Dra. Wanessa Carla Gazzoni (UNISAL/Campinas)

Prof. Dr. Levy Boccato (FEEC/UNICAMP)

A Ata de Defesa, com as respectivas assinaturas dos membros da Comissão Julgadora, encontra-se no SIGA (Sistema de Fluxo de Dissertação/Tese) e na Secretaria de Pós-Graduação da Faculdade de Engenharia Elétrica e de Computação.

# AGRADECIMENTOS

Sou imensamente grato aos meus pais pelo apoio e suporte durante a trajetória deste projeto de aprendizagem e engrandecimento pessoal e intelectual que o mestrado representa.

Agradeço ao meu orientador, professor Romis Attux, a quem devo gratidão, respeito e admiração por ter me dado a oportunidade de estudar e desenvolver-me como pesquisador. A generosidade em orientar-me em um campo promissor da ciência cujas dificuldades nos serviram para aprimorar tal trabalho, e, a determinação de me guiar no caminho da ciência e pesquisa.

Aos meus colegas e amigos do DSPCom, minha estima e apreço pelos momentos de reflexão, devaneios, estudo e de alegria.

Finalmente, agradeço aos funcionários da UNICAMP e da FEEC, especialmente à equipe de Coordenadoria de Pós-Graduação da FEEC, pela qualidade do trabalho e ensino, pela tenacidade e continuidade que realizam em manter em funcionamento nossas atividades dentro do espaço universitário.

Também agradeço ao CNPq, que tornou possível que eu me dedicasse integralmente à minha formação acadêmica.

O presente trabalho foi realizado com o apoio do CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brasil.

# RESUMO

A teoria de aprendizagem profunda por restrição de informação, IBDL (do inglês, *Information Bottleneck theory of Deep Learning*), afirma que uma rede neural artificial profunda (DNN), através dos conceitos da teoria da informação, pode ser interpretada como uma cadeia de *Markov*, e, através da desigualdade de processamento de informação (DPI) pode-se analisar a representação latente formada na rede ao longo do processo de treinamento, através da informação mútua destas camadas utilizando o plano de informação (IP).

Esta tese tem como objetivo investigar a aplicação desta teoria em problemas de regressão, uma vez que a literatura presente até o momento se faz escassa em aplicações desta natureza. Para tal, estabelecemos uma tarefa de regressão formalizada através do problema de separação de fontes supervisionada, onde buscamos a reconstrução dos sinais de fonte. Para desenvolver a análise do problema é necessária a estimação das medidas de informação como entropia e informação mútua, este processo é realizado utilizando o estimador discreto, que se baseia na estimação através de distribuições de probabilidade obtidas por meio de histogramas (discretização – *binning*), um método simples, extremamente eficiente computacionalmente e que vem sendo utilizado com certa frequência pelos trabalhos da área.

Através dos resultados observados nos planos de informação e das projeções latentes da rede, estabelecemos algumas relações sobre o fenômeno de compressão e expansão da informação mútua que descreve tais representações segundo a teoria IBDL. Além disto, analisamos o impacto das não-linearidades utilizadas em redes profundas na dinâmica de treinamento da rede e na formação destas representações. Os resultados indicam que as redes neurais quando aplicadas em problemas de regressão, seguem as relações estabelecidas pela DPI segundo a formulação do IBDL, bem como apresentaram convergem para os limites teóricos também estabelecidos na formulação do problema. Associamos esta convergência à forma como a rede neural opera em termos de capacidade de processamento, utilizando uma parcela ou a totalidade da capacidade disponível de sua estrutura: esse comportamento influencia na formação da representação latente criada durante o processo de treinamento em conjunto com as não linearidades utilizadas. Por fim, constatamos a viabilidade do estudo das redes neurais aplicadas a problemas de regressão utilizando os conceitos estabelecidos na literatura sobre a teoria IBDL, largamente aplicada a problemas de classificação até o presente momento. Dessa forma, este trabalho contribui para o enriquecimento da discussão a respeito da intepretação de redes neurais através da teoria da informação.

**Palavras-chave**: Teoria da informação, restrição de informação, aprendizagem profunda, rede neural profunda, plano de informação, regressão, separação de fontes.

# ABSTRACT

The Information Bottleneck theory of Deep Learning (IBDL) states that the deep neural network (DNN), through the concepts of information theory, generates a successive Markov chain and through de data processing inequality (DPI) the training process of a neural networks and the latent representation formed can be analyzed through the information plane (IP).

This thesis aims to investigate the application of the IBDL on regression problems. For this task we formalized the regression problem as supervised source separation where the objective is to reconstruct one of the sources signals. To develop the analyses, it is necessary to quantify information measures as entropy and mutual information, for that we use the binning method for discretization of the continuous random variables, a simple and efficient method widely applied in the resent literature about this subject.

Through the observed results in the information plane and the inner neurons projection we stablish some relationships about the phenomenon of compression and expansion of the mutual information that describes the latent representation of the deep networks, and the impact of nonlinearities commonly used in deep learning on the dynamics of network training. The results indicate that the neural networks follow the relationships stablished by the DPI as well converge to the theoretical limits. We associate this convergence to the way the neural network operates in terms of processing capacity, using a portion or all the available capacity, this behavior associated with the nonlinearities influences the formation of the latent representation created during the training process. Finally, we verify the feasibility of the extension and application of the IBDL method, widely applied in classification problems to regression problems.

**Keywords**: Information theory, information bottleneck, deep learning, deep neural network, information plane, regression, supervised source separation.

# LISTA DE ILUSTRAÇÕES


FIGURA 2.1 - MODELO MATEMÁTICO DE UM NEURÔNIO ARTIFICIAL DE ROSENBLATT. ..... 17
FIGURA 2.2 - FUNÇÕES DE ATIVAÇÃO. A) SIGMOID. B) TANGENTE HIPERBÓLICA. C) RELU. ..... 20
FIGURA 2.3 - REDE ALIMENTADA ADIANTE DE CAMADA ÚNICA. ..... 21
FIGURA 2.4 - REDE ALIMENTADA ADIANTE MÚLTIPLAS CAMADAS ..... 21
FIGURA 2.5 - ESQUEMA DA RETROPROPAGAÇÃO DO ERRO E DE ATUALIZAÇÃO DOS PESOS ..... 22
FIGURA 3.1 - RELAÇÃO ENTRE ENTROPIAS E INFORMAÇÃO MÚTUA. ..... 28
FIGURA 3.2 - EVOLUÇÃO DO TREINAMENTO DE UMA REDE NEURAL COM CAMADAS INTERNAS NO PLANO DE INFORMAÇÃO (IP) ..... 42
FIGURA 3.3 - PROCESSO DE DISCRETIZAÇÃO DE UMA REDE NEURAL COM DUAS ENTRADAS E UMA CAMADA INTERNA COM DOIS NEURÔNIOS E ÚNICA SAÍDA [2,2,1]. ..... 47
FIGURA 3.4 - EXTRAÇÃO DAS PROBABILIDADES E PROBABILIDADES CONJUNTAS DOS NEURÔNIOS ..... 48
FIGURA 4.1 - REDE NEURAL ARTIFICIAL SEGUNDO A TEORIA DA INFORMAÇÃO. ..... 51
FIGURA 4.2 - PLANO DE INFORMAÇÃO PARA REDES NEURAIS MLP. COMPRESSÃO DE $IX$ A PARTIR DA MARCAÇÃO EM VERDE. SETAS ADICIONAIS REFERENTES AS DUAS FASES DE APRENDIZAGEM EM VERMELHO E DE COMPRESSÃO EM VERDE ..... 52
FIGURA 4.3 - MÉDIA E DESVIO PADRÃO DOS PESOS SINÁPTICOS AO LONGO DO TREINAMENTO. DRIFT PHASE ANTES DA MARCAÇÃO PONTILHADA E DIFUSSION PHASE DEPOIS DA MARCAÇÃO. ..... 53
FIGURA 4.4 - PLANO DE INFORMAÇÃO TANH VS. RELU. A) FUNÇÃO TANH DISCRETIZADA. B) FUNÇÃO RELU DISCRETIZADA. C) FUNÇÃO TANH, CONJUNTO MNIST USANDO O ESTIMADOR KDE. D) FUNÇÃO RELU, CONJUNTO MNIST USANDO O ESTIMADO KDE. ..... 56
FIGURA 4.5 - A) ANÁLISE DOS PESOS PARA A FUNÇÃO TANH. B) ANÁLISE DOS PESOS PARA FUNÇÃO RELU ..... 58
FIGURA 4.6 - A) GRÁFICO SUPERIOR FUNÇÃO TANH COM PARADA ANTECIPADA. B) GRÁFICO SUPERIOR PARADA PERFEITA DA TANH. NA PARTE INFERIOR A RECRIAÇÃO DO EXPERIMENTO DE (SCHWARTZ-ZIV & TISHBY, 2017). ..... 62
FIGURA 5.1 - ESQUEMA DA SEPARAÇÃO DE FONTES E DAS REDES NEURAIS PROPOSTO PARA O TRABALHO ..... 79
FIGURA 5.2 - CONJUNTO DE DADOS UTILIZADOS PARA REGRESSÃO. A) EM LARANJA $S1$ E $S2$ FORMAM UMA VARIÁVEL GAUSSIANA BIDIMENSIONAL, EM AZUL A TRANSFORMAÇÃO DADA PELA MATRIZ A. B) GRÁFICO DE X1 E X2 QUE SÃO A ENTRADA DA REDE POR $S1$ NO EIXO Z DEMONSTRANDO A ROTAÇÃO DOS DADOS ..... 80
FIGURA 5.3 - PLANOS DE INFORMAÇÃO PARA A FUNÇÃO RELU. A) FUNÇÃO DISCRETIZADA SEGUNDO (SAXE, ET AL., 2018). B) FUNÇÃO DISCRETIZADA UTILIZANDO NOSSO MÉTODO ..... 85
FIGURA 5.4 - MÉDIA PERCENTUAL DO NÚMERO DE ATIVAÇÕES QUE SATURAM OS LIMITES TESTADOS PARA DISCRETIZAÇÃO DA FUNÇÃO IDENTIDADE NA SAÍDA DA REDE. O CONJUNTO DE DADOS DE TREINAMENTO POSSUI 1500 AMOSTRAS ..... 86
FIGURA 5.5 - DIAGRAMA DE VENN PARA A DPI. A CAMADA INTERNA DA REDE PODE RETER DIFERENTES ..... 88
FIGURA 6.1 - PLANO DE INFORMAÇÃO PARA DIFERENTES NÃO-LINEARIDADES NA CAMADA INTERNA DA REDE. A MÉDIA DE 20 INICIALIZAÇÕES MOSTRAM CAMINHOS DISTINTOS PARA A INFORMAÇÃO DA CAMADA INTERNA $L_0$. A) SIGMOID. B) RELU. C) TANH. INFORMAÇÃO ESTIMADA EM BITS ..... 89
FIGURA 6.2 - MANIFESTAÇÃO DA COMPRESSÃO NAS FUNÇÕES DE ATIVAÇÃO PARA DIFERENTES INICIALIZAÇÕES. **A**,**B**, E **C** REDES COM A FUNÇÃO RELU. **D**, **E**, E **F** REDES COM FUNÇÃO SIG E **G**, **H**, E **I** REDES COM FUNÇÃO TANH. A COMPRESSÃO É CARACTERIZADA PELA REDUÇÃO DA INFORMAÇÃO $IX; L0$, VIDE

APÊNDICE A. A VARIAÇÃO DA COR DAS CURVAS DE INFORMAÇÃO REPRESENTAM AS ÉPOCAS DE TREINO.....90
FIGURA 6.3 - DETALHAMENTO DO TREINO DE UMA REDE COM A FUNÇÃO RELU. O PROCESSO DE COMPRESSÃO FIG B) E E) É ASSOCIADO VARIAÇÃO DA VARIANCIA DA DISTRIBUIÇÃO GAUSSIANA FORMADA INTERNAMENTE NA REDE FIG A) E D).....91
FIGURA 6.4 – PROCESSO DE COMPRESSÃO EM REDES SIGMOID. A DISTRIBUIÇÃO FORMADA INTERNAMENTE EM UMA REDE FIG A) APRESENTA OS DADOS DISTRIBUIDOS DE FORMA MAIS DISPERSA, APRESENTANDO A MAIOR ENTROPIA, CONSEQUENTEMENTE A DISTRUBIÇÃO FORMADA POSTERIORMENTE TEM ENTROPIA MENOR FIG D), PRODUZINDO A REDUÇÃ DA INFORMAÇÃO MÚTUA $I(X;L_0)$ FIG B) E E).....92
FIGURA 6.5 - PROCESSO DE COMRPESSÃO EM UMA REDE COM A FUNÇÃO TANH. A REDUÇÃO DA INFORMAÇÃO MÚTUA EM DECORRENCIA DA REDUÇÃO DA ENTROPIA DE DISTRIBUIÇÕES BIDIMENCIONAS FIG A) E D), CONSEQUENTEMENTE O COMPORTAMENTO MANIFESTADO EM C) E F).....92
FIGURA 6.6 - LIMITE SUPERIOR DA DPI ATINGIDO POR UMA REDE TREINADA COM A FUNÇÃO RELU. A PROJEÇÃO DA CAMADA INTERNA DA REDE APRESENTA UMA VERSÃO REESCALADA DOS DADOS EM A) CONTENDO TODA A INFORMAÇÃO DISPONÍVEL, OU SEJA, $IX; L0 = HX$ EM B).....94
FIGURA 6.7 - LIMITE SUPERIOR DA DPI ATINGIDO POR UMA REDE TREINADA COM A FUNÇÃO SIG. A PROJEÇÃO DA CAMADA INTERNA FORMA UMA VERSÃO COMPACTA DOS DADOS DE ENTRADA EM A) CONTENDO TODA INFORMAÇÃO DISPONÍVEL , $\mathrm{IX; L0 = HX}$ EM B).....94
FIGURA 6.8 - LIMITE SUPERIOR DA DPI ATINGIDA POR UMA REDE TREINADA COM A FUNÇÃO TANH. A TRANSFORMAÇÃO NÃO LINEAR AFETA A PROJEÇÃO INTERNA DA REDE A) QUE SE ASEMELHA ASO DADOS DE ENTRADA. A INFORMAÇÃO CONTIDA NA CAMADA INTERNA É A MAXIMA DISPONÍVEL, , $IX; L0 = HX$, COMO PODE SER OBSERVADO NA FIGURA B).....94
FIGURA 6.9 - REPRESENTAÇÃO LATENTE PARA FUNÇÃO TANH ONDE A PROJEÇÃO INTERNA É DIFUSA A), ESTA REPRESENTAÇÃO INTERNA CONTÉM TODA A INFORMAÇÃO CONTIDA NOS DADOS, , $IX; L0 = HX$, COMO OBSERVADO EM B).....95
FIGURA 6.10 - REPRESENTAÇÃO LATENTE MÍNIMA ATINGIDA EM UM REDE COM FUNÇÃO RELU, $IX; L0 = IX; Y$. .....95
FIGURA 6.11 - REPRESENTAÇÃO LATENTE QUASE MÍNIMA ATINGIDA EM UMA REDE COM FUNÇÃO SIG., $\mathrm{IX; L0} \cong \mathrm{IX; Y}$.....95
FIGURA A1 - REDES NEURAIS TREINADAS COM A FUNÇÃO RELU NA CAMADA INTERNA. .....105
FIGURA A2 - REDES NEURAIS TREINADAS COM A FUNÇÃO SIGMOID.....106
FIGURA A3 - VARIAÇÃO DE INFORMAÇÃO DE FORMAÇÃO DA REDE SIGMOID REFERENTE A FIGURA A2 D).....107
FIGURA A4 - REDES TREINADAS COM A FUNÇÃO TANGENTE HIPERBÓLICA QUE APRESENTAM COMPRESSÃO....107
FIGURA A5 - REDUÇÃO DA INFORMAÇÃO $IX; L0$ DE C) PARA F) EM DECORRÊNCIA DO BAIXO GRAU DE SIMILARIDADE (DIVERGENTE DE KULLBACK-LEIBLER) ENTRE AS DISTRIBUIÇÕES DOS NEURÔNIOS COM RELAÇÃO A DISTRIBUIÇÃO DOS DADOS $X$. .....108
FIGURA A6 - GANHO DE INFORMAÇÃO $I(L0; Y)$ EM DECORRENCIA DA FORMAÇÃO DA DISTRIBUIÇÃO GAUSSIANA (APROXIMADAMENTE) NO NEURÔNIO $N2$ DA FIGURA D).....108
FIGURA B1 - REDE NEURAL RELU QUE APRESENTA COMPRESSÃO EM $L0$ COM BOA GENERALIZAÇÃO. A) ERRO DE TREINO E TESTE DA REDE. B) PLANO DE INFORMAÇÃO. C) COMPARAÇÃO DA PREDIÇÃO E AS AMOSTRAS ALVO PARA UMA PARCELA DE DADOS.....110
FIGURA B2 - REDE NEURAL RELU QUE NÃO APRESENTA COMPRESSÃO EM $L0$ E TEM BOA GENERALIZAÇÃO. A) ERRO DE TREINO E TESTE DA REDE. B) PLANO DE INFORMAÇÃO. C) COMPARAÇÃO DA PREDIÇÃO E AS AMOSTRAS ALVO PARA UMA PARCELA DE DADOS.....110

FIGURA C1 - REDE TREINADA COM A FUNÇÃO RELU COM REPRESENTAÇÃO INTERMEDIÁRIA, OU SEJA, $I(X;L0)$ NÃO CONVERGIU PARA OS LIMITES $H(X)$ OU $I(X;Y)$. ..... 111
FIGURA C2 - REDE TREINADA COM A FUNÇÃO SIGMOID APRESENTANDO REPRESENTAÇÃO INTERNA INTERMEDIARIA. ..... 111
FIGURA C3 - REDE TREINADA COM A FUNÇÃO TANGENTE HIPERBÓLICA COM REPRESENTAÇÃO INTERNA INTERMEDIARIA. ..... 111

# LISTA DE ABREVIATURAS E SIGLAS

**ANN** - Artificial Neural network

**CDF** - Função de Distribuição Acumulada

**CIFAR** - *Canadian Institute For Advanced Research dataset*

**CNN** - Convolutional Neural Network

**DNN** – Deep Neural Networks

**DPI** - Desigualdade de Processamento de Informação

**EMR** - Minimização do Erro Empírico

**IB** - Information Bottleneck

**KDE** – Kernel Density Estimator

**ML** - Machine Learning

**MLP** - Multi Layer Perceptron

**MNITS** - *Modified National Institute of Standards and Technology dataset*

**PDF** - Função de Densidade de Probabilidade

**PMF** - Função de massa de probabilidade

**QNN** - Quantized Neural Networks

**RC** - Compressão da Representação

**ReLu** - Rectifield Linear Unit

**RNA** - Redes Neurais Artificiais

**RNN** - Recurrent Neural Network

**SAE** - Stacked Autoencoder

**SGC** - Gradiente Estocástico Descendente

**Sig** - Sigmoid

**SNR** – Signal Noise Ration

**SZT** - Schwartz-ziv Tishby *Toy dataset*

**Tanh** - Tangente Hiperbólica

# Sumário


1. Introdução ..... 14
2. Redes Neurais Artificiais ..... 16
   2.1 Introdução ..... 16
   2.2 Modelo Matemático do Neurônio ..... 16
   2.3 Funções de ativação ..... 18
      2.3.1 Função Logística ..... 18
      2.3.2 Tangente Hiperbólica ..... 19
      2.3.3 *Rectified Linear Unit* (*ReLU*) ..... 19
   2.4 Arquiteturas ..... 20
      2.4.1 Redes Alimentadas Adiante – Camada Única ..... 20
      2.4.2 Redes Alimentadas Adiante – Múltiplas Camadas ..... 21
   2.5 Perceptron de Múltiplas Camadas (MLP) ..... 22
      2.5.1 Treinamento e Aprendizagem ..... 22
3. Teoria da Informação ..... 24
   3.1 Introdução ..... 24
   3.2 Entropia, Entropia Conjunta e Entropia Condicional ..... 24
   3.3 Entropia Relativa e Informação Mútua ..... 26
   3.4 Desigualdade de Processamento de Informação (DPI) ..... 29
   3.5 Estatística Suficiente ..... 30
   3.6 Entropia Diferencial ..... 31
   3.7 Relação entre Entropia Diferencial e Discreta ..... 33
   3.8 *Information Bottleneck* (IB) ..... 38
   3.9 Plano de informação (IP) ..... 40
      3.9.1 Leitura e Interpretação do Plano de Informação ..... 41
   3.10 Estimação das medidas de Informação ..... 44
      3.10.1 Discretização (*Binning*) ..... 45
      3.10.2 Método Discreto de Estimação de Medidas de Informação ..... 46
4. Análise de Redes Neurais Artificiais Através da Teoria da Informação ..... 49
   4.1 Hipótese Inicial ..... 49
   4.2 O princípio da compressão ..... 51

4.3 Uma nova perspectiva ..... 54
4.4 O problema da estimação ..... 58
4.5 Caracterização do problema do IBDL: Uma breve reflexão ..... 61
4.6 Trabalhos relacionados e a recente convergência do assunto ..... 65
5. Motivação e Definição do Experimento ..... 74
5.1 O problema de estimação e a entropia diferencial ..... 76
5.2 Configuração do Experimento e *Dataset* ..... 79
5.2 Metodologia de Estimação e Parâmetros ..... 82
5.3 *Information Bottleneck* em Regressão ..... 87
6. Resultados e Análises ..... 89
6.1 O Fenômeno da Compressão em Regressão ..... 89
6.2 Compressão e Generalização: Os limites teóricos e a convergência da rede ..... 93
7. Conclusão ..... 98
7.1 Trabalhos Futuros ..... 100
Referências ..... 101
APÊNDICE ..... 104
Apêndice A ..... 104
Apêndice B ..... 109
Apêndice C ..... 111

# 1. Introdução

A teoria da informação, desenvolvida por Shannon (SHANNON, 1948), estabelece a base teórica das telecomunicações, mas os conceitos por ele criados podem ser generalizados para inúmeras áreas de conhecimento. De fato, a teoria proporcionou ferramental para o desenvolvimento e aperfeiçoamento de algoritmos e estratégias de treinamento em redes neurais (ACHILLE & SOATTO, 2018), árvores de decisão (Quinlan, 1986), agrupamento (*clustering*) (GOKCAY & PRINCIPE, 2002), dentre muitos outros exemplos. Especificamente, na última década, a teoria da informação vem sendo explorada de forma consistente e persistente na área de inteligência artificial, mais precisamente em aprendizagem profunda (do inglês, *deep learning - DL)*, através da teoria de aprendizagem profunda por restrição de informação (*Information Bottleneck Theory of Deep Learning - IBDL*) (TISHBY & ZASLAVSKY, 2015).

A teoria do IBDL estabelece que a maneira como uma rede neural funciona, tendo por base sucessivas camadas ou sequências de processamento, evoca naturalmente uma cadeia de *Markov* (COVER & THOMAS, 2006) de sucessivas representações internas intermediárias, que, juntas, podem formar uma estatística suficiente mínima aproximada. Isto significa que a rede neural pode aprender em sua estrutura (pesos) a representação dos dados de maior complexidade, ou seja, a rede neural aprenderia apenas a informação necessária, do ponto de vista da teoria da informação, contida no conjunto de dados de treinamento, para realizar a tarefa em foco. Com esse ferramental, pode-se analisar a informação preservada ou perdida pela rede durante o processo de treinamento.

Neste trabalho, utilizaremos a teoria IBDL para investigar redes neurais aplicadas em um problema de regressão, o problema de separação de fontes supervisionada (ROMANO, et.al., 2011). Nele, a rede neural tem por objetivo aprender a reconstruir um dos sinais de fonte a partir dos sinais da mistura: neste cenário, é analisado o processo de treinamento das redes através das conjecturas estabelecidas pelo IBDL. Utilizando a informação mútua estimada através da discretização das variáveis latentes, podemos observar o processo de retenção de informação de cada camada ao longo do treinamento e analisar como se relaciona a representação interna da rede e a informação contida nas mesmas, bem como seu fluxo.

Esta análise ainda envolve as particularidades das não-linearidades utilizadas na construção da rede e seus impactos no processo de treinamento, assim como, a caracterização do processo de compressão e/ou expansão da informação retida pela rede. A finalidade principal destas análises é estudar a viabilidade da teoria IBDL em problemas de regressão e as implicações decorrentes das funções de ativação utilizadas na formação da representação interna, bem como, a retenção da

informação pela rede. Por fim, a discussão é elaborada com base nos resultados descritos pela literatura recente em termos comuns aos problemas de classificação e regressão, analisando aspectos gerais das redes neurais através da teoria da informação.

O restante deste trabalho é dividido em 6 capítulos. O Capítulo 2 apresenta formalmente conceitos básicos sobre redes neurais artificiais e seu processo de treinamento. O Capítulo 3 contém os conceitos e fundamentos matemáticos da teoria da informação e estimação. O Capítulo 4 é dedicado à revisão bibliográfica e ao desenvolvimento dos principais aspectos do problema estudado. O Capítulo 5 descreve detalhadamente a definição do experimento e a metodologia utilizada para condução das simulações. No Capítulo 6, são feitas as discussões e análise dos resultados obtidos. E, por fim, o Capítulo 7 traz as principais conclusões e considerações finais do trabalho.

# 2. Redes Neurais Artificiais

## 2.1 Introdução

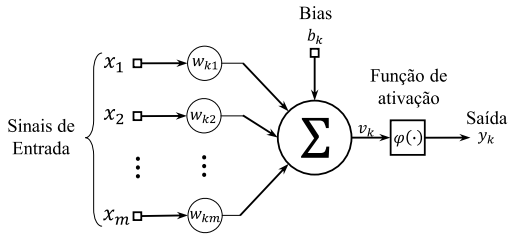
O cérebro humano é conhecido por suas capacidades extraordinárias: abstração, solução de problemas, criatividade, plasticidade e poder de processamento. Essas características fazem com que haja enorme interesse em entender e modelar sua operação, o que constitui um desafio para a comunidade cientifica (HAYKIN, 2001). Na esteira dessas investigações, foram obtidos resultados que lançaram as bases da área de pesquisa em redes neurais artificiais (RNAs). A origem da área remonta ao trabalho de (MCCULLOCH & PITTS, 1943), no qual foi construído um modelo lógico de um neurônio. Nas décadas de 1950 e 1960, a área experimentou um desenvolvimento significativo com trabalhos como os de (ROSENBLATT, 1958), mas, na sequência, houve um período de "inverno" só superado, na década de 1980, com a proposta do algoritmo de retropropagação de erro (em inglês, *error backpropagation*). (RUMELHART, HINTON, & WILLIAMS, 1986). Nos primeiros anos da década de 2010, a área de redes neurais foi impulsionada pelo paradigma de aprendizado profundo (LECUN, BENGIO, & HINTON, 2015), que levou a resultados inéditos em diversos problemas de enorme importância prática, como visão computacional e processamento de linguagem natural (LECUN, BENGIO, & HINTON, 2015). A área continua a atrair diversos pesquisadores, de modo que o estudo das propriedades de redes neurais e de seu treinamento, objeto deste trabalho, tem grande atualidade.

Faremos, na sequência, uma breve discussão dos elementos de redes neurais que são essenciais para a compreensão das análises aqui reportadas.

## 2.2 Modelo Matemático do Neurônio

As Redes Neurais Artificiais (RNAs) são modelos matemáticos inspirados na operação na maneira pela qual operam o cérebro e o sistema nervoso como um todo. A unidade fundamental de uma RNA é o neurônio artificial. O modelo emblemático de neurônio é aquele baseado nas ideias de McCulloch e Pitts (MCCULLOCH & PITTS, 1943), o qual, posteriormente, formou a base do Perceptron de Rosenblatt (ROSENBLATT, 1958). A Figura 2.1 ilustra esse modelo.

Figura 2.1 - Modelo matemático de um neurônio artificial de Rosenblatt.




Fonte: (HAYKIN, 2001).

Além das entradas, existem três elementos que compõem o modelo do neurônio apresentado: os pesos sinápticos, a junção aditiva e a função de ativação. Os pesos sinápticos $w_{kj}$, de certa forma, representam o conhecimento construído pela unidade. Eles interagem diretamente com os sinais de entrada (ou estímulos), produzindo uma resposta de acordo com essa interação. Os pesos sinápticos podem assumir valores positivos ou negativos. Os índices $k$ e $j$ são utilizados para identificar o neurônio e a sinapse, portanto a $j$-ésima sinapse do neurônio $k$.

A junção aditiva é responsável por combinar os sinais de entrada ponderados pelos pesos sinápticos (combinação linear), dando origem ao potencial de ativação $v_k$. A esta combinação linear é somado um valor de ajuste / bias $b_k$, responsável, implicitamente, por transladar a função de ativação em seu eixo, ou seja, alterar o valor do sinal $v_k$ para ajustar o disparo do neurônio, proporcionando maior nível de flexibilidade. O último elemento que compõe o modelo do neurônio é a função de ativação, a qual atua sobre o potencial de ativação, para produzir o sinal de saída da rede. Esta função, classicamente, confere um caráter não-linear aos neurônios, e, em alguns casos, restringe a saída a determinados intervalos.

Matematicamente, o potencial de ativação $v_k$ é definido como:

$$v_k = \sum_{j=1}^{m} x_j w_{kj} + b_k \tag{2.1}$$

onde $m$ é a dimensão do vetor de entrada $\boldsymbol{x}$, respectivamente $x_1, x_2, \ldots, x_m$ são os elementos desse vetor de entrada e $w_{k1}, w_{k2}, \ldots, w_{km}$ são os pesos sinápticos do neurônio $k$.

A equação completa que descreve o neurônio é escrita da seguinte maneira:

$$y_k = \varphi(v_k) = \varphi\left(\sum_{j=1}^{m} x_j w_{kj} + b_k\right) \tag{2.2}$$

onde $\varphi$ é função de ativação.

## 2.3 Funções de ativação

Conforme mencionado, as funções de ativação dos neurônios são, via de regra, responsáveis pelo caráter não-linear de uma rede neural. Por conseguinte, elas afetam de maneira crucial a capacidade de aproximação dessa rede, estabelecendo um passo de projeto de grande importância.

Nesta seção, serão apresentadas algumas das funções de ativação mais comuns, bem como suas definições matemáticas.

### 2.3.1 Função Logística

A função logística (*Sigmoid* ou *Softstep*) é definida como:

$$\varphi(v) = \frac{1}{1 + e^{-v}} \tag{2.5}$$

A função logística, semelhante a tangente hiperbólica mostrada na Figura 2.2 a), entretanto com limites diferentes, possui um comportamento de degrau suave com valores no intervalo $[0,1]$. A função é continua e diferenciável em todo seu domínio, o que é bastante relevante quando se aplicam métodos de otimização baseados nas derivadas da função custo. O caráter de degrau suave entre 0 e 1 também faz dessa função um elemento interessante quando se deseja estimar probabilidades de classificação a partir da saída de uma rede.

### 2.3.2 Tangente Hiperbólica

A função de ativação tangente hiperbólica é definida como:

$$\varphi(v) = \frac{e^{v} - e^{-v}}{e^{v} + e^{-v}} \tag{2.6}$$

A função tangente hiperbólica, mostrada na Figura 2.2 b), possui comportamento similar ao da função logística, mas sua imagem é o intervalo $[-1,1]$, o que pode ser interessante para lidar com dados com simetria relevante com respeito a zero. A função também é continua e diferenciável em todo o domínio.

### 2.3.3 *Rectified Linear Unit* (*ReLU*)

A função linear retificada (*ReLU*, do inglês *Rectifield Linear Unit*) é definida como:

$$\varphi(v) = \begin{cases} 0, se\ v \leq 0 \\ v, se\ v > 0 \end{cases} \tag{2.7}$$

Essa função, mostrada na Figura 2.2 c), ganhou notoriedade a partir das primeiras propostas de redes profundas convolucionais, como (KRIZHEVSKY, SUTSKEVER, & HINTON, 2017). É uma função nula no intervalo $[-\infty, 0)$ e com comportamento linear no intervalo $[0, \infty)$: com isso, introduz-se caráter não-linear com uma função de cálculo simples, sem saturação e com derivada informativa na parte positiva do domínio. As vantagens computacionais se mostraram expressivas no caso de redes com número elevado de camadas.

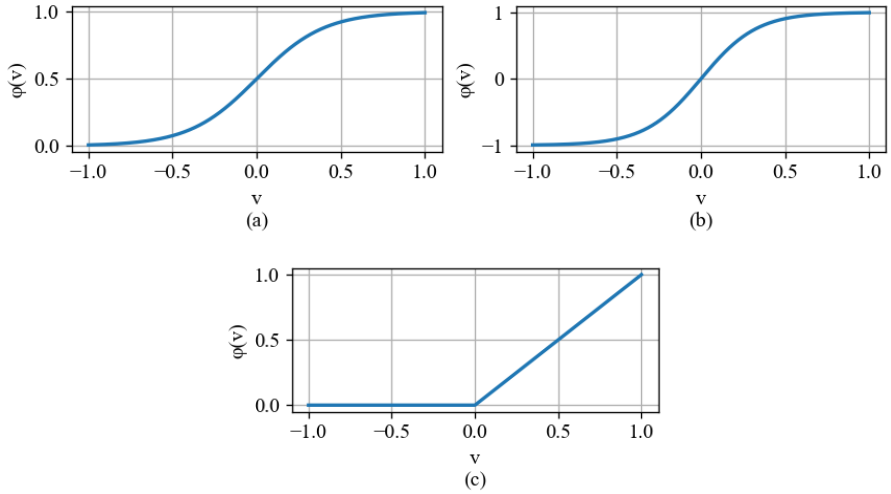



Figura 2.2 - Funções de ativação. a) Sigmoid. b) Tangente Hiperbólica. c) ReLU.

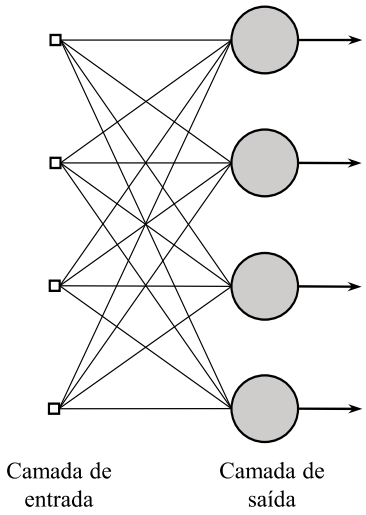
## 2.4 Arquiteturas

A estrutura formada pelos neurônios que constituem uma rede neural pode ser considerada sua arquitetura. Diferentes arquiteturas proporcionam características e comportamento distintos, permitindo uma solução diversificada de tarefas como classificação regressão, tomada de decisão e filtragem. Discutiremos a seguir uma classificação fundamental com base em (HAYKIN, 2001).

### 2.4.1 Redes Alimentadas Adiante – Camada Única

A estrutura mais simples que pode ser formada ao agrupar as unidades neurais seria uma Rede de Camada Única. Caso seja utilizado o modelo neurônio tipo *perceptron* (visto na seção 2.2), fala-se em *Perceptron* de Camada Única. Uma rede desse tipo possui uma camada de entrada e um ou mais nós de saída que formam a camada de saída. A Figura 2.3 mostra um exemplo dessa arquitetura.

Figura 2.3 - Rede alimentada adiante de camada única.

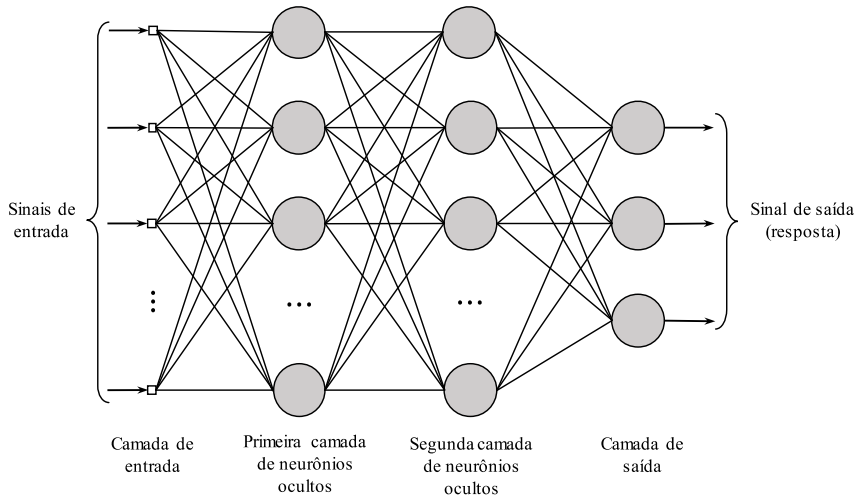



Fonte: (HAYKIN, 2001).

## 2.4.2 Redes Alimentadas Adiante – Múltiplas Camadas

Uma possibilidade mais complexa é compor uma Rede de Múltiplas Camadas alimentada adiante, mostrada na Figura 2.4. Uma estrutura desse tipo contém camadas intermediárias de neurônios entre a entrada e camada de saída. Os sinais da saída da camada de entrada formam o sinal de entrada da camada seguinte, neste caso a primeira camada oculta / intermediária. A informação pode passar por uma ou mais camadas desse tipo até chegar à camada de saída, que, por sua vez, gera a resposta da rede. Essas camadas internas / ocultas proporcionam à rede capacidade de extrair informações complexas e estatísticas de ordem elevada presentes nos dados do problema.

Figura 2.4 - Rede alimentada adiante múltiplas camadas.




Fonte: (HAYKIN, 2001).

## 2.5 Perceptron de Múltiplas Camadas (MLP)

Uma rede neural com uma única camada (além da camada de entrada) tem uma capacidade limitada de processamento, mas, se mais camadas entre a camada de entrada e a camada de saída forem adicionadas, ocorre um significativo acréscimo da capacidade de processamento, e, consequentemente, surge a possibilidade de trabalhar com problemas mais complexos.

Uma rede com neurônios tipo *perceptron* e, ao menos uma camada intermediária não-linear forma uma rede com a capacidade de aproximação universal conhecida como *Multi-Layer Perceptron* (MLP).

As redes neurais MLP utilizadas neste trabalho são do tipo *fully-connected*, ou seja, são redes densas que possuem todas as conexões entre camadas (essas redes também são chamadas de redes densas). A estrutura da rede MLP é ilustrada na Figura 2.4.

### 2.5.1 Treinamento e Aprendizagem

Uma parte importante deste trabalho envolve diretamente a análise da estrutura e do processo de treinamento da rede, bem como seu comportamento interno quando interage com o sinal de entrada.

O treinamento supervisionado (HAYKIN, 2001) da rede neural, consiste no ajuste dos parâmetros livres da rede - os pesos sinápticos - com o intuito de fazer com que a rede aproxime o sinal produzido $\hat{y}$ da resposta desejada $d$. O método para o ajuste dos parâmetros da rede, classicamente, baseia-se na correção do erro através do algoritmo de retropropagação do erro (*error back-propagation*) (HAYKIN, 2001).

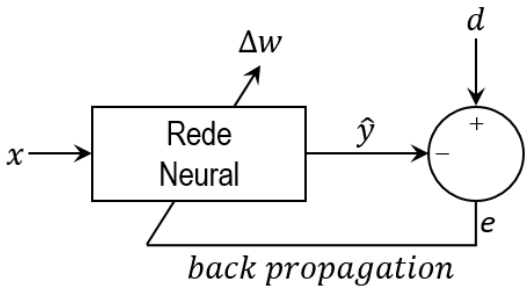



Figura 2.5 - Esquema da retropropagação do erro e de atualização dos pesos.

O processo de treinamento de uma rede consiste em duas etapas (HAYKIN, 2001): A primeira etapa é conhecida como “passo para frente”, nesta etapa os pesos sinápticos não são

modificados, as amostras $x$ são apresentadas a rede e então são calculados os sinais de ativação dos neurônios de forma individual, assim como o erro posteriormente na última camada. Na segunda etapa, conhecida como "passo para trás", o processo se inicia na camada de saída da rede, nessa etapa o sinal de erro (calculado) é retropropagado para dentro da rede até a camada de entrada. Nesse processo de retropropagação do erro, calcula-se o gradiente local de cada neurônio, e, através desse processo recursivo atualiza-se os pesos $w$ da rede através da regra delta, oriunda do cálculo diferencial. Em particular, essa regra delta possui coeficientes que ajustam a magnitude da correção dos pesos da rede influenciando na velocidade de treinamento da rede.

O processo de treinamento é realizado de forma iterativa. A cada iteração, um ou mais padrões são apresentados à rede e todas as saídas de todos os neurônios são computadas. Depois, com o erro calculado, é possível retroceder na rede e reajustar os pesos, assim se faz até a função de custo ser minimizada e a rede aprendido suficientemente bem para realizar a tarefa desejada.

# 3. Teoria da Informação

## 3.1 Introdução

A teoria da informação, proposta por Claude Shannon na década de 1940, estabelece conceitos que formaram a base da explosão das comunicações digitais (SHANNON, 1948). O estudo da relação entre probabilidade de erro e taxa de transmissão dos dados, bem como da compressão máxima que esses dados podem sofrer sem que se comprometa seu conteúdo, podem ser considerados dois elementos centrais da teoria. Os conceitos desenvolvidos por Shannon foram, ao longo dos anos seguintes, desenvolvidos e utilizados em áreas que vão do estudo de portfólios de ações (COVER & THOMAS, 2006) à análise do genoma humano (BRANDÃO, et al., 2015). Conta-se entre essas áreas também a aprendizagem de máquina (*Machine Learning – ML*) (Goodfellow, Bengio, & Courville, 2016).

Os princípios da teoria da informação podem ser utilizados em ML de diversas formas: na formulação de funções custo (e.g. entropia cruzada), na integração de métodos de poda em arvores de decisão, na definição de parâmetros de desempenho de estimação como o divergente de *Kullback-Leibler* etc.

Neste capítulo, serão discutidos conceitos que formam a base dessa teoria, e que, ademais, são necessários para o entendimento adequado do conteúdo abordado neste trabalho. Seguiremos com o desenvolvimento das principais definições em teoria da informação tendo como base (COVER & THOMAS, 2006).

## 3.2 Entropia, Entropia Conjunta e Entropia Condicional

A entropia, na teoria da informação, é uma medida de incerteza de uma variável aleatória, ou seja, é a mediada da quantidade média de informação necessária para descrer tal variável (COVER & THOMAS, 2006). Consideremos uma variável aleatória discreta $X$, o conjunto $\mathcal{S}$ dos possíveis resultados a ela associados, e uma função de massa de probabilidade (PMF) $p(x) = Pr\{X = x\}$, onde $x \in \mathcal{S}$; a entropia de $X$ é definida como segue:

$$H(X) = -\sum_{x \in \mathcal{S}} p(x) \log p(x) \tag{3.1}$$

onde o logaritmo da função pode assumir qualquer base. Geralmente, são utilizadas a base dois (medida de entropia em *bits*) e a base neperiana (a entropia é medida em *nats*). A entropia também pode ser descrita como sendo o negativo do valor esperado do logaritmo, da probabilidade:

$$\begin{aligned} -\mathbb{E}_p[\log_b p(x)] &= -\sum p(x) \log_b p(x) \\ &= H(X) \end{aligned} \tag{3.2}$$

Como consequência, $H(x) \geq 0$, pois vale que $0 \leq p(x) \leq 1$. Deve-se frisar que se assume que $0 \log 0 = 0$ (COVER & THOMAS, 2006).

Esta definição de entropia é válida apenas para o caso de uma única variável aleatória. Para lidar com mais variáveis, a extensão fundamental é no sentido do uso da entropia conjunta, que depende da função de massa de probabilidade conjunta.

Dado um par de variáveis aleatórias $X$ e $Y$ com função de massa de probabilidade conjunta $p(x, y) = Pr\{X = x, Y = y\}$, a entropia conjunta é definida como:

$$H(X,Y) = -\sum_{x \in \mathcal{S}_X} \sum_{y \in \mathcal{S}_Y} p(x,y) \log p(x,y) \tag{3.3}$$

De maneira análoga à mostrada na Equação (3.2), a entropia conjunta pode ser escrita como:

$$\begin{aligned} -\mathbb{E}_p\big[\log\big(p(x,y)\big)\big] &= -\sum_{x \in \mathcal{S}_X} \sum_{y \in \mathcal{S}_Y} p(x,y) \log p(x,y) \\ &= H(X,Y) \end{aligned} \tag{3.4}$$

Pode-se interpretar a entropia conjunta como uma expressão da incerteza associada a duas ou mais variáveis aleatórias consideradas conjuntamente.

Outra medida de entropia importante é a entropia condicional, que, em termos simples, expressa a incerteza sobre uma variável aleatória $Y$ dado que se conhece uma variável aleatória $X$:

$$H(Y|X) = -\sum_{x\in\mathcal{S}_X}\sum_{y\in\mathcal{S}_Y} p(x,y)\log p(y|x) \tag{3.5}$$

A entropia conjunta e a entropia condicional podem ser relacionadas pela seguinte equação:

$$H(X,Y) = H(X) + H(Y|X) \tag{3.6}$$

## 3.3 Entropia Relativa e Informação Mútua

Outra definição importante a ser discutida é o conceito da entropia relativa, ou distância[1] de *Kullback-Leibler* (COVER & THOMAS, 2006)entre duas funções de massa de probabilidade (PMF). A grandeza pode ser entendida como uma medida de semelhança entre duas distribuições de probabilidade ou uma medida de validade de uma delas como estimativa da outra.

A entropia relativa entre duas distribuições de probabilidade $p(x)$ e $p(y)$ é definida como:

$$D(p\|q) = \sum_{x\in\mathcal{S}_X} p(x)\log\frac{p(x)}{q(x)} \tag{3.7}$$

e novamente $0\log\frac{0}{0} = 0$.

É importante ressaltar que $D(p\|q) \geq 0$ (a igualdade é válida sse $p = q$), como pode ser demonstrado a seguir:

$$-D(p\|q) = -\sum_{x\in\mathcal{S}_X} p(x)\log\frac{p(x)}{q(x)} \tag{3.8}$$

$$= \sum_{x\in\mathcal{S}_X} p(x)\log\frac{q(x)}{p(x)} \tag{3.9}$$

$$\leq \log\sum_{x\in\mathcal{S}_X} p(x)\frac{q(x)}{p(x)} \tag{3.10}$$

$$= \log\sum_{x\in\mathcal{S}_X} q(x) \tag{3.11}$$

$$\leq \log\sum_{x\in\mathcal{S}_X} q(x) \tag{3.12}$$

$$= \log 1 \tag{3.13}$$

$$= 0 \tag{3.14}$$

[1] A rigor, o termo "distância" não poderia ser empregado, pois a definição não obedece aos requisitos de simetria e validade da desigualdade triangular.

Com as definições e conceitos estabelecidos até o momento, é possível estabelecer o conceito de informação mútua. A informação mútua mede, em essência, a quantidade de informação que uma variável aleatória contém em relação a outra variável aleatória, ou seja, o que ela pode informar a respeito da outra variável, ou, ainda, a redução de incerteza sobre uma variável dado o conhecimento de outra.

Dadas duas variáveis aleatórias $X$ e $Y$, com função de massa de probabilidade conjunta $p(x, y)$ e com probabilidades marginais $p(x)$ e $p(y)$, a informação mútua (STONE, 2013) é definida como:

$$I(X;Y) = \sum_{x \in \mathcal{S}_X} \sum_{y \in \mathcal{S}_Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{3.15}$$

Considerando a definição da distância de *Kullback-Leibler* $D(p||q)$, tem-se:

$$I(X;Y) = D\big(p(x,y)||p(x)p(y)\big) \tag{3.16}$$

A informação mútua também pode ser expressa em função da entropia $H(X)$ e da entropia condicional $H(X|Y)$. Reescrevendo a relação de probabilidade condicional como:

$$p(x|y) = \frac{p(x,y)}{p(y)} \tag{3.17}$$

E substituindo na Equação (3.15), tem-se:

$$I(X;Y) = \sum_{x \in \mathcal{S}_X} \sum_{y \in \mathcal{S}_Y} p(x,y) \log \frac{p(x|y)}{p(x)} \tag{3.18}$$

$$= -\sum_{x \in \mathcal{S}_X} \sum_{y \in \mathcal{S}_Y} p(x,y) \log p(x) + \sum_{x \in \mathcal{S}_X} \sum_{y \in \mathcal{S}_Y} p(x,y) \log p(x|y) \tag{3.19}$$

$$= -\sum_{x \in \mathcal{S}_X} p(x) \log p(x) - \left( -\sum_{x \in \mathcal{S}_X} \sum_{y \in \mathcal{S}_Y} p(x,y) \log p(x|y) \right) \tag{3.20}$$

$$= H(X) - H(X|Y) \tag{3.21}$$

Por simetria, vale também:

$$I(X;Y) = H(Y) - H(Y|X) \tag{3.22}$$

A informação mútua também pode ser escrita em função das entropias das variáveis e da entropia conjunta como segue:

$$H(X,Y) = H(X) - H(Y|X) \tag{3.23}$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{3.24}$$

Como se vê, a informação mútua é a redução da incerteza de $Y$ quando se conhece $X$, ou seja, o que $X$ informa sobre $Y$, e vice-versa. A relação entre essas medidas pode ser intuitivamente compreendida através do diagrama *Venn* da Figura 3.1.

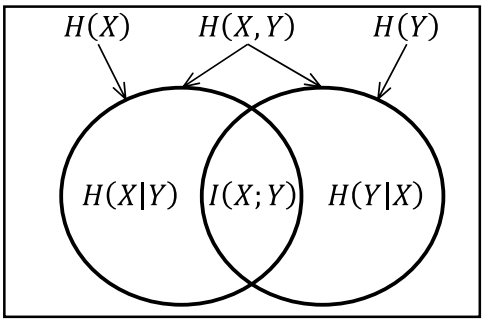



Figura 3.1 - Relação entre entropias e informação mútua.

Algumas considerações importantes devem ser feitas neste ponto. A informação mútua nunca é negativa (STONE, 2013), portanto:

$$I(X;Y) \geq 0 \tag{3.25}$$

Informação mútua nula implica em independência estatística entre as variáveis aleatórias, e uma informação mútua igual à entropia de $X$ ou $Y$ significa que as duas variáveis são idênticas, ou seja, que são versões da mesma variável. Matematicamente, isto pode ser demonstrado da seguinte maneira:

$$I(X;Y) = H(X) - H(X|Y) \tag{3.26}$$

$$= H(Y) - H(Y|X) \tag{3.27}$$

$$I(X;X) = H(X) - H(X|X) = H(X) \tag{3.28}$$

Se $X$ e $Y$ são idênticas, temos:

$$H(Y) = I(X;Y) = H(X) = I(X;X) \tag{3.29}$$

Outro ponto importante é que condicionamento a outra variável reduz a entropia:

$$0 \leq I(X;Y) = H(X) - H(X|Y) \tag{3.30}$$

$$0 \leq H(X) - H(X|Y) \tag{3.31}$$

$$H(X|Y) \leq H(X) \tag{3.32}$$

A igualdade ocorre apenas quando não há dependência estatística entre $X$ e $Y$. Do ponto de vista conceitual, vê-se que o conhecimento de uma variável aleatória jamais aumenta a incerteza sobre outra variável. No pior caso, se as variáveis forem independentes, a incerteza permanece a mesma.

## 3.4 Desigualdade de Processamento de Informação (DPI)

A Desigualdade de Processamento de Informação (COVER & THOMAS, 2006) será de suma importância para o estudo do fluxo de informação em redes com múltiplas camadas.

Considere três variáveis aleatórias $X$, $Y$ e $Z$. Elas formam uma cadeia de *Markov* nessa mesma ordem $(X \rightarrow Y \rightarrow Z)$ se a distribuição condicional de $Z$ depende apenas de $Y$ e é condicionalmente independente de $X$. Além disso, $X$, $Y$ e $Z$ formam uma cadeia de *Markov* $(X \rightarrow Y \rightarrow Z)$ se a função de massa de probabilidade conjunta puder ser escrita da seguinte forma:

$$p(x,y,z) = p(x)p(y|x)p(z|y) \tag{3.33}$$

Uma consequência direta da cadeia de *Markov*, é que $(X \rightarrow Y \rightarrow Z)$ formam nessa ordem uma cadeia se e somente se $X$ e $Z$ são condicionalmente independentes dado $Y$. Isto implica em independência condicional como segue:

$$p(x,z|y) = \frac{p(x,y,z)}{p(y)} \tag{3.34}$$

$$= \frac{p(x,y)p(z|y)}{p(y)} \tag{3.35}$$

$$= p(x|y)p(z|y) \tag{3.36}$$

A partir dessas definições pode-se definir a Desigualdade de Processamento de Dados e demonstrar que nenhuma manipulação de $Y$ pode melhorar a informação que $Y$ contêm a respeito de $X$.

A DPI estabelece que se $(X \to Y \to Z)$ então $I(X;Y) \geq I(X;Z)$. Isto significa que, se uma variável aleatória $Y$ sofrer um processamento determinístico $f(Y)$, gerando uma variável $Z = f(Y)$, esse processamento não aumentará a informação já presente em $Y$ sobre uma terceira variável $X$. Ou seja, realizar algum processamento sobre uma determinada variável pode preservar a informação contida na mesma ou diminuí-la / "destruí-la", mas nunca aumentar / "criar" informação. A demonstração é realizada através da regra da cadeia, Equação (3.6), em conjunto com a expansão da informação mútua da seguinte maneira:

$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z) \tag{3.37}$$

$$= I(X;Y) + I(X;Z|Y) \tag{3.38}$$

Uma vez que $X$ e $Z$ são condicionalmente independentes dado $Y$, isto implica em $I(X;Z|Y) = 0$. Dado que a informação mútua $I(X;Z|Y) \geq 0$, Equação (3.25), temos:

$$I(X;Y) \geq I(X;Z) \tag{3.39}$$

A igualdade é alcançada se e somente se $I(X;Z|Y) = 0$, ou seja, $(X \to Y \to Z)$ formam uma cadeia de *Markov*. A ideia também é válida para o caso de $Z = g(Y)$ onde a relação $I(X;Y) \geq I\big(X;g(y)\big)$ se $\big(X \to Y \to g(y)\big)$, ou seja, funções determinísticas de $Y$ também não aumentam a informação sobre $X$.

## 3.5 Estatística Suficiente

A partir da definição da DPI, podemos definir os conceitos de Estatística Suficiente e Estatística Suficiente Mínima (COVER & THOMAS, 2006). Para isto, é necessário o entendimento de

"estatística" nesse contexto. Suponha uma família de funções de massa de probabilidade $\{f_\theta(x)\}$ indexadas por $\theta$, e.g. funções gaussianas com diferentes parâmetros. Suponha também que $X$ seja uma amostra de uma distribuição nessa família de funções. Seja $T(X)$ qualquer função da amostra (estatística), por exemplo a média das amostras, então, $\theta$, $X$ e $T(X)$ formam uma cadeia de *Markov* nessa mesma ordem $\big(\theta \to X \to T(X)\big)$. Pela DPI, temos então que:

$$I\big(\theta; T(X)\big) \leq I(\theta; X) \tag{3.40}$$

Caso a igualdade seja assegurada, isto significa que nenhuma informação é perdida ao longo da cadeia.

A estatística $T(X)$ é dita suficiente em relação a $\theta$ se $X$ contém toda informação sobre $\theta$. Elaboremos mais a ideia. $T(X)$ é uma estatística suficiente da família $\{f_\theta(x)\}$ se $X$ for independente de $\theta$ dado $T(X)$ para qualquer distribuição $\theta$, ou seja, $(\theta \to T(X) \to X)$ formam uma cadeia de *Markov*. Isto é equivalente a assegurar a igualdade da Equação (3.45) da DPI para todas as distribuições $\theta$, tal que:

$$I(\theta; X) = I\big(\theta; T(X)\big) \tag{3.41}$$

Portanto, estatísticas suficientes preservam a informação mútua. Quando a estatística suficiente $T(X)$ é uma função de todas as outras estatísticas suficientes $U(X)$, então ela é uma estatística suficiente mínima, e pode ser entendida através da DPI como:

$$\theta \to T(X) \to U(X) \to X \tag{3.42}$$

Como consequência uma estatística suficiente mínima tem a máxima compressão da informação sobre $\theta$ na amostra.

## 3.6 Entropia Diferencial

Até o momento, consideram-se, de alguma forma, apenas medidas de informação de variáveis aleatórias discretas. Entretanto, as variáveis aleatórias contínuas são de grande importância em nosso trabalho, e, por esse motivo, alguns conceitos serão revistos para abrangê-las.

Seja $X$ uma variável aleatória contínua com função de distribuição cumulativa (CDF) $F(X) = Pr(X \leq x)$. Seja $f(x) = F'(x)$ a derivada dessa função. Se $\int_{-\infty}^{\infty} f(x)dx = 1$, então $f(x)$ é a Função de Densidade de Probabilidade (PDF) de $X$, e o conjunto onde $f(x) \geq 0$ é denominado conjunto suporte $\mathcal{S}$ de $X$. Neste caso, a entropia diferencial $h(x)$ pode ser definida como:

$$h(x) = -\int_{\mathcal{S}_X} f(x) \log f(x)\, dx \tag{3.43}$$

A entropia diferencial conjunta (STONE, 2013) de um conjunto de variáveis aleatórias $X_1, X_2, \ldots, X_n$ com densidade $f(x_1, x_2, \ldots, x_n)$, é definida como:

$$h(X_1, X_2, \ldots, X_n) = -\int f(\boldsymbol{x}) \log f(\boldsymbol{x})\, dx \tag{3.44}$$

onde $\boldsymbol{x}$ é um vetor de n variáveis.

Se $X$ e $Y$ possuem uma função de densidade conjunta $f(x, y)$, a entropia diferencial condicional $h(X|Y)$ pode ser definida da seguinte maneira:

$$h(X|Y) = -\int f(x, y) \log f(x|y)\, dxdy \tag{3.45}$$

E, analogamente à Equação (3.6), podemos escrever a entropia condicional contínua como:

$$h(X|Y) = h(X, Y) - h(Y) \tag{3.46}$$

Vale ressaltar que a entropia diferencial pode assumir valores negativos e até tendendo a infinito, portanto, é necessário cuidado ao interpretar essas equações.

A entropia relativa ou a divergência de *Kullback-Leibler* para variáveis contínuas com densidades $f(x)$ e $g(x)$ é definida como:

$$D\big(f(x) \| g(x)\big) = \int f(x) \log \frac{f(x)}{g(x)}\, dx \tag{3.47}$$

Para o caso contínuo, o divergente só é finito caso o $\mathcal{S}_f \subset \mathcal{S}_g$. Da mesma maneira a informação mútua $I(X;Y)$ é definida como:

$$I(X;Y) = \int f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy \tag{3.48}$$

e, consequentemente, também pode-se escrever:

$$I(X;Y) = D\big(f(x,y) \| f(x)f(y)\big) \tag{3.49}$$

Por fim, a informação mútua pode ser escrita em função das entropias:

$$\begin{aligned} I(X;Y) &= h(X) - h(X|Y) && (3.50)\\ &= h(Y) - h(Y|X) && (3.51)\\ &= h(X) + h(Y) - h(X,Y) && (3.52) \end{aligned}$$

## 3.7 Relação entre Entropia Diferencial e Discreta

Um método bastante comum para estimar e manipular distribuições de probabilidade de variáveis contínuas é a discretização ou *binning*. Este processo permite que se busque uma representação aproximada de densidades contínuas de um ponto de vista discreto. Todavia o processo tem consequências e implicações que devem ser levadas em consideração.

Para evitar erros conceituais, nesta seção, serão discutidas as relações entre variáveis contínuas e discretas e como elas se dão, com a finalidade de compreender como é possível o uso da discretização para o caso contínuo das variáveis aleatórias (COVER & THOMAS, 2006).

Suponha uma variável aleatória contínua $X$ com função cumulativa de distribuição $F(x) = Pr(X \leq x)$, se $F(x)$ é contínua a variável aleatória $x$ é dita contínua. Seja $f(x) = F'(x)$ quando a derivada é definida, se $\int_{-\infty}^{\infty} f(x) = 1$, então $f(x)$ é chamada de função de densidade de probabilidade para $X$. E o conjunto onde $f(x) > 0$ é chamado de conjunto suporte de $X$. Considere a variável aleatória X com densidade $f(x)$ Suponha ainda que o domínio de $X$ seja dividido em intervalos iguais (*bins*) de tamanho $\Delta$. Assumindo que a função de densidade seja contínua dentro dos intervalos.

Então pelo teorema do valor médio em cálculo, existe um valor $x_i$ em cada intervalo que a seguinte proposição é válida:

$$f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)\,dx \tag{3.53}$$

Considere então que agora a variável aleatória $X$ passe a ser uma variável discretizada $X_\Delta$, que pode ser definida como segue:

$$X_\Delta = x_i \; se \; i\Delta \leq X < (i+1)\Delta \tag{3.54}$$

Então a probabilidade quando $X_\Delta = x_i$ pode ser calculada como:

$$p_i = \int_{i\Delta}^{(i+1)\Delta} f(x)\,dx = f(x_i)\Delta \tag{3.55}$$

Agora é possível elaborar a entropia da variável discretizada da seguinte maneira:

$$H(X_\Delta) = -\sum_{-\infty}^{\infty} p_i \log p_i \tag{3.56}$$

$$= -\sum_{-\infty}^{\infty} \Delta f(x_i) \log\big(\Delta f(x_i)\big) \tag{3.57}$$

$$= -\sum \Delta f(x_i) \log f(x_i) - \sum \Delta f(x_i) \log \Delta \tag{3.58}$$

$$= -\sum \Delta f(x_i) \log f(x_i) - \log \Delta \tag{3.59}$$

Uma vez que $\sum f(x_i)\Delta = \int f(x) = 1$. Se $f(x)\log f(x)$ é integrável pelo método de Riemann (uma condição que garante que o limite é bem definido (WILCOX & MYERS, 1994) ), o primeiro termo da Equação (3.59) se aproxima da integral de $-f(x)\log f(x)$ quando $\Delta \to 0$ conforme a própria definição de integrabilidade de Riemann, provado a seguir:

Se a densidade $f(x)$ de uma variável aleatória $X$ integrável por Riemann, então:

$$H(X_\Delta) + \log \Delta \to h(f) = h(X), \qquad \Delta \to 0 \tag{3.60}$$

onde $h(X)$ é a entropia diferencial.

Ou seja, a entropia de uma variável contínua discretizada em $n$-*bits* é aproximadamente $h(X) + n$, que significa que, em média, é necessária essa quantia para descrever $X$ com precisão de $n$-*bits* (COVER & THOMAS, 2006).

Parte importante deste trabalho é a discretização de variáveis contínuas. Considere a seguinte situação: dada uma PDF $f(x)$, deseja-se encontrar uma PMF que seja aproximadamente igual à PDF de tal maneira que a entropia, informação mútua e o divergente de *Kullback-Leibler* da distribuição discretizada sejam próximos das quantidades referentes à distribuição contínua. É possível alcançar valores bem próximos do valor exato dessas medidas com a discretização correta (LEE & JO, 2021).

O processo de discretização de uma variável contínua é muito semelhante ao processo de integração de Riemann como já foi descrito anteriormente. Através da manipulação das relações estabelecidas pode-se aproximar a entropia continua através de sua discretização como segue.

Podemos reescrever a Equação (3.55) como:

$$f(x_i) = \frac{p_i}{\Delta_i} \tag{3.61}$$

Considerando que $\Delta$ é fixo, e, portanto, $\Delta_i = \Delta, \forall i$, cria-se uma relação em que a probabilidade da variável contínua se relaciona com a probabilidade da variável discreta através do delta de integração $\Delta$, onde $\Delta = dx$.

Substituindo a Equação (3.61) na Equação (3.43), e fazendo $dx = \Delta$, tem-se:

$$h(x) = -\int f(x)\log\big(f(x)\big)\,dx \tag{3.62}$$

$$\approx -\sum f(x)\log\big(f(x)\big)\,\Delta \tag{3.63}$$

$$\cong -\sum \frac{p_i}{\Delta}\log\left(\frac{p_i}{\Delta}\right)\Delta \tag{3.64}$$

$$= -\sum p_i \log\left(\frac{p_i}{\Delta}\right) \tag{3.65}$$

$$= -\sum p_i \log(p_i) + \sum p_i \log(\Delta) \tag{3.66}$$

No segundo termo da Equação (3.66), $\sum p_i = 1$, logo a entropia diferencial $h(x)$ pode ser reescrita em função da entropia discreta $H(x)$ como:

$$h(x) \cong H(x) + \log(\Delta) \tag{3.67}$$

Desta forma, a Equação (3.67) estabelece uma aproximação entre a entropia contínua e discreta, entretanto, é necessário fazer uma observação deveras importante. Comparando a Equação (3.59) com a Equação (3.67) sob a ótica do impacto da discretização nas estimativas, temos; à medida que $\Delta \to 0$, $\log(\Delta) = -\infty$, e a relação de igualdade entre a entropia diferencial e discreta da Equação (3.67) passa a não ser coerente. Ou seja, significa que ela só é válida para um número finito de intervalos de discretização, com isto pode-se alcançar valores próximos do esperado. Para exemplificar a ponderação feita, imagine uma variável aleatória que segue a distribuição normal $x \sim N(\mu, \sigma^2)$: a entropia diferencial dessa gaussiana é definida como ( (COVER & THOMAS, 2006) Exemplo 8.1.2):

$$h(x) = \frac{1}{2}(1 + \log(2\pi\sigma^2)) \tag{3.68}$$

Suponha então que a discretização de $x$ seja realizada com um $\Delta$ tão pequeno que cada amostra de $x$ esteja contida em seu próprio intervalo, isto significa que a probabilidade de cada elemento da discretização é $p_i \cong 1/n$, onde $n$ é número de amostras da distribuição decorrente da escolha do $\Delta$. Logo a entropia discreta de x é dada como:

$$H(x) = -\sum_{i=1}^{n} \frac{1}{n} \log\left(\frac{1}{n}\right) \tag{3.69}$$

$$= -\sum_{i=1}^{n} \frac{1}{n} \log(1) - \sum_{i=1}^{n} \log(n) \tag{3.70}$$

$$= -\sum_{i=1}^{n} \log(n) \tag{3.71}$$

Substituindo a Equação (3.71) na Equação (3.67), tem-se:

$$h(x) \approx -\sum \log(n) + \log(\Delta) \tag{3.72}$$

Entretanto, o primeiro termo da Equação (3.72) é a entropia de uma distribuição uniforme discreta, ou seja, matematicamente a relação estabelecida pela Equação (3.67) faz sentido, mas, na prática, ou seja, ao discretizar a variável $x$ com um $\Delta$ extremamente pequeno resulta em entropias totalmente diferentes da descrita pela Equação (3.68), pois a variável foi discretizada de maneira excessiva. Este exemplo será retomado mais adiante neste trabalho.

Através do raciocínio desenvolvido até o momento, é possível correlacionar a PDF de uma variável aleatória contínua para uma PMF, e aproximar as medidas da teoria da informação entre essas variáveis desde que haja cautela.

Até o momento, foi estabelecida uma relação matemática que vincula a entropia discreta à entropia contínua. Agora vamos estender o raciocínio para entropia conjunta, informação mútua e divergente de *Kullback-Leibler*.

Para o caso da entropia conjunta a relação contínua / discreta se dá conforme a seguir:

$$h(x,y) = \iint f(x,y)\log f(x,y)\,dxdy \tag{3.73}$$

$$\cong -\sum\sum \frac{p(x,y)}{\Delta_x\Delta_y}\log\left(\frac{p(x,y)}{\Delta_x\Delta_y}\right)\Delta_x\Delta_y \tag{3.74}$$

$$= -\sum\sum p(x,y)\log p(x,y) + \sum\sum p(x,y)\log \Delta_x\Delta_y \tag{3.75}$$

$$= H(X,Y) + \log\left(\Delta_x\Delta_y\right) \tag{3.76}$$

A multiplicação entre $\Delta_x$ e $\Delta_y$ se deve ao fato de que a probabilidade $p(x,y)$ é determinada por um pequeno volume no espaço quando se trata de variáveis bidimensionais, logo, $\Delta_x\Delta_y$ compõe a área da base desse volume. Naturalmente, é possível estender a ideia para $n$ dimensões.

A relação entre a entropia conjunta contínua e a entropia discreta está mostrada na Equação (3.76). Esta equação, em conjunto com a Equação (3.67), possibilita estender o mesmo raciocínio envolvido nas suas formulações para criar uma aproximação da informação mútua contínua da seguinte maneira:

$$I(x;y) = h(x) + h(y) - h(x;y) \tag{3.77}$$

$$\cong H(X) + \log(\Delta_X) + H(Y) + \log(\Delta_Y) - [H(X;Y) + \log(\Delta_X\Delta_Y)] \tag{3.78}$$

reescrevendo $\log\Delta_x + \log\Delta_y = \log\left(\Delta_x\Delta_y\right)$ e substituindo na Equação (3.78), temos:

$$I(x;y) \cong H(X) + H(Y) + \log\left(\Delta_x\Delta_y\right) - \left[H(X,Y) + \log\left(\Delta_x\Delta_y\right)\right] \tag{3.79}$$

$$= H(X) + H(Y) + \log\left(\Delta_x\Delta_y\right) - H(X,Y) - \log\left(\Delta_x\Delta_y\right) \tag{3.80}$$

$$= H(X) + H(Y) - H(X,Y) \tag{3.81}$$

Como se vê, a informação mútua contínua se relaciona diretamente com a informação mútua discreta. Isso é igualmente valido quando a informação é calculada pelo divergente, ou seja, $I(X;Y) = D\big(f(x,y)\|f(x)f(y)\big)$.

Anteriormente, quando foram introduzidas a Equação (3.67) e a Equação (3.76), a aproximação contínua-discreta envolvia o termo referente à "correção da transformação" $\log \Delta$. Ao aplicar estas equações para deduzir a mesma transformação para a informação mútua, é nítida a falta do termo de transformação. Obviamente, a aproximação da informação mútua depende indiretamente da discretização, ou seja, da escolha correta e coerente do $\Delta$. Portanto, a Equação (3.81) deve ser reescrita como:

$$I_{\Delta}(x;y) \cong H(X) + H(Y) - H(X,Y) \tag{3.82}$$

## 3.8 *Information Bottleneck* (IB)

O problema de restrição de informação (do inglês, *Information Bottleneck* (IB) (TISHBY, PEREIRA, & BIALEK, 1999)) é descrito como um problema que busca encontrar um equilíbrio entre acurácia e complexidade da representação de uma variável aleatória. Ele consiste em encontrar um mapeamento de $X$ para $Y$ que contenha o máximo de informação a respeito de $Y$. Este mapeamento é conhecido como uma representação de $X$, definida como $\hat{X}$.

De forma simples, significa encontrar um mapeamento que capture a informação relevante contida nos dados, essa informação é definida como sendo a informação mútua entre as variáveis do problema, ou seja, $I(X;Y)$. Pode-se entender informação relevante como sendo a informação necessária para a determinar $Y$. Por exemplo: Imagine uma fotografia de um cachorro em um parque, repleto de arvores e brinquedos. A foto neste caso é representada como sendo a variável $X$. Existe uma quantidade média de *bits* necessário para descrever o conteúdo da foto, que é dado pela entropia $H(X)$. Suponha ainda que o cachorro é representado pela variável $Y$. A informação relevante neste exemplo é descrita como sendo a informação mútua $I(X;Y)$, ou seja, dentro da foto $H(X)$ existe uma parcela de informação que é importante para identificar o animal.

Todos os elementos a mais nesta fotografia (arvores e brinquedos) são considerados informação irrelevante ( $H(X) - I(X;Y)$), ou seja, não contribuem para a descrição/identificação do cachorro. Logo, pelo problema do IB, deseja-se encontrar, um mapeamento de $X$, ou seja, $\hat{X}$, que representa parte de $X$ (ou subconjunto de símbolos), que contenha apenas o animal $(Y)$, ou seja, seria equivalente a cortar a fotografia ao redor do animal, e descartar todo o resto da foto. Posteriormente, ao

analisar as redes neurais através da informação contida nas camadas, buscar-se-á analisar justamente essa retenção de informação relevante e irrelevante pelas camadas durante o processo de treinamento.

O mapeamento mais simples possível em que $\hat{X}$ mantém o máximo de informação sobre $Y$, constitui uma estatística suficiente mínima, Seção 3.5. Para que isto ocorra, é necessário que as variáveis envolvidas componham uma cadeia de *Markov* nessa ordem $Y \to X \to \hat{X}$. Consequentemente as relações da DPI são aplicáveis.

Obter tal representação comprimida necessita que a informação entre $X$ e $\hat{X}$ seja minimizada segundo a desigualdade de processamento de dados, ou seja, a $I(X;\hat{X})$ que representa a compressão de $X$, mas sob a restrição de maximizar a informação de $Y$, $I(\hat{X};Y)$ (TISHBY & ZASLAVSKY, 2015).

Este problema pode ser formulado como um problema de minimização do lagrangiano como segue:

$$\mathcal{L}(p(\hat{x}|x)) = I(X;\hat{X}) - \beta I(\hat{X};Y) \tag{3.83}$$

onde o parâmetro $\beta$ opera como um regulador entre a complexidade da representação $I(X;\hat{X})$ e o quanto ela preserva de informação relevante $I(\hat{X};Y)$ de $Y$.

Entretanto, a estatística suficiente mínima exata só existe para famílias específicas de distribuições e.g. exponenciais. Para solucionar esse problema, foi proposto em (TISHBY, PEREIRA, & BIALEK, 1999) uma relaxação dessa formulação, permitindo que esse mapeamento não seja exato, e sim estocástico, ou seja, permitindo que exista uma estatística suficiente mínima aproximada obtida através das distribuições de probabilidade condicionais $P(T|X)$. Sendo assim, considera-se que a representação $\hat{X}$ capture o máximo de informação possível de $Y$, mas não necessariamente sua totalidade.

Em (TISHBY, PEREIRA, & BIALEK, 1999) é demonstrado que a solução para este problema variacional do IB descrito pela Equação (3.84) consiste em solucionar um conjunto de equações autocontidas para algum valor de $\beta$:

$$p(\hat{x}|x) = \frac{p(\hat{x})}{Z(x;\beta)} e^{-\beta D(p(y|x)\|p(y|\hat{x}))} \tag{3.84}$$

$$p(y|\hat{x}) = \sum_x p(y|x)p(x|\hat{x}) \tag{3.85}$$

$$p(\hat{x}) = \sum_x p(x)p(\hat{x}|x) \tag{3.86}$$

onde $Z(x;\beta)$ é um fator de normalização. A solução dessas equações pode ser obtida pelo algoritmo Arimoto-Blahut (TISHBY, PEREIRA, & BIALEK, 1999).

Para este trabalho é apenas necessário contemplar a ideia que caracteriza o problema IB, para mais detalhes existe o trabalho citado (TISHBY, PEREIRA, & BIALEK, 1999). O problema do IB é muito semelhante ao problema da taxa de distorção que também pode ser encontrado em (COVER & THOMAS, 2006).

## 3.9 Plano de informação (IP)

O plano de informação (IP), introduzido em (TISHBY & ZASLAVSKY, 2015), é um plano cartesiano no qual se representa a evolução da informação mútua das camadas ao longo do processo de treinamento das redes. Ele foi proposto para analisar a representação $p(T|X)$ e $p(Y|T)$ interna das redes neurais através da informação mútua entre as camadas e os dados de entrada $X$ e entre as camadas e os dados de saída $Y$, ou seja, a informação contida nas camadas, $I_X = I(X;T_i)$ e $I_Y = I(T_i;Y)$. O plano tem como objetivo analisar a dinâmica de retenção da informação contida nos dados, a informação relevante e irrelevante da entrada $I_X$ e a quantidade de informação retida a respeito dos dados de saída $I_Y$ preservada por essa representação interna (nas camadas) ao longo do treinamento de uma rede neural.

O estudo no IP, através da informação mútua, se mostra, segundo Tishby e Zaslavsky, interessante, por permitir uma análise qualitativa das redes neurais. Baseado em duas propriedades importantes (SCHWARTZ-ZIV & TISHBY, 2017) que são a invariância a re-parametrizações da informação mútua, ou seja, $I(X;Y) = I\big(\psi(X);\phi(Y)\big)$ para qualquer função inversível $\psi$ e $\phi$, em conjunto as propriedades da desigualdade de processamento de informação da cadeia de *Markov* que as camadas da rede formam, a análise das redes neurais no plano de informação se mostra interessante. Para qualquer variável $T$ (camada) representada por suas distribuições conjuntas em relação a entrada e a saída, ou seja $P(T|X)$ e $P(Y|T)$ tem-se um mapeamento único de $T$ para um ponto no plano de informação de coordenadas $\big(I(X;T), I(T;Y)\big)$. Quando aplicada a cadeia de Markov formada pelas $\eta$-camadas da rede, cada uma dessas camadas é mapeada para $\eta$-pontos conectados no plano de informação, formando as trajetórias de informação, que seguem as relação da DPI como mencionado.

Entretanto, camadas relacionadas através dessas funções inversíveis $\psi$ e $\phi$ podem ser mapeadas para os mesmos pontos, ou seja, redes neurais com estruturas distintas podem exibir trajetórias de informação no plano de informação iguais, como consequência não é possível utilizar o plano de informação para distinguir ou determinar estruturas de redes distintas. Porém como

mencionado, o principal uso do plano de informação é estudar qualitativamente a representação do conhecimento contido no conjunto de dados formada internamente nas redes neurais.

### 3.9.1 Leitura e Interpretação do Plano de Informação

Nesta seção, será discutida a leitura e interpretação do plano de informação, uma vez que, vários resultados do presente trabalho são apresentados em relação aos planos desenvolvidos durante os experimentos. Primeiramente, é necessário entender os elementos presentes no gráfico do plano de informação apresentado. Usaremos como exemplo ilustrativo o plano de informação da Figura 4.2.

Neste gráfico, há dois eixos: o eixo-***x***, que mede a informação mútua $I_x = I(X;T)$, ou seja, que expressa o que cada camada retém de informação disponível nos dados de entrada; o eixo-***y***, por outro lado, mede a informação mútua $I_y = I(T;Y)$, que é a aprendizagem da rede de forma geral. Conforme o erro de treinamento diminui, analogamente, pela teoria da informação, a informação sobre a saída desejada da rede aumenta. Note que o máximo de informação $I_y$, que a rede pode obter é a informação mutua $I(X;Y)$ disponível nos dados. O gráfico ainda possui uma barra que determina, na forma de um gradiente de cores, as respectivas épocas de treinamento da rede para cada camada.

As camadas da rede neural são representadas no plano de informação pelas letras $T$ ou $L$, a depender do autor, sendo que cada índice de $T_n$ representa uma das camadas da rede. O gráfico é lido da direita para a esquerda, portanto, $T_0$ identifica a primeira camada interna da rede, $T_1$ identifica a segunda camada interna da rede e assim sucessivamente até a camada de saída da rede neural $T_n$. Cada camada é representada por um segmento de pontos, descrito por $(I_X, I_Y)$, que formam uma trajetória (caminhos de informação, ou seja, as curvas ***T***), que mostram a evolução da representação interna da rede ao longo do processo de treinamento, como ilustra a Figura 3.2.

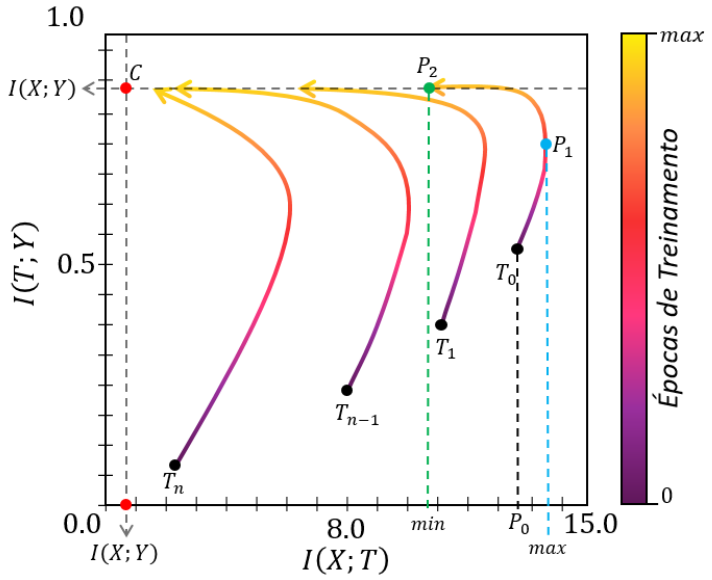



Figura 3.2 - Evolução do treinamento de uma rede neural com camadas internas no Plano de Informação (IP).

A evolução da representação interna da rede diz respeito a como as camadas aprendem, ou seja, qual a retenção de informação que as camadas apresentam ao longo do treinamento. Para entender esse processo, algumas marcações foram adicionadas, facilitando o entendimento do gráfico, e, consequentemente do processo de análise do plano de informação.

Os pontos na base das curvas identificam a informação $(I_X, I_Y)$ que cada camada da rede apresenta na inicialização da rede, ou seja, na época 0 de treinamento. Esses pontos iniciais podem variar de "posição" no gráfico a depender da inicialização da rede: geralmente, os gráficos analisados são resultado da média de diversas redes neurais treinadas, utilizando diversas inicializações de pesos. Com isto, os gráficos no plano de informação muitas vezes apresentam comportamento suave e bem delineados. Conforme o treinamento ocorre, dois comportamentos nas curvas devem ser analisados. Primeiramente, no eixo-***y***, deve se observar o processo de treinamento à luz do erro de treinamento: toda rede neural que for treinada corretamente, ou seja, convergindo e generalizando, deve apresentar incremento da informação $I_y$. Uma rede neural que não é corretamente treinada, ou seja, não apresenta redução no erro de treinamento, não apresenta ganhos de informação $I_y$. O limite máximo de informação possível que as camadas podem absorver é representado pela linha horizontal pontilhada $I(X;Y)$. Para que tenha sentido a comparação das medidas de informação das camadas, é necessário também que seja respeitada a desigualdade de processamento de informação Equação (4.1), segundo a teoria do IBDL ao final do processo de treinamento. Ou seja, a informação das camadas mais próximas da entrada $(T_0)$ devem sempre ser maiores que suas sucessoras $(T_0 \geq T_1 \geq \cdots \geq T_n)$ tanto para $I_x$ quanto para $I_y$ quando a rede está treinada.

O segundo ponto a observar é a evolução das camadas em relação ao eixo-***x***, a variação da informação $I_x = I(X;T)$ é o que mede a compressão ou expansão da informação da camada. Esses conceitos estão diretamente relacionados com à informação irrelevante descrita na Seção 3.8. Quando a informação $I_x$ aumenta, ou seja, a curva se move para a direita no eixo-***x*** do plano de informação, temos a expansão de $I_x$, que significa que a camada analisada está absorvendo informação (irrelevante) presente em $X$. Irrelevante pois, teoricamente, tudo aquilo que é aprendido além da própria $I(X;Y)$ pode ser considerado como informação irrelevante do ponto de vista da teoria da informação, entretanto, a rede irá absorver o que for necessário para uma predição correta de $Y$, ou seja, a estatística aproximada. Por sua vez, a diminuição de $I_x$, curva movendo-se para a esquerda no plano de informação, representa a compressão de $I_x$. Isto significa que a rede neural está descartando a informação irrelevante que ela aprendeu sobre $X$, ou seja, descartando o que não à ajuda na predição de $Y$.

Para ilustrar esse processo de expansão e compressão, são apresentados três pontos, $P_0$, $P_1$ e $P_2$, no caminho de informação da primeira camada interna da rede $T_0$, Figura 3.2. Cada ponto identifica um momento do treinamento da rede. $P_0$ é o ponto de informação inicial da curva/camada, ou seja, na inicialização da rede. Conforme o processo de treinamento é iniciado, a camada $T_0$, começa a aprender sobre a lei que relaciona $x$ e $y$, ou seja, começa a aprender sobre o mapeamento $y = f(x)$. Da inicialização da rede $P_0$ a $P_1$, após várias épocas de treinamento, observa-se uma *expansão* da informação $I(X;T)$ da camada, eixo-***x***, que representa a retenção de informação contida em $X$ (conjunto de treinamento), e, naturalmente observa-se o incremento da informação $I_y$ no eixo-***y***, que significa que a rede está aprendendo a determinar $Y$, ou seja, o erro de treinamento esta reduzindo. Do ponto $P_1$, de máxima informação $I_x$, para o ponto $P_2$, de mínima informação $I_x$, tem-se a redução da informação sobre $X$, ou seja, a *compressão* da informação, ou ainda, o descarte da informação (irrelevante) que não contribui para a determinação correta de $Y$. O ponto $P_2$ por fim é o ponto de convergência da camada, ou seja, após o processo de treinamento da rede, a camada $T_0$ reteve uma determinada quantidade de informação disponível em $X$ necessária para determinar corretamente $Y$. Isto é identificado observando que a camada $T_0$ atingiu o limite teórico da informação no eixo-***y***, descrita pela reta horizontal pontilhada no gráfico, que determina a informação relevante do problema $I(X;Y)$.

O ponto em vermelho definido como "$C$", na , representa o ponto de maior complexidade que as camadas podem atingir, ou seja, o mapeamento mais simples possível que a rede pode aprender (estatística suficiente mínima) contendo o máximo de informação sobre Y (eixo-***y***) e o mínimo de informação sobe $X$ (eixo-***x***). Quanto menos informação $I_x$ uma camada possui maior é sua complexidade, pois o seu mapeamento é mais simples (mínimo) possível, e quando mais informação $I_x$

a camada possui menos complexo é seu mapeamento, ou seja, não é um mapeamento mínimo, pode-se entender como um mapeamento que possui redundâncias desnecessárias.

É importante ressaltar que as curvas no plano de informação devem seguir as relações descritas pela formulação do problema IBDL, Equação (4.1) e (4.2) da Seção 4.1, bem como a própria teoria IBDL que será descrita na mesma seção. Outro aspecto importante a ser dito é com relação a dinâmica das curvas presentes no plano de informação, a Figura 3.2 ilustra de maneira didática um plano de informação semelhante ao apresentado pela Figura 4.2, de um plano de informação referente a uma rede neural treinada em um problema de classificação binário de duas classes. Outra possibilidade semelhante está presente na Figura 4.6, o plano de informação referente a outras estruturas de redes neurais é discutido ao longo do Capítulo 4, o leitor deve ter em mente que os planos de informação dependem diretamente da estrutura utilizada e do problema utilizado (dados) e, portanto, pode assumir diversas dinâmicas nas curvas de informação.

Assim como, os pontos iniciais de informação de cada camada dependem da inicialização dos pesos, ou seja, é possível que para certas inicializações a rede já apresente um certo grau de conhecimento sobre o problema, que geralmente é muito baixo. Esse ponto de partida também influencia na trajetória das camadas no plano de informação.

De forma geral o plano de informação deve ser analisado principalmente em relação ao eixo-x, pois é onde a análise de compressão ocorre, assim como a análise da complexidade da representação (mapeamento) criada internamente nas camadas.

Nesta seção, introduzimos o plano de informação e os conceitos envolvidos em sua construção e interpretação, na Seção 4, discutir-se-á com maiores detalhes como esta ferramenta pode ser utilizada em alguns problemas, assim como sobre a interpretabilidade deste gráfico e suas consequências diante de sua aplicação.

## 3.10 Estimação das medidas de Informação

Uma parte crucial deste trabalho envolve a análise da informação mútua e entropia das variáveis que são pertinentes ao nosso estudo: para estimar essas medidas é necessário o uso de algumas ferramentas para estimação. Neste contexto existem duas formas abordar o problema, uma forma mais simples e muito utilizada faz uso de histogramas para pré-processar as variáveis antes de obter as medidas de informação; a outra é utilizar métodos mais robustos e sofisticados que não envolvem probabilidades, ou seja, as medidas de informação são determinadas diretamente dos dados.

Nesta seção, serão desenvolvidos e formalizados os problemas que envolvem o processo de discretização e estimação das medidas de informação apresentadas ao longo deste capítulo. Os métodos utilizados neste trabalho são discutidos nesta mesma Seção.

### 3.10.1 Discretização (*Binning*)

O uso de histogramas para aproximação de distribuições de probabilidade de variáveis aleatórias contínuas em discretas (SCOTT, 2015), entre outras aplicações, é amplamente difundido por sua simplicidade, (PEEPLES, XU, & ZARE, 2021), (SIZINTSEV, DERPANIS, & HOGUE, 2008). Dois fatores intimamente interligados para a construção do histograma são: o número de intervalos (*bins*) $\mathcal{N}_b$ que compõem o histograma e o tamanho dos intervalos ou largura (*bin width*) $\hbar$.

Através da relação entre eles controla-se, ou através da largura dos intervalos ou pelo número de intervalos, a resolução do histograma. Este representa a distribuição de probabilidade do evento observado. Entende-se por resolução a capacidade do histograma em refletir comportamentos reais e presentes na distribuição de probabilidade analisada.

Um histograma com baixa resolução, ou seja, com poucos intervalos, não representa de maneira adequada a distribuição de probabilidade, e dessa maneira pode induzir ao erro quando se realiza qualquer tipo de análise desses dados. Por outro lado, um histograma de resolução excessiva, ou seja, um número elevado de intervalos, em relação à quantidade de dados, também levará a uma distorção. A seguir serão definidos alguns conceitos e métodos para realizar o processo de discretização.

**Discretização Uniforme**: A discretização uniforme possui a largura $\hbar$ igual para todos os intervalos que compõem o histograma. É definida como:

$$\hbar = \frac{L_{max} - L_{min}}{\mathcal{N}_b} \tag{3.87}$$

onde $L_{max}$ é o valor máximo de $x$, $L_{min}$ é o valor mínimo de $x$, $\mathcal{N}_b$ é o número de intervalos e $\hbar$ é a largura do intervalo. Naturalmente pode-se trabalhar tanto com a largura quanto o número de intervalos para determinar a melhor resolução do histograma.

**Discretização Não Uniforme:** A discretização não-uniforme é um método em que a largura do intervalo $\hbar$ não é uniforme no domínio de discretização, como ocorre no método do *Baysian-blocks* (SCARGLE, et. al., 2013) que utiliza a inferência bayesiana e o princípio estatístico de estimação por

ponto de mudança (do inglês *change-point determination*) para determinar a largura dos intervalos. Este método é utilizado para análise de séries temporais.

Além dos conceitos referentes ao tamanho e número de intervalos temos ainda conceitos que se referem à dinâmica do processo de determinação desses parâmetros, que podem ser determinados de forma estática ou dinâmica (adaptativa):

**Discretização Estática:** A discretização estática é uma abordagem na qual os parâmetros $L_{max}$, $L_{min}$ e $\hbar$ ou $\mathcal{N}_b$ são estáticos ao longo do processo de discretização. Ou seja, uma vez que é definido o espaço de discretização (domínio) $[L_{max}, L_{min}]$ e determinado o número de intervalos ou a largura nenhum destes parâmetros são alterados, ou seja, este procedimento padroniza a escala de trabalho quando utilizada para analisar eventos dinâmicos onde é observado variações nas distribuições de probabilidade.

**Discretização Adaptativa:** A discretização adaptativa tem lugar quando a largura $\hbar$ do intervalo dentro do domínio de discretização é ajustada de maneira automática com base em algum parâmetro estatístico, baseada em algum método de busca ou de divergência de KL. Dentre os métodos adaptativos temos alguns baseados em estatística como o método de Scott (SCOTT, 1979) que incorpora em sua formulação o desvio padrão dos dados para determinar o valor de $\hbar$ e a regra de Doane (DOANE, 1976) que utiliza o terceiro momento central de distorção para determinar o número de intervalos $\mathcal{N}_b$. Já nos métodos baseados em busca temos como exemplo o método de Shimazaki e Shinomoto (SHIMAZAKI & SHINOMOTO, 2007) que envolve buscar o valor ótimo de intervalos baseado na minimização de uma função custo que em sua formulação leva em conta a média e a variância das amostras contidas e cada intervalo de discretização.

### 3.10.2 Método Discreto de Estimação de Medidas de Informação

A estimação de medidas de informação como entropia e informação mútua envolvendo variáveis aleatórias discretas pode ser feita através da aplicação direta das definições desenvolvidas no Capítulo 3. Para estimação dessas medidas de informação para variáveis contínuas, é necessário o uso de ferramentas que nos possibilite estimar as densidades de probabilidade associadas a variável estudada.

Uma das maneiras mais comuns utilizadas é o processo de discretização por histogramas (SCOTT, 2015) descrito na Seção 3.10.1. O processo de discretização de uma variável contínua exige que se determinem os limites inferior e superior $[L_{min}, L_{max}]$ e determinar o número de intervalos $\mathcal{N}_b$ a fim de criar um histograma e posteriormente obter as probabilidades através da normalização das

frequências obtidas. Com as probabilidades estimadas é possível utilizar as equações do Capítulo 3 para calcular as medidas de informação como entropia e informação mútua.

O objeto de estudo deste trabalho é o processo de treinamento de redes neurais sob a ótica da teoria da informação, a estimação dessas medidas envolve diretamente o uso das probabilidades de certas variáveis e das probabilidades conjuntas entre elas. O processo para obtenção dessas probabilidades é descrito a seguir:

Considere o modelo de rede descrito pela Figura 3.3.

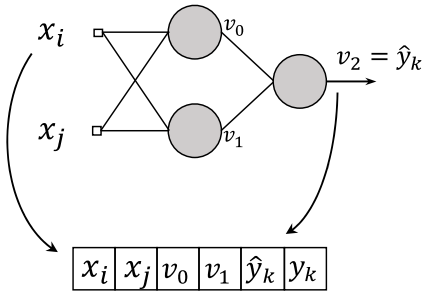



Figura 3.3 - Processo de discretização de uma rede neural com duas entradas e uma camada interna com dois neurônios e única saída [2,2,1].

Ao longo do treinamento da rede, pares de dados $(x_i, x_j)_k$ do conjunto de dados de entrada $X$ que contém $k$ amostras, são apresentados à rede. A interação dos dados com os pesos sinápticos e com a função de ativação produz, na primeira camada interna da rede, as saídas de seus respectivos neurônios $v_o$ e $v_1$. Essas saídas, por sua vez, interagem com os pesos da camada de saída e produzem $v_2 = \hat{y}_k$.

Estes valores são então convertidos para variáveis discretas a partir do processo de discretização. Para discretizar cada elemento que compõe o vetor de dados primeiramente é necessário determinar o domínio de discretização. Para o conjunto de dados $X$ segue que os valores mínimo e máximo do domínio de discretização são determinados pelos próprios valores das respectivas dimensões de $X$. Para os valores das ativações $v_0$, $v_1$ e $v_2 = \hat{y}$ os limites de discretização são determinados pela função de ativação utilizada em cada camada, por exemplo, a função sigmoid tem seus limites entre 0 e 1, portanto a discretização se dá neste intervalo. Já para o conjunto de dados $Y$, o domínio de discretização é exatamente o mesmo que o da camada de saída da rede para que se possa comparar de forma correta as distribuições de probabilidade de $\hat{Y}$ com $Y$.

A escolha do número de intervalos pode se dar de maneira arbitrária como em muitos artigos que serão discutidos no Capítulo 4, ou pode ser adaptativa, caso se escolha utilizar algum dos métodos mencionados na Seção 3.10.1. A metodologia para a escolha do número de *bins* será discutida no Capítulo 5.

Com esses valores um vetor de dados é construído assim como ilustrado pela Figura 3.3. Cada amostra de dados do conjunto $X$ apresentado a rede ao longo de cada época gera um vetor e ao final de cada época uma matriz é formada. O próximo passo é transformar esses vetores em uma distribuição de probabilidade, que é relativamente simples. Cada vetor é interpretado com uma única entidade, ou seja, cada sequência de valores que compõe os vetores representam uma coordenada no plano $\mathbb{R}^n$ onde $n$ é o tamanho do vetor. Portanto, cada vetor é contabilizado em termos de frequência, ou seja, quantas vezes esta mesma sequência se manifesta em uma época de treinamento com as *k* amostras do conjunto $X$ de dados de entrada apresentadas a rede. Desta forma, basta normalizar a frequência de cada vetor único pelo número total de frequências para obter as respectivas probidades de cada vetor, e assim, obter as distribuições de probabilidade.

De forma resumida e ilustrativa esse processo está representado na Figura 3.4.

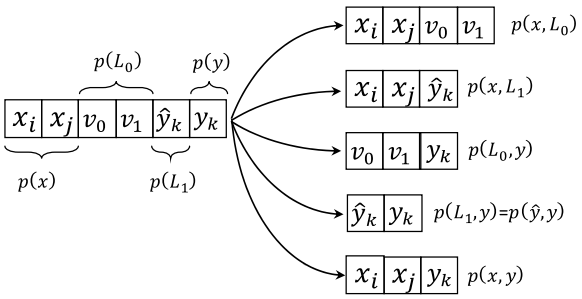



Figura 3.4 - Extração das probabilidades e probabilidades conjuntas dos neurônios.

De posse das distribuições de probabilidade de $X$, $Y$ e $L_i$ (que representa a saída da $i$-ésima camada da rede), basta calcular as entropias e informação mútua que posteriormente são utilizadas para construção do plano de informação.

Outra maneira de obter essas medidas é através de uma estimação direta, que consiste em utilizar métodos mais sofisticados que a discretização para calcular as medidas de entropias e informação mútua diretamente dos dados, e não de uma distribuição de probabilidades como descrito anteriormente. Para tal, podem ser utilizados métodos como o *Multivariate Extension of Matrix-Based Rényi's α-Order Entropy Functional* (YU, et.al., 2020).

# 4. Análise de Redes Neurais Artificiais Através da Teoria da Informação

Neste capítulo, serão discutidos aspectos do problema do *fluxo de informação* em redes neurais, que serão importantes para a análise subsequente, que terá por foco o fenômeno de "gargalo de informação" (ou, restrição de informação) (TISHBY, PEREIRA, & BIALEK, 1999) (IB, do inglês *Information Bottleneck*). Ao longo dos últimos anos, o assunto recebeu atenção em estudos em redes neurais voltadas para problemas de classificação, e a literatura tem se enriquecido com trabalhos que mostram a complexidade deste problema. A discussão do problema em si e das variáveis que o compõem serão explicadas mediante a análise desta literatura recente, e desenvolvida com o intuito de mostrar ao leitor os principais aspectos do problema de gargalo de informação aplicado a redes neurais IBDL (TISHBY & ZASLAVSKY, 2015) (do inglês, *Information Bottleneck Theory of Deep Learning*) e como este problema tem sido explorado.

## 4.1 Hipótese Inicial

Em 2015, Tishby e Zaslavsky desenvolveram um estudo sobre o treinamento de Redes Neurais Profundas (DNN) através da teoria da informação (*Information Bottleneck Theory of Deep Learning*- IBDL) (TISHBY & ZASLAVSKY, 2015). Os autores argumentam que uma rede neural, no treinamento supervisionado, teria naturalmente a capacidade de capturar e representar a informação relevante contida no conjunto de dados através de sua estrutura de maneira eficiente, ou seja, extrair uma estatística suficiente mínima aproximada. Com isso, o processo de treinamento pode ser entendido, sob o aspecto da teoria da informação, como a busca de um compromisso entre compressão de informação relevante e capacidade de predição da rede.

Para desenvolver essa tese, realiza-se uma análise do processo de treinamento através da teoria do gargalo de informação (IB). O problema do IB, análogo ao problema de taxa distorção (COVER & THOMAS, 2006), consiste em encontrar o mapeamento de $X$ para $Y$ que contenha o máximo de informação relevante sobre $Y$.

Da informação total contida nos dados de entrada $X$ de uma rede, apenas uma parcela é, de fato, significativa (informativa): aliás, isso é facilmente visualizado ao se observar a dimensionalidade da entrada e da saída na maioria dos problemas. Como Tishby e Zaslasvsky explicam, geralmente a entrada de uma rede possui alta dimensionalidade (vetor com muitos elementos), e, portanto,

considerado como uma representação de baixo nível, uma vez que se supõe que $X$ possui muita informação irrelevante em decorrência da alta dimensionalidade. A saída $Y$, por sua vez, possui baixa dimensionalidade (vetor com poucos elementos, por exemplo rótulos binários), porém contém muita informação, ou seja, carrega consigo a informação necessária para distinguir uma classe de outra: esta diferença de dimensionalidade indica que grande parte do conteúdo do conjunto $X$ não contém informações relevantes sobre $Y$, ou seja, as características realmente importantes contidas em $X$ estão dispersas e seriam difíceis de serem extraídas. Por exemplo, como saber quais elementos de uma foto são bons para determinar a qual classe ela pertence. Parte do sucesso das redes neurais em extrair essa informação se deve ao seu funcionamento, baseado em camadas que formam uma sequência em etapas de processamento para a construção de características.

O estudo das redes MLP, segundo a abordagem proposta, sustenta-se exatamente na forma como as redes operam. Cada camada interna da rede processa a informação da camada anterior, formando uma cadeia de *Markov* segundo a hipótese dos autores, como é ilustrado pela Figura 4.1. Como consequência direta dessa condição, as propriedades da DPI, discutidas na Seção 3.4, são válidas: com isso a informação sobre $Y$ perdida em cada camada durante esse sucessivo processamento da rede não pode ser recuperada pelas camadas seguintes, exatamente como é mostrado pela Equação (3.39). A igualdade só é alcançada quando cada camada forma uma estatística suficiente da entrada, isto significa dizer que cada camada deve conter a informação mais relevante e mais compacta possível (alta complexidade) a respeito de $X$. Expandindo a Equação (3.39) para todas as camadas da rede, os autores propõem o estudo das redes neurais através do conjunto de inequações a seguir:

$$I(X;Y) \geq I(T_1;Y) \geq \cdots \geq I(T_n;Y) \geq I\left(\hat{Y};Y\right) \tag{4.1}$$

$$H(X) \geq I(X;T_1) \geq \cdots \geq I(X;T_n) \geq I\left(X;\hat{Y}\right) \tag{4.2}$$

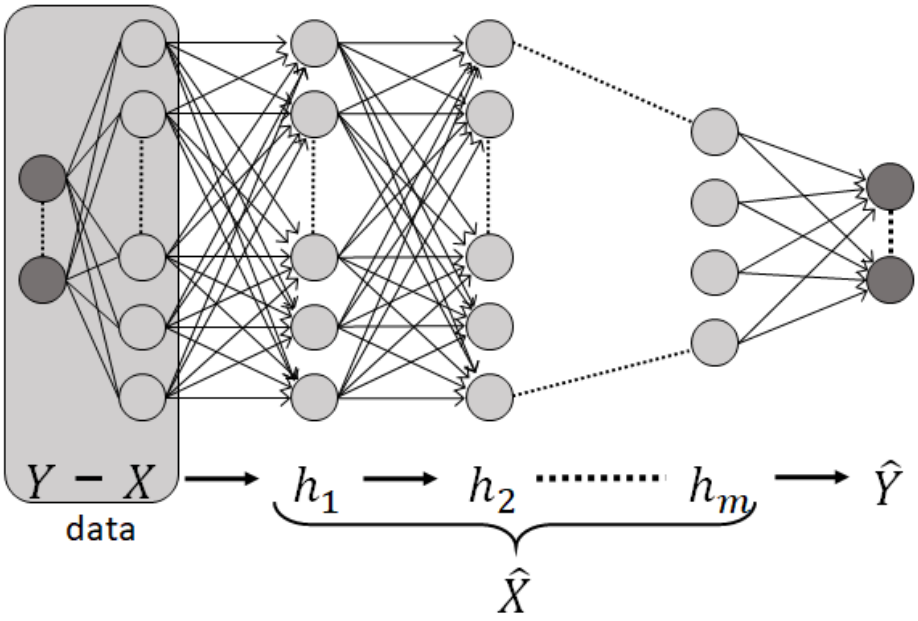
onde $X$ representa o conjunto de dados de entrada, $T_n$ as camadas, ou seja, a representação interna da rede, $Y$ o conjunto de dados da saída e $\hat{Y}$ a predição da rede.

De maneira geral, uma rede com a estrutura mais simples possível para lidar com a tarefa abordada, teria a capacidade de extrair a informação latente contida nos dados em sua forma mais complexa, ou seja, suas representações internas contêm a representação de $X$ mais compacta com o máximo de informação relevante sobre $Y$ ou aproximadamente a estatística suficiente mínima.

Como mencionado, através dos conceitos da teoria da informação é possível realizar o estudo dessa informação para analisar a eficiência da representação interna da rede ao longo do

processo de treinamento, através do plano de informação introduzido pelos próprios autores em (TISHBY & ZASLAVSKY, 2015).

Figura 4.1 - Rede Neural Artificial segundo a teoria da informação.




Fonte: (TISHBY & ZASLAVSKY, 2015).

## 4.2 O princípio da compressão

Com base no trabalho de Tishby e Zaslavsky (2015), Schwartz-Ziv & Tishby (2017) aprofundam o estudo das redes neurais pelo plano de informação em busca de uma análise mais detalhada da dinâmica do processo de treinamento das redes MLP e da representação interna da rede ao longo desse processo.

Segundo os autores, o estudo no plano de informação é pertinente por apresentar algumas propriedades especificas. Em primeiro lugar, a informação mútua é invariante a reparametrizações invertíveis, ou seja, para as representações internas da rede que preservam informação, existem representações equivalentes mesmo quando os neurônios individualmente trabalham com características totalmente diferentes. Por essa razão, o plano de informação é construído por um par de coordenadas $I_X = I(X;T)$ e $I_Y = I(T;Y)$, e as representações internas são definidas pelo encoder $P(T|X)$ e pelo decoder $P(\hat{Y}|T)$. A análise dessas medidas permite medir e comparar a eficiência das representações internas nas redes para diferentes estruturas de rede.

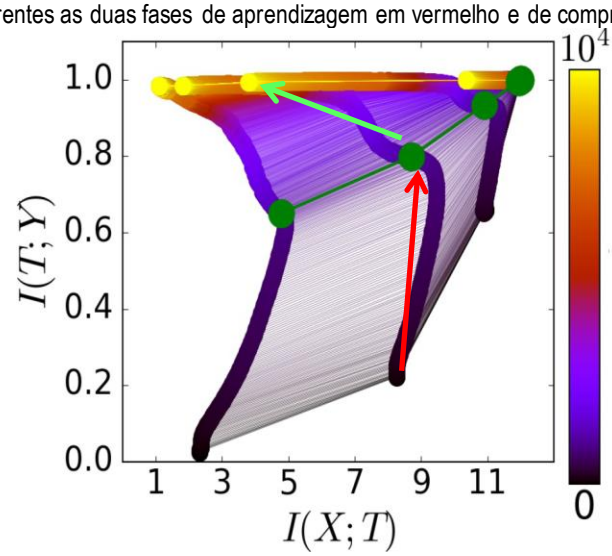
Entretanto, representações internas ótimas, ou seja, estatísticas suficientes mínimas exatas só existem para distribuições de probabilidade bem particulares que pertencem a família de distribuições

exponenciais. Em (TISHBY, PEREIRA, & BIALEK, 1999) o problema do IB é relaxado permitindo que o mapeamento seja estocástico, ou seja, o problema do IB pode ser formulado pelas probabilidades $p(t|x)$, $p(t)$ e $p(y|t)$, possibilitando que tal mapeamento capture o máximo de informação possível mas não necessariamente toda ela, portanto, permitindo alcançar uma estatística suficiente mínima aproximada.

Tais representações, que formam uma cadeia de *Markov*, são alcançadas através do treinamento da rede, portanto, da minimização da função custo através do ajuste dos pesos sinápticos. Para o desenvolvimento do trabalho os autores utilizaram uma rede MLP treinada através do gradiente descendente estocástico (SGD) pelo algoritmo de retorpropagação, discutido na Seção 2.5.1, com a função custo da entropia cruzada.

Durante o processo de treinamento cada camada da rede é tratada como uma única distribuição de probabilidade, ou seja, ao longo do treinamento - épocas - cada camada é caracterizada pela sua informação mútua $(I_X, I_Y)$, decorrente das probabilidades $p(t)$, $p(t|x)$ e $p(\hat{y}|t)$, ou seja, sua representação interna. Cada uma delas possui um caminho distinto no plano de informação, a Figura 4.2 ilustra esses diferentes caminhos.

Figura 4.2 - Plano de Informação para redes neurais MLP. Compressão de $I_X$ a partir da marcação em verde. Setas adicionais referentes as duas fases de aprendizagem em vermelho e de compressão em verde.



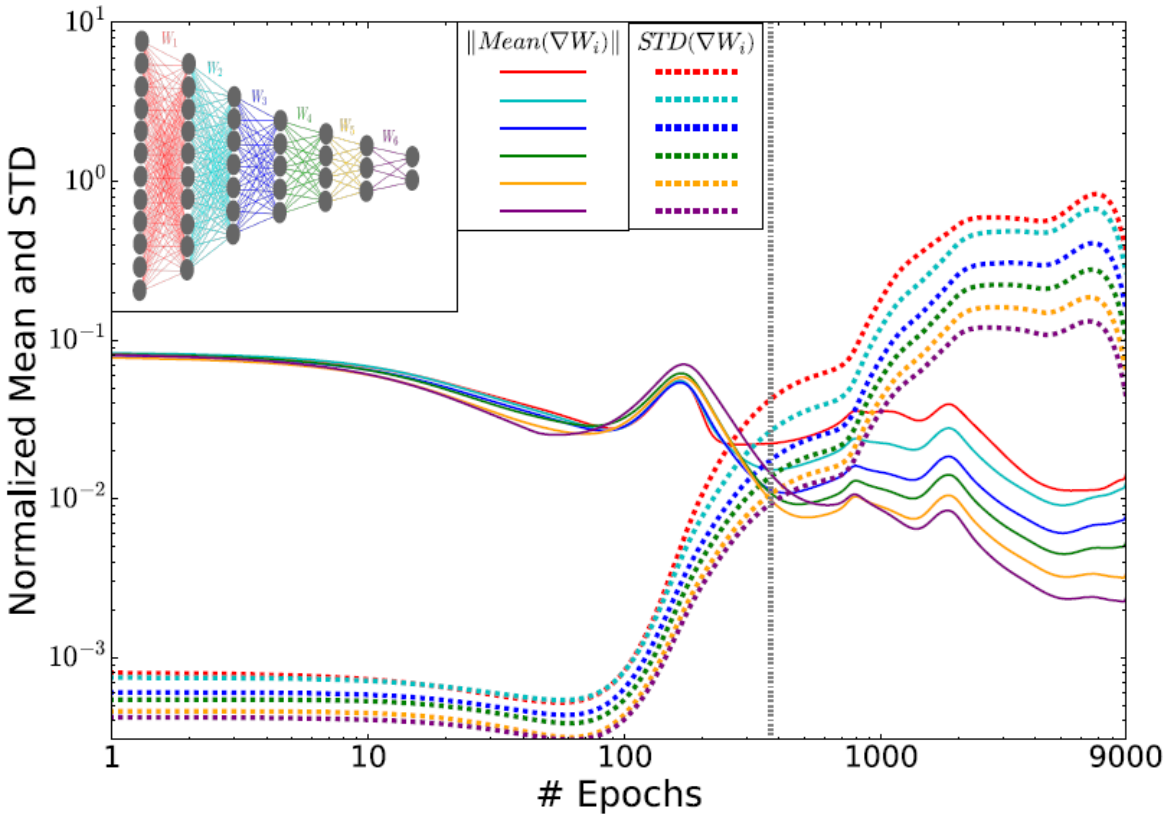

Fonte: (SCHWARTZ-ZIV & TISHBY, 2017).

Ao observar e analisar o plano de informação, com as curvas traçadas por cada camada durante o processo de treinamento, os autores notaram um comportamento peculiar. No primeiro momento percebe-se que há um incremento da informação a respeito dos dados de saída $I_Y$ (seta vermelha), o que é naturalmente esperado, uma vez que a rede está aprendendo como estimar $Y$, segundo os autores isto se deve também ao uso da entropia cruzada como função custo, em conjunto

com este incremento de $I_Y$ observa-se uma expansão modesta da informação a respeito dos dados de entrada, descrita por $I_X$. Entretanto, no segundo momento nota-se uma diminuição considerável da informação em relação aos dados de entrada $I_X$ (seta verde), ou seja, uma compressão da representação interna das camadas em relação aos dados de entrada, contudo, não existe nenhum tipo de regularização durante o processo de treinamento que justifique tal comportamento.

A partir dessas observações, os autores, em busca de uma explicação voltaram suas atenções para o gradiente dos pesos ao longo do treinamento, Figura 4.3. Ao analisar a média e o desvio padrão do gradiente juntamente com o plano de informação os autores levantam a hipótese de que a otimização pelo SGD possuía duas fases bem definidas e distintas.

Figura 4.3 - Média e desvio padrão dos pesos sinápticos ao longo do treinamento. Drift phase antes da marcação pontilhada e difussion phase depois da marcação.




Fonte: (SCHWARTZ-ZIV & TISHBY, 2017).

A primeira fase do gradiente é chamada *drift phase*, esta fase é responsável pelo aumento da informação $I_Y$ em todas as camadas uma vez que o erro de treinamento é reduzido rapidamente. Esta fase é caracterizada por ser bem curta e por possuir uma média elevada com baixo desvio padrão - ruído, a ela está associada, de maneira correspondente, a denominação *Empirical Error Minimization* (ERM) no plano de informação.

A segunda fase do gradiente é chamada de *difussion phase*, ela é, segundo os autores, responsável pela compressão da informação $I_X$, ou seja, a fase de *Representation Compression* (RC) no plano de informação. Esta segunda fase tem um comportamento distinto da primeira: a média do gradiente é muito pequena e o ruído é predominante. Essas flutuações se comportam como um ruído gaussiano que, por sua vez, adiciona ruído aos pesos sinápticos da rede, e faz com que os pesos evoluam como um processo de *Wiener* - movimento Browniano – sob o erro de treinamento ou restrição da informação $I_Y$.

Esse processo ruidoso pode ser descrito segundo as equações de *Focker-Planck*, e possui distribuição estacionária que maximiza a entropia da distribuição dos pesos sob a restrição do erro de treinamento. De maneira indireta, portanto, o gradiente maximiza a entropia condicional $H(X|T)$, ou seja, por consequência minimiza a informação mútua $I(X;T)$. O nome dado a este processo em que a entropia é maximizada através da adição de ruído é relaxação estocástica.

As duas fases podem ser vistas analogamente como: 1) uma fase de aprendizagem do conhecimento - ERM - e 2) uma fase de refinamento do conhecimento - RC -, ou seja, inicialmente a rede aprende de maneira geral sobre a relação entre $X$ e $Y$ e depois aperfeiçoa aquele conhecimento tornando-o mais condensado, mas não menos informativo. Segundo a teoria da informação, trata-se de aumentar a complexidade da representação.

Com base nos dados obtidos os autores chegaram a algumas conclusões. A mais contundente é de que a generalização da rede estaria condicionada à complexidade da representação interna, ou seja, uma rede treinada apresentando uma boa generalização estaria associada a representações internas complexas. Algumas conclusões a respeito da estrutura da rede também são pontuadas. A adição de camadas internas reduz drasticamente a quantidade de épocas necessárias para que a rede alcance uma generalização boa. Foi notado também que a fase de compressão tende a ser mais curta e mais rápida quando uma camada processa a representação de outra camada, ou seja, as camadas mais próximas da saída da rede são as que evoluem com maior rapidez.

## 4.3 Uma nova perspectiva

O trabalho desenvolvido por Schwartz-Ziv & Tishby (2017), apesar de ser bastante interessante e promissor foi desenvolvido tendo como base um experimento muito específico: um problema de classificação binário utilizando apenas a tangente hiperbólica como função de ativação interna da rede nos testes e estimando a informação mútua através da discretização uniforme das camadas internas da rede. Essa exploração inicial do problema de redes neurais como uma cadeia de

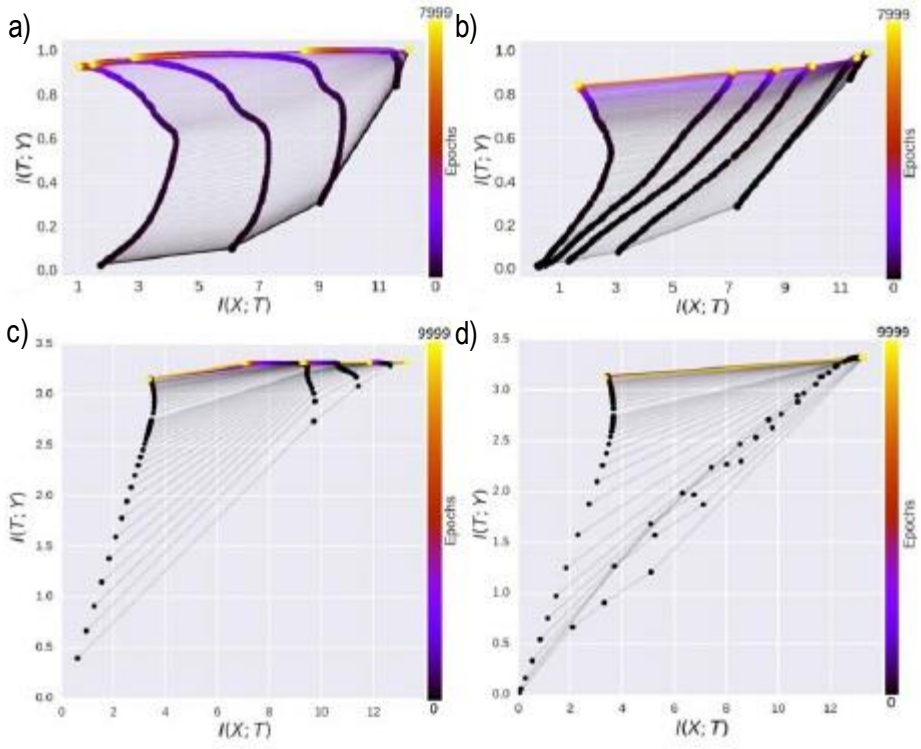
*Markov* apresentava algumas questões em aberto, como a falta de análises com outras funções de ativação, outros tipos de problemas e conjuntos de dados, métodos diferentes de treinamento e otimização e diferentes estimadores de entropia.

Em (SAXE, et al., 2018), os autores notaram que esses pontos eram pertinentes e necessitavam de detalhamento. Uma exploração mais completa dos elementos envolvidos no problema certamente auxiliaria a endossar a hipótese que Schwartz-Ziv e Tishby levantaram em seu trabalho, ou, abriria um novo leque de possibilidades para se explorar.

O ponto inicial de (SAXE, et al., 2018) consistiu em analisar o plano de informação para outra função de ativação nas camadas internas da rede utilizando o mesmo problema. Para este experimento, os autores utilizam a função *ReLU* e discretizaram sua saída em 100 intervalos equidistantes entre os valores mínimo e máximo atingidos por ela ao longo de todo o treinamento, com o uso do método SGD para ajuste dos pesos. Ao replicar o experimento de Schwartz-Ziv e Tishby com esta configuração, o plano de informação não demonstrou a existência das duas fases descritas anteriormente. Em específico, não houve a fase compressão no plano de informação, identifica-se apenas um aumento da informação ao longo do treinamento Figura 4.4.

Para verificar esse resultado, os autores utilizaram ainda um estimador robusto - *Kernel Density Estimator* (KDE) - para realizar os cálculos da informação mútua: os testes foram realizados com uma rede treinada com o conjunto de dados MNIST. Os resultados permaneceram similares, ou seja, ausência de compressão, apenas apresentando um incremento de informação. A Figura 4.4 mostra os resultados obtidos: Em "a", temos a reprodução do experimento utilizando a função tangente hiperbólica como ativação das camadas internas, a compressão é presente (redução de $I(X;T)$). Por sua vez, "b" corresponde ao IP obtido utilizando a função *ReLU:* a fase de compressão está ausente nas camadas internas, apenas presente na camada de saída que utiliza a função sigmoide. Em "c" e "d", foram utilizadas a função tangente hiperbólica e *ReLU* respectivamente, mas com o conjunto MNIST e o estimador KDE, não apresentando a fase de compressão em nenhuma camada interna.

Figura 4.4 - Plano de informação Tanh vs. ReLU. a) Função Tanh discretizada. b) Função ReLU discretizada. c) Função Tanh, conjunto MNIST usando o estimador KDE. d) Função ReLU, conjunto MNIST usando o estimado KDE.




Fonte: (SAXE, et al., 2018) .

Para estudar o comportamento destoante que a *ReLU* apresentou, os autores fizeram um experimento com um modelo simples de rede $[1, 1, 1]$ uma entrada, um neurônio interno e uma saída, portanto, $\hat{y} = f(\omega_1 X)$ onde $f(\cdot)$ é a não linearidade. Com isso a informação mútua da camada em relação aos dados de entrada pode ser calculada diretamente pela Equação (3.1). Para este teste utilizaram uma distribuição Gaussiana $X \sim \mathcal{N}(0,1)$ como entrada, e a função tangente hiperbólica discretizada em 30 intervalos.

Neste modelo mínimo, a saída da rede é apenas uma Gaussiana reescalada, ou seja, a escala da gaussiana formada na saída da rede depende exclusivamente dos pesos, que podem levar a saída da rede ao regime de saturação da função de ativação a depender da magnitude dos pesos. Para a tangente hiperbólica, a informação mútua apresentou compressão, e, para a *ReLU* a informação é apenas crescente. Isto porque para pesos próximos de zero a atividade neural está localizada na parte linear da tangente hiperbólica, mas, quando o peso aumenta, reescalando a Gaussiana, a informação cresce juntamente com a magnitude dos pesos. Para pesos elevados, a tangente hiperbólica atinge a saturação, e, como resultado da discretização, a distribuição é localizada em apenas dois intervalos, como resultado, é obtido algo em torno de 1 *bit* de entropia.

Esse valor é resultado de uma distorção da distribuição de probabilidade da camada interna, que é vista como a compressão na tangente hiperbólica. Para a *ReLU,* metade da distribuição é nula, e

a outra metade é uma distribuição gaussiana reescalada, e, por isso, a entropia só aumenta conforme os pesos aumentam.

Os autores mostram, com esses resultados, que a fase de compressão é consequência de dois fatores em conjunto.

O primeiro está associado à função utilizada nas camadas internas, as funções que apresentam dupla saturação como a tangente hiperbólica e outras, demonstram a fase de compressão quando atingem a saturação. O segundo fator está associado ao processo utilizado para calcular a informação mútua, que pode acentuar ou induzir a fase de compressão. Ou seja, para estimar a informação mútua Schwartz-Ziv e Tishby (2017) utilizaram um processo de discretização - *binning* – das funções de ativação da rede. Esse procedimento é bastante utilizado para estimar distribuições de probabilidade, mas ele pode não ser adequado, como é mostrado pelos próprios autores Saxe et al. (2018).

Por esta razão, os autores utilizaram um estimador de informação mais robustos que o estimador discreto (*binning*), utilizando o KDE para estimar as medidas de informação, foi constatada a presença da fase de compressão para redes com função tangente hiperbólica camadas internas da rede, Figura 4.4 c), e a ausência desta compressão para redes com a função *ReLU*, Figura 4.4 d). Foi concluído através desse teste, portanto, que a compressão identificada no plano de informação estaria associada a saturação das funções de ativação e não ao ruído do gradiente como suposto por Schwartz-Ziv e Tishby.

Especificamente para estudar este ponto referente ao gradiente estocástico e sua influência nas fases de treinamento observadas através do plano de informação, os autores realizaram testes utilizando treinamento por lotes (*batch*). Os resultados no plano de informação das redes treinadas por lotes são exatamente iguais aos resultados das redes treinadas de forma estocástica, ou seja, compressão na tangente hiperbólica e ausência desta mesma fase para *ReLU*. Além disso, eles analisaram a relação sinal-ruído (SNR) do gradiente tanto para tangente hiperbólica como para *ReLU*. A SNR é definida como a razão entre a média do gradiente com o desvio padrão do gradiente ao longo do treinamento.

$$m = \left\| \left\langle \frac{\partial E}{\partial W} \right\rangle \right\|_F \tag{4.3}$$

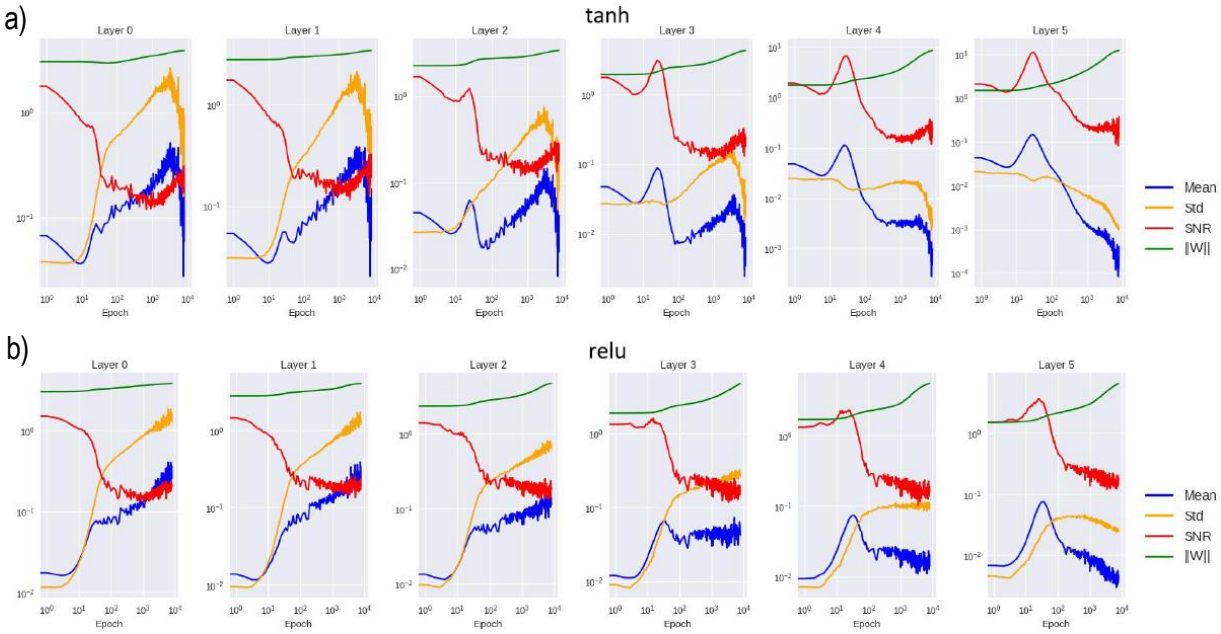
$$s = \left\| std\left( \frac{\partial E}{\partial W} \right) \right\|_F \tag{4.4}$$

$$SNR = \frac{m}{s} \tag{4.5}$$

onde $\langle\cdot\rangle$ representa a média, $std(\cdot)$ o desvio padrão de cada elemento da matriz e $\|\cdot\|_F$ representa a norma de *Frobenius*.

Ao analisar o SNR das funções citadas, os autores notaram que as duas fases presentes no gradiente *drift phase* e *difussion phase*, que supostamente eram responsáveis pela compressão dos dados, estavam presentes tanto na tangente hiperbólica como na *ReLU*. O comportamento é representado como um degrau ao longo do treinamento, como mostra a Figura 4.5.

Figura 4.5 - a) Análise dos pesos para a função tanh. b) Análise dos pesos para função ReLU.




Fonte: (SAXE, et al., 2018).

Reforça-se, portanto, a hipótese de que o gradiente não é responsável pelo comportamento de compressão, ou pelas duas fases descritas em (SCHWARTZ-ZIV & TISHBY, 2017), uma vez que o comportamento do gradiente dos pesos se comporta de forma igual para ambas as funções utilizadas.

Com isto, os autores questionam grande parte dos resultados apresentados por Schwartz-Ziv e Tishby (2017) sobre a existência da compressão nas redes neurais e sua relação com a generalização das redes principalmente pela demonstração da ausência da compressão em funções que não são duplamente saturadas.

## 4.4 O problema da estimação

Um dos pontos mais relevantes acerca do problema do IB aplicado as redes neurais e reiterado por diversos autores (SAXE, et al., 2018), (SCHIEMER & YE, 2019), (GOLDFIELD, et al., 2019)

e (GEIGER, 2020) é com relação ao problema de estimação da informação mútua associada as variáveis discretas e continuas para casos determinísticos. Uma rede neural é definida como determinística quando sua representação latente $L$ é em função da entrada $X$, ou seja, $L = f(X)$, ao contrário, caso exista alguma incerteza associada ao processamento de $f$ que envolva alguma aleatoriedade temos uma rede estocástica.

Uma das formas de calcular a informação mútua $I(X; L)$ é definida a seguir:

$$I(L; X) = H(L) - H(L|X) \tag{4.6}$$

Para uma variável $L$ discreta, temos que a entropia discreta $H(L)$ é determinada como:

$$H(L) = -\sum_{i=1}^{N} P_i(L) \log P_i(L) \tag{4.7}$$

Logo, para uma rede neural determinística com variável discreta, a incerteza $H(L|X) = 0$, pois uma vez que sabemos $X$ e sabemos $f(X)$ e determinamos $L$, sem incertezas, pois o processo é determinístico e não estocástico, logo, $I(L; X) = H(L)$. Entretanto, para o caso de uma variável contínua nesse mesmo contexto, temos:

$$h(L) = -\int p_l(L) \log p_l(L)\, dl \tag{4.8}$$

Para variáveis contínuas é possível obter valores negativos de entropia e até infinitos. Como, por exemplo, para o caso em que $p_l$ é distribuída segundo a função delta, logo $h(L|X) = -\infty$. Como consequência da entropia condicional infinita negativamente, neste exemplo, a informação mútua $I(X; L)$, calculada de forma análoga à Equação (4.6), é infinita para variáveis contínuas $I(X; L) = \infty$ (SAXE, et al., 2018).

Muitos trabalhos como (SAXE, et al., 2018), (GOLDFIELD, et al., 2019), (GEIGER, 2020) e (AMJAD & GEIGER, 2020) enfatizam esse problema. Quando $X$ é uma variável contínua, a probabilidade $p(x)$ será igualmente continua, consequentemente, a distribuição da camada $L$ também será contínua, com isto $I(L; X) = \infty$ independentemente da função de ativação. Para o caso em que $X$ é discreta a informação mútua é uma constante $I(L; X) = H(L)$. Ou seja, em teoria é praticamente irrelevante tentar calcular a informação mútua da rede neural uma vez que são constantes como muitos autores enfatizam (SAXE, et al., 2018), (SCHIEMER & YE, 2019), (GOLDFIELD, et al., 2019) e (GEIGER, 2020).

Para contornar o problema decorrente da continuidade da função, os autores discutem a necessidade da adição de ruído a estas variáveis internas para que não haja entropias infinitas, uma possibilidade é a adição de um pequeno ruído $r$ independente. Essa mudança de variável $T_r = L + r$ tem informação mútua finita pois a entropia condicional $H(T_r|X) = H(r)$ não é nula, logo:

$$I(T_r; X) = H(T_r) - H(r) \tag{4.9}$$

Desta forma, o ruído opera como um limitador garantindo que não seja possível obter valores infinitos de entropia.

Outra forma de lidar com o problema é através da discretização das variáveis tal qual (SCHWARTZ-ZIV & TISHBY, 2017) realizaram. Esse procedimento transforma a variável continua em discreta, garantindo que as estimações sejam finitas, pois o próprio procedimento é limitante.

O grande problema associado as abordagens descritas acima é que tanto o ruído quanto a restrição que a discretização impõe as variáveis, ambos para manterem a informação mútua finita, não são propriedades inerentes a rede neural, ou seja, são adaptações para tentar medir a complexidade das representações internas da rede.

Para contornar essa questão existem duas possibilidades. A primeira é tornar o ruído parte da rede, ou seja, o ruído deve ser estrutural (*Noisy DNN*) sendo injetado nos pesos durante o treinamento (ACHILLE & SOATTO, 2018), (ACHILLE, PAOLINI, & SOATTO, 2019) ou na saída dos neurônios (GOLDFIELD, et al., 2019), ou quando estimada, (SAXE, et al., 2018). A segunda envolve tornar a discretização estrutural, ou seja, a rede neural inteira é discreta como (LORENZEN, IGEL, & NIELSEN, 2021) que utilizam uma *Quantized NN* (HUBARA, et. al., 2017).

Em (AMJAD & GEIGER, 2018) o problema associado a estimação da informação também é discutido, o autor explica como o processo pode ser falho ao treinar as redes neurais usando o funcional IB, sendo resultado de dois fatores. 1) O próprio IB sendo um problema de otimização mal posto como discutido anteriormente em relação a condição pré-estabelecida nas redes determinísticas. Levando então a um problema de estimação, ou seja, a informação medida e visualizada pelo plano de informação estaria dizendo mais sobre os estimadores que propriamente sobre a representação interna da rede. Este último ponto é explorado incessantemente por diversos trabalhos que serão citados posteriormente. 2) A informação mútua apresenta características de invariância[2], e consequentemente imune a transformações determinísticas (DPI), o que a tornaria uma medida insuficiente para caracterizar a

[2] Uma das propriedade que a informação mútua possui é $I(X;Y) \geq I\big(X; f(Y)\big)$ para qualquer função determinística $f$.

complexidade da representação latente das redes por si só. Ou seja, por mais complexa que a representação interna seja em decorrência de um suposto processamento denso da camada interna, a informação mútua medida não seria diferente de uma camada que realizasse um processamento mais simples, como consequência o plano de informação não seria capaz de capturar a evolução da complexidade da representação latente da rede como é estabelecido pelo IBDL, a princípio, pelo menos em problemas de classificação.

Uma discussão bastante relevante acerca do problema do IB para casos determinísticos também é bem discutida em (KOLCHINSKY, TRACEY, & KUYK, 2019) e (AMJAD & GEIGER, 2020).

## 4.5 Caracterização do problema do IBDL: Uma breve reflexão

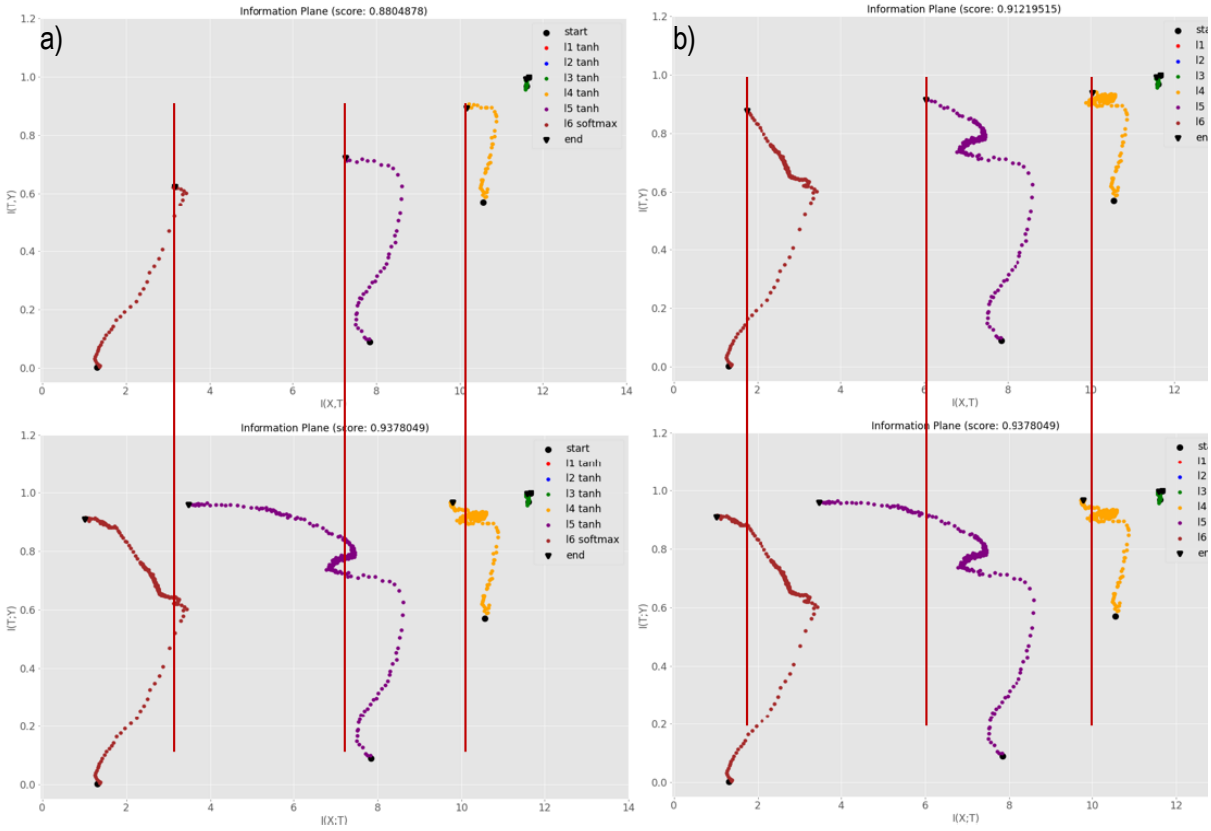
Como visto, nas seções anteriores, a teoria do IBDL (do inglês, *Information Bottleneck Theory of Deep Learning*) consiste em analisar as redes neurais através da teoria da informação, que tem como principal objetivo analisar a qualidade da representação latente formada nas redes neurais durante o treinamento. Foram apresentados, até o momento, três pontos interligados acerca do problema, são eles:

1. A dificuldade em demonstrar a ligação entre compressão e generalização das redes.
2. O problema em identificar a fase de compressão para as diversas funções de ativação utilizadas em redes neurais.
3. A estimação coerente e correta das medidas de informação e seu impacto para o problema.

Pela hipótese de (SCHWARTZ-ZIV & TISHBY, 2017), uma rede treinada, e que generaliza, tem em seus pesos uma representação interna complexa, ou seja, o ponto de melhor desempenho da rede é atingido no ponto de máxima compressão da representação. Entretanto, foi demonstrado em (SAXE, et al., 2018) que uma rede com a função *ReLU* não apresenta sinais de compressão, e, mesmo assim, tem equivalente capacidade de classificação, ou seja, processa corretamente os dados e apresenta generalização. Uma outra forma de estudar o problema foi proposto em (SCHIEMER & YE, 2019), onde os autores para testar a hipótese da compressão-generalização, realizam um experimento utilizado a técnica de parada antecipada (do inglês, *early stopping)*, uma técnica bastante difundida e utilizada para evitar sobretreino (do inglês, *overfitting)*. O conceito da parada antecipada se baseia em interromper o treinamento assim que a rede atinge estagnação ou crescimento no erro de validação,

indicando claro risco de sobreajuste. Ao testar a parada antecipada, o plano de informação resultante mostra que o ponto onde a rede já apresenta uma generalização adequada está localizado no início da fase de compressão, Figura 4.6 a), se comparada com um treinamento completo, ou seja, até o final das épocas estabelecidas.

Figura 4.6 - a) Gráfico superior função Tanh com parada antecipada. b) Gráfico superior parada perfeita da Tanh. Na parte inferior a recriação do experimento de (SCHWARTZ-ZIV & TISHBY, 2017).




Fonte: (SCHIEMER & YE, 2019).

Ou seja, a rede neural treinada utilizando a técnica de parada antecipada como critério de interrupção do processo de treinamento, demonstra, através da Figura 4.6, que a rede ao apresentar um erro de generalização adequado apresenta uma representação interna de baixa complexidade quando comparada a uma rede treinada de forma completa. O que significa que uma rede neural para ter bom desempenho não necessariamente precisa apresentar compressão no plano de informação.

Entretanto, testes complementares foram realizados com o conjunto de dados MNIST onde os autores não obtiveram resultados semelhantes aos descritos, ou seja, o ponto no plano de informação referente a rede treinada utilizando o método de parada antecipada, não é diferente de uma rede treinada completamente. Ou seja, a complexidade da representação interna em ambos os casos fora semelhante, segundo os autores, esses resultados conflitantes não permitiram que esta hipótese pudesse ser

generalizada para qualquer rede com qualquer tipo de conjunto de dados. Porém, os resultados obtidos são suficientes para introduzir mais um elemento que corrobora com a narrativa de que uma rede não necessariamente precisa apresentar compressão no plano de informação para ter um bom desempenho. Segundo Scheimer e Ye, os resultados observados com a parada antecipada indicam na realidade que a compressão pode ter uma forte relação com a complexidade dos dados de treinamento.

Além disto, os autores demonstram em outros testes que as funções de ativação de uma camada interna podem induzir o comportamento das funções na camada seguinte, ou seja, é demonstrado que uma tangente hiperbólica na camada seguinte a uma camada composta pela *ReLU*, pode não apresentar compressão. Em (WICKSTRØM, et al., 2019), o mesmo comportamento é notado em uma rede convolucional utilizando a parada antecipada. Portanto, a depender da função utilizada, pode ou não haver a fase de compressão.

Como foi discutido anteriormente, a manifestação da compressão é consequência da dupla saturação de algumas funções. A grande maioria dos trabalhos dos últimos anos são realizados utilizando problema de classificação, com a saída da rede sendo binária (*one-hot encoding*), portanto, o comportamento dos sinais de saída das camadas dentro da rede tendem a se concentrar nas extremidades das funções utilizadas, gerando sinais na saturação (SAXE, et al., 2018) e (SCHIEMER & YE, 2019), consequentemente a compressão observada. Ou seja, o comportamento de compressão é fortemente associado ao problema de classificação. Além do fato de que esse comportamento pode ser potencializado de acordo com a estratégia utilizada para discretização ou simplesmente não existir, como foi mostrado em (SAXE, et al., 2018) ao utilizar estimadores robustos como o KDE ou como redes mistas como em (SCHIEMER & YE, 2019). O que nos leva ao terceiro ponto levantado.

Quando as funções de ativação são discretizadas utilizando *binning*, em que os limites são determinados pelas próprias funções, como em (SCHWARTZ-ZIV & TISHBY, 2017), ou, no caso da *ReLU* $[0, max]$ onde o $max$ é definido pelo máximo das ativações em todas as épocas como em (SAXE, et al., 2018) o resultado é conflitante. A fase de compressão ocorre apenas na tangente hiperbólica. Entretanto, quando é utilizado o método de *binning* adaptativo (CHELOMBIEV, HOUGHTON, & O'DONNEL, 2019), *binning* e KDE adaptativo ou o método *binning+hash* (NOSHAD, ZENG, & HERO, 2019), a compressão é presente tanto para a tangente hiperbólica quanto para a *ReLU*, em desacordo com a narrativa sobre a fase de compressão existir apenas para funções duplamente saturadas.

Em (SAXE, et al., 2018), para a discretização, é utilizado zero como limite inferior e como limite superior o máximo da *ReLU* em todo o treinamento. De certa forma, esse procedimento pode comprimir as distribuições analisadas de cada camada caso o limite da distribuição seja inferior ao limite máximo estabelecido, pois esta distribuição seria representada em poucos intervalos. Como consequência, o histograma resultante não necessariamente representa as reais distribuições de cada

camada ao longo do treinamento, ou seja, teoricamente, a discretização pode não capturar o comportamento estatístico da distribuição das ativações, mesmo que isto leve a uma entropia incoerente como consequência da discretização a função principal do histograma é representar com fidelidade a distribuição e as propriedades estatísticas presentes (SCOTT, 2015). Nesse caso em específico, como o estudo é sobre a compressão da representação interna, a estratégia que foi utilizada não identifica o processo de compressão, pois, a metodologia adotada pode ter sido inadequada uma vez que não contempla da maneira correta a discretização, ou seja, pode ter havido a concentração de distribuições em poucos intervalos, levando a uma resolução inadequada daquela distribuição naquele momento, quando, na realidade, se a distribuição fosse realizada em um suporte mais adequado, seriam observadas as variações e nuanças que resultariam em oscilações nas entropias, o que pode levar a compressão observada no plano de informação mesmo na função *ReLU*, como foi observado por outros autores.

Discretizar uma variável em um domínio maior ou menor que o domínio real da variável pode resultar em um histograma não representativo, assim como o número inadequado de intervalos podem alterar a percepção da distribuição.

Conforme o número de intervalos aumenta, tendo como início um único intervalo, e de forma crescente até um número infinito de intervalos, o histograma, que é a distribuição discretizada, tem um comportamento bem claro. No início, a resolução é baixa, com todos os pontos concentrados em apenas um ou poucos intervalos: portanto, existe a perda de informação (baixa entropia). Conforme o número de intervalos aumenta, a resolução também aumenta, atingindo um ponto onde a resolução é ótima, ou seja, o histograma representa com fidelidade a distribuição original e a entropia resultante é muito próxima da esperada, ou seja, sem perda de informação. Aumentando ainda mais o número de intervalos (infinito), a resolução é tão elevada que cada amostra da distribuição está contida em seu próprio intervalo, resultando em uma distribuição uniforme, onde a entropia é máxima (Equação 3.71), portanto, tanto o histograma como a entropia estão em desacordo com a distribuição real.

Invariavelmente, a estimação das medidas de informação define certos aspectos do problema. O uso de estimadores robustos, em teoria, deveria apenas servir para medir corretamente as informações e entropias, e não em influenciar o comportamento identificado na dinâmica do plano de informação. Ou seja, ao partir do princípio que a hipótese esteja correta, e que, de fato uma rede neural deva apresentar a fase de compressão como resultado de um bom treinamento, consequentemente, apresentando boa generalização, em teoria, independentemente dos estimadores utilizados, as duas fases do treinamento deveriam estar presentes. Mesmo que a compressão completa da representação não seja necessária para uma boa generalização, e que apenas o indicio das duas fases sejam necessários, assim como a garantia da DPI, com isto o plano de informação deveria apresentar as fases

do treinamento para todos os estimadores, e estes deveriam refletir tal comportamento mesmo que com acentuações diferentes nessa dinâmica, ou seja, se houver compressão deveria estar presente em todos os estimadores e eventualmente mais ou menos acentuados a depender do estimador.

Entretanto, é visto que essa premissa não se mantém, o que nos leva a dois problemas. Primeiro, exatamente como (AMJAD & GEIGER, 2018) mencionam o plano de informação estaria descrevendo mais sobre os estimadores que propriamente sobre a complexidade da representação interna. Segundo, não está sendo medida a complexidade da representação interna da rede como imaginado.

Para o primeiro caso, onde o plano de informação descreve mais sobre os estimadores que a própria representação, tem-se um problema referente a escolha do estimador, ou seja, de alguma forma existe uma sensibilidade ou propriedade dos estimadores que os permitem identificar certos comportamentos estatísticos como essa compressão, em (YU & PRÍNCIPE, 2019) e (WICKSTRØM, et al., 2019) é mencionam que a informação mútua estimada não necessariamente apresenta todas as propriedades da definição probabilística da informação mútua. O que nesse caso seria uma boa explicação para a divergência entre os trabalhos em provar a hipótese levantada por (SCHWARTZ-ZIV & TISHBY, 2017). Para o segundo caso, em que teoricamente não está sendo medida a complexidade da representação interna da rede, surgem as questões, a) O que está sendo medido? e b) Por que os estimadores não medem as mesmas coisas?

Como explicado na Seção 4.4, segundo a teoria do problema de estimação, é elaborado que a informação mútua dentro da rede deveria ser constante ou infinita, mas foi demonstrado (SCHIEMER & YE, 2019) que isto não ocorre. É possível então que de fato a informação não esteja sendo medida como esperado.

Ademais, os autores mostram que a compressão não é necessária para uma boa generalização como discutido até o momento. Além desses pontos, muitos trabalhos relatam ainda inconsistências nos resultados como a violação da DPI em função dos estimadores, e, em relação a compressão, que, até o momento, se apresenta altamente situacional, ou seja, diretamente relacionados a inicialização dos pesos (CHELOMBIEV, HOUGHTON, & O'DONNEL, 2019) e também aos estimadores utilizados.

## 4.6 Trabalhos relacionados e a recente convergência do assunto

Como visto na seção anterior, existem muitas questões que precisam ser esclarecidas. A discussão acerca do problema IBDL, nos últimos anos, tem ocorrido: diversos estudos foram realizados

abordando o assunto com objetivos distintos, porém, semelhantes, especificamente com relação às fases de treinamento descritas em (SCHWARTZ-ZIV & TISHBY, 2017) e a relação compressão-generalização. A seguir, expomos brevemente trabalhos realizados nos últimos anos com alguns detalhes a respeito de suas abordagens e resultados.

Em (SCHWARTZ-ZIV & TISHBY, 2017), foi utilizada uma rede MLP treinada com o conjunto de dados binário SZT. Foi identificada a fase de compressão nas camadas internas utilizando a tangente hiperbólica como função de ativação pela estimação através do *binning*.

Em (SAXE, et al., 2018), foi utilizada uma rede MLP treinada com os conjuntos de dados SZT e MNIST. Não foi identificada a fase de compressão nas camadas internas com a função *ReLU* através da estimação por *binning* e KDE.

Em (CHELOMBIEV, HOUGHTON, & O'DONNEL, 2019), foi utilizada uma rede MLP treinada com o conjunto de dados SZT. Foi identificada a fase de compressão com a função *ReLU* nas camadas internas quando estimada a informação mútua através do *binning* adaptativo e com uma versão adaptativa do KDE em certas situações.

Em (NOSHAD, ZENG, & HERO, 2019), utilizando uma rede MLP e uma rede CNN treinadas com o conjunto de dados MNIST foi identificada compressão utilizando a função *ReLU* utilizando o estimador EDGE (*binning* + hash).

Em (GABRIÉ, et al., 2018), utilizando uma rede MLP treinada com um conjunto de dados sintético, foi identificada a fase de compressão utilizando a função *ReLU*, mas existem ressalvas a respeito da compressão em geral e quando elas ocorrem.

Em (SCHIEMER & YE, 2019), com uma rede MLP treinada com os conjuntos de dados SZT e MNIST não houve a fase de compressão utilizando a função tangente hiperbólica nas camadas internas em certas situações.

Em (YU & PRÍNCIPE, 2019), diferentemente dos demais trabalhos utilizou a estrutura *autoencoder* treinada com conjuntos MNIST, *Fashion*-MNIST e FERG-DB, foram encontrados sinais de compressão, em certas situações para a *ReLU* utilizando o estimador *Multivariate Matrix-based Rényis α-entropy*.

Em (WICKSTRØM, et al., 2019), utilizando uma DNN (VGG16) treinada com o conjunto de dados CIFAR-10, foram encontrados sinais de compressão utilizando o estimador *kernel tensor-based estimator of Rényi's entropy*.

Em (YU, et al., 2020), utilizando uma CNN treinada com os conjuntos de dados MNIST e Fashion-MNIST, foram encontrados sinais de compressão em um IP modificado utilizando o estimador *Multivariate Matrix-based Rényis α-entropy*.

Em (LORENZEN, IGEL, & NIELSEN, 2021), utilizando uma rede *Quantised Neural Network* (QNN) treinada com o conjunto de dados SZT, não foi identificada a fase de compressão na função *ReLU*.

Em (GEIGER, 2020), faz-se uma compilação destes e outros trabalhos classificados em relação a estrutura da rede, *datasets*, método de treinamento (SGD, ADAM) e estimadores, analisando quais apresentam compressão e se estão correlacionadas a generalização.

Inúmeros trabalhos foram desenvolvidos, com o objetivo de estudar a representação interna da rede, utilizando diversas estruturas e estimadores. De maneira geral, até o momento, não existe um consenso a respeito do problema nem sobre a metodologia a ser utilizada. Mas alguns pontos comuns nos trabalhos são claros sob certos aspectos.

Em (SCHIEMER & YE, 2019), a discussão sobre o que realmente está sendo medido como informação mútua e o papel da discretização nessa questão é bastante explorado. Devido ao fato de a natureza do *dataset* utilizado ser inteiramente discreta, as informações mútuas das camadas deveriam ser constantes, isto é, $I(X;Y) = I(T;Y)$, $I(X;T) = H(X)$ e $I(T|X) = 0$ para redes com a tangente hiperbólica como função de ativação. Para entender o papel da discretização na estimação da informação, essas medidas foram analisadas ao longo do treinamento segundo algumas hipóteses. Inicialmente, é estudado o impacto da variação do tamanho dos intervalos na discretização. Quanto menor o tamanho dos intervalos mais as curvas no plano de informação se deslocam em direção a uma única linha no eixo y ($I_Y$), localizada em 1 bit, refletindo a tendência das medidas de informação de serem constantes como mencionado. Entretanto, a informação deveria ser sempre constante ao longo do treinamento, e a estimação por *binning* não segue de acordo, isso se deve ao fato de que o procedimento de discretização implica em perda de informação, pois uma vez que se transformam dois valores ou classes distintas em um único índice (intervalo) da discretização, essas duas classes se tornam indistinguíveis.

Essa é a razão em tornar os intervalos cada vez menores, para que a estimação e identificação das classes sejam mais precisas, e como consequência as medidas no IP se aproximam das constantes já estabelecidas.

Contudo, segundo os autores, é necessário ainda explicar a dinâmica presente no plano de informação descrito como compressão já que as medidas não são sempre constantes. Como visto anteriormente, as informações mútuas entre a camada e os dados de entrada da rede são calculadas como $I(X;T) = H(X) - H(X|T)$, mas $H(X)$ é constante durante o processo de treinamento e a variação de $I(X;T)$ só é possível se $H(X|T)$ oscilar, entretanto, se $X = f(T)$ então $H(X|T) = 0$, ou seja, não há mapeamento direto entre $X$ e $T$, o que existe é apenas um aumento de $H(X|T)$, que leva à redução da informação mútua (compressão). O mesmo ocorre para os dados de saída pois é impossível

distinguir classes apenas com o índice do intervalo ao qual ela está localizada, o que leva a perda de informação entre a camada e os dados de saída.

O que ocorre é que, ao analisar a distribuição das ativações discretizadas em três momentos do treinamento, no início, no meio e no fim, o autor nota que o que se mede são os intervalos nulos, ou seja, intervalos do histograma sem amostras e não a informação da distribuição. No início do treinamento, as ativações estão distribuídas de maneira aleatória, mas seguindo a tendência da distribuição dos pesos, na metade do treinamento, as ativações estão dispersas ao longo de todo o contradomínio da função, e, no final do treinamento, elas estão concentradas apenas nos valores 0 e 1. Essas distribuições são então comparadas com os respectivos momentos no plano de informação. À medida que a camada apresenta compressão, o número de intervalos vazios aumenta se concentrando (agrupamento, (SAXE, et al., 2018), apêndice E) apenas nas extremidades.

Esse comportamento é induzido pela função de ativação da saída da rede que tem apenas valores 0 e 1, em consequência do uso dos rótulos binários (*one-hot-encoding)* para o problema de classificação, portanto, os intervalos da discretização com amostras terão apenas frequências nesses valores, e, por tal razão, é necessário apenas 1 bit (limite teórico para o problema explorado) de informação para determinar o estado da saída da rede. Consequentemente, a função de saída induz todas as camadas da rede a concentrarem as ativações em poucos intervalos a fim de produzir apenas as saídas 0 e 1 da rede, entretanto, o processo perde força conforme adentra a rede.

Para o caso da saída, a informação é perdida por existirem diferentes classes no mesmo intervalo, no início do treinamento a informação mútua é baixa por conta da inicialização aleatória dos pesos, que leva à classificação errada das classes, resultando em múltiplas classes no mesmo intervalo, ou seja, dado o índice do intervalo na camada de saída da rede não é possível distinguir essas diferentes classes. No decurso do treinamento, a rede aprende a distinguir as classes entre 0 e 1 elevando a informação mútua até 1 bit.

Para o caso da *ReLU,* o raciocínio é ligeiramente diferente. Devido ao fato de não haver valores negativos na *ReLU,* não existem ativações com valores negativos como na tangente hiperbólica: muitas das ativações são nulas e outras positivas distribuídas de maneira aleatória, portanto, os intervalos da discretização terão menos ativações distribuídas (baixa entropia) e muitas concentradas próxima de zero o que leva a uma baixa informação mútua. Ao longo do treinamento, mais ativações nulas vão sendo distribuídas nos intervalos da discretização, que é entre 0 e o algum valor da *ReLU*, diminuindo os intervalos sem amostras, levando ao aumento da informação, refletido no plano de informação como apenas um incremento de informação nas camadas, portanto, sem apresentar a "compressão". Ou seja, o que ocorre é que no início do treinamento com a *ReLU* a maioria das ativações estão concentradas em torno do valor 0 em apenas um intervalo da discretização. Ao longo do

treinamento ocorre um espalhamento dessas ativações nulas, pois elas deixam de ser nulas em decorrência do ajuste dos pesos e o processo de treinamento e passam a ter valores entre $[0, 1]$ e no final do treinamento apenas 0 e 1, aumentando a informação mútua.

Esse comportamento de agrupamento (clusterização) é o que é capturado pela discretização por *binning*, e não propriamente a complexidade da representação como suposto. Uma outra maneira que os autores utilizaram para demonstrar essa hipótese foi medindo a informação das camadas para uma rede com uma camada de gargalo ("*bottleneck*") $[12, 3, 2, \underline{12}, 2, 2]$ (o primeiro e o último valor são respectivamente o número de entradas da rede -12 e o número de saídas da rede -2. Os demais índices são o número de neurônios de cada camada interna da rede. A camada de gargalo está subscrita).

Nesse experimento proposto pelos autores, a camada de gargalo apresentou compressão superior à camada posterior, ou seja, no plano de informação L3 está à esquerda de L4 quando deveria ser o contrário, o que segundo a DPI deveria ser impossível, mas o resultado do plano de informação é meramente resultado de muitas ativações estarem no mesmo intervalo, sendo considerado outro forte indício de que a estimação por discretização não mede informação, mas sim o agrupamento (*clustering*) das ativações.

A hipótese a respeito de a compressão da informação $I_X$ estar associada ao agrupamento geométrica da representação interna da rede não é discutida só por (SCHIEMER & YE, 2019) como descrito, ela é partilhada por outros trabalhos.

Utilizando uma Noisy-DNN (estocástica) (GOLDFIELD, et al., 2019) ao estudar o fluxo de informação entre as camadas da rede em um experimento bastante rigoroso, demonstra-se que o fenômeno de compressão observado por inúmeros autores é a manifestação do processo de agrupamento da representação interna da rede devido a saturação das funções de ativação, mas não restrito somente a este elemento. Em uma de suas análises, é estudada a diferença entre as distancias intraclasse e interclasse das representações internas da rede para o problema de classificação. A formação de clusters é clara durante o processo de treinamento, e tal fenômeno é caracterizado na forma de compressão no plano de informação.

Para estudar a fundo esse fenômeno, os autores mensuram o agrupamento através da quantização $H\big(Bin(T_n)\big)$. Ao utilizar a largura dos intervalos da discretização por *binning* como parâmetro de ajuste na reprodução do experimento de (SCHWARTZ-ZIV & TISHBY, 2017), os autores demonstram que esse processo reflete o grau de agrupamento no espaço das representações internas da rede. No início do treinamento, por consequência da inicialização aleatória dos pesos, existem muitas ativações distribuídas ao longo do contradomínio da função de ativação (tanh), ou seja, a distribuição de

probabilidade formada nas camadas internas é bastante espalhada. Ao decorrer do treinamento as ativações se acumulam em poucos intervalos, formando grupos, ou seja, clusters. Essa dinâmica de agrupamento das ativações, segundo os autores, é o que a medida através da $H\big(Bin(T_n)\big)$, refletindo, portanto, o nível de agrupamento da representação interna.

Além disto, os autores notam quer o comportamento da medida $H\big(Bin(T_n)\big)$ ao longo do treinamento, comparado a informação mútua $I(X;T_n)$, apesar de diferentes, as duas medidas quantificam de maneira similar o grau de agrupamento das representações internas.

Em outro teste, utilizando uma rede convolucional com o dataset MNIST também através da análise das distâncias entre classes e intraclasses, foi observada a formação de clusters durante o processo de treinamento, mas não houve compressão no plano de informação devido à alta dimensionalidade da rede convolucional.

Toda esta discussão acerca do agrupamento das representações internas decorre, principalmente, da oscilação da informação mútua observada no plano de informação durante o treinamento das redes neurais. Segundo a literatura apresentada até o momento, a formulação do problema do gargalo de informação (*information bottleneck)* em redes determinísticas (a princípio apenas em problemas de classificação), explicado na Seção 4.4 e discutido até o momento, apresenta medidas de entropias e informação mútua como medidas que deveriam ser constantes ao longo do treinamento. A oscilação da informação mútua observada tem promovido a discussão do problema como demonstrado nesta seção. Entretanto, é sem sombra de dúvidas difícil entender por completo, esta ideia de reduzir uma distribuição de probabilidade, mesmo que obtida através de histogramas, a uma medida de agrupamento utilizando a teoria da informação que tem base probabilística.

Dentro do contexto da discussão desenvolvida neste trabalho é muito importante citar outros dois trabalhos em que a manifestação da compressão tem outras fontes além da saturação das funções ou do método de discretização utilizados que contribuem com o enriquecimento da discussão do problema. Em (YU & PRÍNCIPE, 2019) a rede utilizada para estudo foi um *Stacked Autoencoder* (SAE), sendo que a existência da fase de compressão está diretamente associada ao tamanho da camada gargalo $S_{bn}$ do autoencoder. Para uma camada de gargalo com tamanho maior que a dimensão intrínseca do conjunto de treinamento ($S_{bn} > d_{tr}$), é observado um aumento da informação mútua seguida de uma diminuição (compressão); quando a camada de gargalo é menor que a dimensão intrínseca do conjunto de treinamento ($S_{bn} < d_{tr}$) a informação mútua apresenta apenas um aumento constante até atingir um determinado limite.

Em (YU, et. al., 2020) os autores estudam o fluxo de informação em uma rede convolucional. Eles introduzem uma nova análise das medidas de informação chamada de decomposição parcial da

informação (PID), que envolve a decomposição da informação em outras medidas como informação única e redundante, e assim é realizada a análise do plano de informação com essas novas medidas. Dados os padrões de entrada e dois mapas de características a informação mútua entre eles é descrita como $I(X;\{T_1,T_2\})$, e pode ser decomposta em quatro componentes: Sinergia $Syn(X;\{T_1,T_2\})$ que mede a informação de $X$ dado a combinação dos mapas $T_1$ e $T_2$, redundância $Rdn(X;\{T_1,T_2\})$ que mede a informação de $X$ compartilhada pelos dois mapas e a informação única $Unq(X;T_1)$ e $Unq(X;T_2)$ que mede a informação de $X$ que pode ser apenas obtida a partir de $T_1$ ou $T_2$.

Os mapas mencionados são representações dos filtros presentes em cada camada convolucional da rede; cada mapa, portanto, caracteriza uma propriedade especifica dos dados de entrada da rede. Com isto, a informação a respeito de $X$ está contida nesses filtros $T_1, T_2, \dots, T_n$, e, conforme o número de filtros aumente mais informação é capturada, chegando ao limite que é a própria entropia dos dados. Essa informação é decomposta como mencionado e apenas a informação útil e sinérgica são utilizadas para análise da rede. Utilizando os conjuntos de dado MNIST e Fashion-MNIST, em suas análises os autores notam que ao substituir no plano de informação $I(X;T)$ e $I(T;Y)$ pelas médias de suas respectivas informações única e sinérgica, ou seja, neste plano de informação modificado, é observado o fenômeno de compressão da informação.

Entretanto, estimar todas essas medidas é complicado e os resultados dependem muito do problema que a rede está sendo aplicada além do fato de que as redes convolucionais têm comportamento diferente no plano de informação que as redes MLP utilizadas em muitos trabalhos, o que dificulta a comparação ou análise da rede convolucional com os demais trabalhos.

Por fim, (GEIGER, 2020) faz um estudo bastante detalhado sobre a literatura do assunto, levantando pontos comuns que estes trabalhos apresentam a fim de determinar uma direção comum a que todos os trabalhos possam convergir. Uma das principais considerações feitas pelo autor relativiza a plausibilidade de se comparar estimadores distintos, pois as medidas estimadas de informação mútua $\hat{I}(X;L)$ e $\hat{I}(Y;L)$ dependem consideravelmente da escolha dos estimadores. Para uma análise consistente do plano de informação de diferentes estimadores a análise dos resultados deve levar em consideração esse fator e a sensibilidade dos resultados a ele. Resultados obtidos através do processo de discretização e o KDE, por exemplo, devem ser analisados pela perspectiva geométrica, ou seja, agrupamento (*clustering*) como discutido amplamente, pelo menos em problemas de classificação. Estimadores de informação não necessariamente incorporam todas as propriedades probabilísticas previstas pela definição de informação (YU, et. al., 2020) e (YU & PRÍNCIPE, 2019), portanto, comparar resultados de diferentes estimadores levam a resultados muito distintos e a análises superficiais. Além disso, mesmo que a informação mútua seja finita, ainda assim, existe a necessidade de haver um conjunto de dados consideravelmente grande para que a estimação dessas medidas seja precisa.

Outro ponto levando pelo autor é em relação à violação da desigualdade de processamento de informação (DPI) em experimentos numéricos decorrente dos estimadores. A razão pela qual isto ocorre é pelo fato de os trabalhos estarem analisando a informação estimada e não a real. Além disso, pela restrição do tamanho do conjunto de dados não necessariamente as distribuições de probabilidade refletem as propriedades reais que deveriam. Existe ainda o fato que a desigualdade entre medidas de informação estimadas $\hat{I}(X;L)$, para diferentes camadas, podem ser resultado da dimensão das camadas internas da rede. Camadas internas com poucos neurônios têm pouco "potencial de gerar mapeamentos" se comparadas com camadas maiores. Isso, possivelmente, ajuda a explicar o comportamento visto em (SCHWARTZ-ZIV & TISHBY, 2017), em que a informação $\hat{I}(X;L)$ é decrescente em uma rede neural cuja estrutura utilizada apresenta uma redução gradual no número de neurônios da camada inicial até a camada de saída $[12, 10, 7, 5, 4, 3, 2]$, onde o primeiro índice é a camada de entrada e o último índice é a camada de saída, os demais índices são o número de neurônios de cada camada interna da rede.

Mas, diferentemente de Schwartiziv-Ziv e Tishby (2017), (SCHIEMER & YE, 2019) mostram que a informação mútua entre duas camadas sucessivas pode apresentar comportamentos conflitantes com a DPI devido a diferença de dimensão dessas camadas. Como $\hat{I}(X;L_3) < \hat{I}(X;L_4)$ onde $L_3$ é uma camada com apenas dois neurônios e $L_4$ uma camada com doze neurônios, segundo a DPI, independentemente da dimensão da camada o correto seria $\hat{I}(X;L_3) > \hat{I}(X;L_4)$, ou seja, observou-se uma violação da DPI em função da dimensionalidade das camadas, o que naturalmente é um resultado conflitante com o princípio informativo, no sentido da teoria da informação, suporto pôr (SCHWARTZ-ZIV & TISHBY, 2017) em relação as medidas de informação observadas nos resultados.

O autor ressalta ainda que muitos trabalhos apresentam divergência de resultados em decorrência dos estimadores e, ao final do trabalho, chega a algumas conclusões. Em redes neurais determinísticas a informação mútua entre o conjunto de dados e as camadas $I(X;L)$ é infinita, e que qualquer aproximação finita dessa medida é extremamente dependente do método utilizado para estimação, e, por consequência, a comparação de diferentes estimadores através do plano de informação é inviável. A prova cabal desse fato é a violação da desigualdade de processamento de informação identificada em diversos trabalhos.

O autor continua dizendo que a compressão identificada no plano de informação não tem origem na teoria da informação, ou seja, aquela estabelecida pela teoria de Shannon, e que, na realidade, os estimadores estão mais alinhados ao conceito geométrico de compressão, identificando que as redes neurais determinísticas mapeiam o conjunto de dados de entrada em clusters no espaço das representações, e que funções duplamente saturadas potencializam esse comportamento. Para redes

estocásticas – ruidosas - os estimadores conseguem identificar tanto o comportamento geométrico quanto o comportamento teorético da informação.

Com relação à informação mútua $\hat{I}(Y;L)$ medida pelos estimadores, o autor conclui que, na maior parte da literatura, a fase de aprendizagem EMR ou fase de ajuste (*fitting phase)* na qual a informação aumenta consideravelmente de fato está presente. Uma rede com boa generalização necessariamente tem que ter uma $I(Y;L)$ elevada como a própria DPI sugere, mas uma $I(Y;L)$ elevada não é garantia de boa generalização, assim como uma rede com boa generalização não garante $\hat{I}(Y;L)$ elevada, por exemplo, é observado em muitos trabalhos que o sobre ajuste (*overfitting)* eleva $\hat{I}(Y;L)$ consideravelmente, e que essa quantidade se mostra diferente quando se compara o conjunto de treino com o conjunto de teste. A redução de $\hat{I}(X;L)$, na fase de compressão também não é mandatória ou determinante para o desempenho da rede, e não é um comportamento tão observado como suposto pelas teorias iniciais explicadas no início desta seção. Como consequência desses fatores a garantia da DPI ou de que a representação interna $L$ seja uma estatística suficiente mínima aproximada não são necessárias para que redes determinísticas treinadas apresente boa generalização.

Por último, Geiger ainda faz uma última consideração: apesar de a compressão no plano de informação ter alcançado até o momento uma explicação racional advinda da compressão geométrica, ainda assim, não há obrigatoriedade na existência de tal fenômeno para que uma rede apresente boa generalização.

# 5. Motivação e Definição do Experimento

No Capítulo 4, foram discutidos vários aspectos presentes na literatura acerca do problema de IB aplicado a redes neurais para problemas de classificação. A principal discussão que envolve os trabalhos gira em torno da manifestação do fenômeno de compressão no plano de informação, que está diretamente associado a alguns fatores. Segundo a literatura tal fenômeno pode ter as seguintes fontes:

1. A causa da compressão estaria relacionada a funções duplamente saturadas (SAXE, et al., 2018).
2. A causa da compressão é consequência do método utilizado para estimar as distribuições de probabilidade - neste caso, a discretização potencializa ou gera tal fenômeno (SAXE, et al., 2018), (GEIGER, 2020).
3. A natureza do problema de classificação com vetor binário de rótulo (*one-hot-encoding*) promoveria naturalmente a segmentação dos dados e isto se manifesta em forma de compressão no plano de informação (GEIGER, 2020), (GOLDFIELD, et al., 2019), (SCHIEMER & YE, 2019).

Cuidemos, em primeiro lugar, do item 1. A função *ReLU* é utilizada para demonstrar essa hipótese em (SAXE, et al., 2018), em contraste com a função tangente hiperbólica utilizada em (SCHWARTZ-ZIV & TISHBY, 2017). Entretanto, existem trabalhos (CHELOMBIEV, HOUGHTON, & O'DONNEL, 2019), (NOSHAD, ZENG, & HERO, 2019) nos quais se atesta a compressão na função *ReLU*, mostrando que o fenômeno de compressão não é necessariamente associado à não-linearidade utilizada. Por se tratar de um elemento comum à discussão, adotaremos três funções de ativação comumente utilizadas: a função sigmoid, a tangente hiperbólica e a *ReLU*.

Tratemos agora do item 2. Em acréscimo ao item 1, existe uma discussão bastante viva acerca do estimador e sua influência nos resultados: em (SAXE, et al., 2018), discute-se que a estratégia adotada para discretização pode potencializar ou suprimir o fenômeno de compressão. Entretanto, com o uso de estimadores mais elaborados (CHELOMBIEV, HOUGHTON, & O'DONNEL, 2019) e.g. estimadores adaptativos, (NOSHAD, ZENG, & HERO, 2019), com o estimador EDGE (*hash* + *bin*), e (WICKSTRØM, et al., 2019) com o estimador de *Renyi*, houve a manifestação da compressão, afastando a ideia da compressão ser exclusividade das funções de ativação utilizadas e / ou do estimador empregado. Diversos trabalhos utilizam o estimador discreto (*binning*) para desenvolverem seus

experimentos (CHELOMBIEV, HOUGHTON, & O'DONNEL, 2019); (CHENG, et. al., 2019); (GEIGER, 2020); (SCHWARTZ-ZIV & TISHBY, 2017); (SCHIEMER & YE, 2019); (SAXE, et al., 2018), pela simplicidade do método, pelos resultados prévios e pela discussões já feitas. Como uma abordagem inicial e exploratória do problema de regressão sob a ótica da IBDL nos moldes apresentados por este trabalho, mostra-se adequado a utilização do estimador discreto uma vez que se tenha conhecimento de suas implicações no problema, logo o adotaremos para tal. Entretanto, não há nos trabalhos mencionados nenhuma discussão sobre qualquer metodologia para a escolha dos parâmetros do estimador. Por essa razão levantaremos alguns pontos necessários a esta discussão em nosso trabalho (Seção 5.2).

Por fim, consideremos o item 3. O problema de classificação é o foco de toda a literatura discutida ou mencionada neste trabalho, e a questão da causa da compressão, nesta aplicação em específico, vem apresentando convergência nos últimos anos a uma ideia comum. A natureza do problema de classificação estaria, muitas vezes, relacionada a uma compressão geométrica da representação interna (GEIGER, 2020); (GOLDFIELD, et al., 2019); (SCHIEMER & YE, 2019); (YU & PRÍNCIPE, 2019), induzida pela forma com que o problema é tratado (*one-hot-encoding*), e que estimadores que envolvem algum parâmetro de largura como *binning*, *KDE* se mostram adequados para identificar esse agrupamento nas representações latentes. No plano de informação esse agrupamento se manifesta na forma de compressão (redução de $I(X;L)$).

Entretanto, existem dois pontos cruciais que divergem dessa colocação para o problema de regressão. Primeiramente, histogramas são utilizados para aproximar distribuições com propósitos estatísticos (SCOTT, 2015) e na própria digitalização de sinais (COVER & THOMAS, 2006), tendo como premissa a preservação das características da distribuição e dos sinais, logo, não necessariamente é apenas adequada para identificar o tal agrupamento mencionado, mas com potencial para de fato descrever a distribuição e possivelmente as medidas de informação associadas ao problema de regressão.

Em segundo lugar, a natureza dos problemas de regressão não envolve práticas como os vetores binários de rótulo (*one-hot-encoding)*, na qual determinamos a saída desejada com vetores binários: o que temos é, de fato, uma distribuição que caracteriza o sinal aprendido. Com isto, dificilmente existirá a possibilidade de trabalhar com saídas com valores fixos que possam induzir o comportamento das camadas internas como descrito para os problemas de classificação. Contudo, estão ausentes estudos em regressão nos mesmos moldes proposto por este trabalho, o que torna desafiador estabelecer o ponto de partida ideal. Esta falta de trabalhos pode estar relacionada a natureza do problema e suas dificuldades / implicações. Porém, a definição do problema IBDL estabelecido por Tishby tem se mostrado solida dada as diversas estruturas utilizadas até o momento para serem

generalizadas, a ponto de eventualmente também serem aplicáveis a problemas de regressão, como será discutido ao discorrer deste trabalho.

Este capítulo discute, em detalhes, pontos importantes relacionados aos problemas de regressão e estimação, bem como a configuração do experimento realizado. O capítulo está organizado da seguinte forma: na Seção 5.1 é discutido o problema de estimação associado a variáveis contínuas e suas implicações; na Seção 5.2 é detalhada a estrutura de rede utilizada e seus parâmetros associados e o problema de regressão escolhido para o trabalho; na Seção 5.2 definimos a metodologia utilizada para discretização das variáveis aleatórias e por fim na Seção 5.3, definimos de forma mais direta alguns dos objetivos pretendidos para o nosso trabalho.

## 5.1 O problema de estimação e a entropia diferencial

O cerne deste trabalho consiste em analisar a dinâmica de treinamento das redes neurais artificiais através da teoria da informação. Parte crucial do processo de estudo e análise consiste em observar a evolução da informação das representações internas dessas redes: para isto, é necessário a estimação destas medidas.

Retomando parte da discussão feita na Seção 4.4, a estimação de informação mútua em redes neurais artificiais determinísticas[3] é um problema mal posto (GEIGER, 2020); (GOLDFIELD, et al., 2019); (SAXE, et al., 2018) e (AMJAD & GEIGER, 2020) (Teorema 1) sob os seguintes aspectos. Segue que a informação mútua $I(L;X)$ pode ser calculada como:

$$I(L;X) = H(L) - H(L|X) \tag{5.1}$$

Se considerarmos que a variável $L$ é discreta então $H(L)$ é calculada como:

$$H(L) = -\sum_{i=1}^{N} p_i(l) log\, p_i(l) \tag{5.2}$$

Para uma rede com variável discreta, a incerteza $H(L|X) = 0$, logo $I(L;X) = H(L)$. Entretanto se $L$ é uma variável contínua, temos:

[3] Uma rede é dita determinística quando sua representação latente $L$ é em função da entrada $X$, ou seja, quando $L = f(X)$.

$$h(l) = -\int p_t(l) log\, p_t(l) dl \tag{5.3}$$

Para variáveis contínuas, podemos mostrar que a $I(l;y) = \infty$. Supondo que $p(l|x)$ tenha distribuição delta, logo temos que entropia condicional é dada por:

$$h(l|x) = -\int p_t(l) log\, p_t(l) dt = -\infty \tag{5.4}$$

Com isto, temos que a informação mútua $I(l;y)$ é infinita. Entretanto, quase sempre as variáveis utilizadas em redes neurais são contínuas, e, consequentemente, as distribuições internas também o são. Porém, como (CHELOMBIEV, HOUGHTON, & O'DONNEL, 2019) e (SAXE, et al., 2018) dizem, para observar a dinâmica do treinamento das redes neurais pela informação mútua, é necessário que essas medidas sejam finitas. Mesmo que se usem estimadores que meçam a informação estimada $\hat{I}(X;L)$ infinita, isto não seria interessante, pois se estaria revelando mais sobre a capacidade do estimador do que sobre a própria evolução da informação $I(X;L)$ (AMJAD & GEIGER, 2018); (SAXE, et al., 2018).

Uma das maneiras de contornar esse problema é através da adição de ruído nas camadas da rede. Existem algumas formas de realizar esse procedimento, como discutido na Seção 4.4: uma dessas maneiras foi utilizada em (SCHWARTZ-ZIV & TISHBY, 2017), (SAXE, et al., 2018), (CHELOMBIEV, HOUGHTON, & O'DONNEL, 2019) e (SCHIEMER & YE, 2019), e corresponde a discretizar a variável contínua em intervalos (*bins*), $\hat{L} \cong bin(l)$. O ruído está embutido no processo de discretização que é utilizado para realizar a aproximação da função de densidade de probabilidade da variável aleatória. Esta adição de ruído através da discretização permite a aproximação de $H(L)$ como uma variável discreta, tornando as medidas finitas e permitindo sua estimação.

Como discutido no início desta seção, a necessidade de trabalhar com variáveis discretas é essencial para tornar as medidas de informação finitas, permitindo assim o estudo da dinâmica das redes neurais ao longo do treinamento. O processo de transformação de variáveis contínuas em discretas não é um problema trivial, e envolve uma série de questões teóricas e práticas. Na Seção 3.7, vimos que a relação entre entropia discreta e contínua para um número finito de intervalos pode ser descrita como:

$$h(x) \cong H(x) + \log \Delta \tag{5.5}$$

Vimos, posteriormente, na mesma seção, que o raciocínio por trás da Equação (5.5) pode ser estendido para informação mútua continua para um valor adequado de $\Delta$, de tal forma que temos:

$$I_{\Delta}(x;y) \cong H(X) + H(Y) - H(X,Y) \tag{5.6}$$

Ou seja, a Equação (5.6) estabelece que a informação mútua contínua pode ser aproximadamente igual à informação mútua discreta das variáveis aleatórias. Com isto, criamos uma relação em que a escolha correta do Δ pode gerar uma aproximação adequada de uma variável contínua em uma variável discreta segundo as medidas de informação. Entretanto, existe um detalhe pertinente neste procedimento: a escolha de um $\Delta_{ótimo}$ resultará em estimativas de $I_{\Delta}$ e $h^{\Delta}$ muito próximas às medidas reais, para o caso em que uma variável continua possuir entropia diferencial negativa com a escolha correta do Δ temos que $|\log \Delta| > H(X)$ podendo atingir então o valor negativo (pela Equação 5.5) de entropia como ocorre em alguns casos contínuos. Do ponto de vista de estimação, é uma aproximação fidedigna, uma vez que a aproximação discreta é coerente medida da variável contínua. Entretanto, o presente trabalho tem como fundamento a análise da dinâmica das redes neurais através da DPI - vide Seção 3.4 e Seção 4.1, ao estimar valores de $H(x)$ (entropia discreta) próximos dos valores reais de uma variável contínua, nos casos onde esta entropia é negativa (particularidade de distribuições contínuas), a estimativa da entropia $h(x)$ (entropia contínua) entra em conflito com as equações da desigualdade de processamento de informação do problema IBDL Equação (4.1) e 4.2). Supondo que a formulação do problema IBDL é validada tanto para variáveis discretas quanto para contínuas, e, o fato de que, a DPI é provada válida para ambos os casos, não se pode violar tais definições de forma leviana e simplista, logo, há de se considerar que apenas a relação de hierarquia formada pela informação das camadas seja válida.

Este conflito merece uma discussão bem mais profunda e detalhada envolvendo a continuidade de variáveis aleatórias e uma eventual reestruturação do problema IBDL que não fazem parte do escopo deste trabalho, mas pode-se contornar este problema adotando as variáveis aleatórias discretas da forma tradicional, ou seja, discretizando as variáveis com o objetivo de tornar as medidas de informação finitas como (SAXE, et al., 2018) descreve, permitindo a estimação dessas medidas. Em contrapartida, perdemos a fidelidade de nossas estimações, pois agora não estamos trabalhando com a relação estabelecida pela Equação (5.6) nem buscando uma aproximação da entropia diferencial e sim lidando com a tradicional entropia discreta de Shannon, Equação (3.1). Ao utilizar variáveis discretas, trabalhamos em acordo com as relações da DPI Equação (4.1) e (4.2), como foi estabelecido por (TISHBY, PEREIRA, & BIALEK, 1999) e (TISHBY & ZASLAVSKY, 2015). Infelizmente, existe um ônus associado a esse procedimento, uma vez que ao optar por usá-lo, mudamos o referencial de escalas das nossas medidas, ou seja, a entropia e informação mútuas medidas não são e nem podem ser

comparadas aos valores reais das variáveis contínuas. Como (GEIGER, 2020) diz, os resultados obtidos devem ser comparados de forma cuidadosa - é prudente analisar experimentos que utilizam configurações e estimadores semelhantes.

Portanto, com as variáveis discretas, podemos estimar as medidas de informação, pois agora são finitas em decorrência da discretização, e, além disto, é possível trabalhar com a DPI sem violações teóricas severas relacionadas à entropia diferencial e suas propriedades, permitindo o estudo da dinâmica do treinamento das redes neurais através do IB.

## 5.2 Configuração do Experimento e *Dataset*

Para estudar a teoria da informação aplicada a redes neurais artificiais, IBDL (do inglês, *Information Bottleneck theory of Deep Learning*), em problemas de regressão, escolhemos como objeto de estudo uma versão supervisionada do problema de separação de fontes (Romano et al., 2011). Este problema possui uma ampla gama de aplicações, e, em sua versão não-supervisionada, é um dos pilares da moderna teoria de processamento de sinais. Ele foi escolhido para este trabalho por suscitar um problema de regressão relativamente simples, com solução linearmente atingível.

Consideraremos o problema de separação de fontes supervisionado do ponto de vista da reconstrução do sinal fonte $S_1$ e/ou $S_2$ a partir dos sinais de saída $X_1$ e $X_2$ que são resultado da mistura (composição) dos sinais de fonte. Ou seja, pretende-se recuperar o sinal de origem a partir de sinais compostos. O diagrama do experimento desenvolvido para o presente trabalho é ilustrado pela Figura 5.1.

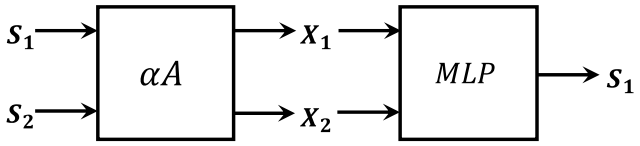



Figura 5.1 - Esquema da separação de fontes e das redes neurais proposto para o trabalho.

Os sinais de fonte (*sources*) $S_1$ e $S_2$, para o nosso problema, são gerados como variáveis gaussianas por se tratar de uma distribuição matematicamente bem caracterizada. Os sinais são criados com média e variância definidas como $S_1, S_2 \sim N(0, 0.12)$. A matriz $A = \begin{bmatrix} 1.0 & 0.7 \\ 0.7 & 1.0 \end{bmatrix}$ responsável pela mistura dos sinais $S_1$ e $S_2$ é construída de forma a ser inversível ($|A| \neq 0$), garantindo a reconstrução dos sinais de fonte através dos sinais $X_1$ e $X_2$ posteriormente. Por fim, o coeficiente $\alpha$ é um fator de escala da mistura que pode ser utilizado em conjunto a matriz $A$ para, eventualmente, normalizar os

sinais $X$ (utilizamos $\alpha = 0.55$). O conjunto de dados possui 2000 amostras das quais, 1500 serão utilizadas para treinamento e 500 para teste.

Tanto $\sigma^2$ quanto $\alpha$ foram determinados de tal forma que $S$ e $X$ estejam contidos dentro do intervalo $[-1,1]$: a razão dessa escolha jaz na padronização do experimento para termos consistência na estimação das medidas de informação. Segundo (CHELOMBIEV, HOUGHTON, & O'DONNEL, 2019), como os valores de informação analisados resultam diretamente do processo de estimação, eles são, portanto, altamente sensíveis a ele. Por tal razão, é interessante, se não necessário ("vital", segundo os autores), manter a consistência do experimento e do processo de estimação, principalmente quando usadas funções que não são saturadas. Garantir que as variáveis sejam bem-comportadas e bem definidas tem este exato propósito: contribuir para a consistência do experimento e, posteriormente, para o procedimento de estimação, além de reduzir os erros sistemáticos. A aparente violação da DPI no plano de informação é bem frequente (GEIGER, 2020): isso decorre de imprecisões no processo de estimação. Para mitigar ao máximo possíveis fontes de erros que possam vir a deteriorar as medidas de informação, foi feita essa padronização do espaço, que, posteriormente, é importante para a estimação via *binning*.

Retomando o objeto de estudo deste trabalho, o experimento consiste, primeiramente, em apresentar o sinal resultante da combinação linear, como ilustrado pela Figura 5.1, à rede neural. Através do treinamento supervisionado, temos como objetivo fazer com que a rede aprenda a relação entre os sinais determinado pela matriz $A$ de tal forma que consiga extrair / reconstruir um dos sinais da fonte, neste caso $S_1$. Geometricamente, a rede neural tem que desfazer a transformação linear que rotaciona e alonga os dados de entrada da fonte, Figura 5.2.

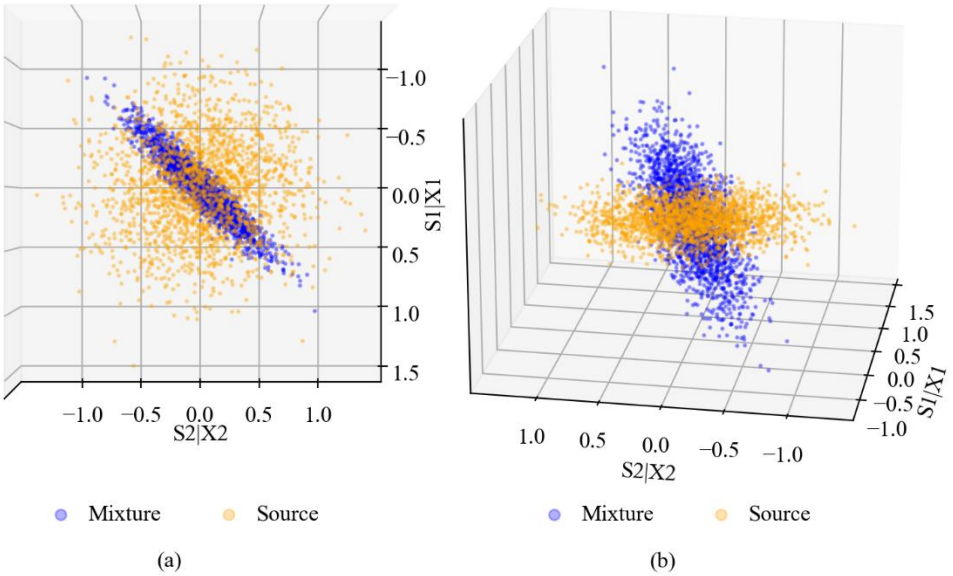



Figura 5.2 - Conjunto de dados utilizados para regressão. a) Em laranja $S_1$ e $S_2$ formam uma variável gaussiana bidimensional, em azul a transformação dada pela matriz A. b) Gráfico de X1 e X2 que são a entrada da rede por $S_1$ no eixo z demonstrando a rotação dos dados.

Para estudar a teoria da informação aplicada a redes neurais artificiais em problemas de regressão, será utilizada uma rede neural do tipo *MultiLayer Perceptron* (MLP), descrita na Seção 2.4.2. Esta rede foi escolhida por ser adequada ao problema proposto para o trabalho, além de ser uma das estruturas mais estudadas no âmbito do IBDL, como discutido no Capítulo 4. Em problemas de regressão/aproximação, é bastante comum o uso de redes denominadas aproximadores / regressores universais, capazes de aproximar com precisão arbitrária uma ampla classe de funções. Tendo em vista que MLPs com uma camada intermediária já são aproximadores universais, a estrutura utilizada em nosso trabalho consiste em uma rede com duas entradas, uma camada oculta com dois neurônios e uma saída, representada como $[2, 2, 1]$. Tanto a dimensão dos dados (e da entrada) quanto a dimensão da camada interna da rede foram definidas como sendo iguais a dois por uma questão de simplicidade, tendo em vista o caráter inicial deste esforço. Ademais, uma rede com esta configuração é suficiente, em tese, para resolver o problema de separação proposto, que é caracterizado por ser um problema linear com solução atingível. Logo, esta rede possui capacidade de processamento suficiente para alcançar o objetivo proposto de regressão linear (HAYKIN, 2001), pois a relação entre os dados é formada por uma lei linear (combinação linear), ou seja, um único neurônio seria capaz de realizar tal tarefa.

Para treinar a rede, utilizamos o método do gradiente descendente com o auxílio do algoritmo de retropropagação de erro (*backpropagation),* discutido na Seção 2.5.1, para o cálculo das derivadas necessárias para tais ajustes ao longo do treinamento. O treinamento da rede, como mencionado, será supervisionado, com gradiente estocástico e passo de adaptação $\eta = 0.001$. Este coeficiente foi selecionado de acordo com testes preliminares, para que proporcionasse à rede uma velocidade de convergência adequada para análise da dinâmica de treinamento. Um coeficiente muito elevado levaria a um treinamento relativamente rápido, não sendo adequado para o nosso objetivo, e o inverso não traria benefícios a não ser prolongar demasiadamente o treinamento sem necessidade.

Para a função custo, escolhemos o erro quadrático médio, que é utilizada comumente em problemas de regressão (Goodfellow, Bengio, & Courville, 2016). Para as funções de ativação da camada, foram selecionadas as funções; tangente hiperbólica ("*tanh.*"), sigmoide ("*sig.*") e *ReLU*, que são funções comuns em DL e são as funções mais estudadas nesse contexto do IBDL, como discutido no Capítulo 4. Como função da ativação da camada de saída, utilizamos unicamente a função "*identidade*", cuja aplicação é usual em problemas de regressão. Serão treinadas 20 redes para cada função utilizada na camada interna com inicializações distintas, totalizando 60 redes em todo o experimento: este procedimento é comum nos trabalhos da área, pois a dinâmica do plano de informação é reflexo da estocasticidade da inicialização (CHELOMBIEV, HOUGHTON, & O'DONNEL, 2019). Diferentes inicializações produzem diferentes planos de informação, e também podem apresentar

dinâmicas distintas para outras estruturas, (FANG, WANG, & YAMAGUCHI, 2018), (SCHIEMER & YE, 2019), (TAPIA & ESTÉVEZ, 2020) e (YU, et. al., 2020). Portanto, a média das redes treinadas é um bom indício da generalização do comportamento encontrado em cada não linearidade utilizada no plano de informação.

O presente trabalho busca explorar de forma inicial um problema de regressão através da teoria IBDL, é relevante ressaltar que, por questão de coerência com a proposta do trabalho, buscou-se trabalhar em um cenário bem definido e com um escopo adequado. Ou seja, neste trabalho para não elevar demasiadamente a complexidade do trabalho e das análises que serão desenvolvidas, não serão estudados certos elementos comuns aos problemas de aprendizagem profunda. Este trabalho não busca explorar inicialmente algoritmos de treinamento robustos como abordado em alguns dos problemas de classificação (CHELOMBIEV, HOUGHTON, & O'DONNEL, 2019), nem técnicas de regularização para evitar o sobre ajuste como parada antecipada (*early-stopping*), também estudado em classificação (SCHIEMER & YE, 2019), porém, esses elementos que compões as boas práticas em aprendizagem profunda devem ser explorados em trabalhos futuros, a fim de complementar e melhorar os estudos sobre os problemas de regressão através da teoria da informação.

## 5.2 Metodologia de Estimação e Parâmetros

No Capítulo 4, vimos que a escolha dos parâmetros e estratégia utilizadas no processo de discretização afetam de maneira direta a informação mútua estimada – consequentemente, isto afeta o que é observado no plano de informação (GEIGER, 2020) e (SAXE, et al., 2018), (CHELOMBIEV, HOUGHTON, & O'DONNEL, 2019).

Como mencionado em (GEIGER, 2020), a violação da DPI pelas medidas estimadas é frequente, consequência do procedimento utilizado. Como (SCHIEMER & YE, 2019) e (GEIGER, 2020) discutem, o resultado do plano de informação está totalmente ligado ao resultado do estimador, e, portanto, são necessários padronização e cuidado com o processo de estimação. Entretanto, na literatura, não há um desenvolvimento consistente sobre a metodologia e / ou atenção necessária sobre a escolha dos parâmetros dos estimadores, sendo esta uma parte importante dos trabalhos. Caso houvesse certo cuidado com este processo poderia haver resultados mais coesos com relação as equações da DPI sem tantas violações. Existe, como discutido no Capítulo 4, uma ampla aplicação e busca por estimadores que apresentem resultados coerentes com as relações estabelecidas pelo problema do IBDL, mas, não há profundidade no trato dos estimadores e seus parâmetros.

Ainda mais, devido à falta de trabalhos aplicados a problemas de regressão semelhantes ao proposto por este, existe a necessidade de um cuidado atento com todo o processo de discretização e estimação das medidas de informação utilizadas na análise do problema. Por esta razão, nesta seção, vamos definir e desenvolver a metodologia utilizada para discretização e seleção dos parâmetros do estimador, que será utilizado para conduzir nossos experimentos posteriormente.

Para dar início a esta elaboração, utilizaremos a discretização uniforme, definida na Seção 3.10. Utilizamos a notação $[a, b, c]$ para definir nosso estimador: esta notação segue a forma de $[X_{Nºbins}, T_{Nºbins}, Y_{Nºbins}]$, ou seja, o primeiro valor $X_{Nºbins}$ representa o número de intervalos utilizados para discretizar o conjunto de dados $X$, o segundo valor $T_{Nºbins}$ representa o número de intervalos utilizados para discretizar as camadas da rede. Por fim o último valor $Y_{Nºbins}$ é o número de intervalos utilizados para discretizar o conjunto de dados $Y$. Todas as medidas de informação do trabalho são realizadas com a base 2, portanto, todas as medidas são em *bits*. A Tabela 5.1 resume alguns detalhes sobre o método de discretização, e alguns pontos são discutidos logo a seguir.

Tabela 5.1: Metodologia de Discretização do Experimento.

| | $[a, b, c]$ |
|---|---|
| $X_\Delta$ e $Y_\Delta$ | $[-1, 1]$ |
| $Tanh$ | $[-1, 1]$ |
| $Sig$ | $[0, 1]$ |
| $ReLU$ | $[0, \overline{max}]$ |
| *Ident* (output) | $[-1.3, 1.3]$ |

Como mencionado anteriormente, a padronização do experimento é importante para trazer estabilidade na estimação das medidas de informação (SCHIEMER & YE, 2019) e (GEIGER, 2020). O conjunto de dados descrito na Seção 5.2 é bem definido: tanto o conjunto $X$ quanto o conjunto $S_1 = Y$ dos dados foram gerados de forma bem comportada dentro de um intervalo padronizado - dessa forma, reduzimos os erros sistemáticos envolvidos no processo de estimação das medidas de informação, e, como descrito na Tabela 5.1, a discretização do conjunto de dados é realizado nos limites naturais dos dados.

Para estimar as medidas de informação, é necessária a discretização das funções de ativação das redes neurais. A discretização das funções *tanh.* e sig. é um processo simples (SCHIEMER & YE, 2019), pois são sempre discretizadas em seu contradomínio, que é limitado. Por razões óbvias, esses limites de discretização são fixos, ou seja, inalterados durante o processo de estimação como mostrado na Tabela 5.1. Já para funções que não são saturadas, como no caso da *ReLU,* o processo

não é tão intuitivo assim. Em (SAXE, et al., 2018), a função *ReLU* é discretizada entre 0 (saturação inferior) e o valor máximo atingido pela camada durante o treinamento $[0, max_{treino}]$, segundo o autor este procedimento é realizado pois todo o espaço explorado pela função deve ser incluso na discretização.

Porém, esta abordagem não é adequada, segundo (CHELOMBIEV, HOUGHTON, & O'DONNEL, 2019): como, em cada época, os limites atingidos na função são diferentes, a informação estimada de cada época dependerá de fatores que não estão relacionados com as propriedades das distribuições em cada uma dessas épocas.

Suponhamos que o valor máximo em uma determinada camada de uma rede neural qualquer atinja um certo valor máximo $m$ durante o treinamento, logo, a função *ReLU* seria discretizada em todas as épocas no intervalo $[0, m]$. Entretanto, caso este valor tenha ocorrido uma única vez ao longo do treinamento, a frequência de incidência deste valor de ativação não é representativa, ou seja, ao definir este valor como "referência" para determinar o valor máximo da discretização da *ReLU* para todo o treinamento, estamos simplesmente ignorando o real comportamento da dispersão dos dados formada nas camadas da rede, uma vez que estamos praticamente considerando apenas um único valor determinado sem qualquer base estatística dos dados.

Como resultado dessa abordagem, muitas das distribuições de probabilidade das camadas podem estar sendo discretizadas em um intervalo maior que o necessário, ou seja, corre-se o risco de comprimir a própria distribuição e com isto perder parte da dinâmica desses dados, algo necessário para uma análise correta. Considerando esses pontos levantados, vamos definir uma estratégia de discretização para a função *ReLU* que leve em consideração algumas propriedades dos dados em questão. O limite superior será definido em função da média dos valores máximos atingidos em cada época: este valor é considerado o valor padrão de discretização. Para as épocas que tiverem valores máximos maiores que este valor padrão, consideraremos o novo valor máximo somente para a época em questão, ou seja, pontualmente. Esta abordagem semi-adaptativa foi considerada uma vez que, mesmo utilizando a média dos valores máximos, ainda assim houve épocas que apresentaram valores superiores ao definido, saturando a função *ReLU,* algo que não pode ocorrer de forma alguma. Logo, com esta abordagem semi-adaptativa, eliminamos por completo qualquer tipo de saturação e não perdemos qualidade nas estimativas; pelo contrário, o resultado observado desta abordagem apresentam uma correção das estimações de informação mútua que violavam a DPI. Como comparativo testamos a forma utilizada pelos autores em (SAXE, et al., 2018) com o valor máximo global versus a nossa abordagem, a Figura 5.3 ilustra esses resultados:

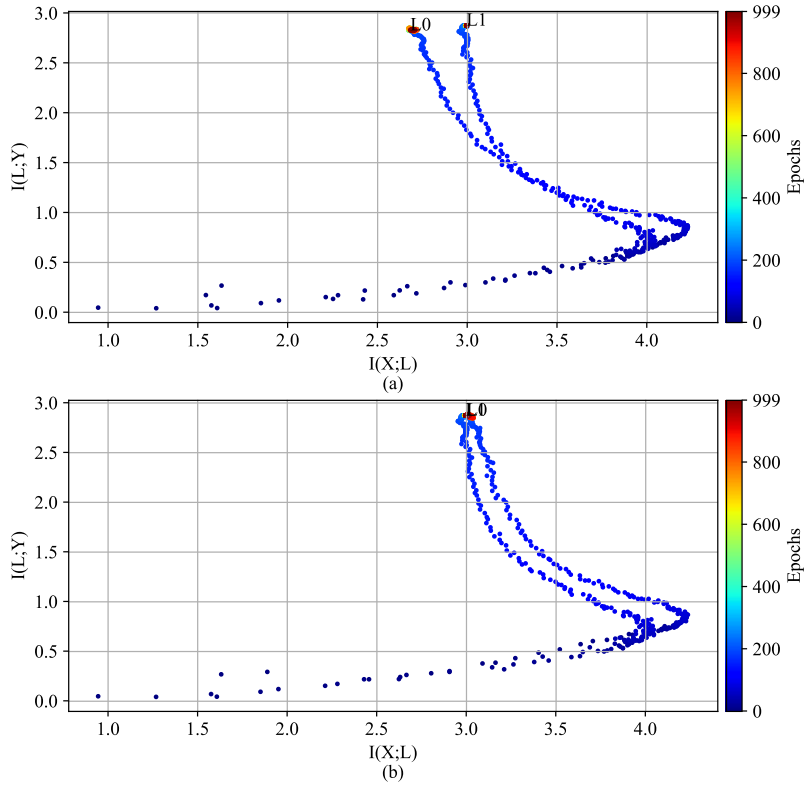



Figura 5.3 - Planos de informação para a função ReLU. a) Função discretizada segundo (SAXE, et al., 2018). b) Função discretizada utilizando nosso método.

Outra parte bastante importante do processo de discretização é referente à discretização da função de saída da rede que utiliza como não-linearidade a função identidade que não possui restrições / saturação. Em (SCHIEMER & YE, 2019), os autores colocam que discretizar funções duplamente saturadas é um processo relativamente simples, pois os limites do domínio de discretização são definidos pelas próprias funções. Entretanto, para funções não saturadas é necessário determinar tais limites uma vez que o processo de estimação utilizado exige esses valores. Para definir os limites da função identidade, utilizamos um procedimento mais direto, como devemos utilizar os mesmos limites de discretização para todas as redes testamos valores de limites que vão de $\pm 1.0$ até $\pm 1.5$ e medimos percentualmente a média da quantidade de amostras que são saturadas em cada época do treinamento. Segundo nossos testes para valores maiores que $\pm 1.3$ as estimativas não sofrem alteração assim como o plano de informação, na realidade o valor 1.1 já é o suficiente para manter a consistência da informação estimada, entretanto, utilizamos um valor um pouco superior por segurança. O resultado dessa medida é ilustrado na Figura 5.4:

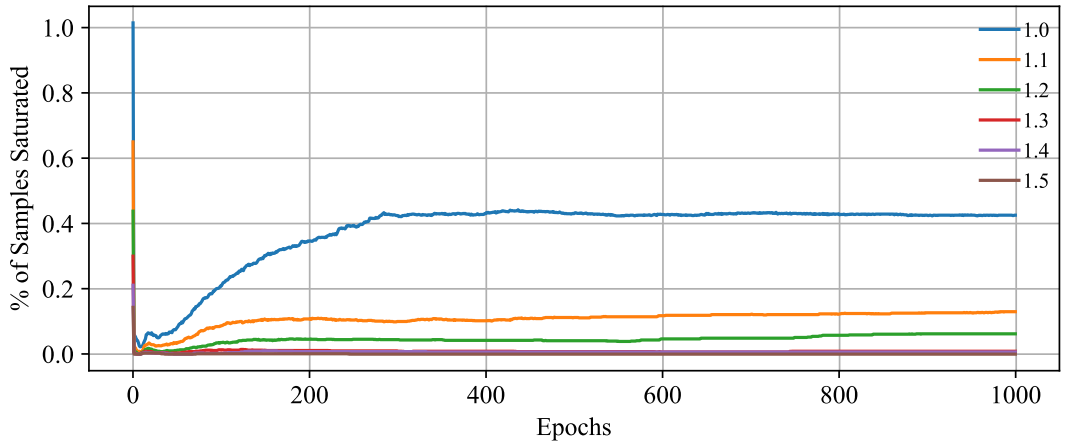



Figura 5.4 - Média percentual do número de ativações que saturam os limites testados para discretização da função identidade na saída da rede. O conjunto de dados de treinamento possui 1500 amostras.

Definidos os limites de discretização para o conjunto de dados e as funções de ativação, realizamos uma busca de parâmetros para nosso estimador. O espaço de busca contempla todas as combinações de parâmetros $[a, b, c]$ variando de 10 a 150 intervalos para cada um dos parâmetros $a$, $b$, e $c$, totalizando 2.800.000 combinações. Cada combinação então é usada para calcular as medidas de informação $I_X = I(X;L)$ e $I_Y = I(L;Y)$ das camadas da rede bem como $H(X)$, $H(Y)$ e $I(X;Y)$. Posteriormente uma rede neural foi escolhida aleatoriamente como *benchmark* para gerar todas essas medidas para cada uma das combinações.

Como mencionado anteriormente, manter as relações da DPI é necessário, uma vez que, com frequência, os resultados reportados na literatura violam tais relações (GEIGER, 2020), para auxiliar na seleção dos parâmetros do estimador levamos em consideração essa necessidade e a fim de evitar violações grosseiras da DPI definimos a seleção dos parâmetros em duas etapas.

A primeira etapa consiste em aplicar dois filtros $I(X;L_0) \geq I(X;L_1)$ e $I(L_0;Y) \geq I(L_1;Y)$ derivados das condições estabelecidas pela própria DPI Equação 4.1 e 4.2, naturalmente considerando um pequeno erro de $\pm 1\%$ nesses filtros pois o processo de estimação não é perfeito. Com isto, são excluídas as combinações que geram estimações que violam a DPI de forma grosseira.

A segunda etapa consiste em utilizar uma função de erro para ordenar as combinações restantes a fim de estruturar e facilitar a escolha dos parâmetros. A função erro utilizada é definida como:

$$e = abs\left(\hat{I}(\hat{Y};Y) - \hat{H}(Y)\right) \tag{5.7}$$

Esta função nada mais é que a relação da auto-informação de variáveis discretas cuja igualdade é definida como $H(Y) = I(Y;Y)$ (COVER & THOMAS, 2006). Este procedimento é realizado para garantir que a comparação entre a informação $I_Y$ da camada de saída e a entropia do conjunto de

dados $H(Y)$ seja comparável, uma vez que a rede treinada exibirá na saída a mesma distribuição que $Y$, ou seja, $\hat{Y} \approx Y$, logo, a relação de auto-informação deve ser respeitada para que os resultados tenham coerência.

Com as combinações resultantes realizamos alguns testes mais gerais incluindo todas as redes neurais treinadas, com o intuito de eliminar aquelas combinações de parâmetros que porventura possam gerar ainda alguma violação da DPI. Aplicando a metodologia definida chegamos à combinação $[26{,}150{,}10]$ que será utilizada para discretizar as variáveis e estimar as medidas de informação.

## 5.3 *Information Bottleneck* em Regressão

Nesta seção, de maneira breve, faremos algumas considerações a respeito do problema de regressão e alguns detalhes que são interessantes para o trabalho.

As relações estabelecidas para o problema do IB (TISHBY, PEREIRA, & BIALEK, 1999), (TISHBY & ZASLAVSKY, 2015) e (SCHWARTZ-ZIV & TISHBY, 2017) se mostram suficientemente sólidas para extrapolarmos sua teoria para o problema de regressão, inclusive existem trabalhos com *auto-encoders* (WICKSTRØM, et al., 2019), (YU & PRÍNCIPE, 2019) e (LEE & JO, 2021) que se assemelham bastante ao problema de regressão em diversos aspectos do treinamento das redes. Portanto, pode ser promissora a aplicação do IB em problemas de regressão da forma que propomos.

Para isto, consideramos algumas hipóteses a respeito do problema IB para serem estudadas neste trabalho, recapitulando as equações fundamentais do problema segundo a DPI temos:

$$I(X;Y) \geq I(L_0;Y) \geq \cdots \geq I\left(L_k = \hat{Y};Y\right) \tag{5.8}$$

$$H(X) \geq I(X;L_0) \geq \cdots \geq I\left(X;L_k = \hat{Y}\right) \tag{5.9}$$

onde $X$ representa o conjunto de dados da entrada da rede, $Y$ representa o conjunto de dados desejados (rótulos), $L$ representa as camadas da rede.

Para o problema de regressão esperamos que as relações descritas pela Equação (5.8) e (5.9) se mantenham, ou seja, que a hierarquia estabelecida para as informações mútua $I(L;Y)$ e $I(X;L)$ seja respeitada. Além disto, cogitamos a hipótese de que a camada interna da rede apresente um comportamento bastante elástico em relação a quantidade de informação retida, ou seja, esperamos observar redes que consigam capturar toda a informação contida nos dadas $X$ ou parte dessa informação, a Figura 5.5 ilustra o comportamento esperado para a camada interna da rede.

Feita essas colocações a respeito do problema de regressão temos como objetivo estudar e analisar os seguintes pontos:

1. Analisar a dinâmica do processo de treinamento e possibilidade do uso do problema do IB e as relações da DPI em problemas de regressão.
2. Verificar se o fenômeno de compressão ocorre em problemas de regressão e caso ocorra entender as possíveis fontes desse comportamento.

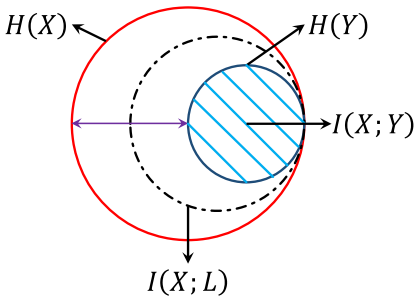



Figura 5.5 - Diagrama de Venn para a DPI. A camada interna da rede pode reter diferentes quantidades de informação sobre $X$.

# 6. Resultados e Análises

Neste capítulo, exibiremos os resultados obtidos aplicando a estratégia de discretização estabelecida no capítulo anterior às redes treinadas, e analisamos os resultados. Utilizando o plano de informação, investigamos a manifestação do fenômeno de compressão em regressão e estabelecemos algumas relações entre a causa da compressão e a viabilidade do uso do IB em regressão. A discussão será estabelecida nas devidas proporções seguindo um paralelo com a literatura disponível até o momento, e os resultados estão divididos em duas seções. A Seção 6.1 discute a manifestação da compressão e as possíveis causas de tal fenômeno; a Seção 6.2 discute o impacto das não linearidades no processo de treinamento, levantando alguns pontos que contribuem com o enriquecimento da discussão estabelecida pelos artigos da área a respeito da teoria do *Information Bottleneck theory of Deep Learning* (IBDL).

## 6.1 O Fenômeno da Compressão em Regressão

Iniciamos nossas análises estudando o fenômeno de compressão no problema de regressão: os resultados médios dos planos de informação para as funções internas da rede são ilustrados através da Figura 6.2. O plano de informação resultante mostra diferentes dinâmicas de informação para a camada interna da rede $L_0$, foco das nossas análises, observamos que para a função s*ig.*, a fase de ajuste (*fitting phase,* Seção 4.1) caracterizada pelo aumento da informação $I_Y$ é predominante. Para as redes com função *ReLU,* observamos que a fase de ajuste ocorre juntamente com a compressão da informação $I_X$, e, para as redes com função *tanh.,* de forma similar, vemos apenas a fase de ajuste.

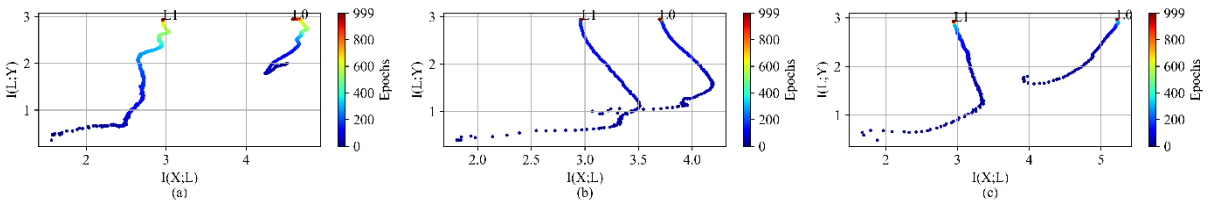



Figura 6.1 - Plano de informação para diferentes não-linearidades na camada interna da rede. A média de 20 inicializações mostram caminhos distintos para a informação da camada interna $L_0$. a) Sigmoid. b) ReLU. c) Tanh. Informação estimada em bits.

Por definição, a compressão é caracterizada pela redução da informação mútua $I(X;L)$ ao longo do treinamento. De forma geral, na média, um comportamento que está associado às redes treinadas com a função *ReLU*, porém, a compressão pode ocorrer em todas as não-linearidades utilizadas (*Tanh*, *Sig*, *ReLU*), como vemos através da Figura 6.2, em diferentes graus de intensidade e com dinâmicas bastante ricas em decorrência das diferentes inicializações (CHELOMBIEV, HOUGHTON, & O'DONNEL, 2019).

Para um compreendimento correto da leitura e interpretação dos gráficos ver Apêndice A.

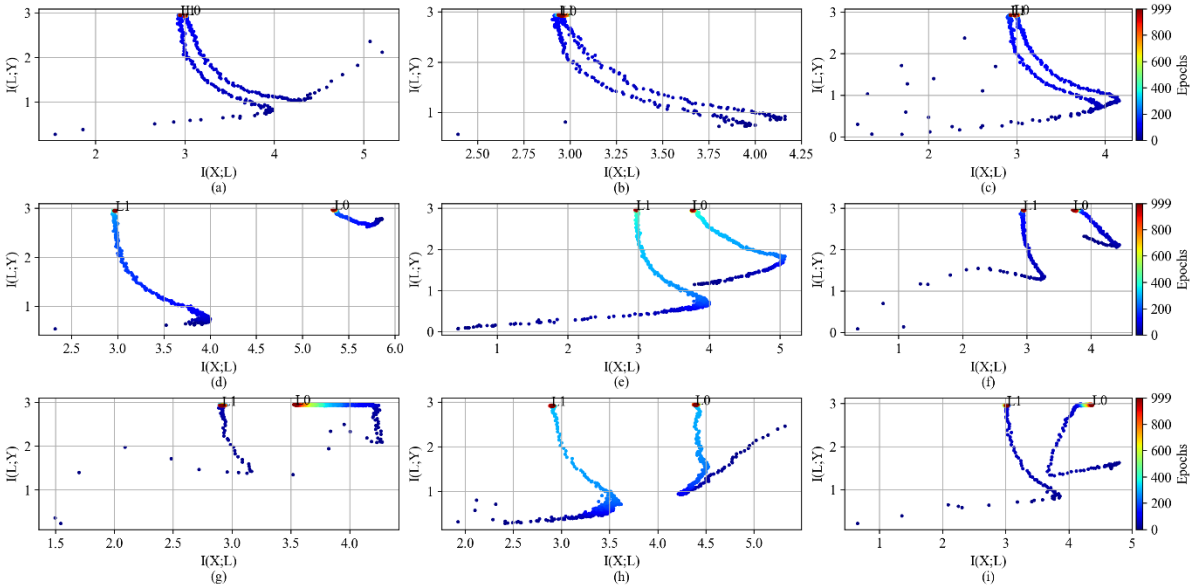



Figura 6.2 - Manifestação da compressão nas funções de ativação para diferentes inicializações. **a**,**b**, e **c** redes com a função ReLU. **d**, **e**, e **f** redes com função Sig e **g**, **h**, e **i** redes com função Tanh. A compressão é caracterizada pela redução da informação $I(X;L_0)$, vide Apêndice A. A variação da cor das curvas de informação representam as épocas de treino.

Identificada a manifestação da compressão, o próximo passo é entender o que promove tal fenômeno. Para investigar a possível fonte da compressão no problema de regressão, vamos analisar o processo de treinamento da rede de forma mais detalhada. Para as redes neurais com a função *ReLU*, ao comparar o instante de maior informação $I(X;L_0)$ anterior ao processo de compressão, observamos que a distribuição gaussiana formada na camada interna possui variância elevada, e conforme a compressão ocorre a distribuição tem sua variância é reduzida, como ilustrado através da Figura 6.4. Para variáveis gaussianas, a entropia é determinada em função da variância da distribuição - Equação (3.74): neste caso, a distribuição formada na camada interna da rede forma uma distribuição gaussiana, uma vez que a parte linear da função *ReLU* não altera a natureza da distribuição dos dados aplicados a entrada da rede. Esta redução de entropia da camada interna por consequência resulta na redução da informação mútua Equação (3.30) gerando a compressão observada.

Entretanto, para o caso das funções sig e tanh a análise da compressão não tão é simples. A distribuição das ativações formada na camada interna sofre um processamento não linear referente as funções utilizadas, o que significa que as distribuições sofrem alterações. Como o exemplo ilustrado na Figura 6.4 cuja distribuição formada no instante de maior informação $I(X;L_0)$, possui a princípio ativações distribuídas de forma mais uniforme produzindo maior entropia, posteriormente ao assumir uma distribuição teoricamente gaussiana que possui menor entropia, se comparada a uma distribuição mais "uniforme" (COVER & THOMAS, 2006), entretanto não podemos afirmar com certeza a natureza das distribuições como consequência da transformação não-linear efetuada pelas funções de ativação. Da mesma forma, a redução da entropia da camada produz a compressão da informação mútua medida como observada no plano de informação.

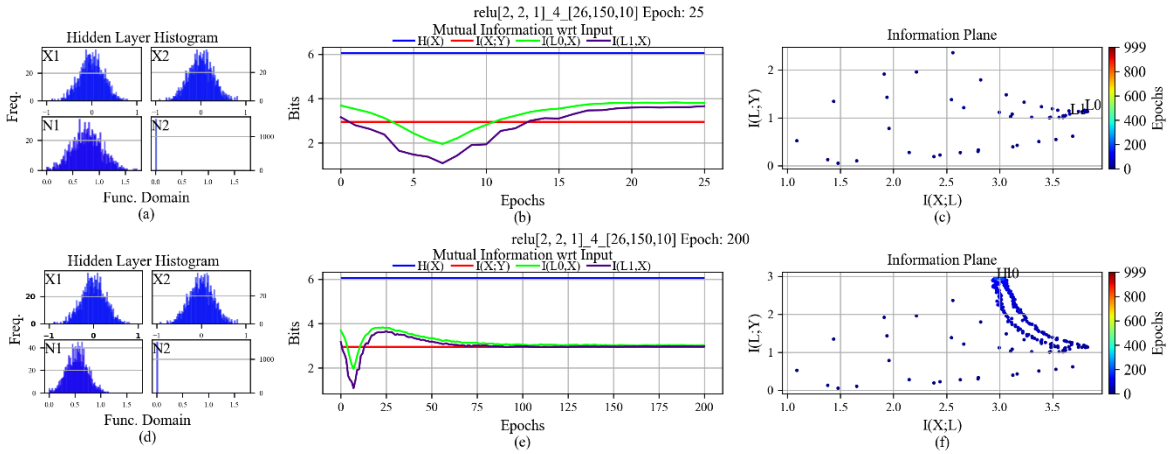



Figura 6.3 - Detalhamento do treino de uma rede com a função ReLU. O processo de compressão fig b) e e) é associado variação da variancia da distribuição gaussiana formada internamente na rede fig a) e d).

A compressão pode ser um processo difícil de analisar, como visto, devido à forma como a transformação não linear pode alterar as distribuições formadas internamente pela rede. Outro exemplo deste impacto é ilustrado na Figura 6.5 para uma rede com função tanh. Vemos, principalmente, que a redução da entropia é consequência da redução da concentração da dispersão dos dados em função do comportamento apresentado pelo neurônio $N_2$, não apenas este fato como a própria alteração da distribuição em $N_1$ também pode ter sua parcela de contribuição na compressão observada.

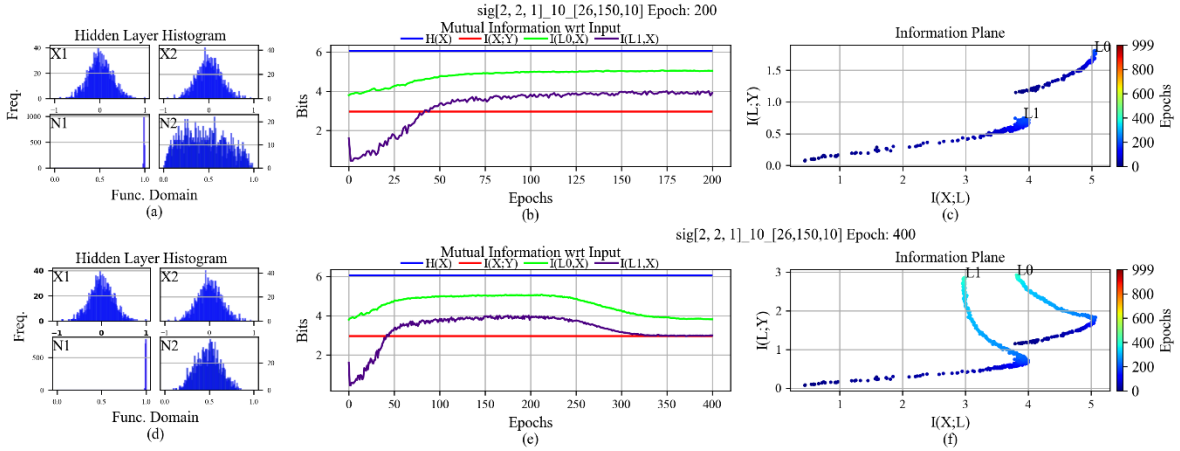



Figura 6.4 – Processo de compressão em redes sigmoid. A distribuição formada internamente em uma rede fig a) apresenta os dados distribuidos de forma mais dispersa, apresentando a maior entropia, consequentemente a distrubição formada posteriormente tem entropia menor fig d), produzindo a redução da informação mútua I(X;L0) fig b) e e).

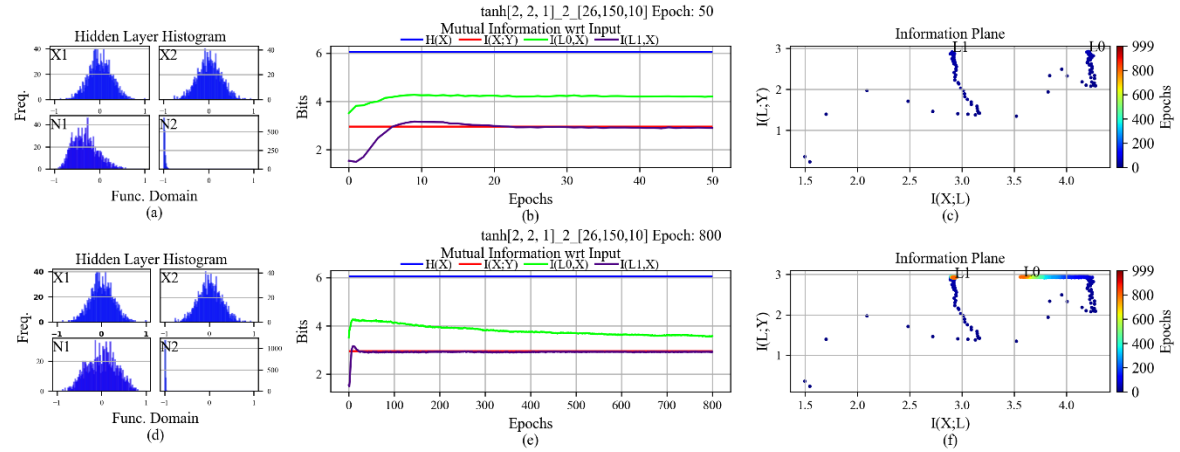



Figura 6.5 - Processo de comrpessão em uma rede com a função tanh. A redução da informação mútua em decorrencia da redução da entropia de distribuições bidimencionas fig a) e d), consequentemente o comportamento manifestado em c) e f).

Diferentemente do que é relatado no problema de classificação a compressão observada não sofre da indução de classe (GOLDFIELD, et al., 2019); (SCHIEMER & YE, 2019) uma vez que não estamos trabalhando com *one-hot-encoding* ou exclusivamente consequência das não-linearidades (SAXE, et al., 2018) uma vez que observamos o fenômeno de compressão em todas as funções utilizadas. Os indícios observados nos resultados obtidos apontam que a fonte de compressão pode ter origem estatística. Com as redes que utilizam a função *ReLU* observamos que a compressão pode estar associada a variância da distribuição gaussiana formada na representação interna da rede, sugerindo que a compressão tenha relação com a distribuição formada (GEIGER, 2020). Entretanto, para as funções sig e tanh as análises não são tão simples, pois a transformação não linear dessas funções é bastante presente alterando substancialmente suas representações internas. Mesmo assim há indícios de que a compressão pode ser originalmente estatística, onde a compressão da informação se mostrou em decorrência da redução da entropia das distribuições formadas na camada interna da rede.

## 6.2 Compressão e Generalização: Os limites teóricos e a convergência da rede

Na seção anterior, vimos pelos resultados observados, que o fenômeno de compressão não parece ter origem nas não linearidades utilizadas. Entretanto, as funções utilizadas influenciam de maneira diferente do que esperávamos na compressão e expansão da informação apresentada no plano de informação. Nesta seção, vamos continuar as análises dos resultados e discutir outro ponto presente na literatura sobre a capacidade de generalização da rede e sua ligação com a compressão.

Quando Tishby propôs a teoria do IBDL (*Information Bottleneck theory of Deep Learning* (TISHBY & ZASLAVSKY, 2015)), foi contemplado pelo autor que a representação latente da rede deveria conter informação suficiente para determinação da saída. De forma que em um treinamento ótimo do ponto de vista da teoria da informação, uma rede neural deveria ser capaz de extrair as propriedades mais relevantes / informativas do conjunto de dados, essa informação retida, segundo a teoria da informação, é uma estatística suficiente mínima aproximada, e, segundo os autores uma rede com a arquitetura mais compacta naturalmente deveria apresentar tal grau de complexidade em sua representação latente. Além disto, os autores afirmam a necessidade da presença de compressão da informação $I(X;L)$ nas camadas ocultas, uma vez que a complexidade da representação (mapeamento complexo) latente é maior que a complexidade $(y = f(x))$ presente no conjunto de dados. Portanto, para uma rede neural apresentar boa generalização, seria necessária compressão da informação $I(X;L)$.

Nossos resultados além da presença da compressão discutida anteriormente, também mostram uma relação bastante interessante entre a representação latente das redes com os limites teóricos estabelecidos pela DPI Equação (5.8) e (5.9). Vamos analisar esses resultados e como eles se relacionam com as não-linearidades e com a capacidade da generalização das redes.

Parte das redes neurais treinadas atingiram o limite superior $I(X;L_0) = H(X)$ da DPI, Equação (5.9), para as diferentes não-linearidades utilizadas, pela DPI isto significa que a camada interna da rede capturou toda informação disponível em $X$ ao longo do treinamento. Do ponto de vista da teoria da informação a representação latente da rede pode ser entendida como uma "cópia" de $X$ no sentido informativo, pois possui toda sua informação. Para investigar e averiguar a coerência das medidas reportadas por nosso estimador, vamos analisar o plano de informação em conjunto da projeção internada da rede, ou seja, a projeção dos neurônios $N_1$ e $N_2$ da camada interna.

Para o caso da rede neural que utiliza a função *ReLU*, Figura 6.6, a projeção interna da rede, Figura 6.6 a), é exatamente uma representação do conjunto de entrada $X$ reescalada e rotacionada, uma vez que sua transformação que a função realiza é linear em sua faixa de operação positiva, ou seja, não altera a natureza dos dados de entrada da rede. Neste caso, a representação interna possui toda a

informação dos dados uma vez que se trata de uma "cópia" geométrica de $X$, consequentemente do ponto de vista da teoria da informação faz sentido a camada interna atingir a igualdade entre a informação mútua e a entropia dos dados $I(X;L_0) = H(X)$. Neste sentido, nosso estimador reportou medidas coerentes entre a informação capturada pela camada interna a respeito da informação disponível na entrada da rede Figura 6.6 b), como é mostrado pela projeção interna. Naturalmente a camada interna também contém toda a informação necessária para determinar $Y$ de forma correta como mostra a Figura 6.6 c).

Esse comportamento observado se repete de formas semelhantes para as demais funções de ativação usadas, como podemos observar na Figura 6.7**Erro! Fonte de referência não encontrada.** para função sigmoid, onde a representação interna contém toda a informação de $X$, Figura 6.7 b), e a projeção interna formada é uma representação mais compacta da entrada da rede por consequência da transformação não linear da função.

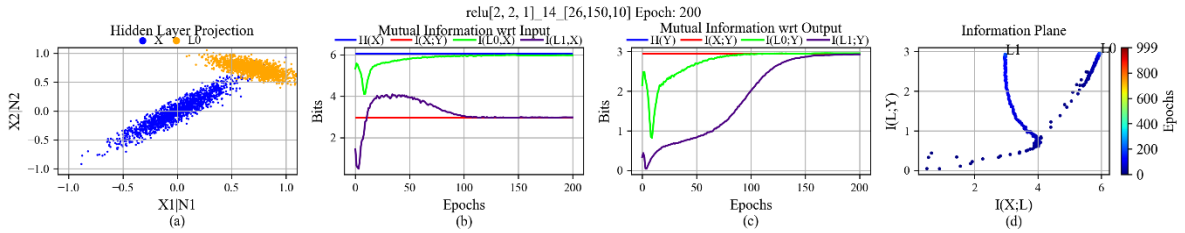



Figura 6.6 - Limite superior da DPI atingido por uma rede treinada com a função ReLU. A projeção da camada interna da rede apresenta uma versão reescalada dos dados em a) contendo toda a informação disponível, ou seja, $I(X;L_0) = H(X)$ em b).

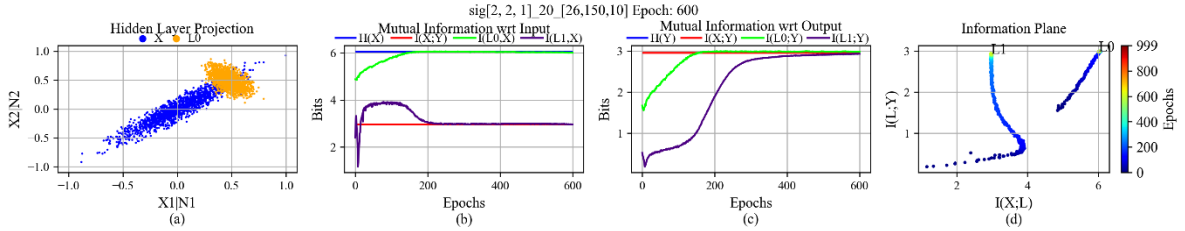



Figura 6.7 - Limite superior da DPI atingido por uma rede treinada com a função Sig. A projeção da camada interna forma uma versão compacta dos dados de entrada em a) contendo toda informação disponível , $I(X;L_0) = H(X)$ em b).

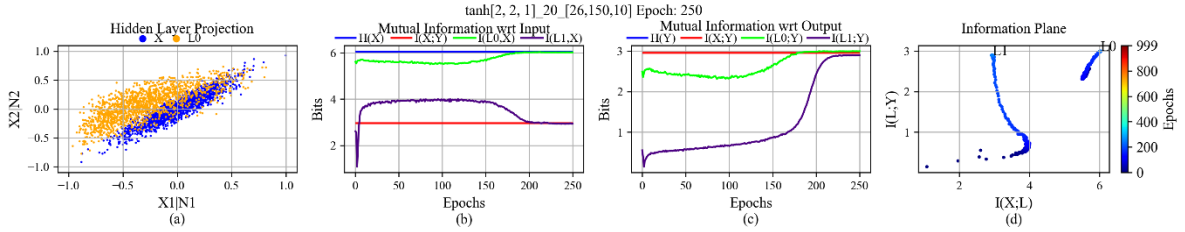



Figura 6.8 - Limite superior da DPI atingida por uma rede treinada com a função Tanh. A transformação não linear afeta a projeção interna da rede a) que se asemelha aso dados de entrada. A informação contida na camada interna é a maxima disponível, , $I(X;L_0) = H(X)$, como pode ser observado na figura b).

Em nossos resultados, a função tanh foi a não-linearidade que mais promoveu impactos na representação da rede no que se refere a projeção interna. Através da Figura 6.8, vemos que a camada

latente possui toda a informação disponível de $X$ Figura 6.8 b), e sua projeção interna é muito semelhante ao conjunto de dados de treinamento Figura 6.8 a). Entretanto, especificamente para as redes treinadas com a função tanh existem casos em que a transformação não linear afeta muito a projeção interna da rede e segundo a estimação essas representações também apresentaram toda a informação referente a $X$, como ilustrada pela Figura 6.9. Neste caso a análise se torna complexa e é difícil estabelecer um sentido entre a projeção interna da rede e a informação estimada.

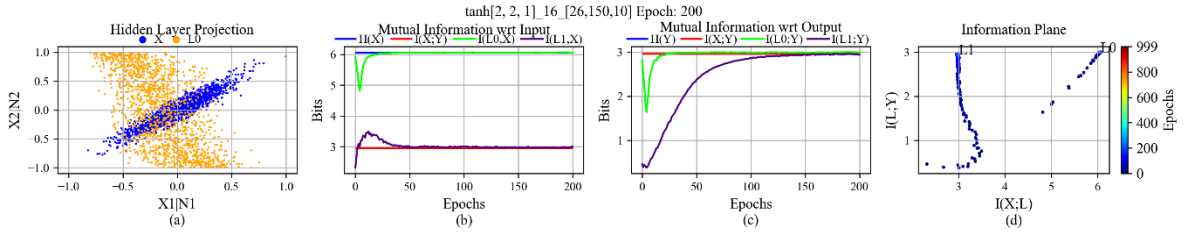



Figura 6.9 - Representação latente para função tanh onde a projeção interna é difusa a), esta representação interna contém toda a informação contida nos dados, , $I(X;L_0) = H(X)$, como observado em b).

Como mencionado também houve redes treinadas que apresentaram convergência da informação mútua para o limite inferior da DPI Equação (5.8), ou seja, $I(X;L_0) = I(X;Y)$, porém ocorreu quase que exclusivamente com as redes treinadas com a função *ReLU*. A Figura 6.10 e a Figura 6.11 ilustram este comportamento para as funções *ReLU* e *Sig*. Nesse ponto, não houve casos de convergência para o limite inferior da DPI em redes que foram treinadas com a função tanh.

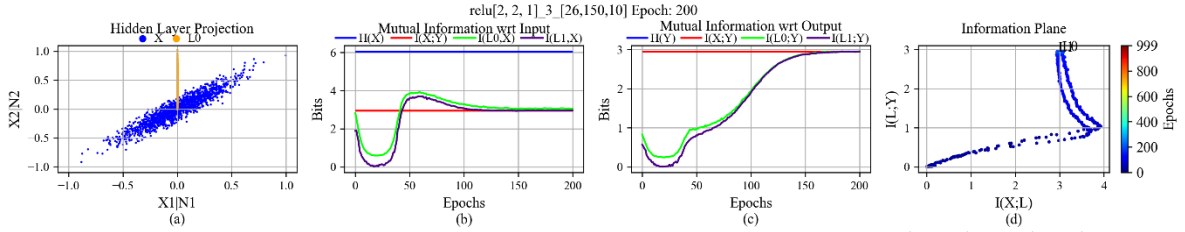



Figura 6.10 - Representação latente mínima atingida em um rede com função ReLU, $I(X;L_0) = I(X;Y)$.

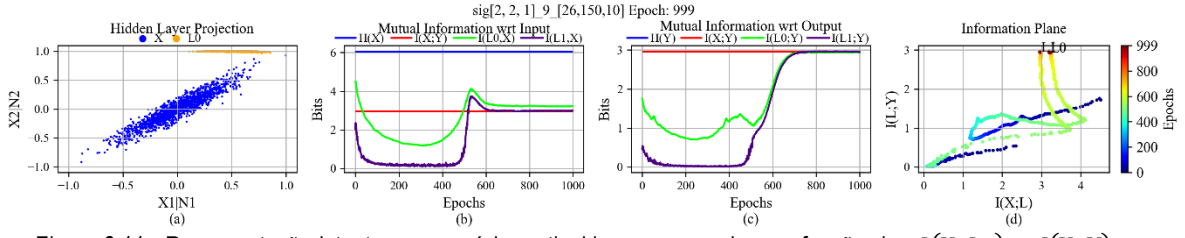



Figura 6.11 - Representação latente quase mínima atingida em uma rede com função sig., $I(X;L_0) \cong I(X;Y)$.

Segundo as formulação da teoria do IBDL estabelecidas por (TISHBY & ZASLAVSKY, 2015) a rede neural atinge a igualdade $I(X;L_0) = I(X;Y)$ quando sua representação interna forma uma estatística suficiente mínima aproximada, segundo os autores isto significa dizer que a rede absorveu apenas a informação essencial (mínimo) disponível em $X$ para estimar corretamente $Y$, ou seja, a própria

informação mútua $I(X;Y)$. Esta representação também é a que apresenta maior complexidade de $X$ (compressão) segundo a teoria do IBDL.

Apesar de as não-linearidades não terem apresentado influência na manifestação da compressão como suposto em (SAXE, et al., 2018), nossos resultados mostram que elas impactam consideravelmente a dinâmica do treinamento segundo a informação medida de suas representações latentes. É interessante ressaltar que todas as funções utilizadas apresentaram compressão ou expansão em seus planos de informação referente a informação $I(X;L_0)$, ou seja, não sendo considerado exclusividade das não linearidades.

Além do impacto causado na dinâmica de treinamento das redes e a quantidade de informação absorvida pela camada interna para as diferentes funções de ativação, foi observado que as saturações dessas funções também contribuem para essa dinâmica. Diferentemente do que é constatado em (SAXE, et al., 2018), a saturação, em nossos testes, auxilia as redes a atingirem a estatística mínima suficiente aproximada. As redes que apresentaram $I(X;L_0) = I(X;Y)$ tiveram em sua camada interna um comportamento interessante no que diz respeito a operação da rede, as redes operam com um dos neurônios saturados (desativados, para a função ReLU pode ter relação com o problema *Dying ReLU*), ou seja, as representações internas são formadas por distribuições unidimensionais que contém toda a informação a respeito de $Y$. Para o caso da função *ReLU* podemos dizer que a representação interna se trata de uma versão de $Y$ rescalada internamente.

Nesse sentido, esse comportamento flexível de operação da camada interna permite a rede a alcançar a representação interna de maior complexidade $I(X;L_0) = I(X;Y)$, que configuraria uma estatística suficiente mínima (COVER & THOMAS, 2006), como é proposto por (TISHBY & ZASLAVSKY, 2015). Em seu trabalho o autor diz que a rede de menor estrutura deve apresentar a representação mais complexa contida em seus pesos. De fato, é o que ocorre, em uma estrutura neural com uma única camada interna com dois neurônios o comportamento identificado reforça essa ideia. A rede simplesmente desativou um dos neurônios, saturando-o, e com isto atinge a máxima complexidade de $L_0$, ou seja, quando atingiu a estatística suficiente mínima operou como se fosse uma estrutura $[2,1,1]$ que no contexto de nossa estrutura representa a estrutura mínima, assim como (TISHBY & ZASLAVSKY, 2015) tinha sugerido.

Entretanto, Tishby também teoriza que uma rede bem treinada e que generaliza bem deve necessariamente apresentar compressão ao longo do treinamento, neste ponto nosso trabalho está alinhado com trabalhos mais recentes (CHELOMBIEV, HOUGHTON, & O'DONNEL, 2019); (GOLDFIELD, et al., 2019); (SAXE, et al., 2018) e (SCHIEMER & YE, 2019). Todas as redes treinas apresentaram boa generalização (Apêndice B) sem que houvesse necessariamente a compressão da informação mútua $I_x$. Ou seja, uma rede bem treinada como mostramos pode conter em sua

representação interna toda a informação contida em $X$ ($I(X;L) = H(X)$) Figura 6.6, apenas a informação útil necessária $I(X;L) = I(X;Y)$ (estatística suficiente mínima aproximada) Figura 6.10 ou uma representação intermediaria Apêndice C, todas elas contendo informação suficiente para determinar $Y$ corretamente. Por conta deste comportamento flexível nem sempre é possível identificar as duas fazes de treinamento como discutido em (SCHWARTZ-ZIV & TISHBY, 2017) Figura 6.2, principalmente quando a rede não apresenta algum grau de compressão como por exemplo as redes treinadas com a função tanh.

# 7. Conclusão

Neste trabalho, realizamos a análise da teoria de aprendizagem profunda por restrição de informação (*Information Bottleneck theory of Deep Lear*ning - IBDL) e de sua viabilidade em aplicações de problemas de regressão, avaliando aspectos como o fenômeno de compressão e o impacto das não-linearidades utilizadas nas redes neurais na dinâmica de aprendizagem e no plano de informação.

Iniciamos nosso trabalho pela elaboração de uma metodologia para determinar os parâmetros do estimador discreto utilizado. Todos os resultados obtidos e observados no plano de informação estão diretamente ligados ao estimador, que tem como objetivo mensurar corretamente as medidas de informação. Para o método escolhido, é necessário realizar a discretização através de histogramas das variáveis aleatórias contínuas, transformando-as em variáveis discretas para posteriormente estimar as informações mútuas utilizadas no plano de informação. A escolha dos parâmetros do estimador tem impacto direto no plano de informação gerado (CHELOMBIEV, HOUGHTON, & O'DONNEL, 2019); (SAXE, et al., 2018), ou seja, se escolhidos de forma incorreta obteremos resultados que violam a DPI, algo muito frequente nos trabalhos da área (GEIGER, 2020). Entretanto, seguindo a metodologia proposta, observamos através dos resultados bastante coerência nas medidas estimadas associadas a alguns comportamentos observados nas projeções internas das redes. Além disso o método discreto de estimação apresentou capacidade de capturar a dinâmica das distribuições de probabilidade e consequentemente transmiti-la para o plano de informação sem tais violações da DPI.

Posteriormente, analisando as não-linearidades e a dinâmica do plano de informação, vimos que não há relação entre o fenômeno de compressão com a escolha dessas funções. Indícios mostram que a compressão da informação $I(X;L)$ pode estar associada a redução da variância das distribuições gaussianas formadas nas camadas internas das redes treinadas com a função ReLU que causam a redução da entropia da camada. De forma semelhante existem indícios desse mesmo comportamento para as redes com função sigmoide, mas, devido à transformação não linear que a função realiza, não é possível afirmar com certeza essa colocação. Além disto, existe a compressão das projeções internas formadas por distribuições bidimensionais que por muitas vezes são extremamente difusas como as vistas em redes que utilizam as funções tanh, sendo necessários estudos mais profundos para um detalhamento correto de que de fato essas compressões em funções não-lineares observada podem ter

origens estatísticas ou se está associada a medidas de espalhamento / agrupamento dos dados como (GEIGER, 2020) e (GOLDFIELD, et al., 2019) sugerem.

Seguimos nossas análises de forma complementar observando como as funções de ativação afetam a dinâmica do treinamento das redes segundo a teoria da informação. Os resultados mostram que as funções de ativação contribuem para a formação da representação latente, de tal forma que as redes que utilizam a função tanh apresentaram maior convergência para o limite superior da DPI $I(X; L_0) = H(X)$, ou seja, foram as redes neurais que mais capturaram na totalidade a informação disponível nos dados. De forma semelhante, as redes com função *ReLU* apresentaram convergência para o limite inferior da DPI $I(X; L_0) = I(X; Y)$ com frequência, ou seja, foram as estruturas que capturaram somente a informação necessária para estimar $Y$ que caracteriza as estatísticas suficientes mínimas aproximadas, e as redes treinadas com a função sig obtiveram equilíbrio nos resultados, ou seja, obtendo representações intermediarias.

Nossos resultados mostram que as redes neurais operam de forma flexível, ou seja, geram internamente representações que podem conter uma quantidade de informação variada em relação a disponível no conjunto de dados de entrada $X$. Ou seja, para uma boa generalização uma rede não necessariamente precisa ter uma representação comprimida em problemas de regressão, de forma similar ao retratado na literatura (GEIGER, 2020) para problemas de classificação. Nos casos em que a rede demostrou absorção total da informação disponível $H(X)$, observamos a formação de representações internas geometricamente iguais aos dos dados de entrada da rede. Quando apresentam representações latentes de complexidade máxima (estatística suficiente mínima aproximada), as redes neurais operaram com a estrutura mínima disponível como suposto por (TISHBY & ZASLAVSKY, 2015), ou seja, as redes apresentaram um dos neurônios saturados (desativados). Este resultado particularmente corrobora com as premissas estabelecidas por Tishby quando foi proposto o *Information Bottleneck Theory of Deep Learning* onde é descrito que a rede deve operar de tal forma, neste sentido não há relatos na literatura deste comportamento nos problemas de classificação estudados. Além disto nossos resultados utilizando nossa metodologia de estimação mostram que a princípio as redes neurais em problemas de regressão seguem as relações da DPI estabelecidas por (TISHBY & ZASLAVSKY, 2015).

Concluímos que estes resultados alcançados contribuem para a área de teoria da informação aplicada a redes neurais. Os resultados exploratórios de nosso trabalho apontam que a extensão da teoria IBDL para problemas de regressão se mostra plausível baseado nas análises levantadas, bem como a existência de indícios de que a origem da compressão em regressão pode ter fontes estatísticas. Isso traz um aporte à teoria estabelecida sobre o assunto em um campo inexplorado como o da regressão de dados sob a ótica da teoria da informação, além de contribuir com aspectos

importantes sobre as não-linearidades e seus impactos na dinâmica de treinamento das redes neurais bem como a relação da estrutura das redes e sua capacidade de processamento.

## 7.1 Trabalhos Futuros

O estudo desenvolvido neste trabalho teve por objetivo principal a aplicação da teoria IBDL em problemas de regressão. Para uma análise inicial do fluxo de informação e o impacto das não linearidades no contexto proposto, foi necessário delimitar o escopo de trabalho. Além disto os resultados alcançados ainda precisam ser trabalhados de forma que se tornem mais sólidos.

Face aos pontos levantados, para os trabalhos futuros consideram-se como direção mais provável e complementares a este trabalho, os seguintes elementos:

- Analisar com mais profundidade as não linearidades utilizadas a fim de tornar sólidos os impactos que elas apresentaram no fluxo de informação.
- Analisar problemas de regressão mais complexos que envolvam relação não lineares entre entrada / saída.
- Estudar o impacto do sobretreinamento nos problemas de regressão no fluxo de informação e na formação da representação latente da rede, assim como, as técnicas que evitam tal comportamento, e.g. parada antecipada (do inglês, *early stopping*) e *dropout*.
- Analisar o impacto no fluxo de informação na rede neural e a formação da representação interna com o uso de algoritmos de treinamento mais robustos com parâmetros adaptativos.

# Referências


ACHILLE, A., & SOATTO, S.. Emergence of Invariance and Disentanglement in Deep Representations. (Y. Bengio, Ed.) *Proceedings of Information Theory and Applications Workshop (ITA),* 2018, p. 1-9. doi:10.1109/ITA.2018.8503149

ACHILLE, A., & SOATTO, S.. Information Dropout: Learning Optimal Representations Through Noisy Computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* v. 40, n. 12, Mar. 2018, p. 2897-2905. doi:10.1109/TPAMI.2017.2784440

ACHILLE, A., PAOLINI, G., & SOATTO, S.. Where is the Information in a Deep Neural Network?, ArXiv abs/1905.12213, May 2019.

AMJAD, R. A., & GEIGER, B. C.. How (Not) To Train Your Neural Network Using the Information Bottleneck Principle. ArXiv, abs/1802.09766, Feb. 2018.

AMJAD, R. A., & GEIGER, B. C.. Learning Representations for Neural Network-Based Classification Using the Information Bottleneck Principle. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* v. 42, n. 9, Sep. 2020, p. 2225-2239. doi:10.1109/TPAMI.2019.2909031

BRANDÃO, M. M., SPOLADORE, L., FARIA, L. C., ROCHA, A. S., SILVA-FILHO, M. C., & PALAZZO, R.. Ancient DNA sequence revealed by error-correcting codes. *Scientific Reports,* v. 5, n. 12051, Jul. 2015. doi:https://doi.org/10.1038/srep12051

CHELOMBIEV, I., HOUGHTON, C. J., & O'DONNEL, C.. Adaptive Estimators Show Information Compression in Deep Neural Networks. *Proceedings of International Conference on Learning Representations (ICLR),* New Orleans, May 2019.

CHENG, H., LIAN, D., GAO, S., & GENG, Y.. Utilizing Information Bottleneck to Evaluate the Capability of Deep Neural Networks for Image Classification. *Entropy,* v. 21, n. 5, May 2019, p. 456. doi:https://doi.org/10.3390/e21050456

COVER, T. M., & THOMAS, J. A.. *Elements of Information Theory,* 2 ed.,. NY, New York: John Willey & Sons, inc., 2006. ISBN: 0471241954.

DOANE, D. P. . Aesthetic Frequency Classifications. *The Amercica Statistician,* v. 30, n. 4, 1976, p. 181-183.

FANG, H., WANG, V., & YAMAGUCHI, M.. Dissecting Deep Learning Networks - Visualizing Mutual Information. *Entropy,* v. 20, n. 8, Aug. 2018, p. 823.

GABRIÉ, M., MANOEL, A., CLÉMENT, L., BARBIER, J., MACRIS, N., KRZAKALA, F., & ZDEBOROVÁ, L.. Entropy and mutual information in models of deep neural networks. *Proceedings on Advances in Neural Information Processing Systems (NeurIPS),* Dec. 2018, p. 1821-1831.

GEIGER, B. C.. On Information Plane Analysis of Neural Network Classifiers - A Review. *IEEE Transactions on Neural Networks and Learning Systems,* Mar. 2020. doi:10.1109/TNNLS.2021.3089037

GOKCAY, E., & PRINCIPE, J. C.. Information Theoretic Clustering. *IEEE Transactions on Pattern Analysis and machine intelligence,* v. 24, n.2, Feb. 2002, p. 158-171. doi:10.1109/34.982897

GOLDFIELD, Z., BERG, E. V., GREENEWALD, K., MELNYK, I., NGUYEN, N., KINGSBURY, B., & POLYANSKIY, Y.. Estimating Information Flow in Deep Neural Networks. *Proceedings of International Conference on Machine Learning (ICML),* Long Beach, California, USA, Jun. 2019, p. 2299-2308.

Goodfellow, I., Bengio, Y., & Courville, A.. *Deep Learning.* MIT Press, 2016. ISBN: 978-0262035613.

HAYKIN, S.. *Redes Neurais: Principios e prática*,2 ed., (P. M. Engel, Trans.) Porto Alegre, RS: Bookman, 2001. ISBN: 9788577800865.

HUBARA, I., COURBARIAUX, M., SOUDRY, D., EL-YANIV, R., & BENGIO, Y. (2017). Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations. *The Journal of Machine Learning Research,* v. 18, n. 1, Jan. 2017, p. 6869-6898.

KOLCHINSKY, A., TRACEY, B. D., & KUYK, S. V.. Caveats for information bottleneck in deterministic dcenarios. *International Conference on Learning Representation (ICLR),* Feb. 2019.

KRIZHEVSKY, A., SUTSKEVER, I., & HINTON, G. E.. ImageNet classification with deep convolutional neural networks. *Communications of the ACM,* v. 60, n. 6, Jun. 2017, p. 84-90. doi:https://doi.org/10.1145/3065386

LECUN, Y., BENGIO, Y., & HINTON, G.. Deep Learning. *Nature,* v. 521, May 2015, p. 436-444. doi:https://doi.org/10.1038/nature14539

LEE, S., & JO, J.. Information flows of divergense autoencoders. *Entropy,* v. 27, n. 7, Jul. 2021, p. 862. doi:10.3390/e23070862

LORENZEN, S., IGEL, C., & NIELSEN, M.. Information Bottleneck: Exact Analysis of (Quantized) Neural Networks. ArXiv, abs/2106.12912, Jun. 2021.

MCCULLOCH, W. S., & PITTS, W.. A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biology,* v. 5, Dec. 1943, p. 115-133. doi:https://doi.org/10.1007/BF02478259

NOSHAD, M., ZENG, Y., & HERO, A. O.. Scalable Mutual Information Estimation Using Dependence Graphs. *Proceedings of IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASP),* May 2019, p. 2962-2966.

PEEPLES, J., XU, W., & ZARE, A.. Histogram Layers for Texture Analyses. *IEEE Transactions on Artificial Intelligence,* v. 3, n. 4, Aug. 2021, p. 541-552. doi:10.1109/TAI.2021.3135804

Quinlan, J. R.. Induction of Decision Trees. *Machine Learning,* v. 1, Mar. 1986, p. 81-106. doi:https://doi.org/10.1007/BF00116251

ROMANO, J. M., ATTUX, R., CAVALCANTE, C. C., & SUYAMA, R.. *Unsupervised signal processing: channel equalization and source separation,* 1 ed.. CRC Press, 2011, ISBN: 978-0849337512.

ROSENBLATT, F.. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review,* v. 65, n. 6, Nov. 1958, p. 386-408. doi:https://doi.org/10.1037/h0042519

RUMELHART, D. E., HINTON, G. E., & WILLIAMS, R. J.. Learning representations by back-propagating errors. *Nature,* v. 232, Oct. 1986, p. 533-536. doi:https://doi.org/10.1038/323533a0

SAXE, A. M., BANSAL, Y., DAPELLO, J., ADVANI, M., KOLCHINSKY, A., TRACEY, B. D., & COX, D. D.. On the Information Bottleneck Theory of Deep Learning. *Proceedings of International Conference on Learning Representations (ICLR),* Vancouver, May 2018.

SCARGLE, J. D., NORRIS, J. P., JACKSON, B., & CHIANG, J.. Studies in Astronomical time series analysis. VI. Bayesian Blocks Representations. *The Astrophysical Journal,* v. 746, n. 2, Feb. 2013. doi:10.1088/0004-637X/764/2/167

SCHIEMER, M., & YE, J.. Revisitin the information plane. *2019*. Retrieved from https://openreview.net/forum?id=Hyljn1SFwr.

SCHWARTZ-ZIV, R., & TISHBY, N.. Opening the black box of Deep Neural Networks via Information. Mar. 2017. Retrieved from arXiv:1703.00810v3 [cs.LG]

SCOTT, D. W.. On Optimal and Data-Based Histograms. (B. Trust, Ed.) *Biometrika, v. 66, n. 3, Dec. 1979*, pp. 605-610.

SCOTT, D. W.. *Multivariate Density Estimation: Theory, Practice, and Visualization*, 2 ed.. Houston, Texas, Estados Unidos da América: Wiley, 2015. ISBN:978-0-471-69755-8.

SHANNON, C. E.. A Mathematical Theory of Communication. *The Bell System Technical Journal,* v. 27, 1948, p. 379-423.

SHIMAZAKI, H., & SHINOMOTO, S.. A Method for Selecting the Bin Size of a Time Histogram. *Neural Computation,* v. 19, n. 6, Jun. 2007(No.6), p. 1503-1527. doi:https://doi.org/10.1162/neco.2007.19.6.1503

SIZINTSEV, M., DERPANIS, K. G., & HOGUE, A.. Histogram-based search: A comparative study. *IEEE Conference on Computer Vision and Pattern Recognition,* 2008, p. 1-8. doi:10.1109/CVPR.2008.4587654

STONE, J. V.. *Information Theory: A Tutorial Introduction*, 1 ed.. 2013, Sebtel Press. ISBN: 978-0956372857.

TAPIA, N. I., & ESTÉVEZ, P. A.. On the information Plane of Autoencoders. *Proceedings of International Joint Conference on Neural Networks (IJCNN),* Jul. 2020, p. 1-8.

TISHBY, N., & ZASLAVSKY, N.. Deep Learning and the Information Bottleneck Principle. *Proceedings of IEEE Information Theory Workshop (ITW),* Apr. 2015, p. 1-5.

TISHBY, N., PEREIRA, F. C., & BIALEK, W.. The information bottleneck method. *Proceedings of the 37th Allerton Conference on Communication, Control and Computation,* v. 49, Jul. 1999, p. 368-377.

WICKSTRØM, K., LØKSE, S., KAMPFFMEYER, M., YU, S., PRÍNCIPE, J., & JENSSEN, R.. Information Plane Analysis of Deep Neural Networks vi Matrix-based Rényi's Entropy and Tensor Kernels. Arxiv, abs/1909.11396, Sep. 2019.

WILCOX, H. J., & MYERS, D. L.. *An Introduction to Lebesgue Integration and Fourier Series.* Courier Corporation, 1994. ISBN: 9780486682938.

YU, S., & PRÍNCIPE, J. C.. Understanding Autoencoders with Information Theoretic Concepts*, Neural Netoworks*, v. 117, p. 104-123, 2019.

YU, S., GIRALDO, L. G., JENSSEN, R., & PRÍNCIPE, J. C.. Multivariate Extension of Matrix-based Rényi's α-Order Entropy Functional. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* v. 42, n. 11, Nov. 2020, p. 2960-2966. doi:10.1109/TPAMI.2019.2932976

YU, S., WICKSTRØM, K., JENSSEN, R., & PRÍNCIPE, J. C.. Understanding Convolutional Neural Networks with Information Theory: An Initial Exploration. *IEEE Transaction on Neural Networks Learning Systems,* Jan. 2020, p. 1-8.

# APÊNDICE

## Apêndice A

A leitura e compreensão correta do plano de informação é de suma importância para o entendimento dos resultados apresentados ao longo do Capítulo 6. Nesta seção serão discutidos e explicados como entender os planos de informação apresentados.

Na Seção 6.1, a Figura 6.1 apresenta o plano de informação para diferentes inicializações de redes, para cada uma das não linearidades utilizadas nos experimentos. A figura em questão será seccionada em relação as suas respectivas não linearidades, de tal forma, que as análises serão realizadas separadamente.

A Figura A1 apresenta apenas redes neurais que utilizaram a função ReLU. A curva que deve ser analisada sempre será referente as camadas internas das redes, portanto, a camada $L_0$ nos planos de informação, que estão acompanhadas de setas para identificar sua dinâmica. Na Figura A1 a), pode-se observar que a camada interna da rede apresenta compressão ao longo de todo o treinamento, pois apresentou redução da informação mútua em relação ao eixo-x, ou seja, $I_x = I(X; L_0)$, com redução de aproximadamente 1,5 *bits* (de 4,5 *bits* para 3 *bits*) de informação.

Neste exemplo, lê-se a dinâmica da camada interna da seguinte forma: A partir da inicialização da rede (seta laranja) por poucas épocas a rede está ajustando o erro grosseiro da rede associado a inicialização dos pesos, com poucos ajustes, porém, proeminentes, chegando então a um ponto de continuidade da curva, onde, de fato o treinamento tem início com o ajuste mais fino dos pesos. Nesta fase de treino (seta vermelha) observa-se que a rede apresenta certo grau de conhecimento sobre $X$, entretanto, ela ainda não aprendeu como relacionar esses dados (da entrada) de forma correta para prever $Y$ corretamente, uma vez que ela possui pouca informação sobre $Y$ (1 *bit*), ou seja, $I_Y = I(L_0; Y)$. Posteriormente, conforme a rede aprende a lei que rege a relação dos dados $y = f(x)$, ela passa a aprender como prever corretamente a saída da rede, ganhando informação $I_Y$, e, concomitantemente descartando informação $I_X = I(X; L_0)$. Conceitualmente, pela teoria do IB, Seção 3.8, a redução da informação mútua $I_X$ tem relação com o descarte de informação irrelevante, ou seja, aquela que não contribui para a predição de $Y$, significando que a rede está aprendendo um mapeamento simples, ou seja, sem redundâncias. Entretanto, como buscamos discutir ao longo do Capítulo 6, a redução da informação mútua $I_X = I(X; L_0)$ está associada a variação estatística da distribuição formada na

camada interna, ou seja, tanto o aumento quando a diminuição dessa informação mútua é consequência da variação da distribuição interna da camada, neste caso, da variância da distribuição gaussiana formada internamente nas redes com a função ReLU.

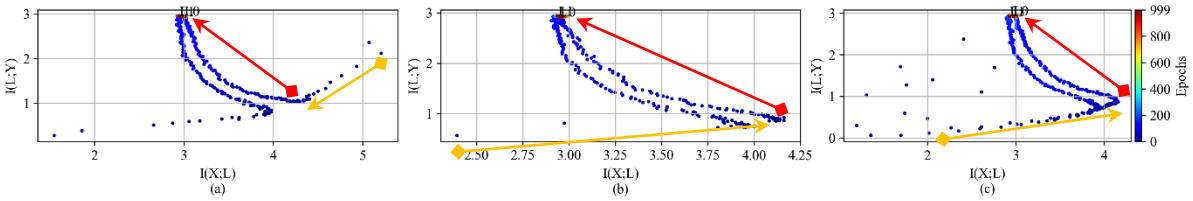



Figura A1 - Redes neurais treinadas com a função ReLU na camada interna.

De forma semelhante, na Figura A1 b), o processo de treinamento leva rapidamente o processo de treinamento para o regime fino de ajustes (seta vermelha), logo, o que se vê são curvas bem delineadas de informação. O gráfico mostra que conforme a rede aprende a prever $Y$, ela apresenta a compressão da informação $I_X$, consequência da redução da variância da distribuição gaussiana da camada $L_0$, isto significa que: a camada interna da rede quando apresenta uma distribuição interna muito diferente da distribuição dos dados de treino $X$, confere a rede pouca capacidade de predição de $Y$ para o problema de regressão. Ou seja, a variação da informação mútua $I_X$, que é de ordem estatística, é, o que se pode chamar de *variação de formação*, ou seja, quando a distribuição interna da rede é diferente da distribuição gaussiana dos dados de entrada $X$, a informação $I(X;L_0)$ apresentará uma variação de compressão ou expansão ao longo do processo de treinamento, até que se forme uma distribuição gaussiana em pelo menos um dos neurônios internos $N_1$ ou $N_2$, e o restante da variação se dá em decorrência do ajuste de escala dessa distribuição a fim de proporcionar a rede a predição correta de $Y$.

Para a função ReLU, que não transforma os dados, ou seja, apenas muda a escala deles, apresenta redução da informação, como mencionado, como consequência da redução da variância da distribuição interna (vide Figura 6.3), outros exemplos desse comportamento serão discutidos adiante.

Por fim, a Figura A1 c) apresenta comportamento e interpretação exatamente igual à aquela descrita para a Figura A1 a), com a diferença de que a rede neural, neste caso, teve uma inicialização de pesos que proporcionou a rede uma informação inicial $I_X = I(X;L_0)$ de 2 *bits* aproximadamente, menor que os 5 *bits* apresentados para a camada interna na Figura A1 a), isto significa que, a distribuição interna formada na inicialização é uma distribuição concentrada em poucos intervalos, portanto com baixa informação $I_X$. Distribuições internas em $L_0$ diferentes da distribuição gaussiana que o conjunto de dados $X$ possui, apresentará informação $I(X;L_0)$ destoante, ou seja, se a distribuição de probabilidade de $L_0$ apresentar baixa entropia $H(L_0)$ (ver Equação 3.24), por exemplo uma distribuição

muito concentrada, a informação mútua $I(X;L_0)$ também será baixa, e apresentara comportamento de expansão no plano de informação ao longo do treinamento até que a distribuição tenha uma similaridade maior para/com uma gaussiana, e, posteriormente, com o ajuste dos pesos, essa distribuição sofrerá modificações que levem a uma representação interna adequada para que a rede aprenda corretamente. Neste momento a rede pode apresentar compressão ou expansão da informação $I_X$, o grande problema é que a imprevisibilidade desta dinâmica está associada a estocasticidade do processo de treinamento, sendo extremamente difícil de prever, vale lembrar ao leitor que a interpretação do plano de informação para o problema de regressão não é simples e intuitiva, pois uma série de fatores influenciam na dinâmica do plano de informação.

Continuando, de forma similar, se a distribuição de $L_0$ apresentar alta entropia, como por exemplo, uma distribuição muito espalhada, o processo de formação da distribuição gaussiana remeterá a uma redução da entropia da camada $H(L_0)$, e posteriormente a redução da informação $I(X;L_0)$, que é o comportamento de compressão descrito ao longo deste trabalho. Novamente, a variação da informação é em decorrência da transformação da distribuição de probabilidade de $L_0$.

Para as redes treinadas com a função sigmoid Figura A2, as variações de $I_X = I(X;L_0)$ também estão relacionadas diretamente com a mudança estatística da distribuição interna formada, vide Figura 6.4. Na Figura A2 d) a camada interna apresenta apenas uma pequena compressão de $I_X$ com um ganho também pequeno em $I_Y$. Internamente esta camada apresenta redução da informação $I_X$ em decorrência da formação da distribuição, ou seja, variação de informação de formação, Figura A3 f) (comparando a) e d) )Figura A3 - Variação de informação de formação da rede sigmoid referente a Figura A2 d)., uma vez que se alcança essa distribuição a rede também apresenta ganho na capacidade de predição de $Y$.

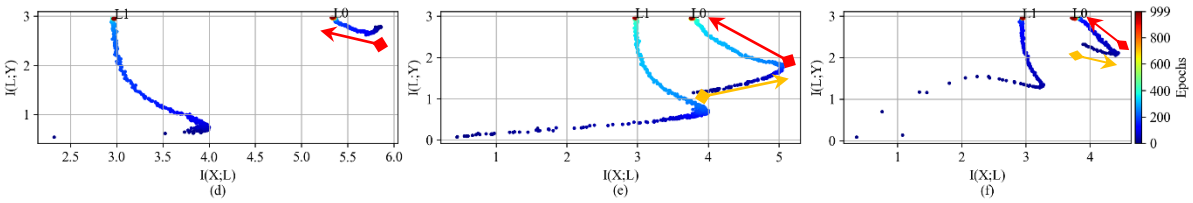



Figura A2 - Redes neurais treinadas com a função sigmoid.

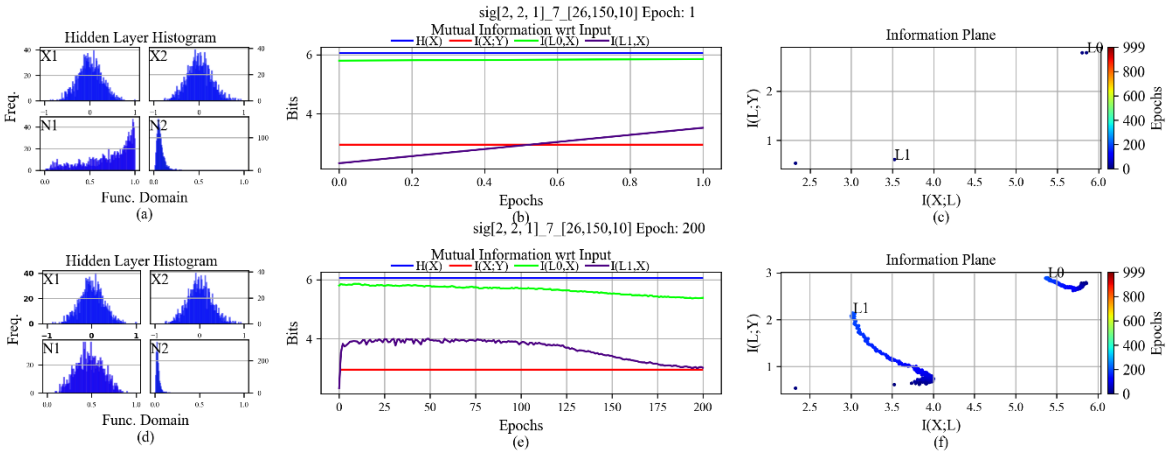



Figura A3 - Variação de informação de formação da rede sigmoid referente a Figura A2 d).

O mesmo processo ocorre para os planos de informação Figura A2 e), referente a Figura 6.4, o primeiro momento do treino (seta laranja) é o aumento da informação é causado pela formação de uma distribuição "uniforme" internamente, que possui entropia máxima, posteriormente a redução da informação é devido a formação da gaussiana (aproximadamente) que possui entropia menor, por definição. Esta variação da informação de formação pode ser observada através da comparação da distribuição e respectivas informações mútuas $I_X$ da Figura A3 com a Figura 6.4.

A variação da informação $I_X$ da Figura A2 f) também é de formação, ou seja, de ordem estatística da distribuição interna. O ganho da informação $I_Y$ aumente conforme a rede ganha capacidade de prever corretamente $Y$.

Por fim, temos as redes que foram treinadas com a função tangente hiperbólica, Figura A4, a variação da informação tanto $I_X$ quanto $I_Y$, representadas pelas setas laranja e vermelha estão relacionadas as variações de informação de formação, ou seja, a informação varia, pois, a distribuição de probabilidade de $L_0$ está sendo formada (gaussiana aproximadamente), conforme a distribuição gaussiana (aproximadamente) vai se formando internamente em um dos neurônios da camada interna a rede ganha capacidade de predição de $Y$, e, a variação de $I_X$ depende da evolução dos pesos ao longo do treinamento, que, diretamente, afetam a distribuição formada.

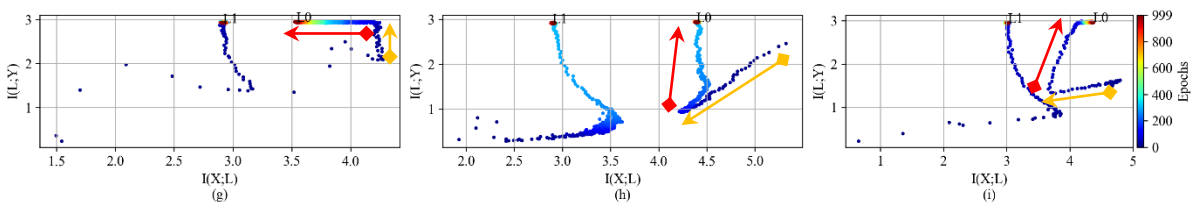



Figura A4 - Redes treinadas com a função tangente hiperbólica que apresentam compressão.

Por exemplo, como mostrado pela Figura A5 e Figura A6, referente a rede da Figura A4 b), ou seja, em um primeiro momento, Figura A5 f), a informação da camada interna se reduz (seta laranja) tanto $I_X$ quanto em $I_Y$, pois a distribuição formada em ambos os neurônios $N_1$ e $N_2$ apresentam baixo grau de similaridade com a distribuição gaussiana dos dados $X$, consequentemente, ela apresenta redução da informação. Posteriormente, na Figura A6 f), vemos que a camada $L_0$ apresenta aumento na informação a respeito de $Y$ (seta vermelha) conforme se forma uma distribuição gaussiana (aproximadamente) no neurônio $N_2$.

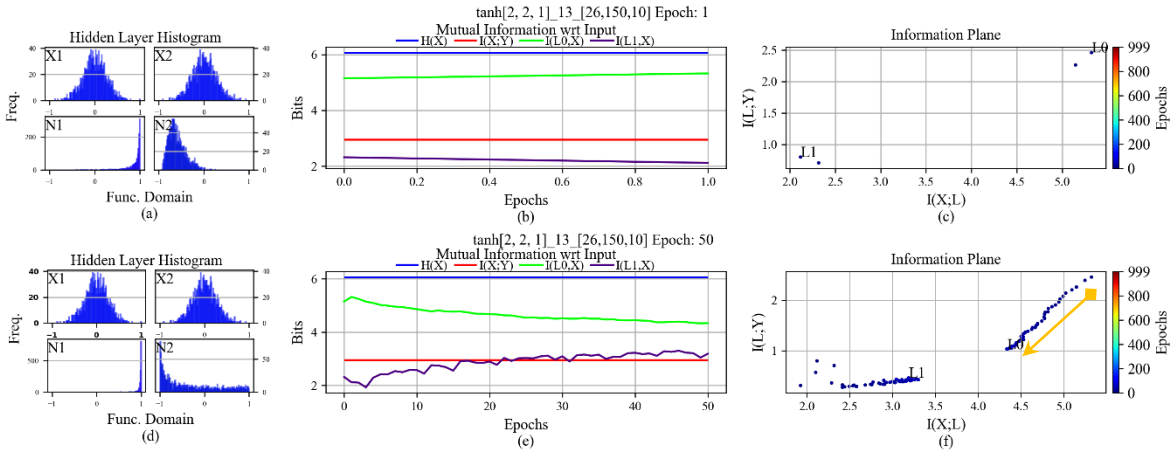



Figura A5 - Redução da informação $I(X; L_0)$ de c) para f) em decorrência do baixo grau de similaridade (divergente de Kullback-Leibler) entre as distribuições dos neurônios com relação a distribuição dos dados $X$.

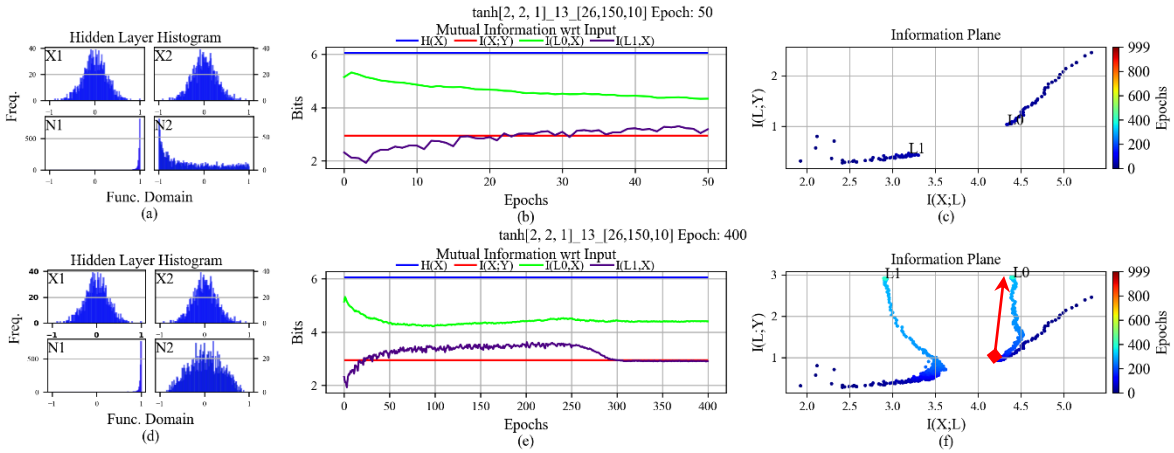



Figura A6 - Ganho de informação $I(L_0; Y)$ em decorrencia da formação da distribuição gaussiana (aproximadamente) no neurônio $N_2$ da figura d).

A inicialização dos pesos das redes afeta o ponto inicial das curvas de informação, entretanto elas tendem a evoluir rapidamente para pontos de informação semelhantes (próximos de 4 bits na grande maioria das vezes), uma vez que os dados de treinamento utilizado para treina-las são os mesmos, posteriormente a evolução das curvas de informação decorrem da aleatoriedade dos pesos e

do processo de atualização deles. O processo de análise e interpretação dos planos de informação não é trivial, a natureza do problema sendo de regressão ou classificação possuem planos de informação diferentes, a estrutura da rede utilizada, MLP, convolucionais etc. também alteram o plano de informação. Por tal razão é necessária uma análise cautelosa e vagarosa dos planos de informação para o correto entendimento da dinâmica apresentada pelas curvas de informação.

## Apêndice B

Em (TISHBY & ZASLAVSKY, 2015) os autores afirmam que uma rede que apresente boa generalização, deve apresentar em seu plano de informação compressão da informação mútua $I_X = I(X; L_0)$. Entretanto, existem na literatura recente (CHELOMBIEV, HOUGHTON, & O'DONNEL, 2019); (GOLDFIELD, et al., 2019); (SAXE, et al., 2018) e (SCHIEMER & YE, 2019) a discussão sobre esse ponto da teoria IBDL, os trabalhos em questão apontam que esta condição não é necessária, ou seja, uma rede neural generalizando bem não precisa apresentar compressão no plano de informação. Neste aspecto o presente trabalho, como mencionado, apresenta resultados que se alinham com esta narrativa.

Para demonstrar que no problema de regressão, proposto pelo presente trabalho, a compressão não está relacionada com a capacidade de generalização da rede, foi selecionada duas redes neurais treinadas com a função ReLU onde ambas as redes apresentam boa generalização.

A Figura B1 apresenta um resumo de treinamento de uma rede ReLU até a época 100, logo que a rede neural atinge um erro de treino e teste adequado. Na Figura B1 a) pode-se observar o erro de treinamento e teste, na Figura B1 b) o plano de informação mostra a compressão da informação da camada interna $I_X = I(X; L_0)$ e pode-se observar através da predição de uma parcela do conjunto de testes que a rede aprendeu corretamente e está generalizando adequadamente. Posteriormente na Figura B2 b), em um resumo do treinamento da rede até a época 125, o plano de informação não apresenta compressão da informação da camada interna $I_X = I(X; L_0)$, e, a rede está apresentando boa generalização, quando observado a predição de uma parcela de dados do conjunto de teste na Figura B2 c). Ou seja, a compressão da informação não é necessária para que a rede apresente boa generalização, a variação da informação é associada a variação da distribuição de probabilidade formada na camada interna $L_0$ da rede, que pode ou não apresentar compressão da informação $I_x$.Figura B1

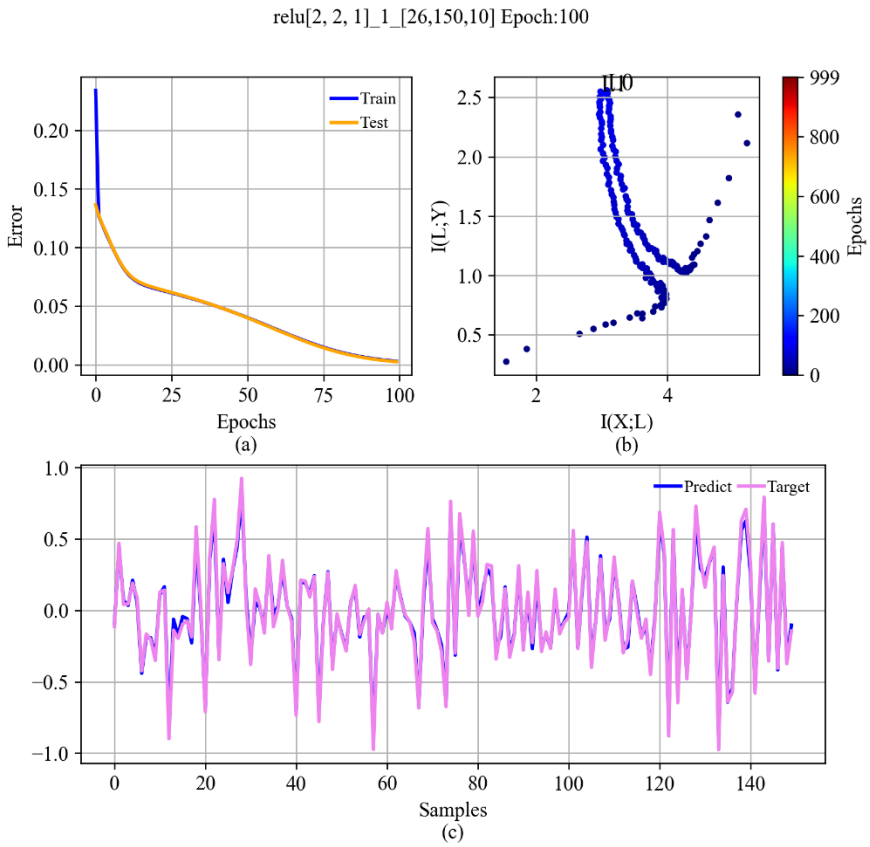



Figura B1 - Rede neural ReLU que apresenta compressão em $L_0$ com boa generalização. a) Erro de treino e teste da rede. b) Plano de informação. c) Comparação da predição e as amostras alvo para uma parcela de dados.

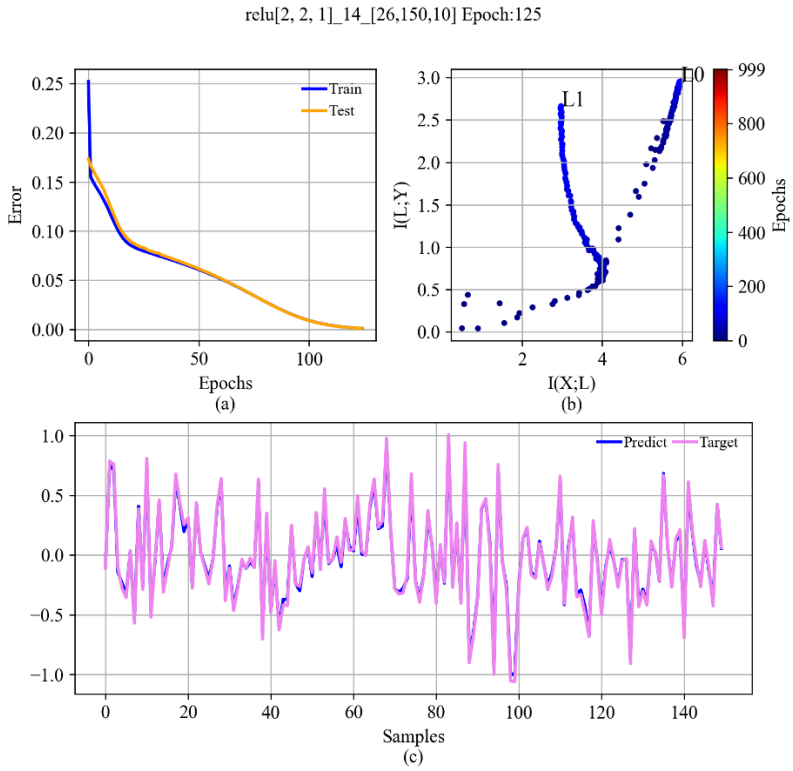



Figura B2 - Rede neural ReLU que não apresenta compressão em $L_0$ e tem boa generalização. a) Erro de treino e teste da rede. b) Plano de informação. c) Comparação da predição e as amostras alvo para uma parcela de dados.

# Apêndice C

Nos resultados obtidos as redes neurais também apresentaram representações internas intermediarias, ou seja, representações internas que capturaram informação entre os limites estabelecidos pelas relações da DPI $H(X) > I(X;L_0) > I(X;Y)$. A grande maioria das redes treinadas com a função sigmoid apresentaram tal comportamento como visto na Figura 6.2. Alguns exemplos são mostrados a seguir:

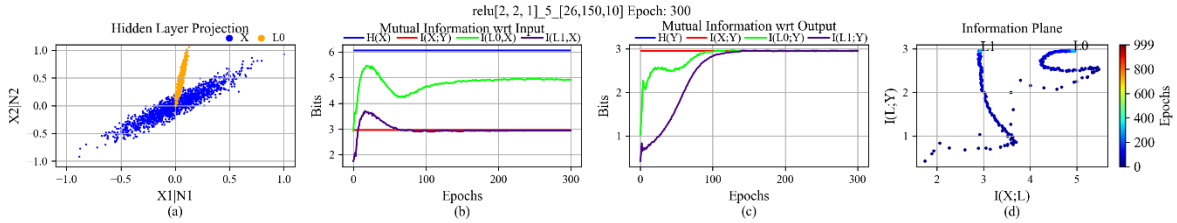



Figura C1 - Rede treinada com a função ReLU com representação intermediária, ou seja, $I(X;L_0)$ não convergiu para os limites $H(X)$ ou $I(X;Y)$.

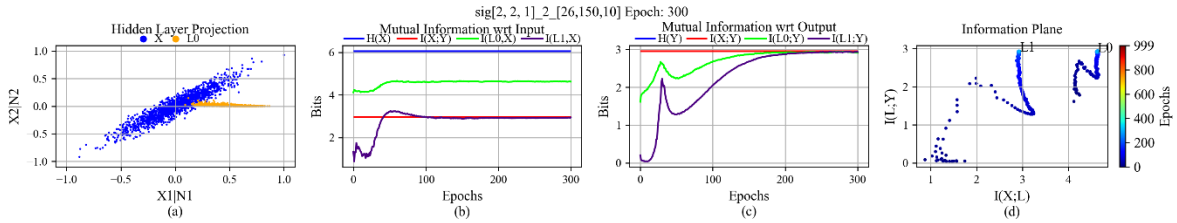



Figura C2 - Rede treinada com a função sigmoid apresentando representação interna intermediaria.

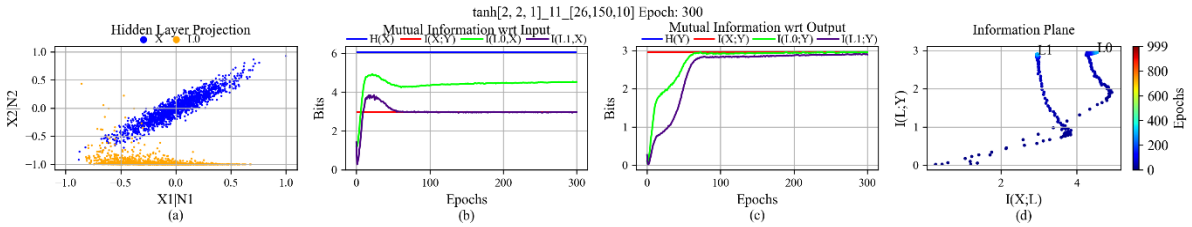



Figura C3 - Rede treinada com a função tangente hiperbólica com representação interna intermediaria.