Mathematical Engineering of Deep Learning

Book Draft

Benoit Liquet, Sarat Moka and Yoni Nazarathy

February 28, 2024

A ma maman,

Benoit Liquet.

To my mother Mariyamma and my wife Toshali,

Sarat Moka.

To Emily, Kayley, and Yarden,

Yoni Nazarathy.

Preface - DRAFT

In the last few years deep learning has seen explosive growth and even dubbed as the "new electricity". The field has shown incredible success in automated applications and predictive tasks. Deep learning models are mathematical in nature, and hence to understand deep learning, one needs to understand the mathematical description of the models. This book aims to provide such an understanding via a concise, accessible, and self contained presentation.

Many deep learning resources focus on programming while making an effort to hide the mathematics. Other resources focus on theoretical results without an attempt to disambiguate the terminology of the field. A third breed of resources puts heavy emphasis on historical progression. Each of these viewpoints is important for a specific purpose, however we conjecture that for a mathematical audience, these are all suboptimal ways to learn about deep learning. Hence we created this book.

Our focus is on the basic mathematical constructs that make up the field. We call this **mathematical engineering of deep learning**. Using the language of equations and algorithms, deep learning objects interface together to make very powerful models. A reader armed with basic familiarity of mathematical notation and knowledge of basic calculus, basic probability, and basic linear algebra can go a long way in understanding deep learning quickly. For this, we use simple mathematical notation to outline the mechanisms used for training, execution, and application of deep neural networks.

Deep learning is certainly not soley about mathematics, as it also requires good software, hardware, and data. However, we aim to present the mathematical technology of deep learning without focusing on the implementation aspects that one would consider if trying to use deep learning frameworks in practice. We also aim to focus on the current state of the art as opposed to the historical progression of the field. Finally, we aim to minimize the focus on the human brain and the loose analogies that one can make between deep artificial neural networks and actual biological neurons. All of these aspects that we downplay are important, but we believe that if in an initial exposure to the field one spends too much time on implementation, history, or bio-neurological analogies, then the simplicity of deep learning is missed.

The book is primarily intended for readers from engineering, signal processing, statistics, physics, econometrics, operations research, quantitative management, pure mathematics, bioinformatics, applied machine learning, or even applied deep learning. A reader with background in one of these domains will be able to get a concentrated and concise description of deep learning. In cases where a mathematical refresher is needed, appendices provide a condensed review, such as for example a review of key aspects of multivariable calculus.

The book can be read sequentially, or alternatively readers may wish to jump between chapters for quick lookup. It is assumed that readers have had exposure to mathematical notation at the level equivalent to at least 3 or 4 university courses. Hence set notation, matrices, basic probability, and calculus are used without apology. However, no explicit knowledge of machine learning, statistics, optimization, or advanced probability is needed or assumed. Our hope is that we strike the right balance so that a mathematically equipped non-expert can easily read the book in a self contained manner.

While the focus of the book is “mathematical engineering”, we fully acknowledge the importance of applications and the ability to use software and hardware effectively. For this you may also use the companion website, https://deeplearningmath.org/, where additional examples, links, and software usage details are provided.

Outline of the Contents

The book has 8 chapters and 2 appendices. Chapters 1 – 4 introduce the field, outline key concepts from machine learning, present an overview of optimization concepts needed for deep learning, and focus on fundamental models and concepts. Chapter 5 is the central chapter introducing fully connected deep neural networks. Chapters 6 and 7 deal with the core models and architectures of deep learning, including convolutional networks, recurrent neural networks, and transformers. Chapter 8 covers additional popular domains such as generative models, reinforcement learning, and graph neural networks. Appendices A and B provide mathematical support. Here is a detailed outlined of the contents.

Chapter 1 – Introduction: In this chapter we present an overview of deep learning, demonstrate key applications, survey the associated ecosystems of high performance computing, discuss big and high-dimensional data, and set the tone for the rest of the book. The chapter discusses key terminology including data science, machine learning, and statistical learning, and with this we place these terms in the context of the book. Key popular datasets such as ImageNet and MNIST digits are also presented together with a description of the deep learning culture that emerged.

Chapter 2 – Principles of Machine Learning: Deep learning can be viewed as a sub-discipline of machine learning and hence this chapter provides an overview of key machine learning concepts and paradigms. The reader is introduced to supervised learning, unsupervised learning, and the general concept of iterative based optimization for learning. The concepts of training sets, test sets, and the like, together with principles of cross validation and model selection are introduced. A key object explored in the chapter is the linear model which can be trained also via iterative optimization. We introduce the most simple gradient descent algorithm and it is later refined in Chapter 4. Gradient descent is used for training almost any deep learning model. We also explore basic unsupervised learning algorithms including K-means clustering, principal component analysis (PCA), and the singular value decomposition (SVD).

Chapter 3 – Simple Neural Networks: In this chapter we focus on logistic regression (sigmoid) for binary classification and the related multinomial regression model (softmax) for multi-class problems. These models are the most popular shallow neural networks. The chapter sets the tone for more complex models by introducing principles of deep learning such as the cross entropy loss and other basic terminology. The chapter also presents a simple non-linear autoencoder architecture and with this introduces general ideas of autoencoders.

Chapter 4 – Optimization Algorithms: The training of deep learning models involves optimization over the learned parameters. Hence a solid understanding of optimization algorithms is required, as well as an understanding of specialized optimization techniques that work well for deep learning models such as the ADAM algorithm. In this chapter we focus on such techniques. We also study the details of various forms of automatic differentiation, a tool that has become critical in deep learning for computing gradients. Other optimization techniques, not always popular in contemporary deep learning, are also surveyed. This includes various first-order and second-order methods.

Chapter 5 – Feedforward Deep Networks: This chapter is the heart of the book where the general feedforward deep neural network, also known as the multi-layer perceptron, is defined and introduced. After introducing the basic architecture and exploring the expressive power of deep neural networks, we dive into the details of training by understanding the backpropagation algorithm for gradient evaluation. We also explore other aspects such as weight initialization, batch normalization, and dropout.

Chapter 6 – Convolutional Neural Networks: Convolutional neural networks are natural models for images and similar spatial data formats. In this chapter we explore the convolution concept and then see it used in the context of deep learning models. The concepts of channels, and general convolutional neural networks are introduced. We then follow with an exploration of common unique architectures that have made significant impact and are still in use today. We also explore a few key tasks associated with images such as object localization and face identification.

Chapter 7 – Sequence Models: Sequence models are critical for data such as text with applications in natural language processing, conversational agents, and translation. In this chapter we get a taste for the key deep learning ideas of the field. We explore recurrent neural networks and their generalizations including long short term memory (LSTM) models and gated recurrent unit (GRU) models. We then explore encoder-decoder architectures building up to the concept of attention where we formalize the attention mechanism. This idea then integrates in transformer models which in many ways are the state of the art models used in large language models (LLM).

Chapter 8 – Specialized Architectures and Paradigms: In this final chapter we survey key ideas of specialized architectures and paradigms which are used for various types of tasks. This includes, generative models, reinforcement learning, and graph neural networks. In terms of generative models we start by diving into the variational autoencoder architecture, a probabilistic deep learning model. We then extend to Markovian hierarchical variational autoencoders of which diffusion models are a special case. We then study generative adversarial networks (GANs) which were the first class of highly powerful deep learning models for realistic looking image generation. The chapter then moves to study reinforcement learning where we first present an overview of basics of Markov decision processes and then hint on how deep reinforcement learning can be implemented. We close with an introduction of graph neural networks. As such, the multitude of ideas in this chapter encompass several paradigms where deep learning models can be modified or joined together for specialized purposes.

With Thanks

We began this project while undertaking instruction at the 2021 AMSI (Australian Mathematical Sciences Institute) summer school. In that course we taught 60 students from all over Australia for 28 lecture hours. See a link to the course material through the book website `https://deeplearningmath.org/`. We thank the students for embarking on the journey with us and further appreciate student comments useful for creating the book. We also mention that without support from our families and loved ones, this book would not be possible. We thank Alan White for supplying the banana for Figure 1.1. We thank various family members for appearing in some of our images.

We especially thank Vishnu Prasath and Ajay Hemanth of Richmond Enterprises PVT LTD for working on many illustrations of the book. The TikZ source code for these illustrations is now open sourced with a link available through the course website. A few of the images in our figures, when mentioned, are taken from other research papers and other sources. We thank the authors for permission to use these images.

We also thank the following people for detailed comments and useful discussions: Teo Nguyen, Thomas Grahm, Vektor Dewanto, and Miriam Redding. In addition, useful comments were received from Marcus Gallagher, Matt Dirks, Adam Bennaceur, Gabriel Bianconi, Kwangsoo Cho, Jerzy Filar, Liam Bluett, Fred Roosta-Khorasani, and Maria Vlasiou. Sarat Moka thanks Celestien Warnaar-Notschaele and Ole Warnaar for friendship and an extensive accommodation period in Brisbane during the extensive Sydney lockdown of 2021.

We hope that you enjoy the book.

Benoit Liquet, Sarat Moka, and Yoni Nazarathy.
February 2024.

Contents

Preface - DRAFT 3

1 Introduction - DRAFT 1
1.1 The Age of Deep Learning . . . 1
1.2 A Taste of Tasks and Architectures . . . 7
1.3 Key Ingredients of Deep Learning . . . 12
1.4 DATA, Data, data! . . . 17
1.5 Deep Learning as a Mathematical Engineering Discipline . . . 20
1.6 Notation and Mathematical Background . . . 23
Notes and References . . . 25

2 Principles of Machine Learning - DRAFT 27
2.1 Key Activities of Machine Learning . . . 27
2.2 Supervised Learning . . . 32
2.3 Linear Models at Our Core . . . 39
2.4 Iterative Optimization Based Learning . . . 48
2.5 Generalization, Regularization, and Validation . . . 52
2.6 A Taste of Unsupervised Learning . . . 62
Notes and References . . . 72

3 Simple Neural Networks - DRAFT 75
3.1 Logistic Regression in Statistics . . . 75
3.2 Logistic Regression as a Shallow Neural Network . . . 82
3.3 Multi-class Problems with Softmax . . . 86
3.4 Beyond Linear Decision Boundaries . . . 95
3.5 Shallow Autoencoders . . . 99
Notes and References . . . 111

4 Optimization Algorithms - DRAFT 113
4.1 Formulation of Optimization . . . 113
4.2 Optimization in the Context of Deep Learning . . . 120
4.3 Adaptive Optimization with ADAM . . . 128
4.4 Automatic Differentiation . . . 135
4.5 Additional Techniques for First-Order Methods . . . 143
4.6 Concepts of Second-Order Methods . . . 152
Notes and References . . . 164

5 Feedforward Deep Networks - DRAFT 167
5.1 The General Fully Connected Architecture . . . 167
5.2 The Expressive Power of Neural Networks . . . 173
5.3 Activation Function Alternatives . . . 180
5.4 The Backpropagation Algorithm . . . 184
5.5 Weight Initialization . . . 192

5.6 Batch Normalization 194
5.7 Mitigating Overfitting with Dropout and Regularization 197
Notes and References 203

6 Convolutional Neural Networks - DRAFT 205
6.1 Overview of Convolutional Neural Networks 205
6.2 The Convolution Operation 209
6.3 Building a Convolutional Layer 216
6.4 Building a Convolutional Neural Network 226
6.5 Inception, ResNets, and Other Landmark Architectures 236
6.6 Beyond Classification 240
Notes and References 247

7 Sequence Models - DRAFT 249
7.1 Overview of Models and Activities for Sequence Data 249
7.2 Basic Recurrent Neural Networks 255
7.3 Generalizations and Modifications to RNNs 265
7.4 Encoders Decoders and the Attention Mechanism 271
7.5 Transformers 279
Notes and References 294

8 Specialized Architectures and Paradigms - DRAFT 297
8.1 Generative Modelling Principles 297
8.2 Diffusion Models 306
8.3 Generative Adversarial Networks 315
8.4 Reinforcement Learning 328
8.5 Graph Neural Networks 338
Notes and References 353

Epilogue - DRAFT 355

A Some Multivariable Calculus - DRAFT 357
A.1 Vectors and Functions in $\mathbb{R}^n$ 357
A.2 Derivatives 359
A.3 The Multivariable Chain Rule 362
A.4 Taylor's Theorem 364

B Cross Entropy and Other Expectations with Logarithms - DRAFT 367
B.1 Divergences and Entropies 367
B.2 Computations for Multivariate Normal Distributions 369

Bibliography 399

Index 401