

The Mathematical Engineering of Deep Learning

Book Draft

Benoit Liquet, Sarat Moka and Yoni Nazarathy

December 6, 2022

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	11
1.4. DATA, Data, data!	15
1.5. Deep Learning as a Mathematical Engineering Discipline	19
1.6. Notation and Mathematical Background	21
Notes, References, and Further Details	23
2. Principles of Machine Learning - DRAFT	25
2.1. Key Activities of Machine Learning	25
2.2. Supervised Learning	30
2.3. Linear Models at Our Core	37
2.4. Iterative Optimization Based Learning	45
2.5. Generalization, Regularization, and Validation	50
2.6. A Taste of Unsupervised Learning	59
Notes, References, and Further Details	69
3. Simple Neural Networks - DRAFT	71
3.1. Logistic Regression in Statistics	71
3.2. Logistic Regression as a Shallow Neural Network	78
3.3. Multi-class Problems with Softmax	82
3.4. Beyond Linear Decision Boundaries	90
3.5. Shallow Autoencoders	94
Notes, References, and Further Details	106
4. Optimization Algorithms - DRAFT	107
4.1. Formulation of Optimization	107
4.2. Optimization in the Context of Deep Learning	114
4.3. Adaptive Optimization with ADAM	122
4.4. Automatic Differentiation	129
4.5. Additional Techniques for First Order Methods	137
4.6. Concepts of Second Order Methods	146
Notes, References, and Further Details	156
5. Feed-Forward Deep Networks - DRAFT	159
5.1. The General Fully Connected Architecture	159
5.2. The Expressive Power of Neural Networks	165
5.3. Activation Function Alternatives	172
5.4. The Back Propagation Algorithm	175
5.5. Weight Initialization	183

Contents

5.6. Batch Normalization	185
5.7. Mitigating Overfitting with Dropout and Regularization	189
Notes, References, and Further Details	194
6. Convolutional Neural Networks - DRAFT	195
6.1. Overview of Convolutional Neural Networks	195
6.2. The Convolution Operation	198
6.3. Building a Convolutional Layer	204
6.4. Building a Convolutional Neural Network	213
6.5. Inception, Resnets, and Other Landmark Architectures	219
6.6. Beyond Classification	221
Notes, References, and Further Details	224
7. Sequence Models - DRAFT	225
7.1. Forms of Sequence Data	226
7.2. Recurrent Neural Networks	226
7.3. Long Short Term Memory Models	236
7.4. Gated Recurrent Unit Models	239
7.5. Encoder-decoder for End to End Translation	239
7.6. Transformers Model	243
7.7. Model Tuning for Sequence Models	249
Notes, References, and Further Details	249
8. Tricks of the Trade	251
8.1. Model and Hyper-parameter Choices	252
8.2. Techniques for Hyper-parameter Choice	255
8.3. Transfer Learning	258
8.4. Dealing With Unbalanced Datasets	262
8.5. Repurposing Models for Sequence Data and Images	262
8.6. Data Augmentation	262
Notes, References, and Further Details	263
9. Generative Models - DRAFT	265
9.1. Generative Modelling	266
9.2. The Generator and Discriminator Game and GAN Architectures	266
9.3. Loss Function Adaptations	279
9.4. Diffusion Models	281
9.5. Assessing Performance and Model Tuning	281
Notes, References, and Further Details	282
10. Deep Reinforcement Learning - DRAFT	283
10.1. Optimal Control Over Time	284
10.2. Markov Decision Processes	285
10.3. Reinforcement Learning via Q-learning	295
10.4. Q-function Approximations With Neural Networks	297
10.5. Deep Reinforcement Learning Paradigms	297
10.6. Implementation Considerations and Applications	297
Notes, references, and further details	298
A. Some Multivariable Calculus - DRAFT	299
A.1. Vectors and Functions in \mathbb{R}^n	299
A.2. Derivatives	300

Contents

A.3. The Multivariable Chain Rule	303
A.4. Taylor's Theorem	305
B. Cross Entropy and Other Expectations with Logarithms - DRAFT	309
B.1. Divergences and Entropies	309
B.2. Interpretation via Information Theory	310
C. Gaussian Processes for Bayesian Optimization - DRAFT	311
C.1. Surrogate models	311
C.2. Gaussian random processes	311
C.3. Integration of Gaussian processes in surrogate models	311
C.4. Choices of kernels	311
Bibliography	313

B. Cross Entropy and Other Expectations with Logarithms - DRAFT

This appendix expands on basic properties of cross entropy, the KL divergence, and basics of information theory. It is not meant to be an extensive review of these areas but rather provides key definitions, properties, and results needed for the content of the book.

B.1. Divergences and Entropies

We first define the relative entropy (KL divergence), cross entropy, and entropy in the context of probability distributions. We then define the Jensen–Shannon divergence. The information theoretic meaning of these quantities and a few related measures is described in Section B.2 below. At this point, we simply make formal use of these definitions.

Assume two probability distributions, $p(\cdot)$, and $q(\cdot)$, over elements in some discrete set \mathcal{X} ¹. That is, for each $x \in \mathcal{X}$ let $p(x)$ or $q(x)$ denote the respective probabilities, and assume that for $x \notin \mathcal{X}$, $p(x)$ and $q(x)$ are both zero. The definitions we present here also extend to continuous, multi-variate, and more general distributions, yet for simplicity we restrict the notation to summations over elements of discrete distributions.

A key measure for the proximity between the distributions $p(\cdot)$ and $q(\cdot)$ is the *Kullback–Leibler divergence*, also shortened as *KL divergence* or called the *relative entropy*. It is denoted $D_{\text{KL}}(p \parallel q)$ and is the expected value of $\log p(X)/q(X)$ where X is a random variable following the probability law $p(\cdot)$. Namely,

$$D_{\text{KL}}(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}.$$

Note that the expression for $D_{\text{KL}}(p \parallel q)$ can be decomposed into the difference of $H(p)$ from $H(p, q)$ via,

$$D_{\text{KL}}(p \parallel q) = \underbrace{\sum_{x \in \mathcal{X}} p(x) \log \frac{1}{q(x)}}_{H(p, q)} - \underbrace{\sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)}}_{H(p)}.$$

Here

$$H(p, q) = - \sum_{x \in \mathcal{X}} p(x) \log q(x) \tag{B.1}$$

is called the *cross entropy* of p and q and

$$H(p) = - \sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{B.2}$$

¹The set \mathcal{X} can be taken as the union of the supports of the distributions.

B. Cross Entropy and Other Expectations with Logarithms - DRAFT

is called the entropy of p . Hence in words, the KL divergence or relative entropy of p and q is the cross entropy of p and q with the entropy of p subtracted. Note that in case where there are only two values in \mathcal{X} , say 0 and 1, where we denote $p(1) = p_1$, and $q(1) = q_1$, we have

$$H(p) = -(p_1 \log p_1 + (1 - p_1) \log(1 - p_1)) \quad (\text{B.3})$$

$$H(p, q) = -(p_1 \log q_1 + (1 - p_1) \log(1 - q_1)). \quad (\text{B.4})$$

Some observations are in order. First observe that $D_{\text{KL}}(p \parallel q) \geq 0$. Further note that in general $D_{\text{KL}}(p \parallel q) \neq D_{\text{KL}}(q \parallel p)$ and similarly $H(p, q) \neq H(q, p)$. Hence as a “distance measure” the KL divergence is not a true metric since it isn’t symmetric over its arguments. Nevertheless, when $p = q$ the KL divergence is 0 and similarly the cross entropy equals the entropy. In addition, it can be shown that $D_{\text{KL}}(p \parallel q) = 0$ only when $p = q$. Hence the KL divergence may play a role similar to a distance metric in certain applications. In fact, in some applications one may consider a sequence $q^{(1)}, q^{(2)}, \dots$ which has decreasing $D_{\text{KL}}(p \parallel q^{(t)})$ approaching 0 as $t \rightarrow \infty$. For such a sequence the probability distributions $q^{(t)}$ approach the target distribution p since the KL divergence converges to 0.

A related measure to the KL divergence which is symmetric in arguments is the *Jensen-Shannon divergence* denoted $\text{JSD}(p \parallel q)$. It is defined by considering a mixture distribution,

$$m(x) = \frac{1}{2}(p(x) + q(x)),$$

and then averaging the KL divergence between each of the distributions and $m(\cdot)$, namely,

$$\text{JSD}(p \parallel q) = \frac{D_{\text{KL}}(p \parallel m) + D_{\text{KL}}(q \parallel m)}{2}. \quad (\text{B.5})$$

The square square root of $\text{JSD}(p \parallel q)$, sometimes called the *Jensen-Shannon distance* is a metric in the mathematical sense.

B.2. Interpretation via Information Theory