

# DEEP LEARNING 101

VU HUNG NGUYEN

A Comprehensive Guide to Deep Learning Theory and Practice





# **Deep Learning 101**

Nguyễn Vũ Hưng – 阮武興

A Comprehensive Guide to  
Deep Learning Theory and Practice

22nd October 2025

License: Creative Commons Attribution 4.0 International (CC BY 4.0)



**Deep Learning 101**

Copyright © 2025 Vu Hung Nguyen

This work is licensed under the Creative Commons Attribution  
4.0 International License.

To view a copy of this license, visit

<http://creativecommons.org/licenses/by/4.0/>

First Edition: 22nd October 2025

Repo:

<https://github.com/vuhung16au/101DeepLearning>

---

## Preface

Dear ML/AI learners,

Welcome to *Deep Learning 101*, a comprehensive guide designed to take you from the fundamentals to advanced concepts in deep learning. This book is crafted for students, researchers, and practitioners who want to build a solid foundation in one of the most transformative fields of our time.

I started this book as study notes for myself, because I couldn't find any book that fit my background. As I delved deeper into the field, I realised that many existing resources were either too basic or too advanced, leaving a gap for learners with a technical foundation seeking comprehensive yet accessible coverage.

This book is a “compressed version” of the deep learning books out there—you’ll find the most important concepts and mathematical formulae presented in a structured, coherent manner. Rather than overwhelming you with every detail, I’ve distilled the essential knowledge that forms the foundation of modern deep learning.

This book assumes you have a fair understanding of mathematics, algorithms, and programming. But don’t worry—you’ll learn along the way if you miss any of them. The early chapters provide mathematical foundations, and examples throughout the book will help reinforce concepts you may need to review.

The target audience includes those who want to learn the basics of deep learning with a focus on mathematics. If you’re interested in research or want to dive a little deeper into the maths and algorithms underlying neural networks, this book will guide you through both theory and practice.

Whether you’re just beginning your journey into machine learning or looking to deepen your understanding of neural networks, this book provides the mathematical foundations, practical insights, and hands-on knowledge you need to succeed.

I hope this resource serves you well in your learning journey.

Best regards,

Vu Hung Nguyen

### Contact Information:

GitHub: <https://github.com/vuhung16au>

LinkedIn: <https://www.linkedin.com/in/nguyenvuhung/>

Website: <https://vuhung16au.github.io/>

---

# Contents

<b>List of Figures</b>	<b>xviii</b>
<b>List of Tables</b>	<b>xix</b>
<b>List of Algorithms</b>	<b>xxi</b>
<b>List of Examples</b>	<b>xxi</b>
<b>List of Remarks</b>	<b>xxi</b>
<b>Acknowledgements</b>	<b>xxiii</b>
<b>Notation</b>	<b>xxv</b>
Difficulty Legend . . . . .	xxvii
<b>1 Introduction</b>	<b>1</b>
1.1 What is Deep Learning? ● . . . . .	1
1.1.1 The Rise of Deep Learning . . . . .	1
1.1.2 Key Characteristics . . . . .	2
1.1.3 Applications . . . . .	2
1.2 Historical Context ● . . . . .	2
1.2.1 The Perceptron Era (1940s-1960s) . . . . .	2
1.2.2 The Backpropagation Revolution (1980s) . . . . .	3
1.2.3 The Second AI Winter (1990s-2000s) . . . . .	3
1.2.4 The Deep Learning Renaissance (2006-Present) . . . . .	3
1.2.5 Key Milestones . . . . .	4
1.3 Fundamental Concepts ● . . . . .	4
1.3.1 Learning from Data . . . . .	4
1.3.2 The Learning Process . . . . .	4
1.3.3 Neural Networks as Universal Approximators . . . . .	5
1.3.4 Representation Learning . . . . .	5
1.3.5 Generalization and Overfitting . . . . .	5
1.4 Structure of This Book ● . . . . .	6
1.5 Prerequisites and Resources ● . . . . .	7
1.5.1 Mathematical Prerequisites . . . . .	7
1.5.2 Programming and Machine Learning Background . . . . .	8
1.5.3 Deep Learning Frameworks . . . . .	8
1.5.4 Additional Resources . . . . .	8
1.5.5 A Note on Exercises . . . . .	9
1.5.6 Getting Help . . . . .	9
Key Takeaways . . . . .	10

---

Exercises . . . . .	10
<b>I Basic Math and Machine Learning Foundation</b>	<b>15</b>
<b>2 Linear Algebra</b>	<b>17</b>
2.1 Scalars, Vectors, Matrices, and Tensors ● . . . . .	17
2.1.1 Scalars . . . . .	17
2.1.2 Vectors . . . . .	17
2.1.3 Matrices . . . . .	18
2.1.4 Tensors . . . . .	18
2.1.5 Notation Conventions . . . . .	19
2.2 Matrix Operations ● . . . . .	20
2.2.1 Matrix Addition and Scalar Multiplication . . . . .	20
2.2.2 Matrix Transpose . . . . .	20
2.2.3 Matrix Multiplication . . . . .	20
2.2.4 Element-wise (Hadamard) Product . . . . .	21
2.2.5 Matrix-Vector Products . . . . .	21
2.2.6 Dot Product . . . . .	22
2.2.7 Computational Complexity . . . . .	22
2.3 Identity and Inverse Matrices ● . . . . .	23
2.3.1 Identity Matrix . . . . .	23
2.3.2 Matrix Inverse . . . . .	24
2.3.3 Properties of Inverses . . . . .	24
2.3.4 Solving Linear Systems . . . . .	25
2.3.5 Conditions for Invertibility . . . . .	25
2.3.6 Singular Matrices . . . . .	25
2.3.7 Pseudo-inverse . . . . .	25
2.3.8 Practical Considerations . . . . .	25
2.4 Linear Dependence and Span ● . . . . .	26
2.4.1 Linear Combinations . . . . .	26
2.4.2 Span . . . . .	26
2.4.3 Linear Independence . . . . .	27
2.4.4 Basis . . . . .	27
2.4.5 Dimension and Rank . . . . .	28
2.4.6 Column Space and Null Space . . . . .	28
2.4.7 Relevance to Deep Learning . . . . .	29
2.5 Norms ● . . . . .	29
2.5.1 Definition of a Norm . . . . .	29
2.5.2 $L^p$ Norms . . . . .	29
2.5.3 Common Norms . . . . .	29
2.5.4 Frobenius Norm . . . . .	30
2.5.5 Unit Vectors and Normalization . . . . .	31
2.5.6 Distance Metrics . . . . .	31
2.5.7 Regularization in Deep Learning . . . . .	32
2.6 Eigendecomposition ● . . . . .	32
2.6.1 Eigenvalues and Eigenvectors . . . . .	32
2.6.2 Finding Eigenvalues . . . . .	32
2.6.3 Eigendecomposition . . . . .	33

---

2.6.4	Symmetric Matrices . . . . .	33
2.6.5	Properties of Eigenvalues . . . . .	33
2.6.6	Positive Definite Matrices . . . . .	34
2.6.7	Applications in Deep Learning . . . . .	34
2.6.8	Computational Considerations . . . . .	34
Key Takeaways . . . . .		35
Exercises . . . . .		35
<b>3</b>	<b>Probability and Information Theory</b>	<b>39</b>
3.1	Probability Distributions ● . . . . .	39
3.1.1	Intuition: What is Probability? . . . . .	39
3.1.2	Visualizing Probability . . . . .	40
3.1.3	Discrete Probability Distributions . . . . .	40
3.1.4	Continuous Probability Distributions . . . . .	40
3.1.5	Joint and Marginal Distributions . . . . .	42
3.2	Conditional Probability and Bayes' Rule ● . . . . .	43
3.2.1	Intuition: Updating Beliefs with New Information . . . . .	43
3.2.2	Conditional Probability . . . . .	43
3.2.3	Independence . . . . .	44
3.2.4	Bayes' Theorem . . . . .	45
3.2.5	Application to Machine Learning . . . . .	45
3.3	Expectation, Variance, and Covariance ● . . . . .	46
3.3.1	Intuition: Characterizing Random Variables . . . . .	46
3.3.2	Expectation . . . . .	46
3.3.3	Variance . . . . .	47
3.3.4	Covariance . . . . .	48
3.3.5	Correlation . . . . .	48
3.4	Common Probability Distributions ● . . . . .	49
3.4.1	Bernoulli Distribution . . . . .	49
3.4.2	Categorical Distribution . . . . .	50
3.4.3	Gaussian (Normal) Distribution . . . . .	50
3.4.4	Exponential Distribution . . . . .	50
3.4.5	Laplace Distribution . . . . .	50
3.4.6	Dirac Delta and Mixture Distributions . . . . .	51
3.5	Information Theory Basics ● . . . . .	52
3.5.1	Self-Information . . . . .	52
3.5.2	Entropy . . . . .	52
3.5.3	Cross-Entropy . . . . .	52
3.5.4	Kullback-Leibler Divergence . . . . .	53
3.5.5	Mutual Information . . . . .	53
3.5.6	Applications in Deep Learning . . . . .	53
Key Takeaways . . . . .		54
Exercises . . . . .		54
<b>4</b>	<b>Numerical Computation</b>	<b>59</b>
4.1	Overflow and Underflow ● . . . . .	59
4.1.1	Intuition: The Problem with Finite Precision . . . . .	59
4.1.2	Floating-Point Representation . . . . .	60
4.1.3	Underflow . . . . .	60

---

4.1.4	Overflow . . . . .	60
4.1.5	Numerical Stability . . . . .	60
4.1.6	Other Numerical Issues . . . . .	61
4.2	Gradient-Based Optimization ● . . . . .	62
4.2.1	Intuition: Finding the Bottom of a Hill . . . . .	62
4.2.2	Gradient Descent . . . . .	62
4.2.3	Jacobian and Hessian Matrices . . . . .	63
4.2.4	Taylor Series Approximation . . . . .	63
4.2.5	Critical Points . . . . .	63
4.2.6	Directional Derivatives . . . . .	64
4.3	Constrained Optimization ● . . . . .	64
4.3.1	Intuition: Optimization with Rules . . . . .	64
4.3.2	Lagrange Multipliers . . . . .	64
4.3.3	Inequality Constraints . . . . .	65
4.3.4	Projected Gradient Descent . . . . .	65
4.3.5	Applications in Deep Learning . . . . .	66
4.4	Numerical Stability and Conditioning ● . . . . .	66
4.4.1	Intuition: The Butterfly Effect in Computation . . . . .	66
4.4.2	Condition Number . . . . .	66
4.4.3	Ill-Conditioned Matrices . . . . .	66
4.4.4	Gradient Checking . . . . .	67
4.4.5	Numerical Precision Trade-offs . . . . .	67
4.4.6	Practical Tips . . . . .	67
	Key Takeaways . . . . .	68
	Exercises . . . . .	68
<b>5</b>	<b>Classical Machine Learning Algorithms</b>	<b>71</b>
5.1	Linear Regression ♦ . . . . .	71
5.1.1	Intuition and Motivation . . . . .	71
5.1.2	Model Formulation . . . . .	72
5.1.3	Ordinary Least Squares . . . . .	72
5.1.4	Regularized Regression . . . . .	72
5.1.5	Gradient Descent Solution . . . . .	73
5.1.6	Geometric Interpretation . . . . .	74
5.2	Logistic Regression ♦ . . . . .	74
5.2.1	Intuition and Motivation . . . . .	74
5.2.2	Binary Classification . . . . .	75
5.2.3	Cross-Entropy Loss . . . . .	75
5.2.4	Gradient Descent for Logistic Regression . . . . .	76
5.2.5	Multiclass Classification . . . . .	76
5.2.6	Categorical Cross-Entropy Loss . . . . .	76
5.2.7	Decision Boundaries . . . . .	77
5.2.8	Regularization in Logistic Regression . . . . .	77
5.2.9	Advantages and Limitations . . . . .	77
5.3	Support Vector Machines ♦ . . . . .	78
5.3.1	Intuition and Motivation . . . . .	78
5.3.2	Linear SVM . . . . .	78
5.3.3	Soft Margin SVM . . . . .	79
5.3.4	Dual Formulation . . . . .	79

---

5.3.5	Kernel Trick . . . . .	79
5.3.6	Kernel Properties . . . . .	80
5.3.7	Advantages and Limitations . . . . .	80
5.3.8	SVM for Regression . . . . .	80
5.4	Decision Trees and Ensemble Methods ♦ . . . . .	81
5.4.1	Intuition and Motivation . . . . .	81
5.4.2	Decision Trees . . . . .	81
5.4.3	Random Forests . . . . .	83
5.4.4	Gradient Boosting . . . . .	83
5.4.5	Advanced Ensemble Methods . . . . .	84
5.4.6	Comparison of Ensemble Methods . . . . .	84
5.4.7	Advantages and Limitations . . . . .	84
5.5	k-Nearest Neighbors ♦ . . . . .	85
5.5.1	Intuition and Motivation . . . . .	85
5.5.2	Algorithm . . . . .	85
5.5.3	Distance Metrics . . . . .	86
5.5.4	Choosing k . . . . .	86
5.5.5	Weighted k-NN . . . . .	87
5.5.6	Computational Considerations . . . . .	87
5.5.7	Advantages and Limitations . . . . .	87
5.5.8	Curse of Dimensionality . . . . .	87
5.5.9	Feature Selection and Scaling . . . . .	88
5.6	Comparison with Deep Learning ♦ . . . . .	88
5.6.1	When Classical Methods Excel . . . . .	88
5.6.2	When Deep Learning Excels . . . . .	88
5.6.3	Performance Comparison . . . . .	88
5.6.4	Hybrid Approaches . . . . .	88
5.6.5	Choosing the Right Approach . . . . .	89
5.6.6	Future Directions . . . . .	89
5.6.7	Conclusion . . . . .	89
	Key Takeaways . . . . .	90
	Exercises . . . . .	90

## II Practical Deep Networks 93

6	Deep Feedforward Networks <span style="float: right;">95</span>	
6.1	Introduction to Feedforward Networks ♦ . . . . .	95
6.1.1	Intuition: What is a Feedforward Network? . . . . .	95
6.1.2	Network Architecture . . . . .	95
6.1.3	Forward Propagation . . . . .	96
6.1.4	Universal Approximation . . . . .	97
6.2	Activation Functions ♦ . . . . .	97
6.2.1	Intuition: Why Do We Need Activation Functions? . . . . .	97
6.2.2	Sigmoid . . . . .	98
6.2.3	Hyperbolic Tangent ( $\tanh$ ) . . . . .	98
6.2.4	Rectified Linear Unit (ReLU) . . . . .	99
6.2.5	Leaky ReLU and Variants . . . . .	99
6.2.6	Swish and GELU . . . . .	99

---

6.3	Output Units and Loss Functions ♦ . . . . .	100
6.3.1	Intuition: Matching Output to Task . . . . .	100
6.3.2	Linear Output for Regression . . . . .	100
6.3.3	Sigmoid Output for Binary Classification . . . . .	101
6.3.4	Softmax Output for Multiclass Classification . . . . .	101
6.4	Backpropagation ♦ . . . . .	101
6.4.1	Intuition: Learning from Mistakes . . . . .	102
6.4.2	The Chain Rule . . . . .	102
6.4.3	Backward Pass . . . . .	104
6.4.4	Computational Graph . . . . .	104
6.5	Design Choices ♦ . . . . .	105
6.5.1	Intuition: Building the Right Network . . . . .	105
6.5.2	Network Depth and Width . . . . .	105
6.5.3	Weight Initialization . . . . .	105
6.5.4	Batch Training . . . . .	106
6.6	Real World Applications . . . . .	106
6.6.1	Medical Diagnosis Support . . . . .	106
6.6.2	Financial Fraud Detection . . . . .	106
6.6.3	Product Recommendation Systems . . . . .	106
6.6.4	Why These Applications Work . . . . .	107
	Key Takeaways . . . . .	107
	Exercises . . . . .	107
<b>7</b>	<b>Regularization for Deep Learning</b> . . . . .	<b>111</b>
7.1	Parameter Norm Penalties ♦ . . . . .	111
7.1.1	Intuition: Shrinking the Model's "Complexity" . . . . .	111
7.1.2	L2 Regularization (Weight Decay) . . . . .	111
7.1.3	L1 Regularization . . . . .	112
7.1.4	Elastic Net . . . . .	112
7.2	Dataset Augmentation ♦ . . . . .	112
7.2.1	Intuition: Seeing the Same Thing in Many Ways . . . . .	112
7.2.2	Image Augmentation . . . . .	112
7.2.3	Text Augmentation . . . . .	113
7.2.4	Audio Augmentation . . . . .	113
7.3	Early Stopping ♦ . . . . .	113
7.3.1	Intuition: Stop Before You Memorize . . . . .	113
7.3.2	Algorithm . . . . .	114
7.3.3	Benefits . . . . .	114
7.3.4	Considerations . . . . .	114
7.4	Dropout ♦ . . . . .	115
7.4.1	Intuition: Training a Robust Ensemble . . . . .	115
7.4.2	Training with Dropout . . . . .	115
7.4.3	Inference . . . . .	115
7.4.4	Interpretation . . . . .	116
7.4.5	Variants . . . . .	116
7.5	Batch Normalization ♦ . . . . .	116
7.5.1	Intuition: Keeping Scales Stable . . . . .	117
7.5.2	Algorithm . . . . .	117
7.5.3	Benefits . . . . .	117

---

7.5.4	Inference . . . . .	118
7.5.5	Variants . . . . .	118
7.6	Other Regularization Techniques ♦ . . . . .	118
7.6.1	Intuition: Many Small Guards Against Overfitting . . . . .	118
7.6.2	Label Smoothing . . . . .	118
7.6.3	Gradient Clipping . . . . .	118
7.6.4	Stochastic Depth . . . . .	119
7.6.5	Mixup . . . . .	119
7.6.6	Adversarial Training . . . . .	120
7.7	Real World Applications . . . . .	120
7.7.1	Autonomous Vehicle Safety . . . . .	120
7.7.2	Medical Imaging Analysis . . . . .	121
7.7.3	Natural Language Processing for Customer Service . . . . .	121
7.7.4	Key Benefits in Practice . . . . .	121
	Key Takeaways . . . . .	121
	Exercises . . . . .	122

<b>8</b>	<b>Optimization for Training Deep Models</b> . . . . .	<b>125</b>
8.1	Gradient Descent Variants ♦ . . . . .	125
8.1.1	Intuition: Following the Steepest Downhill Direction . . . . .	125
8.1.2	Batch Gradient Descent . . . . .	125
8.1.3	Stochastic Gradient Descent (SGD) . . . . .	126
8.1.4	Mini-Batch Gradient Descent . . . . .	126
8.2	Momentum-Based Methods ♦ . . . . .	127
8.2.1	Intuition: Rolling a Ball Down a Valley . . . . .	127
8.2.2	Momentum . . . . .	127
8.2.3	Nesterov Accelerated Gradient (NAG) . . . . .	128
8.3	Adaptive Learning Rate Methods ♦ . . . . .	128
8.3.1	Intuition: Per-Parameter Step Sizes . . . . .	128
8.3.2	AdaGrad . . . . .	129
8.3.3	RMSProp . . . . .	129
8.3.4	Adam (Adaptive Moment Estimation) . . . . .	129
8.3.5	Learning Rate Schedules . . . . .	129
8.4	Second-Order Methods ♦ . . . . .	130
8.4.1	Intuition: Curvature-Aware Steps . . . . .	130
8.4.2	Newton's Method . . . . .	130
8.4.3	Quasi-Newton Methods . . . . .	131
8.4.4	Natural Gradient . . . . .	131
8.5	Optimization Challenges ♦ . . . . .	131
8.5.1	Intuition: Why Training Gets Stuck . . . . .	131
8.5.2	Vanishing and Exploding Gradients . . . . .	131
8.5.3	Local Minima and Saddle Points . . . . .	132
8.5.4	Plateaus . . . . .	133
8.5.5	Practical Optimization Strategy . . . . .	133
8.6	Key Takeaways ♦ . . . . .	134
	Exercises . . . . .	134

---

<b>9 Convolutional Networks</b>	<b>137</b>
9.1 The Convolution Operation . . . . .	137
9.1.1 Definition . . . . .	138
9.1.2 Properties . . . . .	138
9.1.3 Multi-Channel Convolution . . . . .	139
9.1.4 Hyperparameters . . . . .	139
9.2 Pooling . . . . .	140
9.2.1 Max Pooling . . . . .	140
9.2.2 Average Pooling . . . . .	140
9.2.3 Global Pooling . . . . .	140
9.2.4 Alternative: Strided Convolutions . . . . .	141
9.3 CNN Architectures ◆ . . . . .	141
9.3.1 LeNet-5 (1998) . . . . .	141
9.3.2 AlexNet (2012) . . . . .	142
9.3.3 VGG Networks (2014) . . . . .	142
9.3.4 ResNet (2015) . . . . .	143
9.3.5 Inception/GoogLeNet (2014) . . . . .	143
9.3.6 MobileNet and EfficientNet . . . . .	143
9.4 Applications of CNNs ◆ . . . . .	144
9.4.1 Image Classification . . . . .	144
9.4.2 Object Detection . . . . .	144
9.4.3 Semantic Segmentation . . . . .	145
9.5 Core CNN Algorithms ◆ . . . . .	145
9.5.1 Cross-Correlation and Convolution . . . . .	145
9.5.2 Backpropagation Through Convolution . . . . .	145
9.5.3 Pooling Backpropagation . . . . .	145
9.5.4 Residual Connections . . . . .	146
9.5.5 Progressive Complexity: Depthwise Separable Convolutions . . . . .	146
9.5.6 Normalization and Activation . . . . .	146
9.5.7 Other Useful Variants . . . . .	146
9.6 Real World Applications ● . . . . .	147
9.6.1 Medical Image Analysis . . . . .	147
9.6.2 Autonomous Driving . . . . .	147
Key Takeaways . . . . .	147
Exercises . . . . .	148
<b>10 Sequence Modeling: Recurrent and Recursive Nets</b>	<b>151</b>
10.1 Recurrent Neural Networks ◆ . . . . .	151
10.1.1 Motivation . . . . .	151
10.1.2 Why Sequences Matter . . . . .	152
10.1.3 Basic RNN Architecture . . . . .	152
10.1.4 Unfolding in Time . . . . .	152
10.1.5 Types of Sequences . . . . .	153
10.2 Backpropagation Through Time (BPTT) ◆ . . . . .	154
10.2.1 BPTT Algorithm . . . . .	154
10.2.2 Vanishing and Exploding Gradients . . . . .	155
10.2.3 Truncated BPTT . . . . .	155
10.3 Long Short-Term Memory (LSTM) ◆ . . . . .	155
10.3.1 Architecture . . . . .	156

---

10.3.2 Key Ideas . . . . .	157
10.3.3 Advantages . . . . .	157
10.4 Gated Recurrent Units (GRU) ♦ . . . . .	157
10.4.1 Architecture . . . . .	158
10.4.2 Architecture (visual) . . . . .	159
10.4.3 Comparison with LSTM . . . . .	159
10.5 Sequence-to-Sequence Models ♦ . . . . .	159
10.5.1 Encoder-Decoder Architecture . . . . .	159
10.5.2 Attention Mechanism: Attention Is All You Need . . . . .	160
10.5.3 Applications . . . . .	161
10.6 Advanced Topics ♦ . . . . .	162
10.6.1 Bidirectional RNNs . . . . .	162
10.6.2 Deep RNNs . . . . .	163
10.6.3 Teacher Forcing . . . . .	163
10.6.4 Beam Search . . . . .	163
10.7 Real World Applications . . . . .	164
10.7.1 Machine Translation . . . . .	164
10.7.2 Voice Assistants and Speech Recognition . . . . .	164
10.7.3 Predictive Text and Content Generation . . . . .	164
10.7.4 Financial Forecasting and Analysis . . . . .	165
Key Takeaways . . . . .	165
Exercises . . . . .	165

<b>11 Practical Methodology</b>	<b>169</b>
Intuition . . . . .	169
11.1 Performance Metrics ♦ . . . . .	169
11.1.1 Classification Metrics . . . . .	169
11.1.2 Regression Metrics . . . . .	172
11.1.3 NLP and Sequence Metrics . . . . .	173
11.1.4 Worked examples . . . . .	173
11.2 Baseline Models and Debugging ♦ . . . . .	173
11.2.1 Establishing Baselines . . . . .	174
11.2.2 Debugging Strategy . . . . .	174
11.2.3 Common Issues . . . . .	175
11.2.4 Ablation and sanity checks . . . . .	176
11.3 Hyperparameter Tuning ♦ . . . . .	177
11.3.1 Key Hyperparameters (Priority Order) . . . . .	177
11.3.2 Search Strategies . . . . .	177
11.3.3 Best Practices . . . . .	178
11.3.4 Historical notes . . . . .	179
11.4 Data Preparation and Preprocessing ♦ . . . . .	179
11.4.1 Data Splitting . . . . .	179
11.4.2 Normalization . . . . .	180
11.4.3 Handling Imbalanced Data . . . . .	181
11.4.4 Data Augmentation . . . . .	181
11.4.5 Visual aids . . . . .	182
11.4.6 Historical notes . . . . .	182
11.5 Production Considerations ♦ . . . . .	183
11.5.1 Model Deployment . . . . .	183

---

11.5.2 Monitoring . . . . .	184
11.5.3 Iterative Improvement . . . . .	184
11.5.4 Model Drift . . . . .	185
11.5.5 Applications and context . . . . .	186
11.6 Real World Applications . . . . .	186
11.6.1 Healthcare Diagnostic System Deployment . . . . .	186
11.6.2 Recommendation System Development . . . . .	187
11.6.3 Autonomous Vehicle Development . . . . .	187
11.6.4 Key Methodological Principles . . . . .	188
Key Takeaways . . . . .	189
Exercises . . . . .	189
<b>12 Applications</b>	<b>193</b>
Intuition . . . . .	193
12.1 Computer Vision Applications ● . . . . .	194
12.1.1 Image Classification . . . . .	194
12.1.2 Object Detection . . . . .	194
12.1.3 Semantic Segmentation . . . . .	195
12.1.4 Face Recognition . . . . .	196
12.1.5 Image Generation and Manipulation . . . . .	196
12.1.6 Historical context and references . . . . .	197
12.2 Natural Language Processing ● . . . . .	198
12.2.1 Text Classification . . . . .	198
12.2.2 Machine Translation . . . . .	199
12.2.3 Question Answering . . . . .	199
12.2.4 Language Models and Text Generation . . . . .	200
12.2.5 Named Entity Recognition . . . . .	201
12.2.6 Historical context and references . . . . .	201
12.3 Speech Recognition and Synthesis ● . . . . .	202
12.3.1 Automatic Speech Recognition (ASR) . . . . .	202
12.3.2 Text-to-Speech (TTS) . . . . .	203
12.3.3 Speaker Recognition . . . . .	204
12.3.4 Historical context and references . . . . .	204
12.4 Healthcare and Medical Imaging ● . . . . .	205
12.4.1 Medical Image Analysis . . . . .	205
12.4.2 Drug Discovery . . . . .	206
12.4.3 Clinical Decision Support . . . . .	207
12.4.4 Genomics . . . . .	207
12.4.5 Historical context and references . . . . .	208
12.5 Reinforcement Learning Applications ● . . . . .	208
12.5.1 Game Playing . . . . .	209
12.5.2 Robotics . . . . .	209
12.5.3 Autonomous Vehicles . . . . .	210
12.5.4 Recommendation Systems . . . . .	211
12.5.5 Resource Management . . . . .	211
12.5.6 Historical context and references . . . . .	212
12.6 Other Applications ● . . . . .	212
12.6.1 Finance . . . . .	213
12.6.2 Scientific Research . . . . .	213

---

12.6.3 Agriculture . . . . .	214
12.6.4 Manufacturing . . . . .	215
12.6.5 References . . . . .	215
12.7 Ethical Considerations ● . . . . .	216
Key Takeaways . . . . .	217
Exercises . . . . .	217

### III Deep Learning Research 221

<b>13 Linear Factor Models</b> <span style="float: right;">223</span>	
Intuition . . . . .	223
13.1 Probabilistic PCA ♦ . . . . .	223
13.1.1 Principal Component Analysis Review . . . . .	223
13.1.2 Probabilistic Formulation . . . . .	224
13.1.3 Learning . . . . .	224
13.1.4 Historical context and references . . . . .	224
13.2 Factor Analysis ♦ . . . . .	225
13.2.1 Learning via EM . . . . .	225
13.3 Independent Component Analysis ♦ . . . . .	225
13.3.1 Objective . . . . .	225
13.3.2 Non-Gaussianity . . . . .	226
13.4 Sparse Coding ♦ . . . . .	226
13.4.1 Optimization and interpretation . . . . .	226
13.5 Real World Applications . . . . .	226
13.5.1 Facial Recognition Systems . . . . .	227
13.5.2 Audio Signal Processing . . . . .	227
13.5.3 Anomaly Detection in Systems . . . . .	227
13.5.4 Practical Advantages . . . . .	227
Key Takeaways . . . . .	228
Exercises . . . . .	228

<b>14 Autoencoders</b> <span style="float: right;">231</span>	
Intuition . . . . .	231
14.1 Undercomplete Autoencoders ♦ . . . . .	231
14.1.1 Architecture . . . . .	231
14.1.2 Training Objective . . . . .	232
14.1.3 Undercomplete Constraint . . . . .	232
14.2 Regularized Autoencoders ♦ . . . . .	232
14.2.1 Sparse Autoencoders . . . . .	232
14.2.2 Denoising Autoencoders (DAE) . . . . .	232
14.2.3 Contractive Autoencoders (CAE) . . . . .	233
14.3 Variational Autoencoders ♦ . . . . .	233
14.3.1 Probabilistic Framework . . . . .	233
14.3.2 Evidence Lower Bound (ELBO) . . . . .	233
14.3.3 Reparameterization Trick . . . . .	233
14.3.4 Generation . . . . .	233
14.3.5 Notes and references . . . . .	234
14.4 Applications of Autoencoders ♦ . . . . .	234
14.4.1 Dimensionality Reduction . . . . .	234

---

14.4.2 Anomaly Detection . . . . .	234
14.4.3 Denoising . . . . .	235
14.4.4 References . . . . .	235
<b>14.5 Real World Applications . . . . .</b>	<b>235</b>
14.5.1 Image and Video Compression . . . . .	235
14.5.2 Denoising and Enhancement . . . . .	235
14.5.3 Anomaly Detection . . . . .	236
14.5.4 Why Autoencoders Excel . . . . .	236
14.5.5 Autoencoders Compared to other NN Algorithms . . . . .	236
Key Takeaways . . . . .	237
Exercises . . . . .	237
<b>15 Representation Learning</b>	<b>239</b>
Intuition . . . . .	239
Representation Function . . . . .	239
15.1 What Makes a Good Representation? ♦ . . . . .	239
15.1.1 Desirable Properties . . . . .	240
15.1.2 Manifold Hypothesis . . . . .	240
15.1.3 Notes and references . . . . .	240
15.2 Transfer Learning and Domain Adaptation ♦ . . . . .	240
15.2.1 Transfer Learning . . . . .	240
15.2.2 Domain Adaptation . . . . .	240
15.2.3 Few-Shot Learning . . . . .	241
15.3 Self-Supervised Learning ♦ . . . . .	241
15.3.1 Pretext Tasks . . . . .	241
15.3.2 Benefits . . . . .	241
15.4 Contrastive Learning ♦ . . . . .	241
15.4.1 Core Idea . . . . .	242
15.4.2 SimCLR Framework . . . . .	242
15.4.3 MoCo (Momentum Contrast) . . . . .	242
15.4.4 BYOL (Bootstrap Your Own Latent) . . . . .	242
15.4.5 Applications . . . . .	242
15.5 Real World Applications . . . . .	243
15.5.1 Transfer Learning in Computer Vision . . . . .	243
15.5.2 Natural Language Processing . . . . .	243
15.5.3 Cross-Modal Representations . . . . .	243
15.5.4 Impact of Good Representations . . . . .	243
Key Takeaways . . . . .	244
Exercises . . . . .	244
<b>16 Structured Probabilistic Models for DL</b>	<b>247</b>
Intuition . . . . .	247
16.1 Graphical Models ★ . . . . .	247
16.1.1 Motivation . . . . .	247
16.1.2 Bayesian Networks . . . . .	248
16.1.3 Markov Random Fields . . . . .	248
16.2 Inference in Graphical Models ★ . . . . .	248
16.2.1 Exact Inference . . . . .	248
16.2.2 Approximate Inference . . . . .	248

---

16.3 Deep Learning and Structured Models ★ . . . . .	249
16.3.1 Structured Output Prediction . . . . .	249
16.3.2 Structured Prediction with Neural Networks . . . . .	249
16.3.3 Neural Module Networks . . . . .	249
16.3.4 Graph Neural Networks . . . . .	249
16.4 Real World Applications . . . . .	250
16.4.1 Autonomous Vehicle Decision Making . . . . .	250
16.4.2 Medical Diagnosis and Treatment . . . . .	250
16.4.3 Natural Language Understanding . . . . .	250
16.4.4 Social Media Network Analysis . . . . .	250
16.4.5 Value of Structured Models . . . . .	251
Key Takeaways . . . . .	251
Exercises . . . . .	251
<b>17 Monte Carlo Methods</b>	<b>255</b>
Intuition . . . . .	255
17.1 Sampling and Monte Carlo Estimators ★ . . . . .	255
17.1.1 Monte Carlo Estimation . . . . .	255
17.1.2 Variance Reduction . . . . .	256
17.2 Markov Chain Monte Carlo ★ . . . . .	256
17.2.1 Markov Chains . . . . .	256
17.2.2 Metropolis-Hastings Algorithm . . . . .	256
17.2.3 Gibbs Sampling . . . . .	257
17.2.4 Hamiltonian Monte Carlo . . . . .	257
17.3 Importance Sampling ★ . . . . .	257
17.4 Applications in Deep Learning ★ . . . . .	258
17.5 Real World Applications . . . . .	258
17.5.1 Financial Risk Management . . . . .	258
17.5.2 Climate and Weather Modeling . . . . .	259
17.5.3 Drug Discovery and Design . . . . .	259
17.5.4 Practical Advantages . . . . .	259
Key Takeaways . . . . .	260
Exercises . . . . .	260
<b>18 Confronting the Partition Function</b>	<b>263</b>
Intuition . . . . .	263
18.1 The Partition Function Problem ★ . . . . .	263
18.1.1 Why It's Hard . . . . .	264
18.1.2 Impact . . . . .	264
18.2 Contrastive Divergence ★ . . . . .	264
18.2.1 Motivation . . . . .	264
18.2.2 CD-k Algorithm . . . . .	265
18.3 Noise-Contrastive Estimation ★ . . . . .	265
18.3.1 Key Idea . . . . .	265
18.3.2 NCE Objective . . . . .	265
18.3.3 Applications . . . . .	266
18.3.4 Notes and references . . . . .	266
18.4 Score Matching ★ . . . . .	266
18.5 Real World Applications . . . . .	266

---

18.5.1	Recommender Systems at Scale . . . . .	267
18.5.2	Natural Language Processing . . . . .	267
18.5.3	Computer Vision . . . . .	267
18.5.4	Practical Solutions . . . . .	268
Key Takeaways . . . . .		268
Exercises . . . . .		269
<b>19</b>	<b>Approximate Inference</b>	<b>271</b>
Intuition . . . . .		271
19.1	Variational Inference ★ . . . . .	271
19.1.1	Evidence Lower Bound (ELBO) . . . . .	271
19.1.2	Variational Family . . . . .	272
19.1.3	Coordinate Ascent VI . . . . .	272
19.1.4	Stochastic Variational Inference . . . . .	272
19.2	Mean Field Approximation ★ . . . . .	273
19.2.1	Fully Factorized Approximation . . . . .	273
19.2.2	Update Equations . . . . .	273
19.2.3	Properties . . . . .	273
19.3	Loopy Belief Propagation ★ . . . . .	274
19.3.1	Message Passing . . . . .	274
19.3.2	Beliefs . . . . .	274
19.3.3	Exact on Trees . . . . .	274
19.3.4	Loopy Graphs . . . . .	274
19.4	Expectation Propagation ★ . . . . .	275
19.5	Real World Applications . . . . .	275
19.5.1	Autonomous Systems . . . . .	275
19.5.2	Personalized Medicine . . . . .	275
19.5.3	Content Recommendation . . . . .	276
19.5.4	Natural Language Systems . . . . .	276
19.5.5	Why Approximation Is Essential . . . . .	276
Key Takeaways . . . . .		277
Exercises . . . . .		277
<b>20</b>	<b>Deep Generative Models</b>	<b>281</b>
Intuition . . . . .		281
20.1	Variational Autoencoders (VAEs) ★ . . . . .	281
20.1.1	Recap . . . . .	281
20.1.2	Conditional VAEs . . . . .	282
20.1.3	Disentangled Representations . . . . .	282
20.2	Generative Adversarial Networks (GANs) ★ . . . . .	282
20.2.1	Core Idea . . . . .	282
20.2.2	Objective . . . . .	283
20.2.3	Training Procedure . . . . .	283
20.2.4	Training Challenges . . . . .	283
20.2.5	GAN Variants . . . . .	283
20.3	Normalizing Flows ★ . . . . .	284
20.3.1	Key Idea . . . . .	284
20.3.2	Change of Variables . . . . .	284
20.3.3	Requirements . . . . .	284

20.3.4 Flow Architectures . . . . .	284
20.3.5 Advantages . . . . .	285
20.4 Diffusion Models ★ . . . . .	285
20.4.1 Forward Process . . . . .	285
20.4.2 Reverse Process . . . . .	285
20.4.3 Training . . . . .	286
20.4.4 Sampling . . . . .	286
20.4.5 The Algorithm . . . . .	286
20.4.6 Advantages . . . . .	286
20.5 Applications and Future Directions ★ . . . . .	287
20.5.1 Current Applications . . . . .	287
20.5.2 Future Directions . . . . .	287
20.5.3 Societal Impact . . . . .	287
20.6 Real World Applications . . . . .	288
20.6.1 Creative Content Generation . . . . .	288
20.6.2 Scientific Discovery . . . . .	288
20.6.3 Data Augmentation and Synthesis . . . . .	289
20.6.4 Personalization and Adaptation . . . . .	289
20.6.5 Transformative Impact . . . . .	289
Key Takeaways . . . . .	290
Exercises . . . . .	290
List of Abbreviations . . . . .	293
<b>Bibliography</b>	<b>299</b>



---

# List of Figures

3.1	Probability distribution for rain prediction . . . . .	40
3.2	Probability mass function for a fair coin . . . . .	41
3.3	Bayes' theorem as a belief update process . . . . .	45
3.4	Standard normal distribution $\mathcal{N}(0, 1)$ : bell-shaped curve with mean 0, std 1. . . . .	50
3.5	Exponential distribution ( $\lambda = 1$ ): starts at max, decreases exponentially. . . . .	51
3.6	Laplace ( $\mu = 0, b = 1$ ): sharp peak with exponential tails, more robust to outliers than Gaussian. . . . .	51
4.1	Representable range for 32-bit floating-point numbers . . . . .	60
4.2	Error amplification for well-conditioned vs ill-conditioned matrices . . . . .	67
5.1	Linear regression finds best line fitting data points, minimizing sum of squared errors. . . . .	72
5.2	L1 vs L2 regularization: L1 drives weights to zero, L2 shrinks them smoothly. . . . .	73
5.3	Sigmoid function maps real numbers to probabilities (0,1). Decision boundary at 0.5. . . . .	74
5.4	Logistic regression finds linear decision boundary separating classes 0 and 1. . . . .	77
5.5	SVM finds hyperplane maximizing margin between classes. Support vectors are closest to boundary. . . . .	78
5.6	SVM with RBF kernel learns non-linear decision boundaries (circular here). . . . .	80
5.7	Decision tree for product purchases: internal nodes are decisions, leaves are predictions. . . . .	81
5.8	Random forests: multiple trees trained on bootstrap samples, final prediction by averaging/voting. . . . .	83
5.9	k-NN finds k nearest neighbors. With k=3, query classified as class 0 (2/3 neighbors). . . . .	85
5.10	Effect of k on k-NN: small k $\rightarrow$ high variance, large k $\rightarrow$ high bias. Optimal k via CV. . . . .	87
6.1	Simple neural network for digit classification: "This is a 3" or "This is a 7". The network learns to distinguish between these two digits by analyzing pixel patterns. . . . .	96
6.2	Feedforward neural network with 2 hidden layers. Circles=neurons, arrows=flow. . . . .	96
6.3	Sigmoid $\sigma(z) = 1/(1 + e^{-z})$ : smooth S-shape $\mathbb{R} \rightarrow (0, 1)$ . . . . .	98
6.4	$\tanh(z)$ : zero-centered S-shape mapping $\mathbb{R} \rightarrow (-1, 1)$ . . . . .	98
6.5	Leaky ReLU, PReLU, and ELU activation functions with $\alpha = 0.001$ . . . . .	99
6.6	Common activations: sigmoid/tanh saturate, ReLU can "die", Swish smooth. . . . .	100
6.7	MSE loss in 1D: convex parabola at target $y$ . Min at $\hat{y}=y$ . . . . .	101
7.1	Illustration of common image augmentations. Variants preserve labels while encouraging invariances. . . . .	113
7.2	Early stopping: validation loss minimum before training loss; best checkpoint saved and restored. . . . .	115
7.3	Dropout training: randomly deactivating neurons ( $\times$ ) creates subnetworks, forcing robustness. . . . .	116
7.4	Dropout as implicit model averaging over many subnetworks. . . . .	117
7.5	Dropout during training: randomly deactivating hidden units encourages redundancy. . . . .	117
8.1	Newton's method rescales gradient by curvature: longer steps in shallow, shorter in steep. . . . .	130
8.2	Natural gradient accounts for local geometry (Fisher metric), directing updates orthogonally. . . . .	131
8.3	Critical points in $f(x, y) = x^4 - 2x^2 + y^2$ : minimum at $(0, 0)$ , saddle points at $(\pm 1, 0)$ . . . . .	132

9.1	Deep CNN: conv layers (orange) maintain spatial structure, FC layers (blue) process global features.	138
9.2	Downsampling via stride 2: fewer spatial samples after a strided convolution compared to pooling.	141
9.3	LeNet-5 architecture: alternating conv and pooling, followed by small fully connected layers.	142
9.4	ResNet basic residual block with identity skip connection.	143
9.5	Inception module: parallel multi-scale branches with $1 \times 1$ bottlenecks.	144
10.1	Unrolled RNN with shared parameters across time steps.	152
10.2	Unrolling an RNN across time: the same parameters $\theta$ are reused at each step.	153
10.3	Visual aid: A compact LSTM cell diagram.	158
10.4	Illustrative GRU flow with update and reset gates.	159
10.5	Encoder – decoder with fixed context vector $c$ . Attention uses step-dependent $c_t$ .	160
10.6	Deep RNN: multiple recurrent layers stacked over time.	163
11.1	Confusion matrix heatmap (example counts). High diagonal values indicate good performance.	170
11.2	ROC curves for two models. Higher AUC indicates better ranking quality.	171
11.3	Precision–recall curves emphasize performance on the positive class under imbalance.	171
11.4	Reliability diagram illustrating calibration. The diagonal is perfect calibration.	172
11.5	Comparison of squared, absolute, and Huber losses.	172
11.6	Typical overfitting: training loss decreases while validation loss bottoms out and rises.	176
11.7	Gradient norms vanishing with depth; motivates normalization and residual connections.	176
11.8	Learning-rate sweep to identify a stable training regime.	177
11.9	Train/Validation/Test split with typical proportions: 70% training, 15% validation, 15% test.	180
11.10	10-fold cross-validation: red=validation (10 each), purple=training (90 each).	180
11.11	Imbalanced dataset example motivating class weights or resampling.	182
11.12	Min-max scaling (purple) vs. standardization (red) schematic.	182
11.13	Simple drift monitor: KS statistic over time with an alert threshold. The purple line shows the Kolmogorov-Smirnov statistic measuring distribution shift between training and production data, while the red dashed line represents the alert threshold (0.2) that triggers when drift becomes significant.	185
12.1	Detector performance (mAP) vs. IoU threshold schematic.	195
12.2	U-Net style encoder-decoder with skip connections.	196
12.3	Transformer fine-tuning for text classification.	198
12.4	Encoder – decoder Transformer schematic with cross-attention.	199
12.5	Schematic spectrogram input for ASR.	203
13.1	Explained variance as a function of latent dimensionality (illustrative).	225
13.2	Schematic illustrating $\ell_1$ -induced sparsity geometry.	226
14.1	Undercomplete autoencoder architecture with bottleneck at latent layer.	232
20.1	GAN architecture showing the adversarial relationship between generator and discriminator.	282

---

# List of Tables

1.1 Major milestones in deep learning history . . . . .	4
5.1 Tree-based methods comparison. . . . .	84
5.2 Classical ML vs deep learning comparison. . . . .	89
5.3 Pros and cons comparison: classical ML vs deep learning. . . . .	90
10.1 Types of sequence models and their characteristics. . . . .	153
11.1 Priority order for hyperparameter tuning in deep learning. . . . .	177
14.1 Comparison of Autoencoders with other neural network algorithms. . . . .	236



---

# List of Algorithms

1	Gradient Boosting Algorithm . . . . .	84
2	Forward Propagation Algorithm . . . . .	96
3	Early stopping meta-algorithm . . . . .	114
4	Backpropagation Through Time (BPTT) . . . . .	154
5	Truncated Backpropagation Through Time (Truncated BPTT) . . . . .	156
6	GRU Forward Pass . . . . .	158
7	Variational Autoencoder Training Algorithm . . . . .	234
8	Diffusion Model Sampling Algorithm . . . . .	286



---

# Acknowledgements

I would like to express my deepest gratitude to the deep learning community for their invaluable contributions to this rapidly evolving field. Special thanks to the pioneers whose groundbreaking work laid the foundation for modern deep learning. I am grateful to my colleagues, students, and collaborators who have provided feedback, suggestions, and encouragement throughout the writing of this book. Your insights have been instrumental in shaping the content and presentation of this material.

This book draws inspiration from the excellent resources available in the deep learning community, including the seminal work by Goodfellow, Bengio, and Courville, as well as the emerging literature on understanding deep learning.

I would also like to thank my family for their unwavering support and patience during the many hours spent writing and revising this manuscript.

Finally, I am grateful to all readers who engage with this material and contribute to the advancement of deep learning through their research, applications, and education.

Any errors or omissions in this book are entirely my own responsibility.

Vu Hung Nguyen  
22nd October 2025



---

# Notation

This book uses the following notation throughout:

## General Notation

- $a, b, c$  — scalars (lowercase italic letters)
- $\mathbf{a}, \mathbf{b}, \mathbf{c}$  — vectors (bold lowercase letters)
- $\mathbf{A}, \mathbf{B}, \mathbf{C}$  — matrices (bold uppercase letters)
- $\mathcal{A}, \mathcal{B}, \mathcal{C}$  — sets (calligraphic uppercase)
- $a_i$  — the  $i$ -th element of vector  $\mathbf{a}$
- $A_{ij}$  or  $\mathbf{A}_{ij}$  — element at row  $i$ , column  $j$  of matrix  $\mathbf{A}$
- $\mathbf{A}^\top$  — transpose of matrix  $\mathbf{A}$
- $\mathbf{A}^{-1}$  — inverse of matrix  $\mathbf{A}$
- $\|\mathbf{x}\|$  — norm of vector  $\mathbf{x}$  (typically  $L^2$  norm)
- $\|\mathbf{x}\|_p$  —  $L^p$  norm of vector  $\mathbf{x}$
- $|x|$  — absolute value of scalar  $x$
- $\mathbb{R}$  — set of real numbers
- $\mathbb{R}^n$  —  $n$ -dimensional real vector space
- $\mathbb{R}^{m \times n}$  — set of real  $m \times n$  matrices

## Probability and Statistics

- $P(X)$  — probability distribution over discrete variable  $X$
- $p(x)$  — probability density function over continuous variable  $x$
- $P(X = x)$  or  $P(x)$  — probability that  $X$  takes value  $x$
- $P(X|Y)$  — conditional probability of  $X$  given  $Y$
- $\mathbb{E}_{x \sim P}[f(x)]$  — expectation of  $f(x)$  with respect to distribution  $P$
- $\text{Var}(X)$  — variance of random variable  $X$
- $\text{Cov}(X, Y)$  — covariance of random variables  $X$  and  $Y$
- $\mathcal{N}(\mu, \sigma^2)$  — Gaussian distribution with mean  $\mu$  and variance  $\sigma^2$

## Calculus and Optimization

- $\frac{dy}{dx}$  or  $\frac{\partial y}{\partial x}$  — derivative of  $y$  with respect to  $x$
- $\nabla_x f$  — gradient of function  $f$  with respect to  $x$
- $\nabla^2 f$  or  $H$  — Hessian matrix (matrix of second derivatives)
- $\arg \min_x f(x)$  — value of  $x$  that minimizes  $f(x)$
- $\arg \max_x f(x)$  — value of  $x$  that maximizes  $f(x)$

## Machine Learning

- $\mathcal{D}$  — dataset
- $\mathcal{D}_{\text{train}}$  — training dataset
- $\mathcal{D}_{\text{val}}$  — validation dataset
- $\mathcal{D}_{\text{test}}$  — test dataset
- $n$  — number of examples in dataset
- $m$  — mini-batch size
- $x$  — input vector or feature vector
- $y$  — target or label
- $\hat{y}$  — prediction or estimated output
- $\theta$  or  $w$  — parameters or weights
- $\mathcal{L}$  — loss function
- $J$  — cost function (sum or average of losses)
- $\alpha$  or  $\eta$  — learning rate
- $\lambda$  — regularization coefficient

## Neural Networks

- $L$  — number of layers in a neural network
- $n^{[l]}$  — number of units in layer  $l$
- $a^{[l]}$  — activations of layer  $l$
- $z^{[l]}$  — pre-activation values of layer  $l$
- $W^{[l]}$  — weight matrix for layer  $l$
- $b^{[l]}$  — bias vector for layer  $l$
- $f$  or  $\sigma$  — activation function
- $g$  — general function or transformation

## Difficulty Legend

- **Beginner** Intuition-first explanations; minimal prerequisites.
- ◆ **Intermediate** Assumes fundamentals; technical details and derivations.
- ★ **Advanced** Research-level topics or heavier mathematics.



# Chapter 1

## Introduction

This chapter introduces the fundamental concepts of deep learning and its historical context, providing a foundation for understanding the field.

### Learning Objectives

1. Deep learning concepts and differences from traditional machine learning
2. Historical development of neural networks and key breakthroughs
3. Major application domains where deep learning has achieved success
4. Key factors that enabled the rise of deep learning in the 21st century
5. Challenges and limitations of deep learning approaches

### 1.1 What is Deep Learning? ●

Deep learning is a subfield of machine learning that focuses on learning hierarchical representations of data through artificial neural networks with multiple layers. These networks, inspired by the structure and function of the human brain, have revolutionized numerous fields including computer vision, natural language processing, speech recognition, and many others.

#### Remark 1.1: What is "Deep" in Deep Learning?

The term "deep" refers to neural networks with multiple hidden layers (typically 3 or more), where each layer learns increasingly abstract representations from the previous layer's output. In contrast, "shallow" networks have only one or two hidden layers and are limited in their ability to learn complex hierarchical patterns, making deep networks essential for solving sophisticated problems that require understanding of multi-level features and relationships.

#### 1.1.1 The Rise of Deep Learning

The resurgence of neural networks, now known as deep learning, represents one of the most significant technological revolutions of the 21st century, transforming how we approach complex problems across virtually every domain of human endeavor. This remarkable comeback was not the result of a single breakthrough, but rather the convergence of multiple enabling factors that created the perfect storm for artificial intelligence advancement. The digital age has produced unprecedented amounts of data, providing the essential fuel needed to train increasingly complex models that can learn sophisticated patterns and representations. Simultaneously, the advent of Graphics Processing Units (GPUs) and specialized

hardware has enabled the training of networks with millions or even billions of parameters, making previously impossible computational tasks not only feasible but practical for widespread deployment.

### 1.1.2 Key Characteristics

Deep learning fundamentally differs from traditional machine learning approaches through its ability to automatically discover and learn complex hierarchical representations directly from raw data, eliminating the need for manual feature engineering that has long been a bottleneck in machine learning applications. Unlike traditional methods that require domain experts to carefully design and select relevant features, deep learning models automatically learn multiple levels of representation, progressing from simple low-level features like edges and textures in images to increasingly abstract high-level concepts like object categories and semantic relationships. This hierarchical learning process mirrors how the human brain processes information, enabling deep networks to capture intricate patterns and dependencies that would be extremely difficult or impossible to engineer manually. The end-to-end learning paradigm allows entire systems to be trained jointly rather than in separate stages, creating more coherent and optimized solutions that can adapt to the specific requirements of complex tasks. Perhaps most importantly, deep learning models demonstrate remarkable scalability, continuing to improve their performance as more data and computational resources become available, making them particularly well-suited for the data-rich environments of modern applications.

### 1.1.3 Applications

Deep learning has achieved remarkable success across an astonishingly diverse range of domains, fundamentally transforming how we approach complex problems and creating entirely new possibilities for human-computer interaction. In computer vision, deep learning has revolutionized image classification, object detection, semantic segmentation, facial recognition, and image generation, enabling applications from autonomous vehicles that can navigate complex environments to medical imaging systems that can detect diseases with superhuman accuracy. Natural language processing has been completely transformed by deep learning, with machine translation systems that can translate between languages in real-time, sentiment analysis tools that can understand emotional context in text, question-answering systems that can provide accurate responses to complex queries, and text generation models that can create coherent and contextually appropriate content. Speech and audio applications have reached new heights with speech recognition systems that can understand natural conversation, speaker identification technologies that can distinguish between individuals, music generation systems that can create original compositions, and audio synthesis tools that can produce realistic human speech. Healthcare has been particularly revolutionized by deep learning, with medical image analysis systems that can detect cancer and other diseases from X-rays and MRI scans, drug discovery platforms that can identify promising compounds, disease prediction models that can assess patient risk, and personalized medicine approaches that can tailor treatments to individual patients. Robotics applications have advanced dramatically with autonomous navigation systems that can operate in complex environments, manipulation systems that can perform delicate tasks, and decision-making algorithms that can adapt to changing conditions. Game playing has reached new frontiers with systems that have achieved superhuman performance in complex games like Go, Chess, and video games, demonstrating the power of deep learning to master strategic thinking and long-term planning. The impact of deep learning extends far beyond these applications, touching virtually every aspect of modern technology and scientific research, from climate modeling and financial analysis to space exploration and materials science.

## 1.2 Historical Context

The history of deep learning is intertwined with the broader history of artificial intelligence and neural networks. Understanding this context helps us appreciate the current state of the field and its future directions.

### 1.2.1 The Perceptron Era (1940s-1960s)

The foundations of neural networks were laid in the 1940s with the work of Warren McCulloch and Walter Pitts, who created a computational model of a neuron. In 1958, Frank Rosenblatt invented the Perceptron, an algorithm for learning a binary classifier.

The Perceptron showed promise but faced significant limitations. In 1969, Marvin Minsky and Seymour Papert's book *Perceptrons* demonstrated that single-layer perceptrons could not solve non-linearly separable problems like XOR, leading to the first "AI winter."

### 1.2.2 The Backpropagation Revolution (1980s)

The field was revitalized in the 1980s with the rediscovery and popularization of the backpropagation algorithm by David Rumelhart, Geoffrey Hinton, and Ronald Williams. This algorithm enabled the training of multi-layer networks, overcoming the limitations of single-layer perceptrons.

Key developments during this period include Convolutional Neural Networks (CNNs) by Yann LeCun, which revolutionized computer vision by introducing the concept of local connectivity and weight sharing, enabling networks to efficiently process spatial data like images while dramatically reducing the number of parameters compared to fully connected networks. Recurrent Neural Networks (RNNs) emerged as a powerful solution for sequential data processing, introducing the concept of memory and temporal dependencies that allowed networks to understand and generate sequences of data, from natural language text to time series predictions. Improved optimization techniques, including better initialization methods, adaptive learning rates, and more sophisticated gradient descent variants, made it possible to train deeper networks more effectively, addressing many of the convergence issues that had plagued earlier attempts at multi-layer neural networks.

### 1.2.3 The Second AI Winter (1990s-2000s)

Despite theoretical advances, neural networks fell out of favor in the 1990s due to several critical limitations that made them impractical for real-world applications. Limited computational resources severely constrained the size and complexity of networks that could be trained, making it impossible to leverage the full potential of multi-layer architectures that had been theoretically proven to be more powerful. The difficulty of training deep networks, particularly the vanishing gradient problem where gradients became exponentially smaller as they propagated backward through layers, made it nearly impossible to train networks with more than a few layers effectively. The success of alternative methods like Support Vector Machines (SVMs), which provided strong theoretical guarantees and often outperformed neural networks on benchmark datasets, further diminished interest in neural network research. The lack of large labeled datasets, which are essential for training complex models, meant that neural networks could not demonstrate their true potential, as they require substantial amounts of data to learn meaningful representations and avoid overfitting.

During this period, the term "deep learning" was coined to distinguish multi-layer neural networks from shallow architectures.

### 1.2.4 The Deep Learning Renaissance (2006-Present)

The modern era of deep learning began around 2006 with several breakthrough papers that fundamentally changed the landscape of artificial intelligence research and applications. Geoffrey Hinton and colleagues introduced Deep Belief Networks (DBNs) in 2006, demonstrating that deep networks could be trained using layer-wise pretraining, a technique that allowed networks to learn meaningful representations even when traditional backpropagation failed. Large-scale GPU computing for neural networks became practical in 2009, dramatically reducing training times and making it feasible to train networks with millions of parameters on consumer hardware. The watershed moment came in 2012 when AlexNet won the ImageNet competition by a large margin, demonstrating the power of deep CNNs trained on GPUs and sparking a renewed interest in neural networks across the entire machine learning community. The period from 2014 to 2016 saw sequence-to-sequence models and attention mechanisms revolutionize natural language processing, enabling breakthroughs in machine translation, text generation, and language understanding. Since 2017, Transformer architectures and large language models like GPT and BERT have achieved unprecedented performance across a wide range of tasks, fundamentally changing how we approach language processing and creating new possibilities for human-computer interaction that were previously unimaginable.

Year	Milestone
1943	McCulloch-Pitts neuron model
1958	Rosenblatt's Perceptron
1986	Backpropagation popularized
1989	LeCun's CNN for handwritten digits
1997	LSTM networks introduced
2006	Deep Belief Networks
2012	AlexNet wins ImageNet
2014	Generative Adversarial Networks (GANs)
2017	Transformer architecture
2018	BERT for NLP
2020	GPT-3 and large language models

Table 1.1: Major milestones in deep learning history

### 1.2.5 Key Milestones

This historical perspective shows that deep learning is built on decades of research, with periods of both enthusiasm and skepticism. The current success is the result of persistent research, technological advances, and the convergence of multiple enabling factors.

## 1.3 Fundamental Concepts

Before diving into the technical details, it is essential to understand several fundamental concepts that underpin deep learning.

### 1.3.1 Learning from Data

At its core, deep learning is about learning from data. Given a dataset  $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$ , where  $\mathbf{x}_i$  represents input features and  $y_i$  represents corresponding targets, the goal is to learn a function  $f : \mathcal{X} \rightarrow \mathcal{Y}$  that maps inputs to outputs.

#### Definition 1.1: Supervised Learning

In supervised learning, we have access to labeled examples where both inputs and desired outputs are known. The model learns to predict outputs for new, unseen inputs.

#### Definition 1.2: Unsupervised Learning

In unsupervised learning, we only have inputs without explicit labels. The model learns to discover patterns, structure, or representations in the data.

#### Definition 1.3: Reinforcement Learning

In reinforcement learning, an agent learns to make decisions by interacting with an environment and receiving rewards or penalties.

### 1.3.2 The Learning Process

The learning process in deep learning is a sophisticated orchestration of multiple components working together to transform raw data into meaningful predictions and insights. The process begins with model definition, where practitioners specify the architecture of the neural network, including the number of layers, types of layers, and activation functions, creating a computational graph that can learn complex mappings from inputs to outputs. A loss function  $\mathcal{L}(\hat{y}, y)$  is then defined to measure the discrepancy between predictions  $\hat{y}$  and true targets  $y$ , providing the mathematical foundation for learning by quantifying

how far the model's predictions are from the desired outcomes. The optimization phase uses sophisticated algorithms, typically gradient descent variants, to adjust the model parameters  $\theta$  to minimize the loss through iterative updates that gradually improve the model's performance. The final step involves evaluation on held-out test data to estimate generalization, ensuring that the model can perform well on new, unseen data rather than just memorizing the training examples. This entire process is guided by the fundamental principle that learning occurs through the minimization of prediction errors, with the model continuously adjusting its parameters to better capture the underlying patterns in the data.

### 1.3.3 Neural Networks as Universal Approximators

One of the remarkable properties of neural networks is their ability to approximate a wide range of functions.

#### Theorem 1.1: Universal Approximation Theorem (informal)

A neural network with a single hidden layer containing a sufficient number of neurons can approximate any continuous function on a compact subset of  $\mathbb{R}^n$  to arbitrary accuracy.

While this theorem provides theoretical justification for using neural networks, in practice, deep networks with multiple layers are often more efficient and effective than shallow but wide networks.

### 1.3.4 Representation Learning

A key advantage of deep learning is automatic feature learning, also known as representation learning, which fundamentally transforms how we approach complex problems by eliminating the need for manual feature engineering. Lower layers of deep networks learn simple, general features like edges, textures, and basic patterns that are common across many different types of data, creating a foundation of fundamental building blocks that can be combined in various ways. Middle layers combine these simple features into more complex patterns and structures, such as object parts, facial features, or grammatical constructs, creating intermediate representations that capture meaningful relationships between the basic features learned in earlier layers. Higher layers learn abstract, task-specific representations that are directly relevant to the particular problem being solved, such as object categories, semantic meanings, or high-level concepts that enable the network to make accurate predictions or generate appropriate outputs. This hierarchical feature learning process is what makes deep networks particularly powerful for complex tasks, as it mirrors the way humans process information by building up from simple features to increasingly complex and abstract representations that capture the essential characteristics of the data.

### 1.3.5 Generalization and Overfitting

A critical challenge in machine learning is ensuring that models generalize well to new data.

#### Definition 1.4: Overfitting

Overfitting occurs when a model learns the training data too well, including noise and spurious patterns, leading to poor performance on new data.

#### Definition 1.5: Underfitting

Underfitting occurs when a model is too simple to capture the underlying patterns in the data, resulting in poor performance on both training and test data.

The goal is to find the right balance between model complexity and generalization ability, often visualized by the bias-variance tradeoff:

$$\text{Expected Error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error} \quad (1.1)$$

Understanding these fundamental concepts provides a solid foundation for exploring the technical details of deep learning in subsequent chapters.

## 1.4 Structure of This Book ●

This book is organized into three main parts, each building upon the previous one to provide a comprehensive understanding of deep learning.

### Part I: Basic Math and Machine Learning Foundation

The first part establishes the mathematical and machine learning foundations necessary for understanding deep learning.

**Chapter 2: Linear Algebra** Covers vectors, matrices, and operations essential for understanding neural network computations.

**Chapter 3: Probability and Information Theory** Introduces probability distributions, expectation, information theory concepts, and their relevance to machine learning.

**Chapter 4: Numerical Computation** Discusses numerical optimization, gradient-based optimization, and computational considerations.

**Chapter 5: Classical Machine Learning Algorithms** Reviews traditional machine learning methods that provide context and motivation for deep learning approaches.

### Part II: Practical Deep Networks

The second part focuses on practical aspects of designing, training, and deploying deep neural networks.

**Chapter 6: Deep Feedforward Networks** Introduces the fundamental building blocks of deep learning, including multilayer perceptrons and activation functions.

**Chapter 7: Regularization for Deep Learning** Explores techniques to improve generalization and prevent overfitting.

**Chapter 8: Optimization for Training Deep Models** Covers modern optimization algorithms and training strategies.

**Chapter 9: Convolutional Networks** Details architectures specifically designed for processing grid-structured data like images.

**Chapter 10: Sequence Modeling** Examines recurrent and recursive networks for sequential and temporal data.

**Chapter 11: Practical Methodology** Provides guidelines for successfully applying deep learning to real-world problems.

**Chapter 12: Applications** Showcases deep learning applications across various domains.

### Part III: Deep Learning Research

The third part delves into advanced topics and current research directions.

**Chapter 13: Linear Factor Models** Introduces probabilistic models with linear structure.

**Chapter 14: Autoencoders** Explores unsupervised learning through reconstruction-based models.

**Chapter 15: Representation Learning** Discusses learning meaningful representations from data.

**Chapter 16: Structured Probabilistic Models** Covers graphical models and their integration with deep learning.

**Chapter 17: Monte Carlo Methods** Introduces sampling-based approaches for probabilistic inference.

**Chapter 18: Confronting the Partition Function** Addresses computational challenges in probabilistic models.

**Chapter 19: Approximate Inference** Explores methods for tractable inference in complex models.

**Chapter 20: Deep Generative Models** Examines modern approaches to generating new data samples.

## How to Use This Book

This book is designed to accommodate different learning paths and experience levels, recognizing that deep learning attracts learners from diverse backgrounds with varying goals and time constraints.

**For Beginners:** Start with Part I to build a strong mathematical and conceptual foundation, then proceed sequentially through Part II to develop practical skills and understanding of core deep learning architectures.

**For Practitioners:** With a solid mathematical background, you may choose to skip or skim Part I and focus directly on Parts II and III, using the foundational material as a reference when needed.

**For Researchers:** Part III provides advanced material relevant to current research directions in deep learning, including cutting-edge topics like generative models and advanced optimization techniques.

**Flexible Learning:** Each chapter is relatively self-contained, allowing you to focus on topics most relevant to your interests or professional needs, whether that's computer vision, natural language processing, or theoretical foundations.

Throughout the book, we balance theoretical rigor with practical insights, providing both mathematical foundations and intuitive explanations that help bridge the gap between abstract concepts and real-world applications. Code examples and exercises (when available) help reinforce concepts and develop practical skills, ensuring that readers can not only understand the theory but also implement and experiment with the ideas presented.

## 1.5 Prerequisites and Resources ●

To get the most out of this book, certain prerequisites are helpful, though not absolutely necessary. This section outlines the assumed background and provides resources for filling any gaps.

### 1.5.1 Mathematical Prerequisites

While we introduce key concepts in Part I, familiarity with fundamental mathematical topics will significantly enhance your understanding and ability to work with deep learning concepts effectively. Linear algebra forms the backbone of neural network computations, and familiarity with vectors, matrices, eigenvalues, and eigenvectors is essential for understanding how information flows through networks and how transformations are applied to data. Calculus knowledge, particularly derivatives, partial derivatives, chain rule, and basic optimization, is crucial for understanding how neural networks learn through gradient-based optimization, which is the fundamental mechanism by which these systems improve their performance. Probability theory provides the mathematical foundation for understanding uncertainty, randomness, and statistical patterns in data, while knowledge of random variables and common distributions helps in modeling the stochastic nature of learning processes and data generation. Statistics concepts like mean, variance, covariance, and basic statistical inference are essential for evaluating model performance, understanding the significance of results, and making informed decisions about model selection and validation. For readers needing to review these topics, we recommend "Linear Algebra Done Right" by Sheldon Axler for a rigorous treatment of linear algebra, "All of Statistics" by Larry Wasserman for comprehensive coverage of statistical concepts, and online resources like Khan Academy and MIT OpenCourseWare for interactive learning and additional perspectives on these fundamental topics.

### 1.5.2 Programming and Machine Learning Background

Basic programming knowledge is essential for implementing and experimenting with deep learning concepts, as the field is inherently practical and requires hands-on experience to fully understand the theoretical principles. Python programming skills, including understanding of basic syntax, data structures, and functions, form the foundation for working with deep learning frameworks and implementing custom solutions. Familiarity with NumPy is particularly valuable, as it provides the array operations and mathematical functions that are fundamental to neural network computations, and most deep learning frameworks build upon NumPy's efficient array processing capabilities. A general understanding of machine learning basics, including supervised learning concepts, training/test splits, and evaluation metrics, provides important context for understanding how deep learning fits into the broader machine learning landscape and helps in making informed decisions about model selection and validation strategies. Recommended resources for building these skills include "Python for Data Analysis" by Wes McKinney for comprehensive coverage of data manipulation and analysis techniques, "Hands-On Machine Learning" by Aurélien Géron for practical machine learning implementation, and scikit-learn documentation and tutorials for understanding traditional machine learning approaches that provide valuable context for deep learning methods.

### 1.5.3 Deep Learning Frameworks

While this book focuses on concepts rather than specific implementations, familiarity with a deep learning framework is valuable for gaining practical experience and understanding how theoretical concepts translate into working code. PyTorch has become particularly popular for research and prototyping due to its dynamic computational graph, intuitive API, and strong integration with the Python ecosystem, making it an excellent choice for those interested in cutting-edge research and experimental work. TensorFlow and Keras are widely used in industry for production systems, offering robust deployment capabilities, extensive tooling for model optimization and serving, and strong support for distributed training across multiple devices and machines. JAX is an emerging framework that combines the flexibility of NumPy with automatic differentiation and JIT compilation, making it particularly attractive for research in optimization and scientific computing. Hugging Face has emerged as a crucial platform for natural language processing, with their 'transformers' library being widely used for state-of-the-art models like BERT, GPT, and T5, providing easy access to pre-trained models and fine-tuning capabilities that have become essential for modern NLP applications. Official documentation and tutorials for these frameworks provide excellent hands-on learning opportunities, and we encourage readers to experiment with multiple frameworks to understand their relative strengths and use cases in different scenarios.

### 1.5.4 Additional Resources

To complement this book and deepen your understanding of deep learning, consider exploring a diverse range of resources that offer different perspectives and learning approaches. Classic textbooks provide comprehensive theoretical foundations, with "Deep Learning" by Goodfellow, Bengio, and Courville offering an authoritative treatment of the field's mathematical foundations and "Pattern Recognition and Machine Learning" by Christopher Bishop providing excellent coverage of probabilistic approaches to machine learning. Online courses offer structured learning experiences with hands-on projects, including Coursera's Deep Learning Specialization by Andrew Ng for a comprehensive introduction, Fast.ai's Practical Deep Learning for Coders for a more applied approach, and Stanford's CS231n and CS224n courses for computer vision and natural language processing respectively. Research papers are essential for staying current with the rapidly evolving field, with ArXiv.org providing access to the latest research, Papers with Code offering implementations of cutting-edge methods, and conference proceedings from NeurIPS, ICML, ICLR, and CVPR showcasing the most important advances in the field. Community resources offer valuable opportunities for discussion and learning, including Distill.pub for interactive explanations of complex concepts, Towards Data Science on Medium for accessible articles and tutorials, Reddit's r/MachineLearning for community discussions and Q&A, and Twitter/X for following leading researchers and staying updated with the latest developments in the field.

### 1.5.5 A Note on Exercises

Throughout this book, we include exercises at the end of chapters (when available) to help reinforce understanding and develop practical skills that are essential for mastering deep learning concepts. We encourage readers to work through exercises actively rather than just reading solutions, as the process of solving problems helps develop intuition and problem-solving skills that are crucial for applying deep learning in real-world scenarios. Implementing concepts in code is particularly valuable for deepening understanding, as it forces you to think through the details of how algorithms work and helps identify potential issues or misunderstandings that might not be apparent from theoretical study alone. Experimenting with variations to explore the behavior of models is an excellent way to develop intuition about how different parameters and architectures affect performance, and this hands-on experimentation often leads to insights that are difficult to gain through theoretical study alone. Collaborating with others and discussing concepts provides valuable opportunities for learning from different perspectives and can help clarify difficult concepts through explanation and discussion. Remember that deep learning is best learned through a combination of theoretical understanding and practical experience, and don't be discouraged if some concepts take time to fully grasp—this is normal and part of the learning process, as the field is inherently complex and requires time to develop the necessary intuition and skills.

### 1.5.6 Getting Help

If you encounter difficulties or have questions while working through this book, there are several strategies that can help you overcome challenges and deepen your understanding. Review the notation section and relevant earlier chapters to ensure you have a solid foundation in the prerequisite concepts, as many difficulties in deep learning arise from gaps in mathematical or conceptual understanding rather than from the complexity of the specific topic being studied. Consult the bibliography for additional perspectives on challenging topics, as different authors may explain concepts in ways that resonate better with your learning style or provide alternative approaches that clarify difficult points. Engage with online communities for discussions and Q&A, as the deep learning community is generally welcoming and helpful, and many practitioners have faced similar challenges and can offer valuable insights and solutions. Implement toy examples to build intuition about how concepts work in practice, as hands-on experimentation often reveals insights that are difficult to gain through theoretical study alone. Be patient with yourself, as deep learning is a rapidly evolving field with many subtleties and nuances that take time to master, and it's normal to struggle with complex concepts before they become clear. With these prerequisites and resources in mind, you are well-equipped to begin your deep learning journey, and we encourage you to approach the material with curiosity and persistence, knowing that the effort invested in understanding these concepts will pay dividends in your ability to work with and contribute to this exciting field.

## Key Takeaways

### Key Takeaways 1

- **Deep learning** is a subset of machine learning that uses neural networks with multiple layers to learn hierarchical representations.
- **Historical milestones** include the perceptron (1950s), backpropagation (1980s), and the deep learning renaissance (2010s) enabled by data, compute, and algorithms.
- **Key success factors** include large datasets, GPU acceleration, improved algorithms, and better understanding of training dynamics.
- **Major applications** span computer vision, natural language processing, speech recognition, game playing, and scientific discovery.
- **Limitations exist:** Deep learning requires substantial data and compute, can be brittle to distribution shifts, and lacks interpretability.

## Exercises

### Easy

*Exercise 1.1 (Historical Milestones).* List three key breakthroughs that enabled the rise of deep learning in the 21st century and explain their significance.

**Hint:**

*Consider computational advances, data availability, and algorithmic innovations.*

*Exercise 1.2 (Deep Learning vs Traditional ML).* Explain the main difference between deep learning and traditional machine learning approaches in terms of feature engineering.

**Hint:**

*Think about automatic feature learning versus manual feature extraction.*

*Exercise 1.3 (Application Domains).* Name three real-world domains where deep learning has achieved significant success and briefly describe one application in each domain.

**Hint:**

*Consider computer vision, natural language processing, and speech recognition.*

*Exercise 1.4 (Neural Network Basics).* What is the fundamental building block of a neural network, and how does it process information?

**Hint:**

*Think about the basic computational unit that receives inputs, applies weights, and produces an output.*

*Exercise 1.5 (Learning Process).* Explain in simple terms what happens during the training of a neural network.

**Hint:**

*Consider how the network adjusts its parameters based on examples and feedback.*

*Exercise 1.6 (Data Requirements).* Why do deep learning models typically require large amounts of data to perform well?

**Hint:**

*Think about the relationship between model complexity and the need for diverse examples.*

*Exercise 1.7 (Computational Power).* What type of hardware is most commonly used to train deep learning models, and why?

**Hint:**

*Consider the parallel processing capabilities needed for matrix operations.*

*Exercise 1.8 (Feature Learning).* How does deep learning differ from traditional programming in terms of how features are identified?

**Hint:**

*Compare manual feature engineering with automatic feature discovery.*

*Exercise 1.9 (Model Layers).* What does the term "deep" refer to in deep learning, and why is depth important?

**Hint:**

*Think about the hierarchical representation of information in multiple layers.*

*Exercise 1.10 (Training vs Inference).* Distinguish between the training phase and the inference phase of a deep learning model.

**Hint:**

*Consider when the model learns versus when it makes predictions.*

*Exercise 1.11 (Success Stories).* Name one famous deep learning achievement (such as AlphaGo, ImageNet, or GPT) and explain why it was significant.

**Hint:**

*Consider breakthroughs that demonstrated the potential of deep learning to the general public.*

*Exercise 1.12 (Exercise Types).* What types of problems are best suited for deep learning approaches?

**Hint:**

*Think about problems involving pattern recognition, complex relationships, or high-dimensional data.*

## Medium

*Exercise 1.13 (Enabling Factors).* Analyse how the availability of large datasets and computational resources (GPUs) together enabled the practical success of deep learning. Why wasn't one factor alone sufficient?

**Hint:**

*Consider the computational requirements of training deep networks and the need for diverse training examples.*

*Exercise 1.14 (Challenges and Limitations).* Identify two major challenges or limitations of current deep learning approaches and propose potential research directions to address them.

**Hint:**

*Think about interpretability, data efficiency, generalisation, or robustness to adversarial examples.*

*Exercise 1.15 (Historical Context).* Compare the "AI winter" periods with the current deep learning boom. What factors made the difference between failure and success?

**Hint:**

*Consider the role of computational resources, data availability, and algorithmic improvements.*

*Exercise 1.16 (Architecture Evolution).* Trace the evolution from perceptrons to modern deep neural networks. What were the key architectural innovations that enabled deeper networks?

**Hint:**

*Think about activation functions, weight initialisation, and training algorithms.*

*Exercise 1.17 (Data and Performance).* Explain the relationship between dataset size, model complexity, and performance in deep learning. Why does this relationship exist?

**Hint:**

*Consider the bias-variance trade-off and the curse of dimensionality.*

*Exercise 1.18 (Computational Scaling).* Analyse how the computational requirements of deep learning have evolved and what this means for accessibility and democratisation of AI.

**Hint:**

*Consider the costs of training large models and the implications for research and industry.*

*Exercise 1.19 (Generalisation Challenges).* Why do deep learning models sometimes fail to generalise well to new data, and what strategies can help improve generalisation?

**Hint:**

*Think about overfitting, domain shift, and regularisation techniques.*

## Hard

*Exercise 1.20 (Theoretical Foundations).* Critically evaluate the universal approximation theorem and its implications for deep learning. What are the practical limitations of this theoretical guarantee?

**Hint:**

*Consider the difference between theoretical possibility and practical feasibility, including training dynamics and optimisation challenges.*

*Exercise 1.21 (Emergent Properties).* Investigate how emergent properties arise in deep neural networks and their implications for AI safety and interpretability. Provide specific examples.

**Hint:**

*Consider phenomena like in-context learning, few-shot learning, and unexpected capabilities that emerge in large language models.*

*Exercise 1.22 (Scaling Laws and Limits).* Analyse the scaling laws in deep learning and discuss the fundamental limits they might reveal about the approach. What alternatives might be necessary?

**Hint:**

*Consider the relationship between model size, data size, and performance, and think about potential bottlenecks or diminishing returns.*

*Exercise 1.23 (Biological Inspiration).* Compare and contrast biological neural networks with artificial neural networks. What insights from neuroscience could improve deep learning architectures?

**Hint:**

*Consider sparsity, plasticity, energy efficiency, and the role of different types of neurons and connections in biological systems.*

*Exercise 1.24 (Ethical and Societal Implications).* Evaluate the broader societal implications of deep learning advances, including issues of bias, fairness, privacy, and the concentration of AI capabilities.

**Hint:**

*Consider both technical solutions and policy frameworks needed to address these challenges.*

*Exercise 1.25 (Future Research Directions).* Propose three novel research directions that could address current limitations of deep learning. Justify why these directions are promising and what challenges they might face.

**Hint:**

*Consider areas like neurosymbolic AI, quantum machine learning, or biologically-inspired architectures.*

*Exercise 1.26 (Mathematical Rigour).* Critically assess the mathematical foundations of deep learning. What are the key theoretical gaps that need to be addressed for a more rigorous understanding?

**Hint:**

*Consider optimisation theory, generalisation bounds, and the theoretical understanding of why deep networks work so well in practice.*



---

## **Part I**

# **Basic Math and Machine Learning Foundation**



# Chapter 2

## Linear Algebra

This chapter covers the linear algebra foundations essential for understanding deep learning algorithms. Topics include vectors, matrices, eigenvalues, and linear transformations.

### Learning Objectives

1. Vector and matrix operations including addition, multiplication, and transposition
2. Linear transformations and their representation in neural networks
3. Eigenvalues and eigenvectors in optimization and data analysis
4. Matrix decompositions: eigendecomposition and singular value decomposition (SVD)
5. Norms and distances in vector spaces for optimization and loss functions
6. Linear systems and conditions for unique solutions

## 2.1 Scalars, Vectors, Matrices, and Tensors ●

Linear algebra provides the mathematical framework for understanding and implementing deep learning algorithms. We begin with the basic objects that form the foundation of this framework.

### 2.1.1 Scalars

A *scalar* is a single number, in contrast to objects that contain multiple numbers. We typically denote scalars with lowercase italic letters, such as  $a$ ,  $n$ , or  $x$ .

#### Example 2.1

The learning rate  $\alpha = 0.01$  is a scalar. The number of training examples  $n = 1000$  is also a scalar.

In deep learning, scalars are often real numbers ( $a \in \mathbb{R}$ ), but they can also be integers, complex numbers, or elements of other fields depending on the context.

### 2.1.2 Vectors

A *vector* is an array of numbers arranged in order. We identify each individual number in the vector by its position in the ordering. We denote vectors with bold lowercase letters, such as  $\mathbf{x}$ ,  $\mathbf{y}$ , or  $\mathbf{w}$ .

**Definition 2.1: Vector**

A vector  $\mathbf{x} \in \mathbb{R}^n$  is an ordered collection of  $n$  real numbers:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad (2.1)$$

where  $x_i$  denotes the  $i$ -th element of  $\mathbf{x}$ .

**Example 2.2**

A feature vector for a house might be:

$$\mathbf{x} = \begin{bmatrix} 2000 \\ 3 \\ 2 \\ 50 \end{bmatrix} \quad (2.2)$$

representing square footage, number of bedrooms, number of bathrooms, and age in years.

**2.1.3 Matrices**

A *matrix* is a 2-D array of numbers, where each element is identified by two indices. We denote matrices with bold uppercase letters such as  $\mathbf{A}$ ,  $\mathbf{W}$ , or  $\mathbf{X}$ .

**Definition 2.2: Matrix**

A matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$  is a rectangular array of real numbers with  $m$  rows and  $n$  columns:

$$\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{bmatrix} \quad (2.3)$$

where  $A_{ij}$  denotes the element at row  $i$  and column  $j$ .

**Example 2.3**

A matrix of training examples where each row is a feature vector:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{bmatrix} \quad (2.4)$$

Here,  $\mathbf{X} \in \mathbb{R}^{3 \times 3}$  contains 3 examples with 3 features each.

**2.1.4 Tensors**

A *tensor* is an array with more than two axes. While scalars are 0-D tensors, vectors are 1-D tensors, and matrices are 2-D tensors, we typically reserve the term “tensor” for arrays with three or more dimensions.

**Definition 2.3: Tensor**

A tensor  $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_k}$  is a  $k$ -dimensional array where elements are identified by  $k$  indices:  $\mathcal{A}_{i_1, i_2, \dots, i_k}$ .

**Example 2.4**

A batch of color images can be represented as a 4-D tensor:

$$\mathcal{X} \in \mathbb{R}^{B \times H \times W \times C} \quad (2.5)$$

where  $B$  is the batch size,  $H$  and  $W$  are height and width, and  $C$  is the number of color channels (e.g., 3 for RGB).

**Remark 2.1: Tensors in Programming Languages**

In Python, numbers are stored as tensors. Even simple scalars like  $x = 5$  are internally represented as 0-dimensional tensors, while arrays like  $[1, 2, 3]$  are 1-dimensional tensors. This unified representation allows for consistent mathematical operations across different data types and dimensions.

**Remark 2.2: Tensors in Deep Learning Frameworks**

With TensorFlow and PyTorch, tensors are the main data structure for deep learning. These frameworks treat all data as tensors, from simple scalars to complex multi-dimensional arrays, enabling seamless computation across CPUs, GPUs, and specialized hardware. Tensors in these frameworks support automatic differentiation, making gradient computation for backpropagation straightforward and efficient.

**Remark 2.3: Tensor Processing Units (TPUs)**

Google's Tensor Processing Unit (TPU) is specifically designed to accelerate tensor operations, which are the core computations in deep learning. TPUs use a systolic array architecture optimized for matrix multiplication and tensor contractions, the fundamental operations in neural networks. The name "Tensor" in TPU reflects the unit's specialization for high-dimensional array operations, making it particularly effective for training large neural networks with massive tensor computations.

## 2.1.5 Notation Conventions

Throughout this book, we adopt consistent notation conventions that help distinguish between different types of mathematical objects and make the mathematical formulations of deep learning algorithms more readable and intuitive. Scalars are represented using lowercase italic letters such as  $a$ ,  $b$ , and  $x$ , which helps identify them as single numerical values that can represent parameters like learning rates, biases, or individual elements. Vectors are denoted using bold lowercase letters like  $\mathbf{a}$ ,  $\mathbf{x}$ , and  $\mathbf{w}$ , making it clear that these represent ordered collections of numbers that can represent feature vectors, weight vectors, or gradients in neural networks. Matrices are represented using bold uppercase letters such as  $\mathbf{A}$ ,  $\mathbf{X}$ , and  $\mathbf{W}$ , clearly indicating that these are two-dimensional arrays that can represent weight matrices, data matrices, or transformation matrices in neural network layers. Tensors are denoted using calligraphic uppercase letters like  $\mathcal{A}$  and  $\mathcal{X}$ , distinguishing these higher-dimensional arrays that can represent batches of data, multi-channel images, or complex data structures in deep learning applications. Understanding these fundamental objects and their properties is essential for working with the mathematical formulations of deep learning algorithms, as they form the building blocks upon which all neural network computations are constructed.

## 2.2 Matrix Operations ●

Matrix operations form the computational backbone of neural networks. Understanding these operations is crucial for implementing and analyzing deep learning algorithms.

### 2.2.1 Matrix Addition and Scalar Multiplication

Matrices of the same dimensions can be added element-wise:

#### Definition 2.4: Matrix Addition

Given  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$ , their sum  $\mathbf{C} = \mathbf{A} + \mathbf{B}$  is defined as:

$$C_{ij} = A_{ij} + B_{ij} \quad (2.6)$$

for all  $i = 1, \dots, m$  and  $j = 1, \dots, n$ .

#### Definition 2.5: Scalar Multiplication

Given a scalar  $\alpha \in \mathbb{R}$  and a matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$ , the product  $\mathbf{B} = \alpha \mathbf{A}$  is:

$$B_{ij} = \alpha A_{ij} \quad (2.7)$$

### 2.2.2 Matrix Transpose

The transpose is a fundamental operation that exchanges rows and columns.

#### Definition 2.6: Transpose

The transpose of a matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$  is a matrix  $\mathbf{A}^\top \in \mathbb{R}^{n \times m}$  where:

$$(\mathbf{A}^\top)_{ij} = A_{ji} \quad (2.8)$$

Properties of transpose:

$$(\mathbf{A}^\top)^\top = \mathbf{A} \quad (2.9)$$

$$(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top \quad (2.10)$$

$$(\alpha \mathbf{A})^\top = \alpha \mathbf{A}^\top \quad (2.11)$$

$$(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top \quad (2.12)$$

### 2.2.3 Matrix Multiplication

Matrix multiplication is central to neural network computations.

#### Definition 2.7: Matrix Multiplication

Given  $\mathbf{A} \in \mathbb{R}^{m \times n}$  and  $\mathbf{B} \in \mathbb{R}^{n \times p}$ , their product  $\mathbf{C} = \mathbf{AB} \in \mathbb{R}^{m \times p}$  is defined as:

$$C_{ij} = \sum_{k=1}^n A_{ik} B_{kj} \quad (2.13)$$

**Example 2.5**

Consider the multiplication:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 19 & 22 \\ 43 & 50 \end{bmatrix} \quad (2.14)$$

where  $C_{11} = 1 \cdot 5 + 2 \cdot 7 = 19$ .

Matrix multiplication exhibits several important properties that are fundamental to understanding neural network computations. The associative property  $(AB)C = A(BC)$  allows us to group matrix multiplications in different ways without changing the result, which is crucial for optimizing the order of operations in deep neural networks and enables efficient computation strategies. The distributive property  $A(B + C) = AB + AC$  allows us to break down complex matrix operations into simpler components, which is particularly useful in gradient computation and backpropagation algorithms where we need to compute derivatives of matrix expressions. Unlike scalar multiplication, matrix multiplication is generally not commutative, meaning that  $AB \neq BA$  in most cases, which has important implications for the order of operations in neural networks and explains why the order of matrix multiplications in forward and backward passes must be carefully considered.

### 2.2.4 Element-wise (Hadamard) Product

The element-wise product is denoted by  $\odot$  and operates on corresponding elements.

**Definition 2.8: Hadamard Product**

Given  $A, B \in \mathbb{R}^{m \times n}$ , the Hadamard product  $C = A \odot B$  is:

$$C_{ij} = A_{ij}B_{ij} \quad (2.15)$$

This operation is common in neural networks, particularly in activation functions and gating mechanisms.

### 2.2.5 Matrix-Vector Products

When multiplying a matrix by a vector, we can view it as a special case of matrix multiplication:

$$Ax = b \quad (2.16)$$

where  $A \in \mathbb{R}^{m \times n}$ ,  $x \in \mathbb{R}^n$ , and  $b \in \mathbb{R}^m$ .

This operation is fundamental in neural networks, where it represents the linear transformation:

$$b_i = \sum_{j=1}^n A_{ij}x_j \quad (2.17)$$

**Example 2.6**

Let's compute a matrix-vector product using the numbers 11 and 16:

$$\begin{bmatrix} 11 & 3 \\ 2 & 16 \end{bmatrix} \begin{bmatrix} 4 \\ 5 \end{bmatrix} = \begin{bmatrix} ? \\ ? \end{bmatrix} \quad (2.18)$$

Computing each element:

$$b_1 = A_{11}x_1 + A_{12}x_2 = 11 \cdot 4 + 3 \cdot 5 = 44 + 15 = 59 \quad (2.19)$$

$$b_2 = A_{21}x_1 + A_{22}x_2 = 2 \cdot 4 + 16 \cdot 5 = 8 + 80 = 88 \quad (2.20)$$

So the result is:

$$\begin{bmatrix} 11 & 3 \\ 2 & 16 \end{bmatrix} \begin{bmatrix} 4 \\ 5 \end{bmatrix} = \begin{bmatrix} 59 \\ 88 \end{bmatrix} \quad (2.21)$$

**2.2.6 Dot Product**

The dot product (or inner product) of two vectors is a special case of matrix multiplication:

**Definition 2.9: Dot Product**

For vectors  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ , their dot product is:

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^\top \mathbf{y} = \sum_{i=1}^n x_i y_i \quad (2.22)$$

The dot product has geometric interpretation:

$$\mathbf{x} \cdot \mathbf{y} = \|\mathbf{x}\| \|\mathbf{y}\| \cos \theta \quad (2.23)$$

where  $\theta$  is the angle between the vectors.

**Cosine Similarity**

From the geometric interpretation, we can derive the cosine similarity, which measures the angle between two vectors regardless of their magnitude:

$$\cos \theta = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} \quad (2.24)$$

Cosine similarity ranges from -1 to 1:

- $\cos \theta = 1$ : Vectors point in the same direction (perfectly similar)
- $\cos \theta = 0$ : Vectors are perpendicular (no similarity)
- $\cos \theta = -1$ : Vectors point in opposite directions (perfectly dissimilar)

Cosine similarity is particularly useful in machine learning for measuring similarity between feature vectors, as it focuses on the direction rather than the magnitude of the vectors. This makes it robust to differences in scale between features.

**2.2.7 Computational Complexity**

Understanding computational costs is crucial for efficient implementation of deep learning algorithms, as the computational complexity of matrix operations directly impacts training time, memory requirements, and the feasibility of deploying models in production environments. Matrix-matrix multiplication between  $\mathbf{A} \in \mathbb{R}^{m \times n}$  and  $\mathbf{B} \in \mathbb{R}^{n \times p}$  requires  $O(mnp)$  operations,

making it the most computationally expensive operation in neural networks and the primary bottleneck in training large models. Matrix-vector multiplication requires  $O(mn)$  operations, which is significantly more efficient than full matrix multiplication and is commonly used in forward passes through neural network layers. Element-wise operations such as activation functions and Hadamard products require  $O(mn)$  operations, making them relatively inexpensive compared to matrix multiplications but still important for overall computational efficiency. These operations can be efficiently parallelized on modern hardware such as GPUs and TPUs, which is one of the key reasons why deep learning has become practical for large-scale applications, as the parallel nature of matrix operations maps well to the parallel processing capabilities of specialized hardware.

#### Remark 2.4: Matrix Operations on Different Hardware

Matrix operations are handled very differently across CPU, GPU, and TPU architectures, each offering distinct advantages for deep learning workloads. CPUs process matrix operations sequentially using general-purpose cores, making them suitable for small-scale computations and development, but they struggle with the massive parallelization requirements of large neural networks. GPUs excel at matrix operations through their thousands of small, efficient cores designed for parallel processing, providing 10-100x speedups over CPUs for matrix multiplications and enabling the training of deep networks that would be impractical on CPUs alone. TPUs take this specialization even further, using systolic array architectures specifically optimized for tensor contractions and matrix multiplications, delivering superior performance for large-scale training with lower power consumption than GPUs. The choice between these platforms depends on the scale of computation: CPUs for prototyping and small models, GPUs for most deep learning applications, and TPUs for training the largest models where computational efficiency and power consumption are critical factors.

#### Remark 2.5: Real-World Scale: Phi-4-Mini Example

The largest matrix dimension used in the calculation for Phi-4-Mini (a very small LLM, with 3.8B params) is 3,072 x 12,288. This is a huge matrix and you'll understand why Matrix Operations is important when it comes to DL/LLM. Even for what is considered a "small" language model by today's standards, the matrix operations involve tens of millions of parameters, demonstrating the critical importance of efficient matrix computation algorithms and specialized hardware for modern deep learning applications.

## 2.3 Identity and Inverse Matrices ●

Special matrices play important roles in linear algebra and deep learning. The identity matrix and matrix inverses are among the most fundamental.

### 2.3.1 Identity Matrix

The identity matrix is the matrix analog of the number 1.

**Definition 2.10: Identity Matrix**

The identity matrix  $\mathbf{I}_n \in \mathbb{R}^{n \times n}$  is a square matrix with ones on the diagonal and zeros elsewhere:

$$\mathbf{I}_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \quad (2.25)$$

Formally,  $(\mathbf{I}_n)_{ij} = \delta_{ij}$  where  $\delta_{ij}$  is the Kronecker delta:

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \quad (2.26)$$

The key property of the identity matrix is:

$$\mathbf{I}_n \mathbf{A} = \mathbf{A} \mathbf{I}_n = \mathbf{A} \quad (2.27)$$

for any matrix  $\mathbf{A} \in \mathbb{R}^{n \times n}$ .

**2.3.2 Matrix Inverse**

The inverse of a matrix, when it exists, allows us to solve systems of linear equations.

**Definition 2.11: Matrix Inverse**

A square matrix  $\mathbf{A} \in \mathbb{R}^{n \times n}$  is *invertible* (or *non-singular*) if there exists a matrix  $\mathbf{A}^{-1} \in \mathbb{R}^{n \times n}$  such that:

$$\mathbf{A}^{-1} \mathbf{A} = \mathbf{A} \mathbf{A}^{-1} = \mathbf{I}_n \quad (2.28)$$

**Example 2.7**

The matrix  $\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$  has inverse:

$$\mathbf{A}^{-1} = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \quad (2.29)$$

We can verify:  $\mathbf{A} \mathbf{A}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathbf{I}_2$ .

**2.3.3 Properties of Inverses**

If  $\mathbf{A}$  and  $\mathbf{B}$  are invertible, then:

$$(\mathbf{A}^{-1})^{-1} = \mathbf{A} \quad (2.30)$$

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1} \mathbf{A}^{-1} \quad (2.31)$$

$$(\mathbf{A}^\top)^{-1} = (\mathbf{A}^{-1})^\top \quad (2.32)$$

### 2.3.4 Solving Linear Systems

The inverse allows us to solve systems of linear equations. Given  $\mathbf{A}\mathbf{x} = \mathbf{b}$ , if  $\mathbf{A}$  is invertible:

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \quad (2.33)$$

However, computing inverses is expensive ( $O(n^3)$  for dense matrices) and numerically unstable. In practice, we often use more efficient methods like LU decomposition or iterative solvers.

### 2.3.5 Conditions for Invertibility

A matrix  $\mathbf{A} \in \mathbb{R}^{n \times n}$  is invertible if and only if it satisfies several equivalent conditions that are fundamental to understanding when linear transformations can be reversed. The determinant condition  $\det(\mathbf{A}) \neq 0$  provides a direct computational test for invertibility, as the determinant captures the volume scaling factor of the linear transformation and a zero determinant indicates that the transformation collapses the space to a lower dimension. The linear independence condition requires that both the columns and rows of the matrix are linearly independent, meaning that no column or row can be expressed as a linear combination of the others, which ensures that the transformation preserves the full dimensionality of the space. The rank condition  $\text{rank}(\mathbf{A}) = n$  ensures that the matrix has full rank, meaning that all  $n$  dimensions of the input space are preserved in the output space, which is necessary for the transformation to be reversible. The null space condition requires that the null space contains only the zero vector, meaning that the only vector that maps to zero is the zero vector itself, which ensures that the transformation is one-to-one and therefore invertible.

### 2.3.6 Singular Matrices

Matrices that are not invertible are called *singular* or *degenerate*.

#### Example 2.8

The matrix  $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$  is singular because its rows are linearly dependent (the second row is twice the first). Its determinant is  $\det(\mathbf{A}) = 4 - 4 = 0$ .

Singular matrices arise in deep learning when several problematic conditions occur that can significantly impact model performance and training stability. Features that are perfectly correlated create linear dependencies in the data matrix, leading to singular matrices that cannot be inverted and causing numerical instability in optimization algorithms. Overparameterized models with more parameters than training examples can lead to singular weight matrices, particularly in the early layers where the model has not yet learned meaningful representations and the weight matrices may have insufficient rank. Numerical precision issues occur when the computational precision of floating-point arithmetic is insufficient to distinguish between nearly identical values, causing matrices that should be invertible to become singular due to rounding errors, which is particularly problematic in deep networks where small errors can accumulate and propagate through multiple layers.

### 2.3.7 Pseudo-inverse

For non-square or singular matrices, we can use the Moore-Penrose pseudo-inverse  $\mathbf{A}^+$ , which provides a generalized notion of inversion. The pseudo-inverse is particularly useful in least squares problems and is discussed further in later chapters.

### 2.3.8 Practical Considerations

In deep learning implementations, several practical considerations are essential for maintaining numerical stability and computational efficiency when working with matrix inverses and related operations. Avoiding explicit computation of matrix inverses whenever possible is crucial, as direct inversion is computationally expensive with  $O(n^3)$  complexity and can be numerically unstable, particularly for ill-conditioned matrices that are common in deep learning applications. Using numerically

stable algorithms such as QR decomposition and singular value decomposition (SVD) provides more robust alternatives to direct inversion, as these methods are designed to handle numerical precision issues and can provide meaningful results even when matrices are nearly singular. Adding regularization terms such as  $(\mathbf{A}^\top \mathbf{A} + \lambda \mathbf{I})^{-1}$  helps ensure invertibility by adding a small positive value to the diagonal elements, which prevents singular matrices and improves numerical stability while maintaining the mathematical properties of the original problem. Leveraging optimized linear algebra libraries such as BLAS, LAPACK, and cuBLAS is essential for achieving high performance in deep learning applications, as these libraries are highly optimized for specific hardware architectures and can provide significant speedups over naive implementations, particularly when working with large matrices on GPU hardware.

## 2.4 Linear Dependence and Span

Understanding linear independence and span is crucial for analyzing the capacity and expressiveness of neural networks.

### 2.4.1 Linear Combinations

#### Definition 2.12: Linear Combination

A *linear combination* of vectors  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$  is any vector of the form:

$$\mathbf{v} = a_1 \mathbf{v}_1 + a_2 \mathbf{v}_2 + \cdots + a_n \mathbf{v}_n \quad (2.34)$$

where  $a_1, a_2, \dots, a_n$  are scalars called *coefficients*.

#### Example 2.9

If  $\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$  and  $\mathbf{v}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ , then any vector in  $\mathbb{R}^2$  can be written as a linear combination:

$$\begin{bmatrix} x \\ y \end{bmatrix} = x \mathbf{v}_1 + y \mathbf{v}_2 \quad (2.35)$$

### 2.4.2 Span

#### Definition 2.13: Span

The *span* of a set of vectors  $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$  is the set of all possible linear combinations of these vectors:

$$\text{span}(\{\mathbf{v}_1, \dots, \mathbf{v}_n\}) = \left\{ \sum_{i=1}^n a_i \mathbf{v}_i \mid a_i \in \mathbb{R} \right\} \quad (2.36)$$

The span defines all vectors that can be reached by scaling and adding the given vectors.

**Example 2.10**

In  $\mathbb{R}^3$ , the span of  $\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$  and  $\mathbf{v}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$  is the  $xy$ -plane:

$$\text{span}(\{\mathbf{v}_1, \mathbf{v}_2\}) = \left\{ \begin{bmatrix} x \\ y \\ 0 \end{bmatrix} \mid x, y \in \mathbb{R} \right\} \quad (2.37)$$

**2.4.3 Linear Independence****Definition 2.14: Linear Independence**

A set of vectors  $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$  is *linearly independent* if no vector can be written as a linear combination of the others. Formally, the only solution to:

$$a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_n = \mathbf{0} \quad (2.38)$$

is  $a_1 = a_2 = \cdots = a_n = 0$ .

If a set of vectors is not linearly independent, it is *linearly dependent*.

**Example 2.11: Linear Dependence**

The vectors  $\mathbf{v}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ ,  $\mathbf{v}_2 = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$  are linearly dependent because  $\mathbf{v}_2 = 2\mathbf{v}_1$ .

**Example 2.12: Linear Independence**

The standard basis vectors  $\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$  and  $\mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$  are linearly independent.

**2.4.4 Basis****Definition 2.15: Basis**

A *basis* for a vector space  $V$  is a set of linearly independent vectors that span  $V$ . Every vector in  $V$  can be uniquely expressed as a linear combination of basis vectors.

**Example 2.13: Standard Basis**

The standard basis for  $\mathbb{R}^3$  is:

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad (2.39)$$

## 2.4.5 Dimension and Rank

### Definition 2.16: Dimension

The *dimension* of a vector space is the number of vectors in any basis for that space. We write  $\dim(V)$  for the dimension of space  $V$ .

### Definition 2.17: Rank

The *rank* of a matrix  $A$  is the dimension of the space spanned by its columns (column rank) or rows (row rank). For any matrix, column rank equals row rank, so we simply refer to “the rank.”

Properties of rank provide important insights into the behavior of matrices and their applications in deep learning. The inequality  $\text{rank}(A) \leq \min(m, n)$  for  $A \in \mathbb{R}^{m \times n}$  establishes that the rank of a matrix cannot exceed the smaller of its dimensions, which means that a matrix can have at most as many linearly independent columns as it has rows, and vice versa. The rank inequality  $\text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B))$  shows that matrix multiplication cannot increase the rank, which has important implications for the expressiveness of neural network layers, as the rank of the output is limited by the rank of the weight matrices. The invertibility condition states that  $A$  is invertible if and only if  $\text{rank}(A) = n$  (full rank), which connects the concept of rank directly to the invertibility conditions we discussed earlier, as a matrix must have full rank to be invertible, meaning that all its columns must be linearly independent.

### Example 2.14

Calculating Rank of a 2x2 Matrix Consider the matrix  $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ . To calculate its rank, we need to determine how many of its columns are linearly independent. The first column is  $\begin{bmatrix} 1 \\ 3 \end{bmatrix}$  and the second column is  $\begin{bmatrix} 2 \\ 4 \end{bmatrix}$ . These columns are linearly independent because there is no scalar  $c$  such that  $\begin{bmatrix} 2 \\ 4 \end{bmatrix} = c \begin{bmatrix} 1 \\ 3 \end{bmatrix}$  (this would require  $2 = c \cdot 1$  and  $4 = c \cdot 3$ , which gives  $c = 2$  and  $c = 4/3$ , a contradiction). Since both columns are linearly independent, the rank of this matrix is 2, which is the maximum possible rank for a  $2 \times 2$  matrix, making it full rank and therefore invertible.

## 2.4.6 Column Space and Null Space

### Definition 2.18: Column Space

The *column space* (or *range*) of a matrix  $A \in \mathbb{R}^{m \times n}$  is the span of its columns:

$$\text{Col}(A) = \{Ax \mid x \in \mathbb{R}^n\} \quad (2.40)$$

The dimension of the column space is the rank of  $A$ .

### Definition 2.19: Null Space

The *null space* (or *kernel*) of  $A$  is the set of all vectors that map to zero:

$$\text{Null}(A) = \{x \in \mathbb{R}^n \mid Ax = \mathbf{0}\} \quad (2.41)$$

### 2.4.7 Relevance to Deep Learning

These concepts are fundamental to understanding the capacity, expressiveness, and limitations of neural networks, providing crucial insights into how different architectural choices affect the model's ability to learn and represent complex patterns in data. The expressiveness of a neural network layer depends critically on the rank of its weight matrix, as the rank determines the dimensionality of the space that the layer can map inputs to, with higher rank matrices providing more expressive power but also requiring more parameters and computational resources. Linear dependence in features indicates redundant information that can lead to overfitting and poor generalization, as the model may learn to rely on correlated features rather than discovering the underlying patterns that generalize to new data. Dimensionality reduction methods like PCA seek low-dimensional representations by identifying the directions of maximum variance in the data, which can help reduce overfitting and improve computational efficiency while preserving the most important information. Understanding which transformations are possible with given architectures is essential for network design, as the rank and structure of weight matrices determine what kinds of mappings the network can learn, which directly impacts the model's ability to solve specific tasks and the efficiency of the learning process.

## 2.5 Norms ●

Norms are functions that measure the size or length of vectors. They are essential for regularization, optimization, and measuring distances in deep learning.

### 2.5.1 Definition of a Norm

A norm is a mathematical function that provides a consistent and intuitive way to measure the size or length of vectors, serving as the foundation for many important concepts in deep learning including regularization, optimization, and distance metrics. The non-negativity property  $f(\mathbf{x}) \geq 0$  with equality if and only if  $\mathbf{x} = \mathbf{0}$  ensures that norms always return non-negative values and that only the zero vector has zero norm, which makes intuitive sense as the zero vector has no magnitude. The homogeneity property  $f(\alpha\mathbf{x}) = |\alpha|f(\mathbf{x})$  ensures that scaling a vector by a scalar  $\alpha$  scales its norm by the absolute value of that scalar, which means that the norm scales proportionally with the vector's magnitude, preserving the relative size relationships between vectors. The triangle inequality  $f(\mathbf{x} + \mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$  ensures that the norm of the sum of two vectors is never greater than the sum of their individual norms, which captures the intuitive notion that the shortest path between two points is a straight line and prevents the norm from behaving in counterintuitive ways. We typically denote norms using the notation  $\|\mathbf{x}\|$ , which provides a concise and standardized way to refer to the magnitude of vectors in mathematical expressions and algorithms.

### 2.5.2 $L^p$ Norms

The most common family of norms are the  $L^p$  norms.

#### Definition 2.20: $L^p$ Norm

For  $p \geq 1$ , the  $L^p$  norm of a vector  $\mathbf{x} \in \mathbb{R}^n$  is:

$$\|\mathbf{x}\|_p = \left( \sum_{i=1}^n |x_i|^p \right)^{1/p} \quad (2.42)$$

### 2.5.3 Common Norms

#### $L^1$ Norm (Manhattan Distance)

The  $L^1$  norm is the sum of absolute values:

$$\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i| \quad (2.43)$$

**Example 2.15**

For  $\mathbf{x} = \begin{bmatrix} 3 \\ -4 \\ 2 \end{bmatrix}$ , we have  $\|\mathbf{x}\|_1 = 3 + 4 + 2 = 9$ .

The  $L^1$  norm is particularly useful in applications where sparsity is desired, as it encourages solutions with many zero components, making it ideal for feature selection and model compression in deep learning applications.

 **$L^2$  Norm (Euclidean Distance)**

The  $L^2$  norm is the most common norm, corresponding to Euclidean distance:

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2} = \sqrt{\mathbf{x}^\top \mathbf{x}} \quad (2.44)$$

**Example 2.16**

For  $\mathbf{x} = \begin{bmatrix} 3 \\ -4 \\ 2 \end{bmatrix}$ , we have  $\|\mathbf{x}\|_2 = \sqrt{9 + 16 + 4} = \sqrt{29} \approx 5.39$ .

The  $L^2$  norm is the most widely used norm in deep learning, serving as the foundation for many important algorithms and techniques including ridge regularization, gradient descent optimization, and distance-based similarity measures.

The squared  $L^2$  norm is often used in optimization because it has simpler derivatives:

$$\|\mathbf{x}\|_2^2 = \mathbf{x}^\top \mathbf{x} = \sum_{i=1}^n x_i^2 \quad (2.45)$$

 **$L^\infty$  Norm (Maximum Norm)**

The  $L^\infty$  norm is defined as:

$$\|\mathbf{x}\|_\infty = \max_i |x_i| \quad (2.46)$$

This can be viewed as the limit of  $L^p$  norms as  $p \rightarrow \infty$ .

**Example 2.17**

For  $\mathbf{x} = \begin{bmatrix} 3 \\ -4 \\ 2 \end{bmatrix}$ , we have  $\|\mathbf{x}\|_\infty = \max(3, 4, 2) = 4$ .

**2.5.4 Frobenius Norm**

For matrices, the Frobenius norm is analogous to the  $L^2$  norm for vectors.

**Definition 2.21: Frobenius Norm**

For a matrix  $A \in \mathbb{R}^{m \times n}$ :

$$\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n A_{ij}^2} = \sqrt{\text{trace}(A^\top A)} \quad (2.47)$$

The Frobenius norm is used for regularizing weight matrices in neural networks.

### 2.5.5 Unit Vectors and Normalization

A vector with unit norm ( $\|x\| = 1$ ) is called a *unit vector*.

**Definition 2.22: Normalization**

To normalize a vector  $x$ , we divide by its norm:

$$\hat{x} = \frac{x}{\|x\|} \quad (2.48)$$

resulting in a unit vector pointing in the same direction.

Normalization is commonly used in deep learning for improving training stability and convergence, including batch normalization for normalizing activations across mini-batches, layer normalization for normalizing within individual samples, input feature scaling for ensuring consistent input ranges, and weight normalization for controlling the magnitude of network parameters.

### 2.5.6 Distance Metrics

Norms induce distance metrics. The distance between vectors  $x$  and  $y$  is:

$$d(x, y) = \|x - y\| \quad (2.49)$$

Different norms lead to different notions of distance that are useful in various contexts. The  $L^1$  norm leads to Manhattan distance, which measures the sum of coordinate differences and is particularly useful in applications where movement is constrained to grid-like patterns, such as in image processing or when dealing with sparse data where the path between points must follow coordinate axes. The  $L^2$  norm leads to Euclidean distance, which measures the straight-line distance between points and is the most intuitive distance metric for most applications, making it the default choice for similarity measures, clustering algorithms, and optimization problems in deep learning. The  $L^\infty$  norm leads to Chebyshev distance, which measures the maximum coordinate difference and is useful in applications where the worst-case difference between coordinates is more important than the overall distance, such as in error analysis or when dealing with constraints that must be satisfied simultaneously.

**Remark 2.6**

While the Lebesgue measure is not covered in this book, it is an important concept in advanced deep learning. The Lebesgue measure provides a more general framework for measuring sets and integrating functions, which becomes crucial when dealing with high-dimensional spaces, probability measures, and advanced optimization theory in deep learning. Understanding Lebesgue integration is particularly important for rigorous analysis of convergence properties in optimization algorithms and for understanding the theoretical foundations of probability measures in machine learning.

## 2.5.7 Regularization in Deep Learning

Norms are central to regularization techniques that help prevent overfitting and improve generalization in deep learning models.  $L^1$  regularization adds  $\lambda \|\mathbf{w}\|_1$  to the loss function, promoting sparsity by encouraging many weights to become exactly zero, which helps with feature selection and model compression by automatically identifying and removing less important features.  $L^2$  regularization adds  $\lambda \|\mathbf{w}\|_2^2$  to the loss function, preventing large weights by penalizing the squared magnitude of the weight vector, which helps prevent overfitting by keeping the model parameters small and well-behaved. Elastic Net regularization combines both  $L^1$  and  $L^2$  penalties using  $\lambda_1 \|\mathbf{w}\|_1 + \lambda_2 \|\mathbf{w}\|_2^2$ , providing a balanced approach that can achieve both sparsity and smoothness in the learned parameters, making it particularly useful in applications where both feature selection and parameter stability are important.

Understanding norms and their properties is essential for designing effective regularization strategies and analyzing model behavior.

## 2.6 Eigendecomposition

Eigendecomposition is a powerful tool for understanding and analyzing linear transformations, with important applications in deep learning.

### 2.6.1 Eigenvalues and Eigenvectors

#### Definition 2.23: Eigenvector and Eigenvalue

An *eigenvector* of a square matrix  $\mathbf{A} \in \mathbb{R}^{n \times n}$  is a non-zero vector  $\mathbf{v}$  such that:

$$\mathbf{Av} = \lambda \mathbf{v} \quad (2.50)$$

where  $\lambda \in \mathbb{R}$  (or  $\mathbb{C}$ ) is the corresponding *eigenvalue*.

The eigenvector's direction is preserved under the transformation  $\mathbf{A}$ , with only its magnitude scaled by  $\lambda$ .

#### Example 2.18

Consider  $\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 0 & 2 \end{bmatrix}$ . We can verify that  $\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$  is an eigenvector:

$$\mathbf{Av}_1 = \begin{bmatrix} 3 & 1 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \end{bmatrix} = 3\mathbf{v}_1 \quad (2.51)$$

So  $\lambda_1 = 3$  is an eigenvalue.

### 2.6.2 Finding Eigenvalues

To find eigenvalues, we solve the *characteristic equation*:

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0 \quad (2.52)$$

This gives a polynomial of degree  $n$  called the characteristic polynomial, which has  $n$  roots (counting multiplicities) in  $\mathbb{C}$ .

**Example 2.19**

For  $\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ :

$$\det \begin{bmatrix} 2 - \lambda & 1 \\ 1 & 2 - \lambda \end{bmatrix} = (2 - \lambda)^2 - 1 = \lambda^2 - 4\lambda + 3 = 0 \quad (2.53)$$

Solving gives  $\lambda_1 = 3$  and  $\lambda_2 = 1$ .

### 2.6.3 Eigendecomposition

If a matrix  $\mathbf{A} \in \mathbb{R}^{n \times n}$  has  $n$  linearly independent eigenvectors, it can be decomposed as:

$$\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^{-1} \quad (2.54)$$

where  $\mathbf{V}$  is the matrix whose columns are eigenvectors  $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n]$  and  $\Lambda$  is a diagonal matrix of eigenvalues  $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ .

This is called the *eigendecomposition* or *spectral decomposition*.

### 2.6.4 Symmetric Matrices

Symmetric matrices have particularly nice properties.

**Theorem 2.1: Spectral Theorem for Symmetric Matrices**

If  $\mathbf{A}$  is a real symmetric matrix ( $\mathbf{A} = \mathbf{A}^\top$ ), then:

1. All eigenvalues are real
2. Eigenvectors corresponding to different eigenvalues are orthogonal
3.  $\mathbf{A}$  can be decomposed as:

$$\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^\top \quad (2.55)$$

where  $\mathbf{Q}$  is an orthogonal matrix ( $\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}$ ) of eigenvectors.

This decomposition is fundamental in many algorithms, including Principal Component Analysis (PCA).

### 2.6.5 Properties of Eigenvalues

For a matrix  $\mathbf{A} \in \mathbb{R}^{n \times n}$ , several important properties connect eigenvalues to other matrix characteristics. The trace of a matrix equals the sum of its eigenvalues  $\text{trace}(\mathbf{A}) = \sum_{i=1}^n A_{ii} = \sum_{i=1}^n \lambda_i$ , providing a direct relationship between the diagonal elements and the eigenvalues that is useful for understanding the overall behavior of the matrix. The determinant of a matrix equals the product of its eigenvalues  $\det(\mathbf{A}) = \prod_{i=1}^n \lambda_i$ , which connects the volume scaling factor of the linear transformation to the eigenvalues and provides insight into whether the transformation preserves or changes the orientation of the space. If  $\mathbf{A}$  is invertible, the eigenvalues of  $\mathbf{A}^{-1}$  are the reciprocals of the original eigenvalues  $1/\lambda_i$ , which means that the inverse transformation scales vectors by the inverse of the original scaling factors. The eigenvalues of  $\mathbf{A}^k$  are the  $k$ -th powers of the original eigenvalues  $\lambda_i^k$ , which is particularly useful for understanding the long-term behavior of iterative processes and the stability of dynamical systems.

## 2.6.6 Positive Definite Matrices

### Definition 2.24: Positive Definite

A symmetric matrix  $\mathbf{A}$  is *positive definite* if for all non-zero  $\mathbf{x} \in \mathbb{R}^n$ :

$$\mathbf{x}^\top \mathbf{A} \mathbf{x} > 0 \quad (2.56)$$

Equivalently, all eigenvalues of  $\mathbf{A}$  are positive.

### Definition 2.25: Positive Semi-definite

$\mathbf{A}$  is *positive semi-definite* if  $\mathbf{x}^\top \mathbf{A} \mathbf{x} \geq 0$  for all  $\mathbf{x}$ , i.e., all eigenvalues are non-negative.

Positive definite matrices are crucial in optimization, as they ensure that local minima are global minima for quadratic functions.

## 2.6.7 Applications in Deep Learning

Eigendecomposition has several important applications in deep learning that leverage the geometric and algebraic properties of eigenvalues and eigenvectors. Principal Component Analysis (PCA) finds directions of maximum variance by computing eigenvectors of the covariance matrix, which is fundamental for dimensionality reduction and data preprocessing in many machine learning pipelines. In optimization, the Hessian matrix's eigenvalues determine the curvature of the loss surface, with positive definite Hessians indicating convexity and providing insights into the convergence properties of optimization algorithms. Spectral normalization constrains the largest eigenvalue of weight matrices to stabilize training of generative adversarial networks (GANs), preventing the discriminator from becoming too powerful and maintaining training stability. Graph Neural Networks use graph Laplacian eigendecomposition to define spectral graph convolutions, which enable the application of convolutional neural network concepts to irregular graph structures. Understanding the dynamics of recurrent neural networks involves analyzing the eigenvalues of recurrent weight matrices, as these eigenvalues affect gradient flow and stability during training, with eigenvalues greater than 1 leading to exploding gradients and eigenvalues less than 1 leading to vanishing gradients.

## 2.6.8 Computational Considerations

Computing eigendecomposition involves several computational considerations that are important for practical applications in deep learning. Full eigendecomposition requires  $O(n^3)$  operations for dense matrices, making it computationally expensive for large matrices, which is why approximate methods are often preferred in practice. Power iteration for finding the dominant eigenvector requires  $O(kn^2)$  operations for  $k$  iterations, providing a more efficient way to compute the most important eigenvalue and eigenvector when only the dominant direction is needed. Iterative methods such as Lanczos algorithm are particularly useful for sparse matrices, as they can exploit the sparsity structure to reduce computational complexity and memory requirements. For large-scale deep learning applications, we often use approximations such as power iteration when only the top eigenvalues are needed, focus on top- $k$  eigenvalues and eigenvectors when the full spectrum is not required, and employ specialized algorithms for specific matrix structures such as symmetric or sparse matrices to achieve better performance and numerical stability.

Understanding eigendecomposition provides insight into the geometric properties of linear transformations and is essential for many advanced deep learning techniques.

## Key Takeaways

### Key Takeaways 2

- **Vectors and matrices** are fundamental building blocks for representing data and transformations in neural networks.
- **Matrix operations** (multiplication, transposition, inversion) enable efficient computation of forward passes and gradients.
- **Eigendecomposition and SVD** reveal structure in data and are crucial for PCA, matrix factorisation, and understanding dynamics.
- **Norms and distances** provide metrics for measuring similarity, regularisation, and convergence in optimisation.
- **Linear systems** underpin many machine learning algorithms and their solvability determines model identifiability.

## Exercises

### Easy

*Exercise 2.1 (Matrix Multiplication).* Given  $\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix}$  and  $\mathbf{B} = \begin{bmatrix} 1 & 2 \\ 3 & 1 \end{bmatrix}$ , compute  $\mathbf{AB}$ .

**Hint:**

Remember that  $(\mathbf{AB})_{ij} = \sum_k a_{ik} b_{kj}$ .

*Exercise 2.2 (Vector Norms).* Calculate the L1, L2, and  $L_\infty$  norms of the vector  $\mathbf{v} = [3, -4, 0]$ .

**Hint:**

L1 norm is sum of absolute values, L2 norm is Euclidean length,  $L_\infty$  is maximum absolute value.

*Exercise 2.3 (Linear Independence).* Determine whether the vectors  $\mathbf{v}_1 = [1, 0, 1]$ ,  $\mathbf{v}_2 = [0, 1, 0]$ , and  $\mathbf{v}_3 = [1, 1, 1]$  are linearly independent.

**Hint:**

Check if  $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 = \mathbf{0}$  has only the trivial solution  $c_1 = c_2 = c_3 = 0$ .

*Exercise 2.4 (Matrix Transpose Properties).* Prove that  $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$  for any compatible matrices  $\mathbf{A}$  and  $\mathbf{B}$ .

**Hint:**

Use the definition of transpose and matrix multiplication element-wise.

*Exercise 2.5 (Dot Product and Angle).* Calculate the dot product of vectors  $\mathbf{u} = [1, 2, 3]$  and  $\mathbf{v} = [4, -1, 2]$ , and find the angle between them.

**Hint:**

Use  $\mathbf{u} \cdot \mathbf{v} = |\mathbf{u}||\mathbf{v}| \cos \theta$  and the dot product formula.

**Exercise 2.6 (Matrix Addition and Scalar Multiplication).** Given  $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$  and  $\mathbf{B} = \begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix}$ , compute  $2\mathbf{A} + 3\mathbf{B}$ .

**Hint:**

Perform scalar multiplication first, then add the resulting matrices element-wise.

**Exercise 2.7 (Identity Matrix Properties).** Show that  $\mathbf{I}\mathbf{A} = \mathbf{A}\mathbf{I} = \mathbf{A}$  for any square matrix  $\mathbf{A}$  of the same size as the identity matrix  $\mathbf{I}$ .

**Hint:**

Use the definition of the identity matrix where  $I_{ij} = 1$  if  $i = j$  and 0 otherwise.

**Exercise 2.8 (Vector Space Axioms).** Verify that the set of all  $2 \times 2$  matrices forms a vector space under matrix addition and scalar multiplication.

**Hint:**

Check closure, associativity, commutativity, identity element, inverse element, and distributive properties.

**Exercise 2.9 (Cross Product in 3D).** Calculate the cross product of  $\mathbf{a} = [1, 0, 2]$  and  $\mathbf{b} = [3, 1, 1]$ , and verify that the result is perpendicular to both vectors.

**Hint:**

Use the determinant formula for cross product and check that  $\mathbf{a} \cdot (\mathbf{a} \times \mathbf{b}) = 0$ .

**Exercise 2.10 (Matrix Rank).** Find the rank of the matrix  $\mathbf{C} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 1 & 1 & 2 \end{bmatrix}$ .

**Hint:**

Use row reduction to find the number of linearly independent rows or columns.

**Exercise 2.11 (Determinant Properties).** Calculate the determinant of  $\mathbf{D} = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 3 & 2 \\ 0 & 1 & 1 \end{bmatrix}$  using cofactor expansion.

**Hint:**

Expand along the first row:  $\det(\mathbf{D}) = \sum_{j=1}^3 (-1)^{1+j} d_{1j} M_{1j}$ .

## Medium

**Exercise 2.12 (Eigenvalues and Eigenvectors).** Find the eigenvalues and eigenvectors of the matrix  $\mathbf{A} = \begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix}$ .

**Hint:**

Solve  $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$  for eigenvalues, then find eigenvectors by solving  $(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}$ .

*Exercise 2.13 (SVD Application).* Explain how Singular Value Decomposition (SVD) can be used for dimensionality reduction. Describe the relationship between SVD and Principal Component Analysis (PCA).

**Hint:**

Consider which singular values and vectors to keep, and how this relates to variance in the data.

*Exercise 2.14 (Matrix Inversion and Linear Systems).* Solve the system of linear equations using matrix inversion:  $2x+y = 5$  and  $x-3y = -1$ .

**Hint:**

Write as  $\mathbf{Ax} = \mathbf{b}$ , then  $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ .

*Exercise 2.15 (Orthogonal Matrices).* Prove that the columns of an orthogonal matrix are orthonormal vectors. Show that  $\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}$  for an orthogonal matrix  $\mathbf{Q}$ .

**Hint:**

Use the definition of orthogonality and the properties of matrix multiplication.

*Exercise 2.16 (Eigenvalue Decomposition).* Find the eigenvalue decomposition of the symmetric matrix  $\mathbf{S} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$  and verify that  $\mathbf{S} = \mathbf{Q}\Lambda\mathbf{Q}^\top$ .

**Hint:**

For symmetric matrices, eigenvectors are orthogonal and can be normalised to form an orthogonal matrix.

*Exercise 2.17 (Matrix Norms and Condition Number).* Calculate the Frobenius norm and the condition number of the matrix  $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 0.5 & 1 \end{bmatrix}$ .

**Hint:**

Frobenius norm is  $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$ , condition number is  $\kappa(\mathbf{A}) = \|\mathbf{A}\| \|\mathbf{A}^{-1}\|$ .

## Hard

*Exercise 2.18 (Matrix Decomposition for Neural Networks).* Show how the weight matrix in a neural network layer can be decomposed using SVD to reduce the number of parameters. Analyse the computational and memory trade-offs.

**Hint:**

Consider low-rank approximation  $\mathbf{W} \approx \mathbf{U}_k \Sigma_k \mathbf{V}_k^\top$  where  $k < \min(m, n)$ .

*Exercise 2.19 (Krylov Subspace Methods).* Explain how Krylov subspace methods can be used to solve large linear systems efficiently. Compare the computational complexity with direct methods.

**Hint:**

Consider the Arnoldi iteration and how it constructs an orthogonal basis for  $\mathcal{K}_k(\mathbf{A}, \mathbf{b})$ .

*Exercise 2.20 (Tensor Decomposition).* Derive the CP (CANDECOMP/PARAFAC) decomposition for a 3-way tensor and show how it relates to matrix factorisation. Discuss applications in deep learning.

**Hint:**

*Express the tensor as a sum of rank-1 components:  $\mathcal{T} = \sum_{r=1}^R \lambda_r \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$ .*

*Exercise 2.21 (Numerical Stability in Matrix Computations).* Analyse the numerical stability of computing eigenvalues using the QR algorithm. Discuss the role of Householder transformations and Givens rotations.

**Hint:**

*Consider the accumulation of rounding errors and the convergence properties of the QR iteration.*

*Exercise 2.22 (Sparse Matrix Representations).* Design an efficient storage scheme for sparse matrices and implement key operations (matrix-vector multiplication, matrix-matrix multiplication). Analyse the computational complexity.

**Hint:**

*Consider formats like CSR (Compressed Sparse Row) or COO (Coordinate) and their trade-offs in terms of storage and access patterns.*

## Chapter 3

# Probability and Information Theory

This chapter introduces fundamental concepts from probability theory and information theory that are essential for understanding machine learning and deep learning. Topics include probability distributions, conditional probability, expectation, variance, entropy, and mutual information.

### Learning Objectives

1. Probability foundations and intuitive meaning of discrete and continuous distributions
2. Conditional probability and Bayes' theorem in machine learning algorithms
3. Statistical measures: expectation, variance, and covariance of random variables
4. Common probability distributions and their use in machine learning
5. Information content using entropy, cross-entropy, and KL divergence
6. Information theory concepts in loss functions and representation learning

## 3.1 Probability Distributions ●

Probability distributions are mathematical functions that describe how probabilities are distributed across different possible outcomes of a random variable, providing the fundamental framework for modeling uncertainty in data and making predictions in machine learning and deep learning applications.

### 3.1.1 Intuition: What is Probability?

Imagine you're playing a game of dice where you roll a standard six-sided die with numbers 1 through 6. Before rolling, you know that each face has an equal chance of appearing, meaning that if you roll the die many times, each number will appear approximately one-sixth of the time. This "chance" is what we call probability - a number between 0 and 1 that quantifies how likely an event is to occur, where 0 means impossible and 1 means certain. In machine learning, we face uncertainty everywhere, from data uncertainty about whether the next customer will click on an ad, to model uncertainty about how confident our neural network is in its prediction, to parameter uncertainty about what the best value is for our model's weights. Probability distributions are mathematical tools that help us model and work with this uncertainty systematically, providing a rigorous framework for making decisions under uncertainty and quantifying the confidence we can have in our predictions and model parameters.

### 3.1.2 Visualizing Probability

Consider a simple example: predicting whether it will rain tomorrow. We might say there's a 30% chance of rain, which means that if we could repeat tomorrow 100 times, rain would occur about 30 times, the probability of rain is 0.3, and the probability of no rain is 0.7. This intuitive understanding of probability helps us visualize how probability distributions work - they assign numerical values to different outcomes, showing us not just what can happen, but how likely each outcome is to occur.

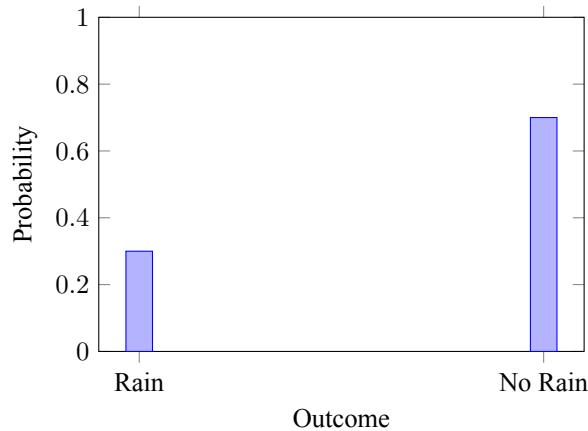


Figure 3.1: Probability distribution for rain prediction

Probability theory provides a mathematical framework for quantifying uncertainty. In deep learning, we use probability distributions to model uncertainty in data, model parameters, and predictions.

### 3.1.3 Discrete Probability Distributions

Discrete probability distributions deal with outcomes that can be counted and listed explicitly, such as coin flips with heads or tails, dice rolls with numbers 1 through 6, or email classification with spam or not spam. These distributions assign probabilities to each possible outcome, where the sum of all probabilities equals 1, representing the fact that one of the possible outcomes must occur.

A discrete random variable  $X$  takes values from a countable set. The **probability mass function** (PMF)  $P(X = x)$  assigns probabilities to each possible value:

$$P(X = x) \geq 0 \quad \text{for all } x \tag{3.1}$$

$$\sum_x P(X = x) = 1 \tag{3.2}$$

#### Example: Fair Coin

For a fair coin, we have:

$$P(X = 0) = 0.5 \quad (\text{Tails}) \tag{3.3}$$

$$P(X = 1) = 0.5 \quad (\text{Heads}) \tag{3.4}$$

### 3.1.4 Continuous Probability Distributions

Continuous probability distributions deal with variables that can take any value within a continuous range, such as the height of people which can be 170 cm, 170.5 cm, 170.52 cm, and so on, or temperature which can be 22.3°C, 22.34°C, 22.341°C, and so on, or neural network weights which can be 0.1234, 0.12345, 0.123456, and so on. Since there are infinitely many possible

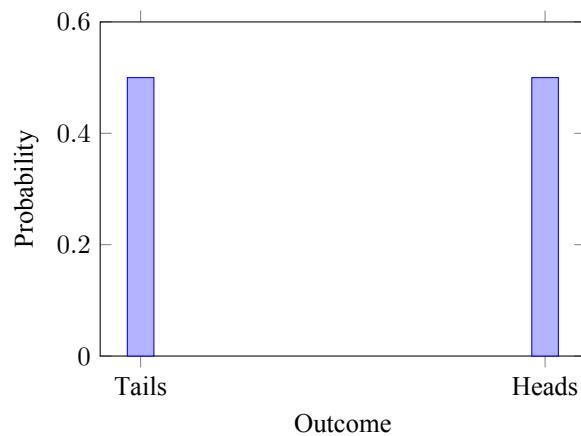


Figure 3.2: Probability mass function for a fair coin

values, we can't assign probabilities to individual points, but instead use density functions that describe how "concentrated" the probability is in different regions of the continuous space.

A continuous random variable can take any value in a continuous range. We describe it using a **probability density function** (PDF)  $p(x)$ :

$$p(x) \geq 0 \quad \text{for all } x \tag{3.5}$$

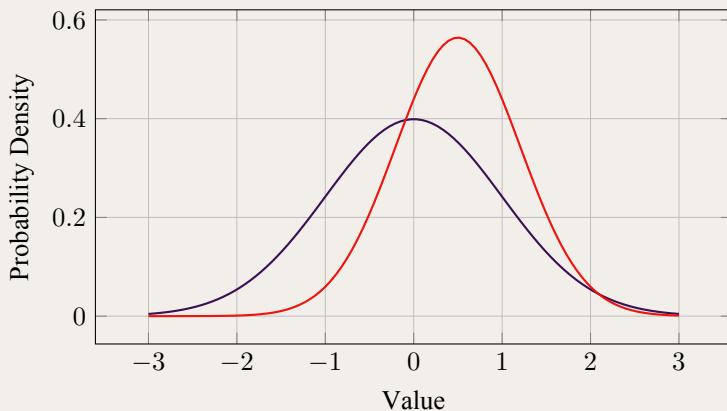
$$\int_{-\infty}^{\infty} p(x) dx = 1 \tag{3.6}$$

The probability that  $X$  falls in an interval  $[a, b]$  is:

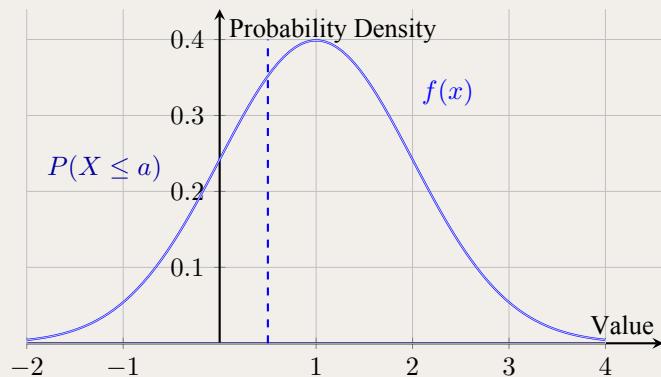
$$P(a \leq X \leq b) = \int_a^b p(x) dx \tag{3.7}$$

**Example 3.1**

Normal Distribution The most common continuous distribution is the **normal (Gaussian) distribution**, which looks like a bell curve:



The area under the curve between any two points gives the probability of the variable falling in that range.



### 3.1.5 Joint and Marginal Distributions

In real-world scenarios, we often deal with multiple variables simultaneously, such as weather prediction involving both temperature and humidity, image classification involving pixel values at different positions, or stock prices involving multiple stocks in a portfolio. The joint distribution tells us about the probability of combinations of values across all variables, while marginal distributions tell us about individual variables when we ignore the others, providing a way to understand both the relationships between variables and the behavior of individual variables in isolation.

**Example 3.2**

Weather Data Consider a simple weather dataset with two variables:

- $X$ : Temperature (Hot/Cold)
- $Y$ : Humidity (High/Low)

	$Y = \text{High}$	$Y = \text{Low}$	<b>Marginal</b>
$X = \text{Hot}$	0.3	0.2	<b>0.5</b>
$X = \text{Cold}$	0.1	0.4	<b>0.5</b>
<b>Marginal</b>	<b>0.4</b>	<b>0.6</b>	<b>1.0</b>

From our weather example:

- $P(X = \text{Hot}) = 0.3 + 0.2 = 0.5$  (marginal probability of hot weather)
- $P(Y = \text{High}) = 0.3 + 0.1 = 0.4$  (marginal probability of high humidity)

For multiple random variables  $X$  and  $Y$ , the **joint distribution**  $P(X, Y)$  describes their combined behavior. The **marginal distribution** is obtained by summing (or integrating) over the other variable:

$$P(X = x) = \sum_y P(X = x, Y = y) \quad (3.8)$$

For continuous variables:

$$p(x) = \int p(x, y) dy \quad (3.9)$$

## 3.2 Conditional Probability and Bayes' Rule ●

This section introduces the fundamental concepts of conditional probability and Bayes' theorem, which are essential for understanding how to update beliefs with new information in machine learning.

### 3.2.1 Intuition: Updating Beliefs with New Information

Imagine you're a doctor trying to diagnose a patient. Initially, you might think there's a 5% chance the patient has a rare disease. But then the patient tells you they have a specific symptom that's present in 80% of people with that disease. How should you update your belief?

This is exactly what **conditional probability** helps us do - it tells us how to update our beliefs when we get new information.

### 3.2.2 Conditional Probability

The conditional probability of  $X$  given  $Y$  is defined as:

$$P(X|Y) = \frac{P(X, Y)}{P(Y)} \quad (3.10)$$

This quantifies how the probability of  $X$  changes when we know the value of  $Y$ . This concept is fundamental to understanding how new information updates our beliefs about uncertain events, allowing us to make more informed decisions based on the evidence we observe.

**Example 3.3**

Medical Diagnosis Let's make this concrete with our medical example:

- $D$ : Patient has the disease (1 = yes, 0 = no)
- $S$ : Patient has the symptom (1 = yes, 0 = no)

From medical records, we know:

- $P(D = 1) = 0.05$  (5% of population has the disease)
- $P(S = 1|D = 1) = 0.8$  (80% of diseased patients have the symptom)
- $P(S = 1|D = 0) = 0.1$  (10% of healthy patients have the symptom)

If a patient has the symptom, what's the probability they have the disease?

Using Bayes' theorem (which we'll derive next):

$$P(D = 1|S = 1) = \frac{P(S = 1|D = 1)P(D = 1)}{P(S = 1)} \quad (3.11)$$

$$= \frac{0.8 \times 0.05}{0.8 \times 0.05 + 0.1 \times 0.95} \quad (3.12)$$

$$= \frac{0.04}{0.04 + 0.095} \quad (3.13)$$

$$= \frac{0.04}{0.135} \approx 0.296 \quad (3.14)$$

So even with the symptom, there's only about a 30% chance the patient has the disease!

### 3.2.3 Independence

Two events are independent if knowing one doesn't change our belief about the other, such as rolling two dice where the result of the first die doesn't affect the second, or they can be dependent like weather and clothing choice where knowing it's raining affects the probability you'll wear a raincoat. Two random variables  $X$  and  $Y$  are independent if  $P(X, Y) = P(X)P(Y)$ , which means that the joint probability equals the product of the individual probabilities, or equivalently,  $P(X|Y) = P(X)$  and  $P(Y|X) = P(Y)$ , indicating that knowing the value of one variable doesn't change our belief about the other.

**Example 3.4**

Independent vs Dependent Variables Consider two scenarios:

**Scenario 1 (Independent):** Flipping two coins

- $P(\text{First coin} = \text{Heads}) = 0.5$
- $P(\text{Second coin} = \text{Heads}) = 0.5$
- $P(\text{Both Heads}) = 0.5 \times 0.5 = 0.25 \checkmark$

**Scenario 2 (Dependent):** Drawing cards without replacement

- $P(\text{First card} = \text{Ace}) = 4/52 = 1/13$
- $P(\text{Second card} = \text{Ace}) = 3/51$  (if first was Ace) or  $4/51$  (if first wasn't Ace)
- The probability of the second card depends on what the first card was

### 3.2.4 Bayes' Theorem

#### Intuition: The Most Important Formula in Machine Learning

Bayes' theorem is like a "belief update machine." It tells us how to revise our initial beliefs (prior) when we observe new evidence, to get our updated beliefs (posterior).

##### Theorem 3.1: Bayes' Theorem

Bayes' theorem is fundamental to probabilistic inference:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \quad (3.15)$$

#### Understanding Each Component

In machine learning terminology:

- $P(X)$  is the **prior** probability - what we believed before seeing the data
- $P(Y|X)$  is the **likelihood** - how likely the data is given our hypothesis
- $P(X|Y)$  is the **posterior** probability - what we believe after seeing the data
- $P(Y)$  is the **evidence** or marginal likelihood - the probability of observing the data

#### Visualizing Bayes' Theorem

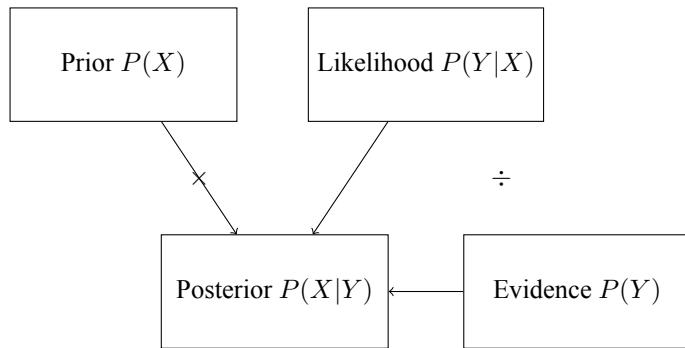


Figure 3.3: Bayes' theorem as a belief update process

The formula can be read as: "Posterior = (Likelihood × Prior) ÷ Evidence"

### 3.2.5 Application to Machine Learning

Bayes' theorem forms the basis of several important machine learning techniques. Bayesian inference provides a framework for updating beliefs about model parameters as new data becomes available, allowing for principled uncertainty quantification and decision-making under uncertainty. Naive Bayes classifiers use the assumption of feature independence to efficiently compute posterior probabilities for classification tasks, making them particularly useful for text classification and spam detection applications. Maximum a posteriori (MAP) estimation combines prior knowledge about parameters with observed data to find the most likely parameter values, providing a principled way to incorporate domain expertise into machine learning models. Bayesian neural networks extend traditional neural networks by treating weights as random variables with probability distributions, enabling uncertainty quantification in predictions and providing more robust estimates of model confidence.

Given data  $\mathcal{D}$  and model parameters  $\theta$ :

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})} \quad (3.16)$$

### 3.3 Expectation, Variance, and Covariance ●

Expectation, variance, and covariance are fundamental statistical measures that characterize the behavior of random variables, providing essential tools for understanding data distributions, relationships between variables, and uncertainty quantification in machine learning and deep learning applications.

#### 3.3.1 Intuition: Characterizing Random Variables

When we have a random variable, we often want to summarize its behavior with a few key numbers that capture the essential characteristics of the distribution. The expected value or mean represents the "center" or "typical" value around which the data is distributed, while variance measures how much the values spread out from the center, indicating the degree of uncertainty or variability in the data. Covariance tells us how two variables move together, revealing whether they tend to increase or decrease simultaneously, which is crucial for understanding relationships between different features or measurements. Think of it like describing a person: the mean height represents the average height of people in a group, the variance in height shows how much heights vary between tall and short people, and the covariance of height and weight reveals whether taller people tend to weigh more, providing insights into the relationships between different characteristics.

#### 3.3.2 Expectation

The **expected value** or **mean** of a function  $f(x)$  with respect to distribution  $P(x)$  is:

For discrete variables:

$$\mathbb{E}_{x \sim P}[f(x)] = \sum_x P(x)f(x) \quad (3.17)$$

For continuous variables:

$$\mathbb{E}_{x \sim p}[f(x)] = \int p(x)f(x) dx \quad (3.18)$$

##### Remark 3.1

Discrete expectation is a special case of continuous expectation. When we have a discrete distribution with probability mass function  $P(x)$ , we can think of it as a continuous distribution using Dirac delta functions:  $p(x) = \sum_i P(x_i)\delta(x - x_i)$ . The integral  $\int p(x)f(x) dx$  then becomes  $\sum_i P(x_i)f(x_i)$ , which is exactly the discrete expectation formula. This unified view helps us understand that both discrete and continuous expectations follow the same fundamental principle of weighted averaging.

**Example 3.5**

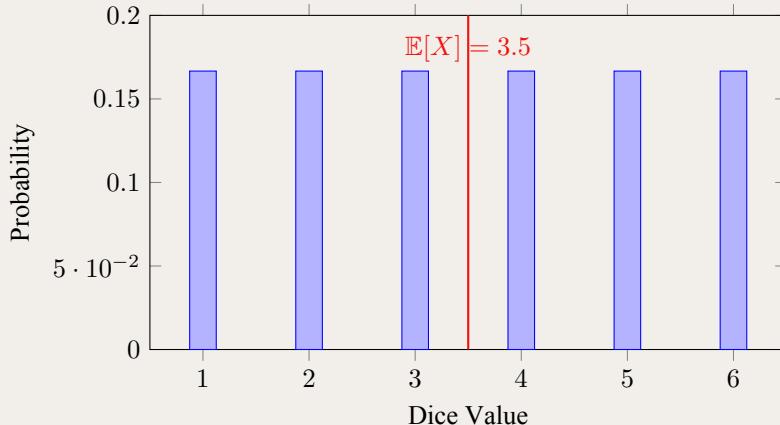
Expected Value of Dice For a fair six-sided die:

$$\mathbb{E}[X] = \sum_{x=1}^6 x \cdot P(X = x) \quad (3.19)$$

$$= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \dots + 6 \cdot \frac{1}{6} \quad (3.20)$$

$$= \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3.5 \quad (3.21)$$

The expected value is 3.5, even though we can never actually roll 3.5!



### 3.3.3 Variance

#### Intuition: Measuring Spread

Variance tells us how "spread out" the values are around the mean. Imagine two dart players with very different playing styles. The first player has low variance in their throws - every dart lands very close to the bullseye, creating a tight cluster of holes around the center. This player is consistent and precise, with their throws showing little variation from the target. The second player has high variance in their throws - their darts are scattered all over the board, some landing far to the left, others to the right, some high, some low. This player is inconsistent and unpredictable, with their throws showing large variation from the target. Variance quantifies this difference in consistency, measuring how much the individual values deviate from the average or expected value.

The **variance** measures the spread of a distribution:

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \quad (3.22)$$

The **standard deviation** is  $\sigma = \sqrt{\text{Var}(X)}$ .

#### Example: Variance of Dice

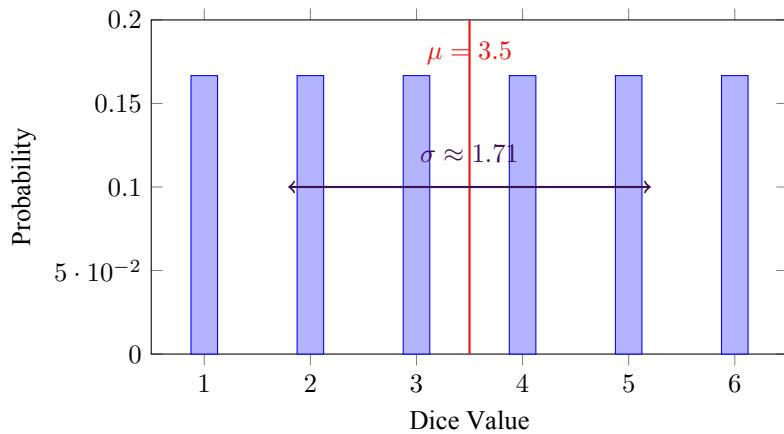
For our fair die:

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \quad (3.23)$$

$$= \left( \frac{1^2 + 2^2 + \dots + 6^2}{6} \right) - (3.5)^2 \quad (3.24)$$

$$= \frac{91}{6} - 12.25 = 15.17 - 12.25 = 2.92 \quad (3.25)$$

So  $\sigma = \sqrt{2.92} \approx 1.71$ .



### 3.3.4 Covariance

Covariance tells us whether two variables tend to move in the same direction or opposite directions, with positive covariance indicating that when one variable goes up, the other tends to go up too, negative covariance indicating that when one goes up, the other tends to go down, and zero covariance indicating no clear relationship. Examples include height and weight having positive covariance because taller people tend to weigh more, price and demand having negative covariance because higher prices usually mean lower demand, and height and IQ having near zero covariance because there's no clear relationship between physical height and cognitive ability.

The **covariance** measures how two variables vary together:

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \quad (3.26)$$

Positive covariance indicates that  $X$  and  $Y$  tend to increase together, while negative covariance indicates they tend to vary in opposite directions.

#### Example 3.6

Height and Weight Consider a small dataset of people:

Height (cm)	Weight (kg)
152	54
165	64
178	73
191	82

The means are  $\mu_X = 171.5$  and  $\mu_Y = 68.25$ . The covariance is:

$$\text{Cov}(X, Y) = \frac{1}{4} \sum_{i=1}^4 (x_i - 171.5)(y_i - 68.25) \quad (3.27)$$

$$= \frac{1}{4} [(-19.5)(-14.25) + (-6.5)(-4.25) + (6.5)(4.75) + (19.5)(13.75)] \quad (3.28)$$

$$= \frac{1}{4} [277.875 + 27.625 + 30.875 + 268.125] = \frac{604.5}{4} = 151.125 \quad (3.29)$$

Positive covariance confirms that taller people tend to weigh more!

### 3.3.5 Correlation

The correlation coefficient normalizes covariance by dividing it by the product of the standard deviations, providing a standardized measure of linear relationship between variables that ranges from -1 to 1. A correlation of 1 indicates a perfect

positive linear relationship, a correlation of -1 indicates a perfect negative linear relationship, and a correlation of 0 indicates no linear relationship, though the variables may still be dependent in non-linear ways.

### Definition 3.1: Correlation Coefficient

The correlation coefficient between random variables  $X$  and  $Y$  is defined as:

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \quad (3.30)$$

where  $\sigma_X$  and  $\sigma_Y$  are the standard deviations of  $X$  and  $Y$  respectively.

The correlation coefficient has several important properties. It is always bounded between -1 and 1, with  $\rho = 1$  indicating perfect positive linear correlation (as one variable increases, the other increases proportionally),  $\rho = -1$  indicating perfect negative linear correlation (as one variable increases, the other decreases proportionally), and  $\rho = 0$  indicating no linear correlation (though the variables may still be related in non-linear ways). The correlation coefficient is invariant to linear transformations of the variables, meaning that scaling or shifting the variables doesn't change their correlation. This makes correlation particularly useful for comparing relationships between variables that may have different scales or units.

## 3.4 Common Probability Distributions

Understanding common probability distributions is essential for deep learning because different distributions model different types of data and uncertainty, from binary outcomes in classification tasks to continuous values in regression problems, and from simple univariate cases to complex multivariate scenarios. These distributions provide the mathematical foundation for modeling the behavior of neural network weights, activation functions, and output predictions, enabling us to make principled decisions about model architecture, loss functions, and regularization strategies. By learning these distributions, we can better understand how to design neural networks that can handle various types of data and uncertainty, from the simple Bernoulli distribution for binary classification to the complex multivariate Gaussian for high-dimensional feature representations.

### 3.4.1 Bernoulli Distribution

Models a binary random variable (0 or 1):

$$P(X = 1) = \phi, \quad P(X = 0) = 1 - \phi \quad (3.31)$$

Used for binary classification problems.

#### Example 3.7

Consider a biased coin that lands heads with probability 0.7. Let  $X$  be the random variable representing the outcome:

- $X = 1$  if the coin lands heads
- $X = 0$  if the coin lands tails

The probability mass function is:

$$P(X = 1) = 0.7 \quad (\text{probability of heads}) \quad (3.32)$$

$$P(X = 0) = 0.3 \quad (\text{probability of tails}) \quad (3.33)$$

The expected value is  $\mathbb{E}[X] = 1 \cdot 0.7 + 0 \cdot 0.3 = 0.7$ , and the variance is  $\text{Var}(X) = 0.7 \cdot 0.3 = 0.21$ .

### 3.4.2 Categorical Distribution

Generalizes Bernoulli to  $k$  discrete outcomes. If  $X$  can take values  $\{1, 2, \dots, k\}$ :

$$P(X = i) = p_i \quad \text{where} \quad \sum_{i=1}^k p_i = 1 \quad (3.34)$$

### 3.4.3 Gaussian (Normal) Distribution

The Gaussian or normal distribution is the most important continuous distribution in deep learning, characterized by its bell-shaped curve and defined by its mean  $\mu$  and variance  $\sigma^2$ . This distribution is particularly important because of the central limit theorem, which states that sums of independent variables approach a Gaussian distribution, making it a natural choice for modeling many real-world phenomena and neural network activations. The Gaussian distribution has several key properties that make it useful in deep learning: it has a single peak at the mean, it's symmetric around the mean, and it has the property that about 68

The multivariate Gaussian with mean vector  $\mu$  and covariance matrix  $\Sigma$  is:

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \quad (3.35)$$

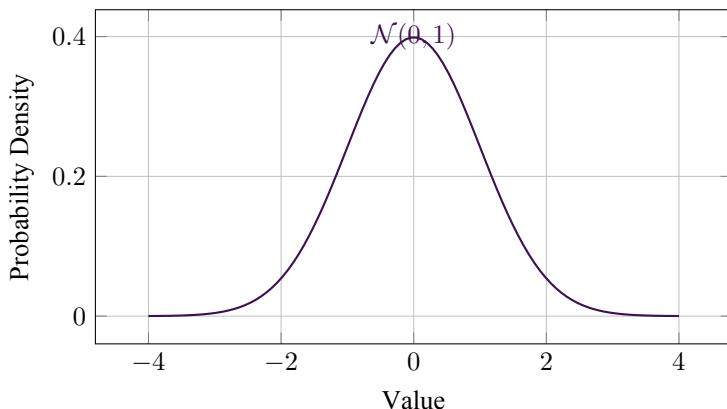


Figure 3.4: Standard normal distribution  $\mathcal{N}(0, 1)$ : bell-shaped curve with mean 0, std 1.

### 3.4.4 Exponential Distribution

Models the time between events in a Poisson process:

$$p(x; \lambda) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0 \quad (3.36)$$

### 3.4.5 Laplace Distribution

Heavy-tailed alternative to Gaussian:

$$\text{Laplace}(x; \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right) \quad (3.37)$$

where  $\mu$  is the location parameter (mean) and  $b$  is the scale parameter (controls the spread). The Laplace distribution is used in robust statistics because it's less sensitive to outliers than the Gaussian distribution, and in L1 regularization because its sharp peak at the mean encourages sparsity by penalizing small weights more heavily than the Gaussian distribution.

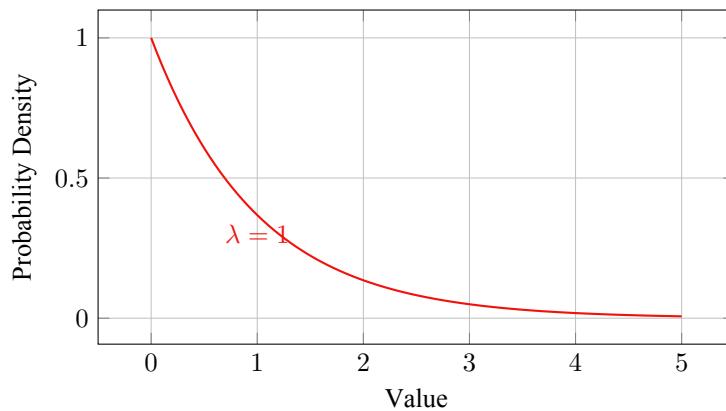


Figure 3.5: Exponential distribution ( $\lambda = 1$ ): starts at max, decreases exponentially.

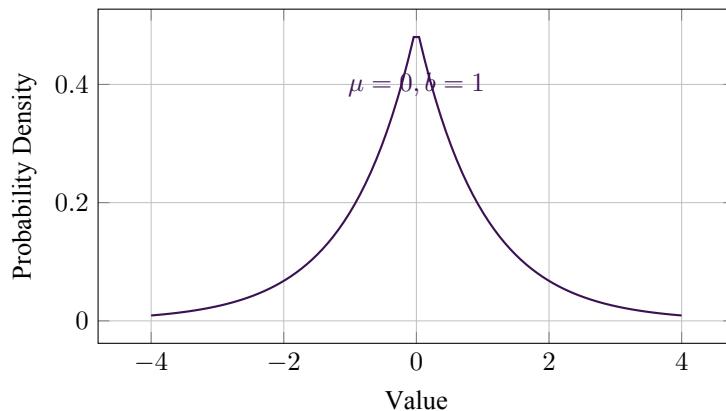


Figure 3.6: Laplace ( $\mu = 0, b = 1$ ): sharp peak with exponential tails, more robust to outliers than Gaussian.

### 3.4.6 Dirac Delta and Mixture Distributions

The **Dirac delta**  $\delta(x)$  concentrates all probability at a single point:

$$p(x) = \delta(x - \mu) \quad (3.38)$$

**Mixture distributions** combine multiple distributions:

$$p(x) = \sum_{i=1}^k \alpha_i p_i(x), \quad \sum_{i=1}^k \alpha_i = 1 \quad (3.39)$$

### Example 3.8: Gaussian Mixture Model (GMM)

A Gaussian Mixture Model is a probabilistic model that represents a probability distribution as a weighted sum of multiple Gaussian distributions, allowing it to model complex, multi-modal data that cannot be captured by a single Gaussian. GMMs are particularly useful in clustering, density estimation, and as building blocks for more complex generative models in deep learning.

The probability density function of a GMM with  $K$  components is:

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (3.40)$$

where:

- $\pi_k$  are the mixing weights (probabilities) satisfying  $\sum_{k=1}^K \pi_k = 1$
- $\boldsymbol{\mu}_k$  and  $\boldsymbol{\Sigma}_k$  are the mean and covariance matrix of the  $k$ -th Gaussian component
- $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$  is the multivariate Gaussian density function

The key insight is that each data point  $\mathbf{x}$  can be thought of as being generated by first selecting one of the  $K$  Gaussian components according to the mixing weights  $\pi_k$ , and then sampling from that selected component. This allows the model to capture complex, multi-modal distributions that arise naturally in real-world data, such as images with multiple object classes or speech signals with different phonemes.

## 3.5 Information Theory Basics

Information theory provides tools for quantifying information and uncertainty, which are crucial for understanding learning and compression.

### 3.5.1 Self-Information

The **self-information** or **surprisal** of an event  $x$  is:

$$I(x) = -\log P(x) \quad (3.41)$$

Rare events have high information content, while certain events have zero information.

### 3.5.2 Entropy

The **Shannon entropy** measures the expected information in a distribution:

$$H(X) = \mathbb{E}_{x \sim P}[I(x)] = -\sum_x P(x) \log P(x) \quad (3.42)$$

For continuous distributions, we use **differential entropy**:

$$H(X) = -\int p(x) \log p(x) dx \quad (3.43)$$

Entropy is maximized when all outcomes are equally likely.

### 3.5.3 Cross-Entropy

The **cross-entropy** between distributions  $P$  and  $Q$  is:

$$H(P, Q) = -\mathbb{E}_{x \sim P}[\log Q(x)] = -\sum_x P(x) \log Q(x) \quad (3.44)$$

In deep learning, cross-entropy is commonly used as a loss function for classification.

### 3.5.4 Kullback-Leibler Divergence

The **KL divergence** measures how one distribution differs from another:

$$D_{KL}(P\|Q) = \mathbb{E}_{x \sim P} \left[ \log \frac{P(x)}{Q(x)} \right] = \sum_x P(x) \log \frac{P(x)}{Q(x)} \quad (3.45)$$

Properties:

- $D_{KL}(P\|Q) \geq 0$  with equality if and only if  $P = Q$
- Not symmetric:  $D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$
- Related to cross-entropy:  $D_{KL}(P\|Q) = H(P, Q) - H(P)$

### 3.5.5 Mutual Information

The **mutual information** between  $X$  and  $Y$  quantifies how much knowing one reduces uncertainty about the other:

$$I(X; Y) = D_{KL}(P(X, Y)\|P(X)P(Y)) \quad (3.46)$$

Equivalently:

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \quad (3.47)$$

Mutual information is symmetric and measures the dependence between variables.

### 3.5.6 Applications in Deep Learning

Information theory concepts are used in several important applications in deep learning. Loss functions such as cross-entropy loss are directly based on information-theoretic principles, measuring the difference between predicted and true probability distributions to guide model training. Model selection techniques like AIC and BIC use information-theoretic principles to balance model complexity with goodness of fit, helping to prevent overfitting and select the most appropriate model architecture. Variational inference minimizes KL divergence between approximate and true posterior distributions, enabling efficient approximation of complex probabilistic models in deep learning. Information bottleneck theory provides a principled framework for understanding how neural networks learn to compress input information while preserving relevant features for the task at hand. Mutual information maximization in self-supervised learning helps models learn meaningful representations by maximizing the mutual information between different views or augmentations of the same data, enabling effective unsupervised representation learning.

## Key Takeaways

### Key Takeaways 3

- **Probability distributions** model uncertainty and enable principled reasoning under incomplete information.
- **Bayes' theorem** provides a framework for updating beliefs with evidence, central to many machine learning algorithms.
- **Expectation and variance** characterise random variables and guide choices of loss functions and model architectures.
- **Common distributions** (Bernoulli, Gaussian, categorical) serve as building blocks for probabilistic models.
- **Information theory** quantifies uncertainty through entropy and divergence, directly connecting to loss functions and regularisation.

## Exercises

### Easy

*Exercise 3.1 (Bayes' Theorem Application).* Given that  $P(\text{Disease}) = 0.01$ ,  $P(\text{Positive Test} \mid \text{Disease}) = 0.95$ , and  $P(\text{Positive Test} \mid \text{No Disease}) = 0.05$ , calculate  $P(\text{Disease} \mid \text{Positive Test})$ .

**Hint:**

Use Bayes' theorem:  $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ . Remember to compute  $P(\text{Positive Test})$  first.

*Exercise 3.2 (Expectation and Variance).* A discrete random variable  $X$  takes values 1, 2, 3, 4 with probabilities 0.1, 0.2, 0.4, 0.3 respectively. Calculate  $\mathbb{E}[X]$  and  $\text{Var}(X)$ .

**Hint:**

$\mathbb{E}[X] = \sum_i x_i P(X = x_i)$  and  $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ .

*Exercise 3.3 (Entropy Calculation).* Calculate the entropy of a fair coin flip and compare it to the entropy of a biased coin with  $P(\text{Heads}) = 0.9$ .

**Hint:**

$\text{Entropy } H(X) = -\sum_i p_i \log_2 p_i$ . Higher entropy means more uncertainty.

*Exercise 3.4 (Independence Test).* Given  $P(A) = 0.3$ ,  $P(B) = 0.4$ , and  $P(A \cap B) = 0.12$ , determine if events A and B are independent.

**Hint:**

Events are independent if  $P(A \cap B) = P(A)P(B)$ .

*Exercise 3.5 (Conditional Probability).* In a deck of 52 cards, what is the probability of drawing a heart given that the card drawn is red?

**Hint:**

*Use the definition of conditional probability:  $P(A|B) = P(A \cap B)/P(B)$ .*

*Exercise 3.6 (Joint Probability).* Given  $P(A) = 0.6$ ,  $P(B) = 0.4$ , and  $P(A|B) = 0.8$ , find  $P(A \cap B)$  and  $P(B|A)$ .

**Hint:**

*Use the multiplication rule:  $P(A \cap B) = P(A|B)P(B)$ .*

*Exercise 3.7 (Probability Distributions).* A random variable  $X$  follows a uniform distribution on  $[0, 2]$ . Find  $P(X > 1.5)$  and the expected value  $E[X]$ .

**Hint:**

*For uniform distribution on  $[a, b]$ , the density is  $f(x) = 1/(b-a)$  for  $x$  in  $[a, b]$ .*

*Exercise 3.8 (Binomial Distribution).* A fair coin is flipped 10 times. What is the probability of getting exactly 7 heads?

**Hint:**

*Use the binomial probability formula:  $P(X = k) = C(n, k)p^k(1 - p)^{(n-k)}$ .*

*Exercise 3.9 (Normal Distribution).* If  $X \sim N(50, 25)$ , find  $P(45 < X < 55)$  using the standard normal distribution.

**Hint:**

*Standardise using  $Z = (X - \mu)/\sigma$ , then use standard normal tables.*

*Exercise 3.10 (Information Content).* Calculate the information content of an event with probability 0.1, and compare it to an event with probability 0.5.

**Hint:**

*Information content is  $I(x) = -\log_2 P(x)$ .*

*Exercise 3.11 (Mutual Information).* Given  $P(X=0) = 0.6$ ,  $P(X=1) = 0.4$ ,  $P(Y=0|X=0) = 0.8$ ,  $P(Y=1|X=0) = 0.2$ ,  $P(Y=0|X=1) = 0.3$ ,  $P(Y=1|X=1) = 0.7$ , calculate  $I(X;Y)$ .

**Hint:**

*Use  $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$ .*

## Medium

*Exercise 3.12 (KL Divergence for Model Comparison).* Explain why Kullback-Leibler (KL) divergence is not symmetric and discuss its implications when comparing probability distributions in machine learning.

**Hint:**

*Consider  $D_{KL}(P||Q)$  versus  $D_{KL}(Q||P)$  and their behaviour when  $P$  or  $Q$  is close to zero.*

**Exercise 3.13 (Cross-Entropy Loss).** Show that minimising cross-entropy loss is equivalent to maximising the log-likelihood for classification tasks. Derive the relationship mathematically.

**Hint:**

Start with the cross-entropy  $H(p, q) = -\sum_i p_i \log q_i$  where  $p$  is the true distribution and  $q$  is the predicted distribution.

**Exercise 3.14 (Maximum Likelihood Estimation).** Given a sample of  $n$  independent observations from a normal distribution  $N(\mu, \sigma^2)$ , derive the maximum likelihood estimators for  $\mu$  and  $\sigma^2$ .

**Hint:**

Write the likelihood function  $L(\mu, \sigma^2)$  and take partial derivatives with respect to  $\mu$  and  $\sigma^2$ .

**Exercise 3.15 (Jensen's Inequality Application).** Use Jensen's inequality to show that the entropy of a mixture of distributions is at least the weighted average of the individual entropies.

**Hint:**

Consider  $H(\sum_i \alpha_i p_i) \geq \sum_i \alpha_i H(p_i)$  where  $\sum_i \alpha_i = 1$  and  $\alpha_i \geq 0$ .

**Exercise 3.16 (Central Limit Theorem).** Explain how the Central Limit Theorem applies to the convergence of sample means and discuss its implications for machine learning.

**Hint:**

Consider the distribution of sample means and how it approaches normality regardless of the original distribution.

**Exercise 3.17 (Concentration Inequalities).** Use Markov's inequality to bound  $P(X \geq 2E[X])$  for a non-negative random variable  $X$ , and compare with Chebyshev's inequality.

**Hint:**

Markov's inequality:  $P(X \geq a) \leq E[X]/a$  for  $a > 0$ .

**Exercise 3.18 (Bayesian Inference).** Given a prior Beta(2, 2) distribution and observing 7 successes in 10 trials, find the posterior distribution and the Bayesian estimate.

**Hint:**

Use the Beta-Binomial conjugate relationship:  $Beta(\alpha, \beta) + Binomial(n, p) \rightarrow Beta(\alpha + k, \beta + n - k)$ .

**Hard**

**Exercise 3.19 (Information Theory in Neural Networks).** Analyse how mutual information between layers in a neural network can be used to understand information flow during training. Discuss the information bottleneck principle.

**Hint:**

Consider  $I(X; Y) = H(X) - H(X|Y)$  and how it relates to representation learning.

*Exercise 3.20 (Variational Inference).* Derive the Evidence Lower Bound (ELBO) for variational inference and explain its relationship to the Kullback-Leibler divergence.

**Hint:**

*Start with  $\log p(x) = \log \int p(x,z) dz$  and use Jensen's inequality with a variational distribution  $q(z)$ .*

*Exercise 3.21 (Information Bottleneck Theory).* Prove that the information bottleneck principle leads to a trade-off between compression and prediction accuracy in representation learning.

**Hint:**

*Consider the Lagrangian  $L = I(X;T) - \beta I(T;Y)$  where  $T$  is the representation and  $\beta$  controls the trade-off.*

*Exercise 3.22 (PAC-Bayes Bounds).* Derive a PAC-Bayes bound for generalisation error in terms of the KL divergence between prior and posterior distributions.

**Hint:**

*Use the change of measure inequality and the union bound over the hypothesis space.*

*Exercise 3.23 (Maximum Entropy Principle).* Show that the maximum entropy distribution under moment constraints is exponential family, and derive the dual optimisation problem.

**Hint:**

*Use Lagrange multipliers to maximise  $H(p)$  subject to  $E[\Phi_i(X)] = \mu_i$  for moment constraints.*

*Exercise 3.24 (Causal Inference and Information Theory).* Analyse the relationship between causal discovery and information-theoretic measures, particularly in the context of conditional independence testing.

**Hint:**

*Consider how mutual information relates to conditional independence:  $X \perp\!\!\!\perp Y | Z$  if and only if  $I(X;Y|Z) = 0$ .*



# Chapter 4

## Numerical Computation

This chapter covers numerical methods and computational considerations essential for implementing deep learning algorithms. Topics include gradient-based optimization, numerical stability, and conditioning.

### Learning Objectives

1. Numerical precision issues and solutions for overflow, underflow, and numerical instability in deep learning
2. Gradient-based optimization and the role of Jacobian and Hessian matrices in optimization landscapes
3. Constrained optimization using Lagrange multipliers and KKT conditions for regularization
4. Numerical stability assessment and gradient checking techniques
5. Practical numerical techniques including log-sum-exp tricks and mixed precision training
6. Common numerical problems in deep learning and their solutions

## 4.1 Overflow and Underflow ●

Overflow and underflow are critical numerical issues that occur when computations produce values outside the representable range of floating-point arithmetic, causing catastrophic failures in deep learning algorithms that rely on precise numerical calculations.

### 4.1.1 Intuition: The Problem with Finite Precision

Imagine you're trying to measure the height of a building using a ruler that only has markings every meter. If the building is 10.7 meters tall, you might record it as 11 meters - you've lost some precision. Now imagine doing millions of calculations with this imprecise ruler, and you can see how small errors can compound into big problems. Computers face a similar challenge because they can only represent a finite number of digits, so they must round numbers, which is like having a ruler with limited markings where some information is always lost. In deep learning, this becomes critical because neural networks perform millions of calculations where small errors accumulate, exponential functions are common and can produce extremely large or small numbers, and gradients can become very small and might round to zero, breaking training. Computers represent real numbers with finite precision, typically using floating-point arithmetic, which leads to rounding errors that can accumulate and cause problems in deep learning algorithms.

### 4.1.2 Floating-Point Representation

The IEEE 754 standard defines floating-point numbers using a fixed number of bits to represent real numbers, with 32-bit floats having a smallest positive number of approximately  $10^{-38}$ , a largest number of approximately  $10^{38}$ , and a machine epsilon of approximately  $10^{-7}$ . This representation allows computers to handle a wide range of numbers but introduces limitations that can cause overflow when numbers exceed the maximum representable value and underflow when numbers become smaller than the minimum representable value.

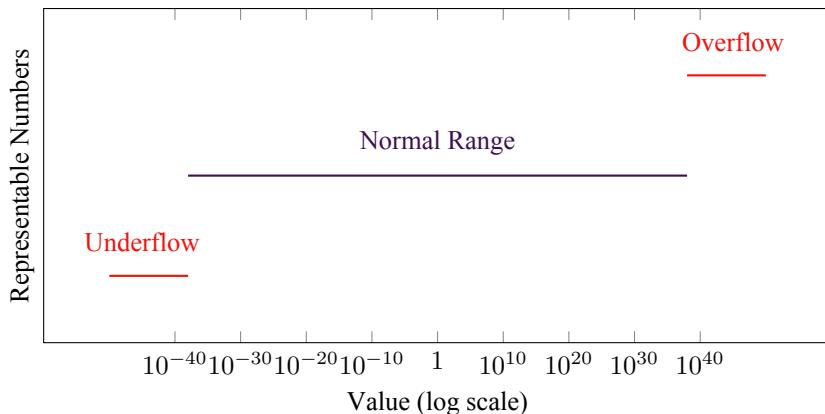


Figure 4.1: Representable range for 32-bit floating-point numbers

### 4.1.3 Underflow

#### Intuition: When Numbers Become Too Small

Think of underflow like trying to measure the width of a human hair with a ruler marked in meters. The hair is so thin that your ruler shows 0 meters - you've lost all information about the actual size.

In computers, **underflow** occurs when numbers become so small that they round to zero. This is like your ruler being too coarse to measure tiny objects.

**Underflow** occurs when numbers near zero are rounded to zero. This can be problematic when we need to compute ratios or logarithms. For example, the softmax function:

$$\text{softmax}(\mathbf{x})_i = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \quad (4.1)$$

can underflow if all  $x_i$  are very negative.

### 4.1.4 Overflow

#### Intuition: When Numbers Become Too Large

Imagine trying to count the number of atoms in the universe using a calculator that can only display 8 digits. When you reach 99,999,999, the next number would be 100,000,000, but your calculator shows "Error" or resets to 0.

**Overflow** occurs when large numbers exceed representable values. In the softmax example, overflow can occur if some  $x_i$  are very large.

### 4.1.5 Numerical Stability

To stabilize softmax, we use the identity:

$$\text{softmax}(\mathbf{x}) = \text{softmax}(\mathbf{x} - c) \quad (4.2)$$

where  $c = \max_i x_i$ . This prevents both overflow and underflow.

Similarly, when computing  $\log(\sum_i \exp(x_i))$ , we use the **log-sum-exp** trick:

$$\log\left(\sum_i \exp(x_i)\right) = c + \log\left(\sum_i \exp(x_i - c)\right) \quad (4.3)$$

### Example 4.1

Softmax Numerical Issues Consider computing softmax for  $\mathbf{x} = [1000, 1001, 1002]$ :

**Naive approach:**

$$\exp(1000) \approx \infty \quad (\text{overflow!}) \quad (4.4)$$

$$\exp(1001) \approx \infty \quad (\text{overflow!}) \quad (4.5)$$

$$\exp(1002) \approx \infty \quad (\text{overflow!}) \quad (4.6)$$

$$\text{softmax}(\mathbf{x}) = [\text{NaN}, \text{NaN}, \text{NaN}] \quad (4.7)$$

**Stable approach:**

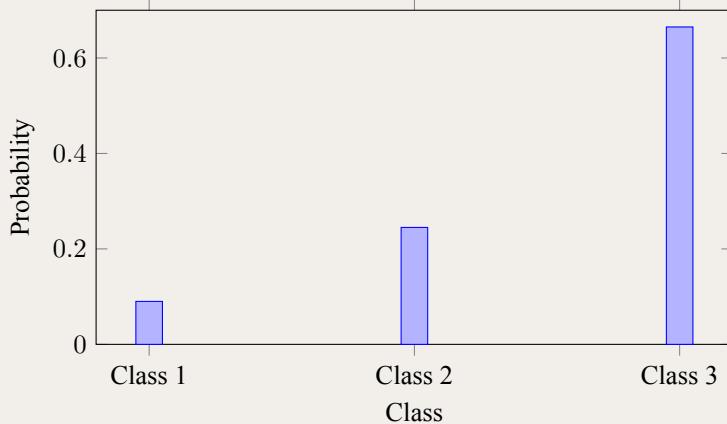
$$c = \max(1000, 1001, 1002) = 1002 \quad (4.8)$$

$$\exp(1000 - 1002) = \exp(-2) \approx 0.135 \quad (4.9)$$

$$\exp(1001 - 1002) = \exp(-1) \approx 0.368 \quad (4.10)$$

$$\exp(1002 - 1002) = \exp(0) = 1.000 \quad (4.11)$$

$$\text{softmax}(\mathbf{x}) = [0.090, 0.245, 0.665] \quad (4.12)$$



### 4.1.6 Other Numerical Issues

Beyond overflow and underflow, several other numerical issues can affect deep learning computations. Catastrophic cancellation occurs when subtracting nearly equal numbers, leading to loss of precision that can significantly impact gradient calculations and model training. Accumulated rounding errors represent another critical concern where small errors compound through many operations, potentially causing the model to converge to suboptimal solutions or fail to train entirely. To address these issues, several solutions can be employed including using higher precision arithmetic with 64-bit floats to reduce rounding errors, implementing algorithmic modifications like the log-sum-exp trick to maintain numerical stability, applying batch normalization to stabilize activations and gradients, and using gradient clipping to prevent exploding gradients that can cause numerical instability during training.

## 4.2 Gradient-Based Optimization

Gradient-based optimization is the fundamental method for training deep learning models, using the mathematical concept of gradients to iteratively find optimal parameters that minimize loss functions through careful navigation of high-dimensional parameter spaces.

### 4.2.1 Intuition: Finding the Bottom of a Hill

Imagine you're hiking in foggy mountains and need to find the lowest point in a valley. You can't see far ahead, but you can feel the slope under your feet. The steepest downward direction tells you which way to walk to get lower. This is exactly what gradient descent does, where the mountain represents the loss function we want to minimize, your position represents the current parameter values, the slope represents the gradient (direction of steepest increase), and your steps represent parameter updates. Most deep learning algorithms involve optimization, which is the process of finding parameters that minimize or maximize an objective function, making gradient descent the cornerstone of neural network training.

### 4.2.2 Gradient Descent

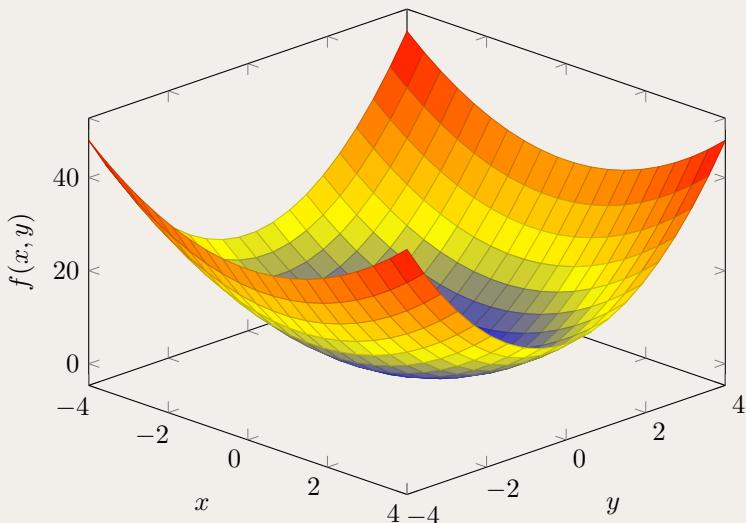
For a function  $f(\theta)$ , **gradient descent** updates parameters as:

$$\theta_{t+1} = \theta_t - \alpha \nabla_{\theta} f(\theta_t) \quad (4.13)$$

where  $\alpha > 0$  is the **learning rate**.

#### Example 4.2

Gradient Descent in 2D Consider minimizing  $f(x, y) = x^2 + 2y^2$  starting from  $(3, 3)$ :



The gradient is  $\nabla f = [2x, 4y]$ . Starting from  $(3, 3)$  with learning rate  $\alpha = 0.1$ :

$$\nabla f(3, 3) = [6, 12] \quad (4.14)$$

$$(x_1, y_1) = (3, 3) - 0.1[6, 12] = (2.4, 1.8) \quad (4.15)$$

$$\nabla f(2.4, 1.8) = [4.8, 7.2] \quad (4.16)$$

$$(x_2, y_2) = (2.4, 1.8) - 0.1[4.8, 7.2] = (1.92, 1.08) \quad (4.17)$$

### 4.2.3 Jacobian and Hessian Matrices

The **Jacobian matrix** contains all first-order partial derivatives. For  $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ :

$$\mathbf{J}_{ij} = \frac{\partial f_i}{\partial x_j} \quad (4.18)$$

The **Hessian matrix** contains second-order derivatives:

$$\mathbf{H}_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j} \quad (4.19)$$

The Hessian characterizes the local curvature of the function. In machine learning and deep learning, Jacobian matrices are essential for backpropagation algorithms to compute gradients efficiently, while Hessian matrices provide information about optimization landscapes and are used in second-order optimization methods like Newton's method and natural gradient descent.

### 4.2.4 Taylor Series Approximation

Near point  $x_0$ , we can approximate  $f(\mathbf{x})$  using Taylor series:

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0)^\top \nabla f(\mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top \mathbf{H}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) \quad (4.20)$$

This provides insight into optimization behavior.

**Example 4.3: Taylor Series for  $f(x) = e^x$  around  $x_0 = 0$**

For the exponential function  $f(x) = e^x$ , we can compute the Taylor series expansion around  $x_0 = 0$ :

Since all derivatives of  $e^x$  are  $e^x$ , and  $e^0 = 1$ , we have:

$$f(0) = 1 \quad (4.21)$$

$$f'(0) = 1 \quad (4.22)$$

$$f''(0) = 1 \quad (4.23)$$

$$f'''(0) = 1 \quad (4.24)$$

$$\vdots \quad (4.25)$$

The Taylor series expansion is:

$$e^x \approx 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \dots \quad (4.26)$$

For small values of  $x$ , we can use the first few terms:

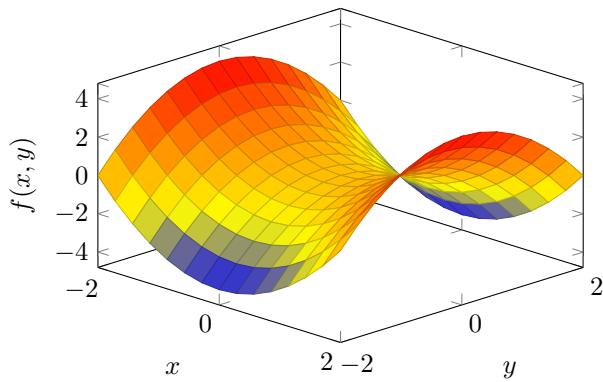
- Linear approximation:  $e^x \approx 1 + x$
- Quadratic approximation:  $e^x \approx 1 + x + \frac{x^2}{2}$
- Cubic approximation:  $e^x \approx 1 + x + \frac{x^2}{2} + \frac{x^3}{6}$

For example,  $e^{0.1} \approx 1 + 0.1 + \frac{0.01}{2} = 1.105$ , which is very close to the true value of approximately 1.1052.

### 4.2.5 Critical Points

Critical points are locations where the gradient is zero, representing different types of terrain features in the optimization landscape. A local minimum is like the bottom of a bowl where you can't go lower in any direction, a local maximum is like the top of a hill where you can't go higher in any direction, and a saddle point is like a mountain pass where you can go down in some directions and up in others. At a critical point,  $\nabla f(\mathbf{x}) = \mathbf{0}$ , and the Hessian matrix determines the nature of the critical

point: a local minimum occurs when the Hessian is positive definite, a local maximum occurs when the Hessian is negative definite, and a saddle point occurs when the Hessian has both positive and negative eigenvalues.



Deep learning often encounters saddle points rather than local minima in high dimensions.

#### 4.2.6 Directional Derivatives

The directional derivative in direction  $\mathbf{u}$  (with  $\|\mathbf{u}\| = 1$ ) is:

$$\frac{\partial}{\partial \alpha} f(\mathbf{x} + \alpha \mathbf{u}) \Big|_{\alpha=0} = \mathbf{u}^\top \nabla f(\mathbf{x}) \quad (4.27)$$

To minimize  $f$ , we move in the direction  $\mathbf{u} = -\frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|}$ .

### 4.3 Constrained Optimization

Constrained optimization extends traditional optimization by incorporating additional requirements or limitations that must be satisfied while finding optimal solutions, providing essential tools for implementing regularization, fairness constraints, and other practical considerations in deep learning applications.

#### 4.3.1 Intuition: Optimization with Rules

Imagine you're trying to find the best location for a new store, but you have constraints such as being within 10 km of the city centre, having parking for at least 50 cars, and a budget that cannot exceed \$1 million. You can't just pick any location - you must follow these rules while still optimizing your objective like maximizing customer traffic. In deep learning, we often have similar constraints including weight constraints to keep weights small and prevent overfitting, probability constraints where outputs must sum to 1 like in softmax functions, and fairness constraints where the model must treat different groups equally. Many problems require optimizing a function subject to constraints, making constrained optimization an essential tool for practical machine learning applications.

#### 4.3.2 Lagrange Multipliers

For equality constraint  $g(\mathbf{x}) = 0$ , the **Lagrangian** is:

$$\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x}) \quad (4.28)$$

At the optimum, both:

$$\nabla_{\mathbf{x}} \mathcal{L} = \mathbf{0} \quad \text{and} \quad \frac{\partial \mathcal{L}}{\partial \lambda} = 0 \quad (4.29)$$

**Remark 4.1**

Lagrange multipliers are crucial in deep learning for implementing regularization techniques like weight decay and dropout, where we optimize the loss function subject to constraints on parameter norms or activation patterns. They also enable the development of constrained optimization algorithms for training neural networks with fairness constraints, ensuring models treat different demographic groups equally while maintaining high performance.

**4.3.3 Inequality Constraints**

For inequality constraint  $g(\mathbf{x}) \leq 0$ , we use the **Karush-Kuhn-Tucker (KKT)** conditions:

$$\nabla_{\mathbf{x}} \mathcal{L} = \mathbf{0} \quad (4.30)$$

$$\lambda \geq 0 \quad (4.31)$$

$$\lambda g(\mathbf{x}) = 0 \quad (\text{complementary slackness}) \quad (4.32)$$

$$g(\mathbf{x}) \leq 0 \quad (4.33)$$

**Remark 4.2**

KKT conditions are essential in deep learning for handling inequality constraints such as bounding neural network weights to prevent exploding gradients, ensuring activation functions stay within valid ranges, and implementing robust optimization with adversarial training where perturbations must remain within specified bounds. They also enable the development of constrained optimization algorithms for training neural networks with fairness constraints, ensuring models treat different demographic groups equally while maintaining high performance.

**4.3.4 Projected Gradient Descent**

For constraints defining a set  $\mathcal{C}$ , **projected gradient descent** applies:

$$\mathbf{x}_{t+1} = \text{Proj}_{\mathcal{C}}(\mathbf{x}_t - \alpha \nabla f(\mathbf{x}_t)) \quad (4.34)$$

where  $\text{Proj}_{\mathcal{C}}$  projects onto the feasible set.

The mathematical operation works as follows: first, we compute the standard gradient descent step  $\mathbf{x}_t - \alpha \nabla f(\mathbf{x}_t)$ , which may move us outside the feasible set  $\mathcal{C}$ . Then, the projection operator  $\text{Proj}_{\mathcal{C}}$  finds the closest point in  $\mathcal{C}$  to this intermediate result, ensuring that  $\mathbf{x}_{t+1} \in \mathcal{C}$ . The projection is defined as:

$$\text{Proj}_{\mathcal{C}}(\mathbf{y}) = \arg \min_{\mathbf{x} \in \mathcal{C}} \|\mathbf{x} - \mathbf{y}\|_2 \quad (4.35)$$

This ensures that each update step remains within the constraint set while still moving in the direction that reduces the objective function.

**Remark 4.3**

Projected gradient descent is essential in deep learning for maintaining constraints during training, such as keeping neural network weights within specified bounds to prevent exploding gradients, ensuring probability outputs sum to 1 in softmax layers, and implementing robust optimization with adversarial training where perturbations must stay within  $\ell_\infty$  balls around input examples.

### 4.3.5 Applications in Deep Learning

Constrained optimization appears in several important applications in deep learning, including weight constraints such as unit norm constraints that help prevent overfitting and improve generalization, projection to valid probability distributions that ensures model outputs are mathematically valid, adversarial training with bounded perturbations that creates robust models by training against carefully crafted adversarial examples, and fairness constraints that ensure models treat different groups equally and avoid discriminatory behavior.

## 4.4 Numerical Stability and Conditioning

Numerical stability and conditioning are crucial considerations in deep learning that determine how sensitive computations are to small errors, affecting the reliability and accuracy of neural network training and inference in real-world applications.

### 4.4.1 Intuition: The Butterfly Effect in Computation

Imagine a house of cards where a tiny breeze can cause the entire structure to collapse. In numerical computation, we have a similar problem where small errors can grow into large ones, especially problematic in deep learning because deep networks have many layers where errors compound, matrix operations can amplify small errors, and gradient computation requires precise derivatives. The condition number tells us how "sensitive" a computation is to small changes, where a high condition number means small input errors become large output errors, making the computation unstable and unreliable.

### 4.4.2 Condition Number

#### Definition 4.1: Condition Number

The condition number of matrix  $A$  is:

$$\kappa(A) = \|A\| \|A^{-1}\| \quad (4.36)$$

#### Definition 4.2: Condition Number for Symmetric Matrices

For symmetric matrices with eigenvalues  $\lambda_i$ :

$$\kappa(A) = \frac{\max_i |\lambda_i|}{\min_i |\lambda_i|} \quad (4.37)$$

High condition numbers indicate numerical instability: small changes in input lead to large changes in output.

#### Example 4.4: Well-Conditioned vs Ill-Conditioned Matrices

Consider two matrices:

$$A_1 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \quad (\text{well-conditioned}) \quad (4.38)$$

$$A_2 = \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix} \quad (\text{ill-conditioned}) \quad (4.39)$$

### 4.4.3 Ill-Conditioned Matrices

In deep learning, ill-conditioned Hessians can make optimization difficult, motivating techniques like batch normalization to stabilize activations, careful weight initialization to avoid poor starting points, adaptive learning rate methods to adjust step sizes dynamically, and preconditioning to improve the conditioning of optimization problems.

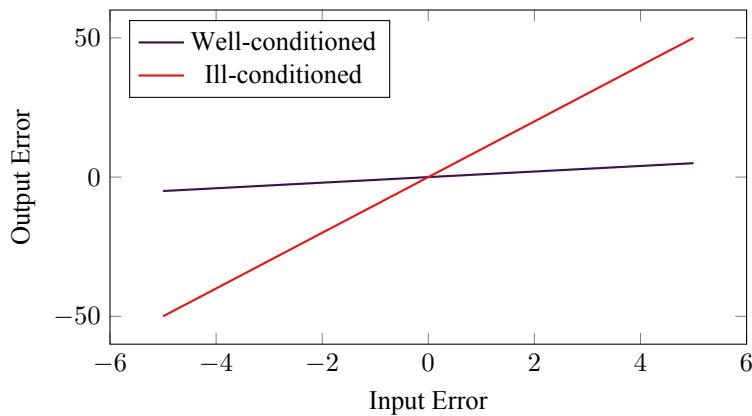


Figure 4.2: Error amplification for well-conditioned vs ill-conditioned matrices

#### 4.4.4 Gradient Checking

To verify gradient computations, we use **finite differences**:

$$\frac{\partial f}{\partial \theta_i} \approx \frac{f(\theta_i + \epsilon) - f(\theta_i - \epsilon)}{2\epsilon} \quad (4.40)$$

Finite differences approximate derivatives by computing the slope of the function between two nearby points, providing a numerical way to verify that analytical gradients are computed correctly. This is computationally expensive but useful for debugging gradient computations in neural networks.

#### 4.4.5 Numerical Precision Trade-offs

Mixed precision training involves storing weights in FP32 for numerical stability while computing activations and gradients in FP16 for computational efficiency, using loss scaling to prevent underflow in the lower precision computations, and achieving 2-3x speedup with minimal accuracy loss by carefully managing the precision trade-offs between numerical stability and computational performance.

##### Remark 4.4: Mixed Precision Training and Quantization

Mixed Precision Training (FP32/FP16/BF16). Activates FP16 mixed precision with ‘fp16=True’ on Hugging Face. Quantization primarily targets inference and memory reduction by converting weights from floating-point (FP32/FP16/BF16) to low-bit integers (INT8, INT4).

#### 4.4.6 Practical Tips

Practical tips for maintaining numerical stability include monitoring gradient norms during training to detect potential instability, using gradient clipping for RNNs to prevent exploding gradients, preferring numerically stable implementations like log-space computations that avoid overflow and underflow, and being aware of precision limits in very deep networks where accumulated errors can become significant.

## Key Takeaways

### Key Takeaways 4

- **Numerical precision** matters: Finite precision arithmetic can cause overflow, underflow, and instability in deep learning computations.
- **Gradient-based optimisation** relies on Jacobian and Hessian matrices to navigate loss landscapes and find optimal parameters.
- **Constrained optimisation** uses Lagrange multipliers and KKT conditions to solve problems with constraints.
- **Numerical stability** is assessed via condition numbers; ill-conditioned problems require careful handling.
- **Practical techniques** like log-sum-exp tricks and gradient checking ensure robust implementations.

## Exercises

### Easy

*Exercise 4.1 (Floating-Point Basics).* Consider a hypothetical 4-bit floating-point system with 1 sign bit, 2 exponent bits, and 1 mantissa bit. What is the smallest positive number that can be represented? What is the largest number?

**Hint:**

Use the IEEE 754 format:  $(-1)^s \times 2^{e-b} \times (1+m)$  where  $s$  is the sign bit,  $e$  is the exponent,  $b$  is the bias, and  $m$  is the mantissa.

*Exercise 4.2 (Softmax Stability).* Compute the softmax of  $\mathbf{x} = [1000, 1001, 1002]$  using both the naive approach and the numerically stable approach. Show your work step by step.

**Hint:**

Use the identity  $\text{softmax}(\mathbf{x}) = \text{softmax}(\mathbf{x} - c)$  where  $c = \max_i x_i$ .

*Exercise 4.3 (Gradient Computation).* Compute the gradient of  $f(x, y) = x^2 + 3xy + y^2$  at the point  $(2, 1)$ .

**Hint:**

The gradient is  $\nabla f = \left[ \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right]$ .

*Exercise 4.4 (Condition Number).* Calculate the condition number of the matrix  $\mathbf{A} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ .

**Hint:**

For a  $2 \times 2$  matrix,  $\kappa(\mathbf{A}) = \frac{\lambda_{\max}}{\lambda_{\min}}$  where  $\lambda_{\max}$  and  $\lambda_{\min}$  are the eigenvalues.

*Exercise 4.5 (Numerical Differentiation).* Implement the forward, backward, and central difference formulas for computing the derivative of  $f(x) = \sin(x)$  at  $x = \pi/4$ . Compare the accuracy with the analytical derivative.

**Hint:**

Use  $f'(x) \approx \frac{f(x+h) - f(x)}{h}$  for forward difference and similar formulas for other methods.

*Exercise 4.6 (Machine Epsilon).* Write a program to find the machine epsilon of your system. What is the smallest number  $\epsilon$  such that  $1 + \epsilon \neq 1$  in floating-point arithmetic?

**Hint:**

*Start with  $\epsilon = 1$  and repeatedly divide by 2 until  $1 + \epsilon = 1$ .*

*Exercise 4.7 (Numerical Integration).* Compare the trapezoidal rule and Simpson's rule for approximating  $\int_0^1 e^{-x^2} dx$ . Use  $n = 4, 8, 16$  subintervals.

**Hint:**

*Trapezoidal rule:  $\int_a^b f(x) dx \approx \frac{h}{2} [f(a) + 2 \sum_{i=1}^{n-1} f(x_i) + f(b)]$ .*

*Exercise 4.8 (Matrix Conditioning).* Given the matrix  $A = \begin{bmatrix} 1 & 1 \\ 1 & 1.0001 \end{bmatrix}$ , compute its condition number and solve  $Ax = b$  where  $b = [2, 2.0001]^T$ .

**Hint:**

*A small change in  $b$  can cause a large change in the solution when the condition number is large.*

## Medium

*Exercise 4.9 (Log-Sum-Exp Trick).* Derive the log-sum-exp trick:  $\log(\sum_{i=1}^n \exp(x_i)) = c + \log(\sum_{i=1}^n \exp(x_i - c))$  where  $c = \max_i x_i$ .

**Hint:**

*Start by factoring out  $\exp(c)$  from the sum.*

*Exercise 4.10 (Lagrange Multipliers).* Find the maximum value of  $f(x, y) = xy$  subject to the constraint  $x^2 + y^2 = 1$  using Lagrange multipliers.

**Hint:**

*Set up the Lagrangian  $\mathcal{L}(x, y, \lambda) = xy + \lambda(1 - x^2 - y^2)$  and solve the system of equations.*

*Exercise 4.11 (Gradient Descent Convergence).* For the function  $f(x, y) = x^2 + 100y^2$ , implement gradient descent with different learning rates. Show that the convergence rate depends on the condition number.

**Hint:**

*The condition number of the Hessian matrix determines the convergence rate.*

*Exercise 4.12 (Newton's Method).* Use Newton's method to find the root of  $f(x) = x^3 - 2x - 5$  starting from  $x_0 = 2$ . Compare with the bisection method.

**Hint:**

*Newton's method:  $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$ .*

*Exercise 4.13 (Eigenvalue Computation).* For the matrix  $\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$ , compute the eigenvalues and eigenvectors using the characteristic equation and verify with numerical methods.

**Hint:**

Solve  $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$  for eigenvalues.

## Hard

*Exercise 4.14 (Numerical Stability of Matrix Inversion).* Consider the matrix  $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 + \epsilon \end{bmatrix}$  where  $\epsilon$  is small. Show that the condition number grows as  $\epsilon \rightarrow 0$ . Implement a numerical experiment to demonstrate this and show how the error in  $\mathbf{A}^{-1}$  grows.

**Hint:**

Use the formula  $\kappa(\mathbf{A}) = \frac{\lambda_{\max}}{\lambda_{\min}}$  and compute the eigenvalues analytically.

*Exercise 4.15 (KKT Conditions Application).* Consider the optimisation problem:

$$\min_{x,y} \quad x^2 + y^2 \tag{4.41}$$

$$\text{subject to} \quad x + y \geq 1 \tag{4.42}$$

$$x \geq 0, y \geq 0 \tag{4.43}$$

Find the optimal solution using the KKT conditions and verify that all conditions are satisfied.

**Hint:**

Set up the Lagrangian with multiple constraints and check the complementary slackness conditions.

# Chapter 5

## Classical Machine Learning Algorithms

This chapter reviews traditional machine learning methods that provide context and motivation for deep learning approaches. Understanding these classical algorithms helps appreciate the advantages and innovations of deep learning.

### Learning Objectives

1. Mathematical foundations of classical machine learning algorithms
2. Comparison of different approaches to classification and regression
3. Optimization of classical algorithms using closed-form and iterative methods
4. Ensemble methods: random forests and gradient boosting
5. Trade-offs between classical methods and deep learning
6. Algorithm selection based on dataset characteristics and constraints
7. Limitations of classical methods that motivated deep learning
8. Regularization techniques to prevent overfitting in classical models

This chapter assumes familiarity with linear algebra, probability theory, and basic optimization concepts from previous chapters.

### 5.1 Linear Regression ◆

**Linear regression** is one of the most fundamental and widely-used machine learning algorithms. It models the relationship between input features and a continuous output by finding the best linear function that minimizes prediction errors.

#### 5.1.1 Intuition and Motivation

Imagine you're trying to predict house prices based on features like size, number of bedrooms, and location. Linear regression assumes that the price can be expressed as a weighted sum of these features plus a base price (bias), where the algorithm learns the optimal weights that best explain the relationship between features and prices in your training data. The key insight is that linear relationships are often sufficient for many real-world problems, and they have several advantages including interpretability where each weight tells us how much the output changes when a feature increases by one unit, computational efficiency with fast training and prediction, and statistical properties with well-understood theoretical guarantees that make the method reliable and predictable.

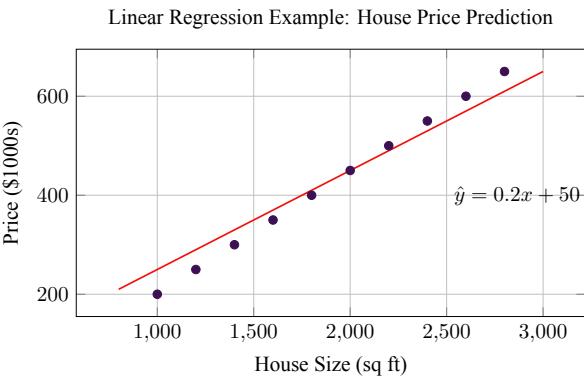


Figure 5.1: Linear regression finds best line fitting data points, minimizing sum of squared errors.

### 5.1.2 Model Formulation

For input  $\mathbf{x} \in \mathbb{R}^d$  and output  $y \in \mathbb{R}$ , linear regression models the relationship as  $\hat{y} = \mathbf{w}^\top \mathbf{x} + b$ , where  $\mathbf{w} \in \mathbb{R}^d$  are the weights (regression coefficients) that determine how much each feature contributes to the prediction,  $b \in \mathbb{R}$  is the bias (intercept term) that represents the base value when all features are zero, and  $\hat{y}$  is the predicted output that represents our model's estimate of the true target value.

### 5.1.3 Ordinary Least Squares

The goal is to find parameters that minimize the prediction error. We use the **mean squared error** (MSE) as our loss function:

$$L(\mathbf{w}, b) = \frac{1}{n} \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2 = \frac{1}{n} \sum_{i=1}^n (y^{(i)} - \mathbf{w}^\top \mathbf{x}^{(i)} - b)^2 \quad (5.1)$$

#### Matrix Formulation

For computational efficiency, we can absorb the bias into the weight vector by adding a constant feature of 1 to each input. Let  $\mathbf{X} \in \mathbb{R}^{n \times (d+1)}$  be the design matrix with an additional column of ones, and  $\mathbf{w} \in \mathbb{R}^{d+1}$  include the bias term.

The closed-form solution (normal equation) is:

$$\mathbf{w}^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \quad (5.2)$$

#### Remark 5.1

The normal equation requires  $\mathbf{X}^\top \mathbf{X}$  to be invertible. This condition is satisfied when the features are linearly independent and we have at least as many training examples as features.

### 5.1.4 Regularized Regression

When we have many features or when features are correlated, the normal equation can become unstable. Regularization helps by adding a penalty term to prevent overfitting.

#### Ridge Regression (L2 Regularization)

**Ridge regression** adds an L2 penalty to the loss function:

$$L(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda \|\mathbf{w}\|^2 \quad (5.3)$$

The solution becomes:

$$\mathbf{w}^* = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y} \quad (5.4)$$

where  $\lambda > 0$  is the regularization strength.

### Example 5.1

For a simple 2D case with features  $x_1$  and  $x_2$ , ridge regression finds:

$$\hat{y} = w_1x_1 + w_2x_2 + b$$

The L2 penalty  $\lambda(w_1^2 + w_2^2)$  encourages smaller weights, leading to a smoother, more stable solution.

### Lasso Regression (L1 Regularization)

**Lasso regression** uses L1 regularization, which promotes sparsity:

$$L(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|_1 \quad (5.5)$$

Unlike ridge regression, lasso can drive some weights to exactly zero, effectively performing automatic feature selection.

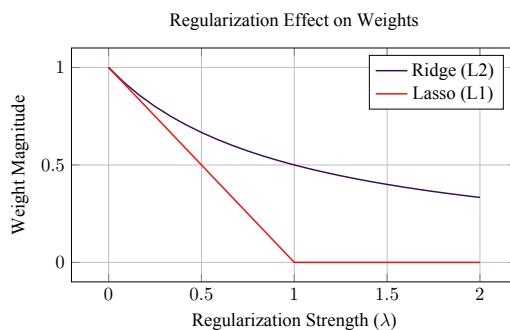


Figure 5.2: L1 vs L2 regularization: L1 drives weights to zero, L2 shrinks them smoothly.

### 5.1.5 Gradient Descent Solution

For large datasets, computing the inverse of  $\mathbf{X}^\top \mathbf{X}$  can be computationally expensive. Gradient descent provides an iterative alternative:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha \nabla_{\mathbf{w}} L(\mathbf{w}_t) \quad (5.6)$$

where the gradient is:

$$\nabla_{\mathbf{w}} L(\mathbf{w}) = \frac{2}{n} \mathbf{X}^\top (\mathbf{X}\mathbf{w} - \mathbf{y}) \quad (5.7)$$

### Stochastic Gradient Descent

For very large datasets, we can use stochastic gradient descent (SGD), which updates weights using only a subset of the data at each iteration:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha \nabla_{\mathbf{w}} L_i(\mathbf{w}_t) \quad (5.8)$$

where  $L_i$  is the loss for a single training example or a small batch.

**Remark 5.2**

The "stochastic" approach uses random sampling of data points instead of the entire dataset, introducing noise that helps escape local minima and provides better generalization by preventing overfitting to the specific training set. This stochasticity also enables faster convergence and more efficient memory usage, making it the preferred method for training large neural networks on massive datasets.

### 5.1.6 Geometric Interpretation

Linear regression can be understood geometrically as finding the projection of the target vector  $\mathbf{y}$  onto the column space of the design matrix  $\mathbf{X}$ . The residual vector  $\mathbf{y} - \mathbf{X}\mathbf{w}^*$  is orthogonal to the column space of  $\mathbf{X}$ .

**Theorem 5.1: Orthogonality Principle**

The optimal solution  $\mathbf{w}^*$  satisfies:

$$\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\mathbf{w}^*) = \mathbf{0}$$

This means the residual vector is orthogonal to all feature vectors.

## 5.2 Logistic Regression ◆

**Logistic regression** is a fundamental classification algorithm that models the probability of class membership using a logistic (sigmoid) function. Despite its name, it's actually a classification method, not a regression method.

### 5.2.1 Intuition and Motivation

Logistic regression extends linear regression to handle classification problems. Instead of predicting continuous values, it predicts probabilities that an input belongs to a particular class. The key insight is to use a sigmoid function to map linear combinations of features to probabilities between 0 and 1.

Think of logistic regression as answering: "Given these features, what's the probability that this example belongs to the positive class?" For example, given a patient's symptoms, what's the probability they have a particular disease?

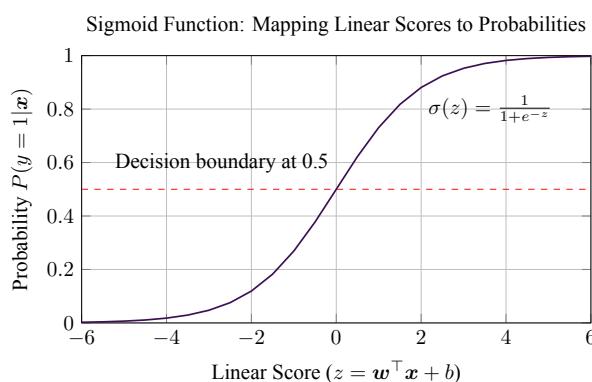


Figure 5.3: Sigmoid function maps real numbers to probabilities (0,1). Decision boundary at 0.5.

## 5.2.2 Binary Classification

For binary classification with classes  $\{0, 1\}$ , logistic regression models the probability of the positive class using the sigmoid function  $\sigma(z) = \frac{1}{1+e^{-z}}$ , which smoothly maps any real number to a probability between 0 and 1. The prediction probability is:

$$P(y = 1|\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b) = \frac{1}{1 + e^{-(\mathbf{w}^\top \mathbf{x} + b)}} \quad (5.9)$$

where the linear combination of features and weights determines the input to the sigmoid function, and the output represents the probability that the input belongs to the positive class.

### Remark 5.3

For categorical classification with more than 2 classes, we extend logistic regression to softmax regression, which uses the softmax function to compute probabilities for each class that sum to 1. This allows us to handle multi-class problems like image classification with 10 digit classes or natural language processing with hundreds of word categories.

## Properties of the Sigmoid Function

The sigmoid function has several important properties:

- **Range:**  $\sigma(z) \in (0, 1)$  for all  $z \in \mathbb{R}$
- **Monotonic:**  $\sigma'(z) = \sigma(z)(1 - \sigma(z)) > 0$  for all  $z$
- **Symmetric:**  $\sigma(-z) = 1 - \sigma(z)$
- **Asymptotic:**  $\lim_{z \rightarrow \infty} \sigma(z) = 1$  and  $\lim_{z \rightarrow -\infty} \sigma(z) = 0$

## 5.2.3 Cross-Entropy Loss

Unlike linear regression, we can't use mean squared error for classification because the sigmoid function is non-linear. Instead, we use the cross-entropy loss (negative log-likelihood):

### Definition 5.1: Cross-Entropy Loss

The cross-entropy loss for binary classification is:

$$L(\mathbf{w}, b) = -\frac{1}{n} \sum_{i=1}^n \left[ y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \right] \quad (5.10)$$

where  $\hat{y}^{(i)} = P(y = 1|\mathbf{x}^{(i)})$ .

### Remark 5.4

Cross-entropy loss is the standard loss function for classification tasks in both classical machine learning and deep learning, providing better gradient properties than mean squared error for probability outputs. In deep neural networks, it's commonly used in the final layer for binary and multi-class classification, enabling efficient backpropagation and stable training across various architectures from simple logistic regression to complex convolutional and transformer networks.

**Definition 5.2: Derivation of Cross-Entropy Loss**

The cross-entropy loss comes from maximum likelihood estimation. For a single example, the likelihood is:

$$L_i = P(y^{(i)} | \mathbf{x}^{(i)}) = (\hat{y}^{(i)})^{y^{(i)}} (1 - \hat{y}^{(i)})^{1-y^{(i)}}$$

Taking the negative log-likelihood:

$$-\log L_i = -y^{(i)} \log \hat{y}^{(i)} - (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})$$

### 5.2.4 Gradient Descent for Logistic Regression

The gradient of the cross-entropy loss with respect to the weights is:

$$\nabla_{\mathbf{w}} L = \frac{1}{n} \sum_{i=1}^n (\hat{y}^{(i)} - y^{(i)}) \mathbf{x}^{(i)} \quad (5.11)$$

The weight update rule becomes:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha \frac{1}{n} \sum_{i=1}^n (\hat{y}^{(i)} - y^{(i)}) \mathbf{x}^{(i)} \quad (5.12)$$

**Remark 5.5**

The gradient has a simple form: it's the average of the prediction errors multiplied by the input features. This makes logistic regression particularly efficient to train.

### 5.2.5 Multiclass Classification

For  $K$  classes, we extend logistic regression to softmax regression (also called multinomial logistic regression), where instead of a single sigmoid function, we use the softmax function to compute the probability of each class:

$$P(y = k | \mathbf{x}) = \frac{\exp(\mathbf{w}_k^\top \mathbf{x} + b_k)}{\sum_{j=1}^K \exp(\mathbf{w}_j^\top \mathbf{x} + b_j)} \quad (5.13)$$

The softmax function normalizes the exponential scores so that they sum to 1, creating a valid probability distribution over all possible classes.

#### Properties of Softmax

The softmax function has several key properties:

- **Probability distribution:**  $\sum_{k=1}^K P(y = k | \mathbf{x}) = 1$
- **Non-negative:**  $P(y = k | \mathbf{x}) \geq 0$  for all  $k$
- **Monotonic:** Higher scores lead to higher probabilities
- **Scale invariant:** Adding a constant to all scores doesn't change probabilities

### 5.2.6 Categorical Cross-Entropy Loss

For multiclass classification, we use the categorical cross-entropy loss:

$$L = -\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K y_k^{(i)} \log \hat{y}_k^{(i)} \quad (5.14)$$

where  $y_k^{(i)}$  is 1 if example  $i$  belongs to class  $k$ , and 0 otherwise (one-hot encoding).

### 5.2.7 Decision Boundaries

Logistic regression creates linear decision boundaries. For binary classification, the decision boundary is the hyperplane where  $P(y = 1|\mathbf{x}) = 0.5$ , which occurs when  $\mathbf{w}^\top \mathbf{x} + b = 0$ .

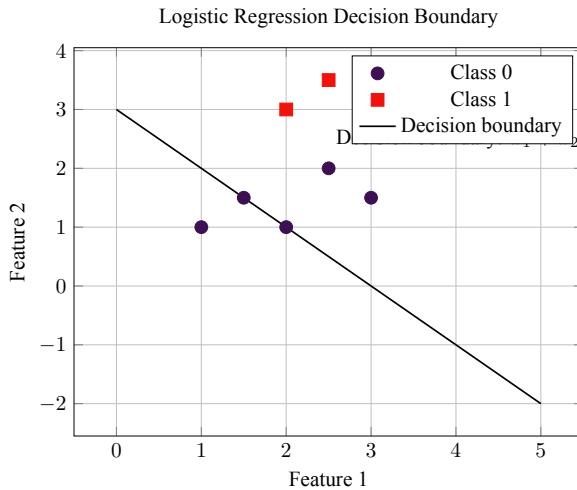


Figure 5.4: Logistic regression finds linear decision boundary separating classes 0 and 1.

### 5.2.8 Regularization in Logistic Regression

Just like linear regression, logistic regression can benefit from regularization to prevent overfitting:

#### L2 Regularization (Ridge)

$$L(\mathbf{w}) = -\frac{1}{n} \sum_{i=1}^n \left[ y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \right] + \lambda \|\mathbf{w}\|^2 \quad (5.15)$$

#### L1 Regularization (Lasso)

$$L(\mathbf{w}) = -\frac{1}{n} \sum_{i=1}^n \left[ y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \right] + \lambda \|\mathbf{w}\|_1 \quad (5.16)$$

#### Remark 5.6

The key difference between L1 and L2 regularization lies in their penalty shapes: L2 uses the sum of squared weights  $\|\mathbf{w}\|^2$  which creates smooth, continuous shrinkage of all weights, while L1 uses the sum of absolute weights  $\|\mathbf{w}\|_1$  which can drive some weights to exactly zero, performing automatic feature selection. L2 is better for preventing overfitting when all features are relevant, while L1 is preferred when you want to identify and remove irrelevant features.

### 5.2.9 Advantages and Limitations

Logistic regression has several advantages including being simple and interpretable with clear mathematical foundations, fast training and prediction due to efficient gradient computation, providing probability estimates that are useful for decision-making, working well with small datasets where complex models might overfit, and making no assumptions about feature distributions. However, it also has limitations including assuming a linear relationship between features and log-odds which may not hold for complex data, being sensitive to outliers that can significantly affect the decision boundary, potentially not working well with highly correlated features that can cause numerical instability, and being limited to linear decision boundaries which may not be sufficient for non-linearly separable data.

## 5.3 Support Vector Machines ◆

**Support Vector Machines** (SVMs) are powerful classification algorithms that find the optimal hyperplane that maximally separates different classes. The key insight is to maximize the margin between classes, leading to better generalization performance.

### Remark 5.7

The name "Support Vector Machine" comes from the critical role that support vectors play in defining the optimal hyperplane: these are the data points closest to the decision boundary that literally "support" the entire structure of the classifier. If you move any other data point, the hyperplane doesn't change, but if you move a support vector, the entire hyperplane shifts, demonstrating that the entire "machine" is defined and supported by these few critical vectors.

### 5.3.1 Intuition and Motivation

Imagine you have two groups of points on a plane that you want to separate with a line, where there are many possible lines that could separate them, but SVM finds the line that maximizes the distance to the nearest points from each class. This "maximum margin" approach leads to better generalization because the decision boundary is as far as possible from both classes, making the model more robust to new data. The key concepts include support vectors which are the training examples closest to the decision boundary and determine its position, the margin which is the distance between the decision boundary and the nearest support vectors, and the maximum margin principle which chooses the hyperplane that maximizes this margin to achieve the best possible separation.

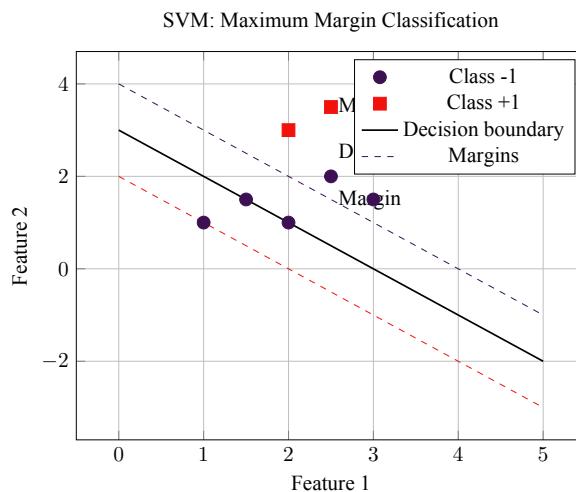


Figure 5.5: SVM finds hyperplane maximizing margin between classes. Support vectors are closest to boundary.

### 5.3.2 Linear SVM

For binary classification with labels  $y \in \{-1, +1\}$ , the decision boundary is:

$$\mathbf{w}^\top \mathbf{x} + b = 0 \quad (5.17)$$

The **margin** is the distance between the decision boundary and the nearest support vectors. For a point  $\mathbf{x}^{(i)}$ , the distance to the hyperplane is:

$$\text{distance} = \frac{|y^{(i)}(\mathbf{w}^\top \mathbf{x}^{(i)} + b)|}{\|\mathbf{w}\|} \quad (5.18)$$

Since we want to maximize the margin, we can set the margin to be  $\frac{2}{\|\mathbf{w}\|}$  by requiring:

$$y^{(i)}(\mathbf{w}^\top \mathbf{x}^{(i)} + b) \geq 1 \quad \forall i \quad (5.19)$$

### Optimization Problem

Maximizing the margin is equivalent to minimizing  $\|\mathbf{w}\|^2$  subject to the constraints:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 \quad (5.20)$$

subject to:

$$y^{(i)}(\mathbf{w}^\top \mathbf{x}^{(i)} + b) \geq 1 \quad \forall i \quad (5.21)$$

This is a quadratic programming problem that can be solved using Lagrange multipliers.

### 5.3.3 Soft Margin SVM

In practice, data is rarely linearly separable, so the soft margin SVM allows some training examples to be misclassified by introducing slack variables  $\xi_i$  that measure how much each point violates the margin constraint. The optimization problem becomes  $\min_{\mathbf{w}, b, \xi} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i$  subject to  $y^{(i)}(\mathbf{w}^\top \mathbf{x}^{(i)} + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$ , where the parameter  $C$  controls the trade-off between large margin (small  $C$  allows more slack and larger margin) and training accuracy (large  $C$  penalizes misclassifications more heavily).

### 5.3.4 Dual Formulation

The SVM optimization problem can be reformulated in its dual form, which reveals the support vectors and enables the kernel trick:

$$\max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y^{(i)} y^{(j)} \mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)} \quad (5.22)$$

subject to:

$$\sum_{i=1}^n \alpha_i y^{(i)} = 0, \quad 0 \leq \alpha_i \leq C \quad (5.23)$$

The decision function becomes:

$$f(\mathbf{x}) = \sum_{i=1}^n \alpha_i y^{(i)} \mathbf{x}^{(i)} \cdot \mathbf{x} + b \quad (5.24)$$

Only examples with  $\alpha_i > 0$  are support vectors.

### 5.3.5 Kernel Trick

For non-linear decision boundaries, we can map inputs to a higher-dimensional space using a kernel function  $k(\mathbf{x}, \mathbf{x}')$  that computes the inner product in the transformed space without explicitly computing the transformation. The kernel trick is used in machine learning to handle non-linear relationships by implicitly mapping data to higher-dimensional spaces where linear separation becomes possible, enabling SVMs to learn complex decision boundaries while maintaining computational efficiency.

#### Common Kernels

**Linear kernel:**

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}' \quad (5.25)$$

**Polynomial kernel:**

$$k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^\top \mathbf{x}' + c)^d \quad (5.26)$$

**RBF (Gaussian) kernel:**

$$k(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2) \quad (5.27)$$

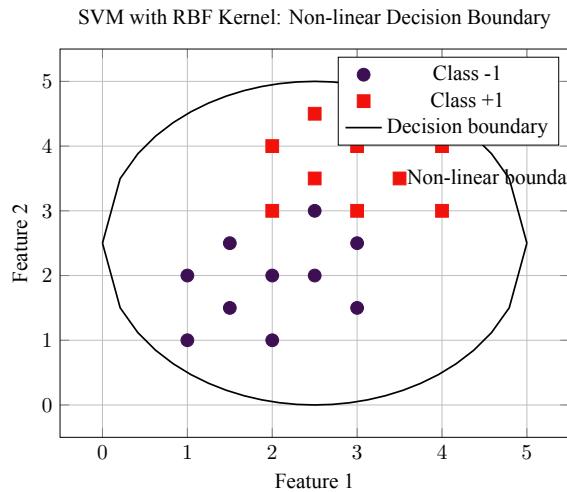


Figure 5.6: SVM with RBF kernel learns non-linear decision boundaries (circular here).

### 5.3.6 Kernel Properties

A function  $k(\mathbf{x}, \mathbf{x}')$  is a valid kernel if and only if it is symmetric where  $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}', \mathbf{x})$ , and positive semi-definite where for any set of points  $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\}$ , the kernel matrix  $K_{ij} = k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$  is positive semi-definite. These properties ensure that the kernel function represents a valid inner product in some feature space, making it suitable for use in kernel-based machine learning algorithms.

### 5.3.7 Advantages and Limitations

SVMs have several advantages including being effective in high-dimensional spaces where the curse of dimensionality affects other methods less, being memory efficient by only storing support vectors rather than all training data, being versatile with different kernel functions that can handle various data types, having a strong theoretical foundation with well-understood generalization bounds, and working well with small to medium datasets where they can achieve good performance. However, they also have limitations including poor performance on large datasets where training time becomes prohibitive, being sensitive to feature scaling which requires careful preprocessing, providing no probabilistic output which limits their use in applications requiring uncertainty quantification, having kernel selection that can be tricky and often requires domain expertise, and being computationally expensive for very large datasets where other methods might be more practical.

### 5.3.8 SVM for Regression

SVM can also be extended to regression problems (Support Vector Regression, SVR). Instead of finding a hyperplane that separates classes, SVR finds a hyperplane that fits the data within an  $\epsilon$ -tube:

$$\min_{\mathbf{w}, b, \xi, \xi^*} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \quad (5.28)$$

subject to:

$$y^{(i)} - \mathbf{w}^\top \mathbf{x}^{(i)} - b \leq \epsilon + \xi_i \quad (5.29)$$

$$\mathbf{w}^\top \mathbf{x}^{(i)} + b - y^{(i)} \leq \epsilon + \xi_i^* \quad (5.30)$$

$$\xi_i, \xi_i^* \geq 0 \quad (5.31)$$

## 5.4 Decision Trees and Ensemble Methods ◆

**Decision trees** are intuitive, interpretable models that make predictions by asking a series of yes/no questions about the input features. While individual trees can be prone to overfitting, combining multiple trees through ensemble methods often leads to much better performance.

### 5.4.1 Intuition and Motivation

Think of a decision tree as a flowchart for making decisions, where to classify whether someone will buy a product, you might ask "Is their income > \$50k?" and if yes, ask "Are they under 30?" or if no, ask "Do they have children?" Each question splits the data into smaller, more homogeneous groups that become easier to classify. The key advantages of decision trees include interpretability where they are easy to understand and explain, no feature scaling requirements since they work with mixed data types, handling missing values by being able to deal with incomplete data, and being non-parametric with no assumptions about data distribution, making them flexible and robust across different problem domains.

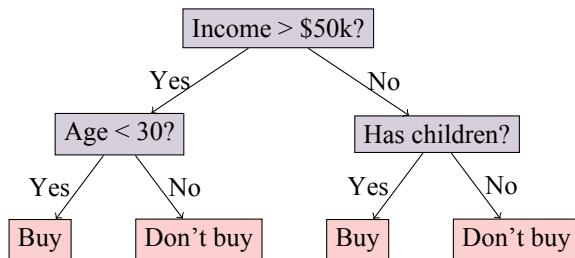


Figure 5.7: Decision tree for product purchases: internal nodes are decisions, leaves are predictions.

### 5.4.2 Decision Trees

A decision tree recursively partitions the input space based on feature values, where the algorithm works by starting with all training examples at the root, finding the best feature and threshold to split on that maximizes information gain, creating child nodes for each split that represent the different outcomes, repeating recursively until a stopping criterion is met such as maximum depth or minimum samples per leaf, and assigning a prediction to each leaf node based on the majority class or mean value of the examples that reach that node.

### Example 5.2: Decision Tree: "Will we go out?"

Consider the decision of whether to go out based on four features:

- **Is it raining?** (Yes/No)
- **Temperature** (Hot/Mild/Cold)
- **Traffic OK?** (Yes/No)
- **Anyone goes with me?** (Yes/No)

A decision tree might look like:

#### 1. Is it raining?

- Yes → **Don't go out** (regardless of other factors)
- No → Continue to next question

#### 2. Temperature?

- Hot → **Don't go out** (too hot)
- Mild/Cold → Continue to next question

#### 3. Traffic OK?

- No → **Don't go out** (traffic is bad)
- Yes → Continue to next question

#### 4. Anyone goes with me?

- Yes → **Go out** (company makes it worthwhile)
- No → **Don't go out** (not worth going alone)

This tree shows how decision trees make sequential decisions, where each question splits the data based on the most important feature at that point, leading to a final prediction.

## Splitting Criteria

The key question is: "Which feature and threshold should we use to split the data?" We want splits that create the most homogeneous child nodes.

**For classification:**

**Gini impurity:**

$$\text{Gini} = 1 - \sum_{k=1}^K p_k^2 \quad (5.32)$$

**Entropy:**

$$\text{Entropy} = - \sum_{k=1}^K p_k \log p_k \quad (5.33)$$

**For regression:**

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y^{(i)} - \bar{y})^2 \quad (5.34)$$

where  $p_k$  is the proportion of class  $k$  examples in a node, and  $\bar{y}$  is the mean target value.

## Information Gain

The **information gain** measures how much a split reduces impurity:

$$IG = \text{Impurity}(\text{parent}) - \sum_j \frac{n_j}{n} \text{Impurity}(\text{child}_j) \quad (5.35)$$

We choose the split that maximizes information gain.

### 5.4.3 Random Forests

Random forests address the overfitting problem of individual trees by combining multiple trees trained on different subsets of the data, where each tree is trained on a bootstrap sample of the original data and considers only a random subset of features at each split, creating diversity among the trees that reduces overfitting and improves generalization performance.

#### Bootstrap Aggregating (Bagging)

Random forests inject diversity through two complementary mechanisms. First, each tree is trained on a *bootstrap sample*—a dataset created by sampling with replacement from the original training set—so trees see different subsets of examples and learn different decision boundaries. Second, at each split within a tree, the algorithm considers only a *random subset of features* rather than all features, forcing trees to rely on different signals and preventing a few strong predictors from dominating every split. At inference time, the forest aggregates individual tree predictions—by averaging for regression or majority voting for classification—so the ensemble reduces variance relative to any single overfit tree while keeping low bias.

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B f_b(\mathbf{x}) \quad (5.36)$$

where  $B$  is the number of trees.

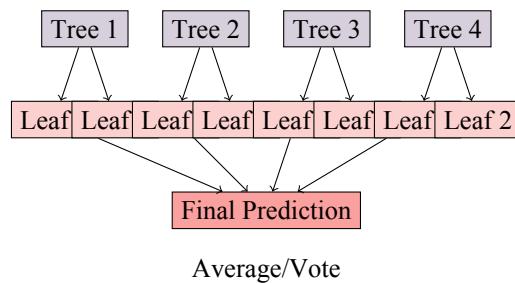


Figure 5.8: Random forests: multiple trees trained on bootstrap samples, final prediction by averaging/voting.

#### Advantages of Random Forests

Random forests substantially reduce overfitting by averaging many de-correlated trees, which lowers variance without greatly increasing bias. They provide built-in *feature importance* estimates that highlight which inputs most influence predictions, helping interpretation and feature selection. Because each tree is trained on a bootstrap sample, the ensemble is naturally robust to outliers and noisy examples that might mislead a single tree. Training is trivially parallelizable since trees are independent, and, like decision trees, random forests require no feature scaling and handle mixed data types gracefully.

### 5.4.4 Gradient Boosting

Gradient boosting builds an ensemble sequentially, where each new tree corrects the errors of the previous ensemble by learning to predict the residuals of the current model, unlike random forests where trees are trained independently. This sequential approach allows the ensemble to focus on the most difficult examples and gradually improve its performance through iterative refinement.

## Algorithm

### Algorithm 1 Gradient Boosting Algorithm

```

1: Initialize  $\hat{y}^{(0)} = \frac{1}{n} \sum_{i=1}^n y^{(i)}$                                      ▷ Start with mean for regression
2: for  $m = 1$  to  $M$  do
3:   Compute residuals:  $r_i^{(m)} = y^{(i)} - \hat{y}^{(m-1)}(\mathbf{x}^{(i)})$  for  $i = 1, \dots, n$ 
4:   Fit tree  $f_m$  to residuals  $\{(\mathbf{x}^{(i)}, r_i^{(m)})\}_{i=1}^n$ 
5:   Update:  $\hat{y}^{(m)} = \hat{y}^{(m-1)} + \nu f_m(\mathbf{x}^{(i)})$  for all  $i$ 
6: end for
7: Return final ensemble  $\hat{y}^{(M)}$ 
```

where  $\nu$  is the learning rate (shrinkage parameter) and  $M$  is the number of boosting iterations.

## Intuition

Gradient boosting works by starting with a simple model like the mean for regression, calculating the errors (residuals) of the current model to identify where it fails, training a new model to predict these errors and correct the mistakes, adding the new model to the ensemble with a small weight to avoid overfitting, and repeating this process until convergence where the ensemble can no longer be improved significantly.

### Example 5.3

For regression with target values [10, 20, 30, 40]:

- 1: Initial prediction:  $\hat{y}^{(0)} = 25$  ▷ mean of targets
- 2: Residuals:  $[-15, -5, 5, 15]$  ▷  $r_i = y^{(i)} - \hat{y}^{(0)}$
- 3: Fit a small tree to residuals  $\{(\mathbf{x}^{(i)}, r_i)\}$
- 4: Update ensemble:  $\hat{y}^{(1)}(\mathbf{x}) = 25 + \nu f_1(\mathbf{x})$
- 5: Recompute residuals and repeat for  $m = 2, \dots, M$

## 5.4.5 Advanced Ensemble Methods

Advanced ensemble methods extend the basic boosting and bagging approaches with sophisticated techniques that improve performance and efficiency. AdaBoost (Adaptive Boosting) is an early boosting algorithm that assigns higher weights to misclassified examples, combines weak learners with weights based on their performance, and focuses on the hardest examples to improve the overall ensemble. Modern gradient boosting implementations like XGBoost and LightGBM add regularization with L1 and L2 penalties to prevent overfitting, pruning to remove splits that don't improve performance, feature subsampling with random feature selection at each split, and efficient implementation optimized for speed and memory usage.

## 5.4.6 Comparison of Ensemble Methods

Method	Bias	Variance	Interpretability
Single Tree	Low	High	High
Random Forest	Low	Medium	Medium
Gradient Boosting	Low	Low	Low

Table 5.1: Tree-based methods comparison.

## 5.4.7 Advantages and Limitations

Decision trees and ensemble methods have several advantages including being interpretable and easy to understand and explain, being flexible by handling mixed data types and missing values, making no assumptions about data distribution, providing

feature importance measures that can identify important features, and being robust with less sensitivity to outliers than linear methods. However, they also have limitations including overfitting where individual trees can overfit easily, instability where small changes in data can lead to very different trees, computational cost where ensemble methods can be slow to train, memory usage where storing many trees requires significant memory, and being less effective with high-dimensional data where performance can degrade with many features.

## 5.5 k-Nearest Neighbors ◆

**k-Nearest Neighbors** (k-NN) is a simple yet powerful non-parametric algorithm that makes predictions based on the similarity to training examples. It's called "lazy learning" because it doesn't build a model during training—all computation happens at prediction time.

### 5.5.1 Intuition and Motivation

The k-NN algorithm is based on a simple principle: "similar things are close to each other." If you want to predict whether someone will like a movie, look at what movies similar people (with similar tastes) liked. If you want to predict house prices, look at prices of similar houses in the neighborhood.

The key insight is that we can make good predictions by finding the most similar examples in our training data and using their outcomes as a guide.

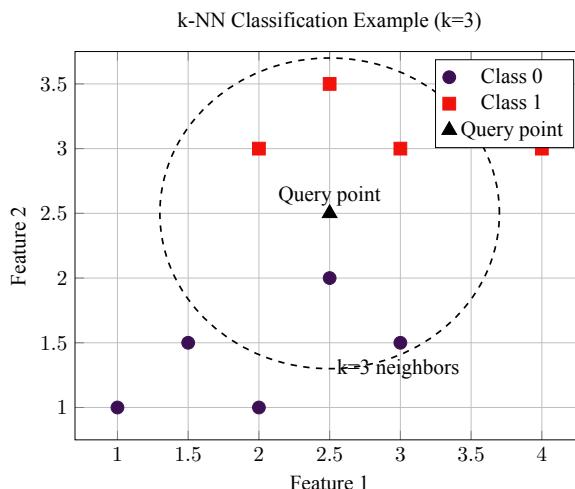


Figure 5.9: k-NN finds k nearest neighbors. With k=3, query classified as class 0 (2/3 neighbors).

### 5.5.2 Algorithm

For a query point  $\mathbf{x}$ , the k-NN algorithm works by finding the  $k$  closest training examples based on a distance metric, then for classification returning the majority class among the  $k$  neighbors, and for regression returning the average of the target values of the  $k$  neighbors. The key ideas of the algorithm include using distance-based similarity to find relevant examples, leveraging the assumption that similar examples have similar outcomes, and making predictions based on local patterns in the data rather than global models.

#### Mathematical Formulation

For classification, the predicted class is:

$$\hat{y} = \arg \max_c \sum_{i=1}^k \mathbb{I}(y^{(i)} = c) \quad (5.37)$$

For regression, the predicted value is:

$$\hat{y} = \frac{1}{k} \sum_{i=1}^k y^{(i)} \quad (5.38)$$

where  $y^{(i)}$  are the target values of the  $k$  nearest neighbors.

### 5.5.3 Distance Metrics

The choice of distance metric significantly affects k-NN performance, where different metrics are used in various machine learning algorithms to measure similarity between data points. Euclidean distance is commonly used in k-NN and clustering algorithms, Manhattan distance is useful in recommendation systems and text analysis, and Minkowski distance provides a flexible framework for different distance measures in various machine learning applications.

#### Euclidean Distance

**Definition 5.3: Euclidean Distance**

$$d(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{i=1}^d (x_i - x'_i)^2} \quad (5.39)$$

This is the most common choice, measuring the straight-line distance between points.

#### Manhattan Distance

**Definition 5.4: Manhattan Distance**

$$d(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^d |x_i - x'_i| \quad (5.40)$$

Also known as L1 distance, it measures the sum of absolute differences. Useful when features have different scales.

#### Minkowski Distance

**Definition 5.5: Minkowski Distance**

$$d(\mathbf{x}, \mathbf{x}') = \left( \sum_{i=1}^d |x_i - x'_i|^p \right)^{1/p} \quad (5.41)$$

This is a generalization where:

- $p = 1$ : Manhattan distance
- $p = 2$ : Euclidean distance
- $p \rightarrow \infty$ : Chebyshev distance (maximum coordinate difference)

### 5.5.4 Choosing k

The choice of  $k$  is crucial and involves a bias-variance trade-off, where small  $k$  values like  $k=1$  result in low bias and high variance with flexible decision boundaries that are sensitive to noise and outliers and may overfit to training data, while large  $k$  values like  $k=n$  result in high bias and low variance with smooth decision boundaries that may miss local patterns and underfit the data. The optimal  $k$  is typically found through cross-validation, balancing the need for local pattern recognition with robustness to noise.

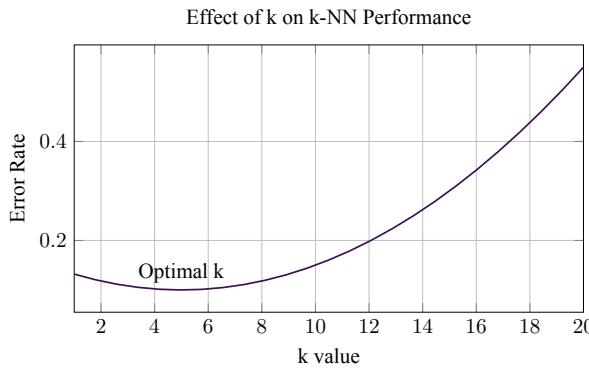


Figure 5.10: Effect of  $k$  on k-NN: small  $k \rightarrow$  high variance, large  $k \rightarrow$  high bias. Optimal  $k$  via CV.

### 5.5.5 Weighted k-NN

Instead of giving equal weight to all  $k$  neighbors, we can weight them by their distance:

$$\hat{y} = \frac{\sum_{i=1}^k w_i y^{(i)}}{\sum_{i=1}^k w_i} \quad (5.42)$$

where  $w_i = \frac{1}{d(\mathbf{x}, \mathbf{x}^{(i)})}$  is the weight based on distance.

### 5.5.6 Computational Considerations

k-NN has several computational characteristics including no training time since it's a lazy learner that defers all computation to prediction time, memory usage that requires storing all training examples, and scalability issues where performance degrades with large datasets. The naive approach has  $O(n \cdot d)$  complexity for each prediction, but optimized approaches using data structures like KD-trees, ball trees, and locality sensitive hashing can significantly reduce this complexity for large-scale applications.

#### Speedup Techniques

Several speedup techniques can be employed to improve the computational efficiency of k-NN algorithms. KD-Trees partition the space into regions using hyperplanes, reducing search time to  $O(\log n)$  in low dimensions, though they become less effective in high dimensions where the curse of dimensionality affects their performance. Ball Trees use hyperspheres instead of hyperplanes to partition the space, making them better suited for high-dimensional data, though they are more complex to implement than KD-trees. Locality Sensitive Hashing (LSH) provides approximate nearest neighbor search capabilities, offering significant speedup for very large datasets, though it may sacrifice some accuracy in exchange for computational efficiency, making it suitable for applications where approximate results are acceptable.

### 5.5.7 Advantages and Limitations

k-NN has several advantages including being simple and easy to understand and implement, making no assumptions about data distribution and working with any data type, being non-parametric with no model to fit, being able to capture complex decision boundaries through local patterns, and being incremental by making it easy to add new training examples. However, it also has limitations including computational cost where it's slow for large datasets, memory usage where it must store all training data, sensitivity to irrelevant features where all features are treated equally, the curse of dimensionality where performance degrades in high dimensions, and lack of interpretability where it's hard to understand why a prediction was made.

### 5.5.8 Curse of Dimensionality

In high-dimensional spaces, k-NN faces the curse of dimensionality where all points become roughly equidistant due to distance concentration, most of the space is empty making it difficult to find meaningful neighbors, and performance degrades with

irrelevant features that can dominate the distance calculations, making the algorithm less effective in high-dimensional spaces.

**Example 5.4**

In a 1000-dimensional space, even if only 10 features are relevant, the 990 irrelevant features can dominate the distance calculations, making k-NN ineffective.

### 5.5.9 Feature Selection and Scaling

To improve k-NN performance, several techniques can be employed including feature selection to remove irrelevant features that can hurt performance, feature scaling to normalize features to the same scale so that all features contribute equally to distance calculations, dimensionality reduction using PCA or other techniques to reduce the curse of dimensionality, and distance weighting to weight features by importance so that more relevant features have greater influence on the similarity calculations.

## 5.6 Comparison with Deep Learning ◆

Understanding the relationship between classical machine learning methods and deep learning is crucial for choosing the right approach for your problem. While deep learning has achieved remarkable success in many domains, classical methods still have important advantages in certain scenarios.

### 5.6.1 When Classical Methods Excel

Classical machine learning methods have several advantages that make them preferable in certain situations, including interpretability and debugging where linear models have coefficients that directly show feature importance, decision trees have rules that are human-readable, SVMs have support vectors that provide insight into decision boundaries, and easier debugging where predictions can be traced step by step. For small to medium datasets, classical methods are less prone to overfitting with fewer parameters, have faster training requiring less computational resources, often perform better with limited data through better generalization, and work well with original data without needing data augmentation. In terms of computational efficiency, classical methods have lower memory requirements without needing to store large networks, faster inference through simple mathematical operations, no GPU requirements allowing them to run on standard hardware, and suitability for real-time applications and embedded systems.

### 5.6.2 When Deep Learning Excels

Deep learning addresses several fundamental limitations of classical methods through automatic feature learning where networks learn relevant features automatically without manual feature engineering, hierarchical representations where lower layers learn simple features and higher layers learn complex combinations, end-to-end learning where a single model handles feature extraction and classification, and adaptive features that adapt to the specific problem. In terms of scalability with data and model size, deep learning has data scalability where performance typically improves with more data, model capacity to handle very large models with millions of parameters, distributed training capabilities that can leverage multiple GPUs/TPUs, and transfer learning where pre-trained models can be fine-tuned for new tasks. For handling complex data types, deep learning excels with images through convolutional networks that excel at computer vision, text through recurrent and transformer networks that handle natural language, audio by processing raw audio signals, and multimodal applications that can combine different data types.

### 5.6.3 Performance Comparison

### 5.6.4 Hybrid Approaches

In practice, the best solutions often combine classical and deep learning methods through feature engineering with deep learning where deep networks are used to extract features and classical methods like SVM and random forest are applied on the extracted

Aspect	Classical ML	Deep Learning	Best Use Case
Interpretability	High	Low	Medical diagnosis, finance
Training Speed	Fast	Slow	Prototyping, small datasets
Inference Speed	Fast	Medium	Real-time applications
Data Requirements	Low	High	Small companies, research
Computational Cost	Low	High	Resource-constrained environments
Feature Engineering	Manual	Automatic	Complex domains

Table 5.2: Classical ML vs deep learning comparison.

features, combining the interpretability of classical methods with the representation power of deep learning. Ensemble methods combine predictions from classical and deep learning models, using classical methods for interpretable components and deep learning for complex pattern recognition. Two-stage approaches use classical methods for initial screening and apply deep learning for final classification, balancing efficiency and accuracy.

### 5.6.5 Choosing the Right Approach

The choice between classical and deep learning methods depends on several factors including data characteristics where small datasets (< 10k examples) often favor classical methods, medium datasets (10k-100k examples) make both approaches viable, large datasets (> 100k examples) typically favor deep learning, high-dimensional data excels with deep learning, and structured data often works well with classical methods. Problem requirements include interpretability needs where classical methods are preferred, real-time inference where classical methods are often faster, complex patterns where deep learning is better, and unstructured data where deep learning is necessary. Resource constraints include limited computational resources favoring classical methods, limited labeled data favoring classical methods or transfer learning, need for quick prototyping favoring classical methods, and production deployment requiring consideration of inference costs.

### 5.6.6 Future Directions

The field continues to evolve with several promising directions including automated machine learning (AutoML) with neural architecture search that automatically designs network architectures, hyperparameter optimization that automatically tunes classical methods, and model selection that automatically chooses between classical and deep learning. Explainable AI includes SHAP values that explain predictions from any model, LIME for local interpretable model-agnostic explanations, and attention mechanisms that help understand what deep networks focus on. Efficient deep learning involves model compression to reduce model size while maintaining performance, quantization to use lower precision arithmetic, and knowledge distillation to transfer knowledge from large to small models.

### 5.6.7 Conclusion

Classical machine learning methods and deep learning are not competing approaches but complementary tools in the machine learning toolkit, where the key is to understand the strengths and limitations of each approach and choose the right tool for your specific problem. The approach should start simple by beginning with classical methods for baseline performance, consider complexity by only using deep learning if classical methods are insufficient, think about requirements by considering interpretability, speed, and resource constraints, combine approaches by using hybrid methods when appropriate, and stay updated as the field continues to evolve rapidly. The goal is not to use the most complex method, but to use the most appropriate method for your specific problem and constraints.

**Remark 5.8: Transformer Architecture Excellence**

Transformer architecture excels compared to traditional deep learning by using self-attention mechanisms that can process entire sequences in parallel, enabling much faster training than sequential RNNs, and by capturing long-range dependencies more effectively through direct attention connections rather than relying on sequential processing. This parallel processing capability and superior sequence modeling make transformers the foundation for modern large language models and state-of-the-art performance in natural language processing tasks.

Aspect	Classical ML	Deep Learning
Interpretability	High	Low
Training Speed	Fast	Slow
Inference Speed	Fast	Medium
Data Requirements	Low	High
Computational Cost	Low	High
Feature Engineering	Manual	Automatic
Best for Small Data	Yes	No
Best for Large Data	No	Yes

Table 5.3: Pros and cons comparison: classical ML vs deep learning.

## Key Takeaways

**Key Takeaways 5**

- **Classical algorithms** like linear regression, logistic regression, and SVMs provide interpretable baselines for machine learning tasks.
- **Trade-offs exist** between model complexity, interpretability, and performance; classical methods excel in low-data regimes.
- **Regularisation** (L1/L2) prevents overfitting and enables feature selection in high-dimensional settings.
- **Ensemble methods** combine weak learners to improve accuracy and robustness through variance reduction.
- **Understanding limitations** of classical methods motivates deep learning's hierarchical representation learning.

## Exercises

### Easy

*Exercise 5.1 (Linear Regression Basics).* Given the dataset  $\{(1, 2), (2, 4), (3, 6), (4, 8)\}$ , find the linear regression model  $\hat{y} = wx + b$  that minimises the mean squared error.

**Hint:**

Use the normal equation  $\mathbf{w}^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$  where  $\mathbf{X}$  includes a column of ones for the bias term.

*Exercise 5.2 (Logistic Regression Decision Boundary).* For a logistic regression model with weights  $\mathbf{w} = [2, -1]$  and bias  $b = 0$ , find the decision boundary equation and classify the point  $(1, 1)$ .

**Hint:**

*The decision boundary occurs where  $P(y = 1|\mathbf{x}) = 0.5$ , which means  $\mathbf{w}^\top \mathbf{x} + b = 0$ .*

*Exercise 5.3 (Decision Tree Splitting).* Given a node with 10 examples: 6 belong to class A and 4 to class B, calculate the Gini impurity and entropy.

**Hint:**

*Gini impurity =  $1 - \sum_k p_k^2$  and entropy =  $-\sum_k p_k \log p_k$  where  $p_k$  is the proportion of class  $k$ .*

*Exercise 5.4 (Regularisation Effect).* Explain why L1 regularisation (Lasso) can drive some weights to exactly zero, while L2 regularisation (Ridge) cannot.

**Hint:**

*Consider the shape of the L1 and L2 penalty functions and their derivatives at zero.*

*Exercise 5.5 (Naive Bayes Assumption).* Explain the naive Bayes assumption and why it's called "naive". Give an example where this assumption might be violated.

**Hint:**

*The assumption is that features are conditionally independent given the class label.*

*Exercise 5.6 (K-Means Initialization).* Compare k-means clustering with random initialization versus k-means++ initialization. Why does k-means++ often perform better?

**Hint:**

*k-means++ chooses initial centroids to be far apart, reducing the chance of poor local minima.*

*Exercise 5.7 (Decision Tree Pruning).* Explain the difference between pre-pruning and post-pruning in decision trees. When would you use each approach?

**Hint:**

*Pre-pruning stops growth early, while post-pruning grows the full tree then removes branches.*

*Exercise 5.8 (Cross-Validation Types).* Compare k-fold cross-validation, leave-one-out cross-validation, and stratified cross-validation. When would you use each?

**Hint:**

*Consider computational cost, variance of estimates, and class distribution preservation.*

## Medium

*Exercise 5.9 (Ridge Regression Derivation).* Derive the closed-form solution for ridge regression:  $\mathbf{w}^* = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$ .

**Hint:**

Start with the ridge regression objective function  $L(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|^2$  and take the gradient with respect to  $\mathbf{w}$ .

*Exercise 5.10 (Random Forest Bias-Variance).* Explain how random forests reduce variance compared to a single decision tree. What happens to bias?

**Hint:**

Consider how averaging multiple models affects the bias and variance of the ensemble.

*Exercise 5.11 (Feature Selection Methods).* Compare filter methods, wrapper methods, and embedded methods for feature selection. Give examples of each.

**Hint:**

Filter methods use statistical measures, wrapper methods use model performance, embedded methods are built into the learning algorithm.

*Exercise 5.12 (Bias-Variance Trade-off).* For a given dataset, explain how increasing model complexity affects bias and variance. Provide a concrete example.

**Hint:**

More complex models typically have lower bias but higher variance.

*Exercise 5.13 (Regularization Effects).* Compare L1 and L2 regularization in terms of their effect on feature selection and model interpretability.

**Hint:**

L1 regularization can drive coefficients to exactly zero, while L2 regularization shrinks them toward zero.

## Hard

*Exercise 5.14 (SVM Kernel Trick).* Prove that the polynomial kernel  $k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^\top \mathbf{x}' + c)^d$  is a valid kernel function by showing it corresponds to an inner product in some feature space.

**Hint:**

Expand the polynomial and show it can be written as  $\phi(\mathbf{x})^\top \phi(\mathbf{x}')$  for some feature map  $\phi$ .

*Exercise 5.15 (Ensemble Methods Theory).* Prove that for an ensemble of  $B$  independent models with error rate  $p < 0.5$ , the ensemble error rate approaches 0 as  $B \rightarrow \infty$ .

**Hint:**

Use the binomial distribution and the fact that the ensemble makes an error only when more than half of the models are wrong.

---

## **Part II**

# **Practical Deep Networks**



# Chapter 6

## Deep Feedforward Networks

This chapter introduces deep feedforward neural networks, also known as multilayer perceptrons (MLPs). These are the fundamental building blocks of deep learning.

### ⌚ Learning Objectives

1. Feedforward neural network architecture and information processing from input to output
2. Activation functions and their role in introducing non-linearity
3. Output layers and loss functions for regression and classification tasks
4. Backpropagation algorithm for efficient gradient computation in neural networks
5. Network architecture design, initialization, and training procedures
6. Practical applications of feedforward networks

### 6.1 Introduction to Feedforward Networks ◆

Feedforward neural networks are the fundamental building blocks of deep learning, consisting of interconnected layers of neurons that process information sequentially from input to output without feedback loops.

#### 6.1.1 Intuition: What is a Feedforward Network?

Imagine you're trying to recognize handwritten digits, where a feedforward neural network is like having a team of experts, each specialized in detecting different features. First layer experts look for basic patterns like edges, curves, and lines, middle layer experts combine these basic patterns to detect more complex shapes like loops, corners, and curves, and final layer experts combine these complex shapes to make the final decision like "This is a 3" or "This is a 7". The key insight is that each layer builds upon the previous one, creating increasingly sophisticated representations that mirror how our own visual system works, from detecting simple edges to recognizing complex objects. A feedforward neural network approximates a function  $f^*$  where for input  $\mathbf{x}$ , the network computes  $y = f(\mathbf{x}; \boldsymbol{\theta})$  and learns parameters  $\boldsymbol{\theta}$  such that  $f \approx f^*$ .

#### 6.1.2 Network Architecture

A feedforward network consists of layers including an input layer that receives raw features  $\mathbf{x}$ , hidden layers that create intermediate representations  $\mathbf{h}^{(1)}, \mathbf{h}^{(2)}, \dots$  through nonlinear transformations, and an output layer that produces predictions  $\hat{y}$  based on the processed information from the hidden layers.

For a network with  $L$  layers:

Input: 4x4 Hidden Layer Output: "3" or "7"

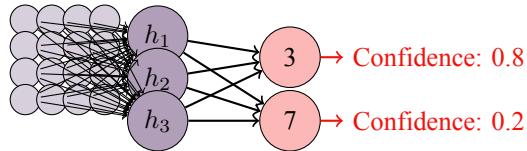


Figure 6.1: Simple neural network for digit classification: "This is a 3" or "This is a 7". The network learns to distinguish between these two digits by analyzing pixel patterns.

Input Layer   Hidden Layer 1   Hidden Layer 2   Output Layer

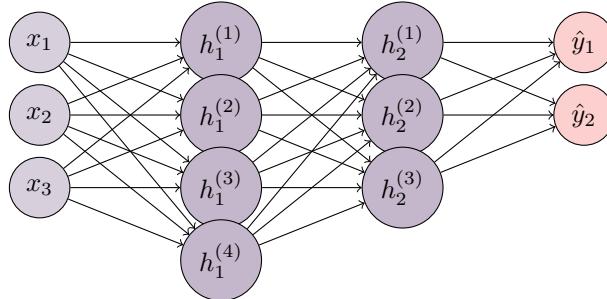


Figure 6.2: Feedforward neural network with 2 hidden layers. Circles=neurons, arrows=flow.

$$\mathbf{h}^{(l)} = \sigma(\mathbf{W}^{(l)} \mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}) \quad (6.1)$$

where  $\mathbf{h}^{(0)} = \mathbf{x}$ ,  $\mathbf{W}^{(l)}$  are weights,  $\mathbf{b}^{(l)}$  are biases, and  $\sigma$  is an activation function.

### 6.1.3 Forward Propagation

The computation proceeds from input to output through a series of matrix multiplications and nonlinear transformations. For each layer  $l$ , the network computes pre-activations:

$$\mathbf{z}^{(l)} = \mathbf{W}^{(l)} \mathbf{h}^{(l-1)} + \mathbf{b}^{(l)} \quad (6.2)$$

followed by activations  $\mathbf{h}^{(l)} = \sigma(\mathbf{z}^{(l)})$  where  $\sigma$  is the activation function. The process continues through all hidden layers until reaching the output layer, where the final prediction  $\hat{y} = \mathbf{h}^{(L)}$  is produced.

---

**Algorithm 2** Forward Propagation Algorithm

---

- 1: **Input:**  $\mathbf{x}$ , weights  $\{\mathbf{W}^{(l)}\}_{l=1}^L$ , biases  $\{\mathbf{b}^{(l)}\}_{l=1}^L$
  - 2: Initialize  $\mathbf{h}^{(0)} = \mathbf{x}$
  - 3: **for**  $l = 1$  to  $L$  **do**
  - 4:     Compute pre-activation:  $\mathbf{z}^{(l)} = \mathbf{W}^{(l)} \mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}$
  - 5:     Apply activation function:  $\mathbf{h}^{(l)} = \sigma(\mathbf{z}^{(l)})$
  - 6: **end for**
  - 7: **Return**  $\hat{y} = \mathbf{h}^{(L)}$
-

**Example 6.1: Simple Forward Pass**

Consider a simple network with 2 inputs, 2 hidden neurons, and 1 output for binary classification, where the input is  $\mathbf{x} = [1, 0.5]$ , weights to hidden layer are  $\mathbf{W}^{(1)} = \begin{bmatrix} 0.5 & -0.3 \\ 0.8 & 0.2 \end{bmatrix}$ , bias is  $\mathbf{b}^{(1)} = [0.1, -0.2]$ , and the activation function is ReLU. This example demonstrates how the network processes information through matrix multiplication and nonlinear transformation to produce meaningful representations.

**Step 1:** Compute pre-activation

$$\mathbf{z}^{(1)} = \mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)} \quad (6.3)$$

$$= \begin{bmatrix} 0.5 & -0.3 \\ 0.8 & 0.2 \end{bmatrix} \begin{bmatrix} 1 \\ 0.5 \end{bmatrix} + \begin{bmatrix} 0.1 \\ -0.2 \end{bmatrix} \quad (6.4)$$

$$= \begin{bmatrix} 0.5 \cdot 1 + (-0.3) \cdot 0.5 + 0.1 \\ 0.8 \cdot 1 + 0.2 \cdot 0.5 + (-0.2) \end{bmatrix} \quad (6.5)$$

$$= \begin{bmatrix} 0.45 \\ 0.7 \end{bmatrix} \quad (6.6)$$

**Step 2:** Apply activation function

$$\mathbf{h}^{(1)} = \text{ReLU}(\mathbf{z}^{(1)}) = \begin{bmatrix} \max(0, 0.45) \\ \max(0, 0.7) \end{bmatrix} = \begin{bmatrix} 0.45 \\ 0.7 \end{bmatrix} \quad (6.7)$$

The hidden layer has learned to represent the input in a transformed space where both neurons are active (positive values).

### 6.1.4 Universal Approximation

The **universal approximation theorem** states that a feedforward network with a single hidden layer containing a finite number of neurons can approximate any continuous function on a compact subset of  $\mathbb{R}^n$ , given appropriate activation functions. However, deeper networks often learn more efficiently.

## 6.2 Activation Functions ◆

Activation functions are nonlinear transformations applied to neuron outputs that introduce the essential non-linearity needed for neural networks to learn complex patterns and relationships in data.

### 6.2.1 Intuition: Why Do We Need Activation Functions?

Imagine you're building a house with only straight lines and right angles, where no matter how many rooms you add, you can only create rectangular spaces, but what if you want curved walls, arches, or domes? You need curved tools! Similarly, without activation functions, neural networks can only learn linear relationships, no matter how many layers you add, where activation functions are the "curved tools" that allow networks to learn non-linear patterns. Think of an activation function as a decision maker that takes a weighted sum of information from other neurons as input, decides how much this neuron should "fire" or contribute to the next layer, and outputs a transformed value that becomes input to the next layer. Activation functions introduce non-linearity, enabling networks to learn complex patterns that would be impossible with linear transformations alone.

## 6.2.2 Sigmoid

The sigmoid function is an S-shaped curve that maps any real number to a value between 0 and 1, making it useful for binary classification and probability estimation:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (6.8)$$

However, the sigmoid function has several limitations including saturating for large absolute values of  $z$  which causes vanishing gradients during backpropagation, not being zero-centered which can slow down learning, and being historically important but less common in hidden layers due to these issues, though it remains useful in output layers for binary classification tasks.

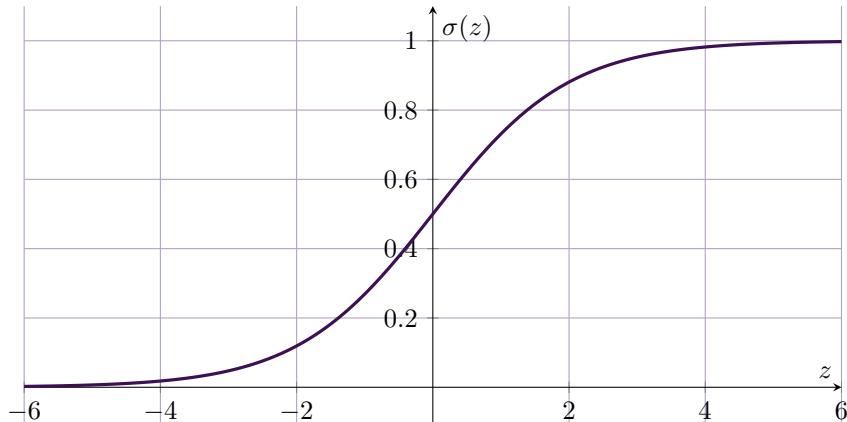


Figure 6.3: Sigmoid  $\sigma(z) = 1/(1 + e^{-z})$ : smooth S-shape  $\mathbb{R} \rightarrow (0, 1)$ .

## 6.2.3 Hyperbolic Tangent (tanh)

The hyperbolic tangent function is a zero-centered S-shaped curve that maps any real number to a value between -1 and 1, making it an improvement over sigmoid for hidden layers:

$$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \quad (6.9)$$

The tanh function is zero-centered which helps with gradient flow and learning dynamics, though it still suffers from saturation at extreme values which can cause vanishing gradients in deep networks, making it better than sigmoid but still not ideal for very deep architectures.

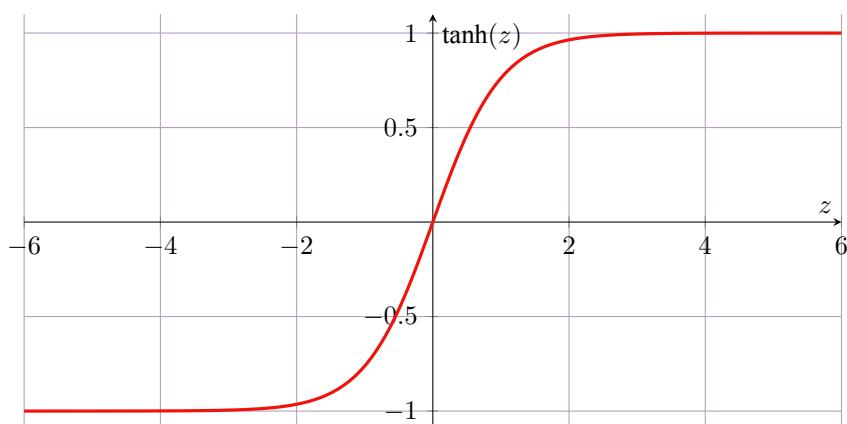


Figure 6.4:  $\tanh(z)$ : zero-centered S-shape mapping  $\mathbb{R} \rightarrow (-1, 1)$ .

## 6.2.4 Rectified Linear Unit (ReLU)

The Rectified Linear Unit (ReLU) function is a simple piecewise linear function that outputs the input directly if it's positive, otherwise outputs zero:

$$\text{ReLU}(z) = \max(0, z) \quad (6.10)$$

ReLU is simple and computationally efficient with no saturation for positive values, making it the most widely used activation function in modern deep learning. However, ReLU can cause "dead neurons" that always output 0 when the input is negative, which can slow down learning, though this issue is often mitigated by proper initialization and other techniques.

## 6.2.5 Leaky ReLU and Variants

### Definition 6.1: Leaky ReLU

$$\text{LeakyReLU}(z) = \max(\alpha z, z), \quad \alpha \ll 1 \quad (6.11)$$

### Definition 6.2: Parametric ReLU (PReLU)

$$\text{PReLU}(z) = \max(\alpha z, z) \quad (6.12)$$

where  $\alpha$  is learned.

### Definition 6.3: Exponential Linear Unit (ELU)

$$\text{ELU}(z) = \begin{cases} z & \text{if } z > 0 \\ \alpha(e^z - 1) & \text{if } z \leq 0 \end{cases} \quad (6.13)$$

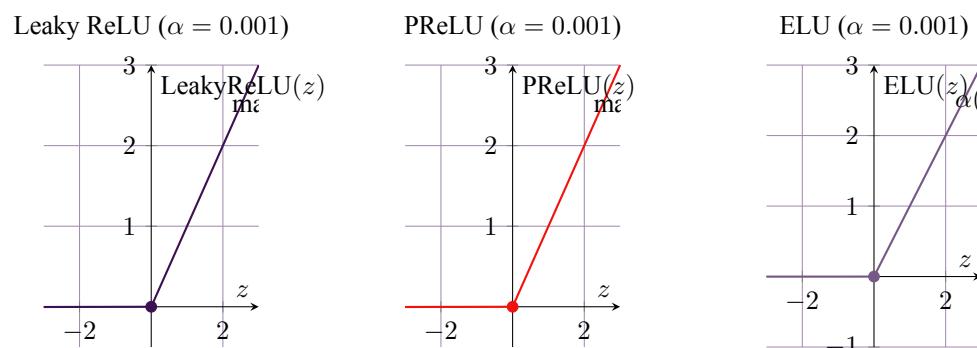


Figure 6.5: Leaky ReLU, PReLU, and ELU activation functions with  $\alpha = 0.001$ .

## 6.2.6 Swish and GELU

### Definition 6.4: Swish Activation Function

The Swish activation function is defined as:

$$\text{Swish}(z) = z \cdot \sigma(z) \quad (6.14)$$

where  $\sigma(z)$  is the sigmoid function. Swish is a smooth, non-monotonic function that combines the benefits of ReLU with smooth gradients, often outperforming ReLU in many tasks.

### Definition 6.5: Gaussian Error Linear Unit (GELU)

The GELU activation function is defined as:

$$\text{GELU}(z) = z \cdot \Phi(z) \quad (6.15)$$

where  $\Phi(z)$  is the Gaussian CDF. GELU is used in modern transformers because it provides smooth gradients and better performance in transformer architectures, particularly in the attention mechanisms and feedforward layers of models like BERT, GPT, and other large language models where the smooth activation helps with training stability and convergence.

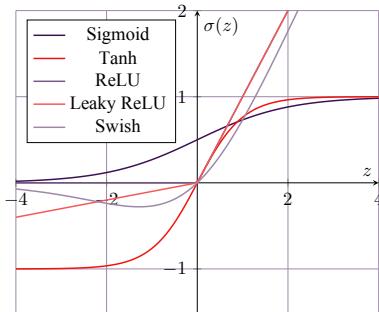


Figure 6.6: Common activations: sigmoid/tanh saturate, ReLU can "die", Swish smooth.

## 6.3 Output Units and Loss Functions ◆

Output units and loss functions are the final components of neural networks that determine how predictions are formatted and how the network learns from errors, with different choices required for different types of machine learning tasks.

### 6.3.1 Intuition: Matching Output to Task

Think of the output layer as the "final decision maker" in your network, where just like different jobs require different tools, different machine learning tasks require different output formats. For regression tasks like predicting prices, you want a real number such as "This house costs \$250,000", for binary classification tasks like spam detection, you want a probability such as "This email is 95% likely to be spam", and for multiclass classification tasks like image recognition, you want probabilities for each class such as "This image is 80% cat, 15% dog, 5% bird". The loss function is like a "teacher" that tells the network how wrong it is, where a good teacher gives clear, helpful feedback that guides learning in the right direction, and the choice of output layer and loss function depends on the specific task requirements.

### 6.3.2 Linear Output for Regression

For regression, use linear output:

$$\hat{y} = \mathbf{W}^\top \mathbf{h} + b \quad (6.16)$$

where  $\mathbf{W}$  are the output weights,  $\mathbf{h}$  is the final hidden layer activation, and  $b$  is the bias term. This produces a continuous real-valued prediction.

with MSE loss:

$$L = \frac{1}{n} \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2 \quad (6.17)$$

where  $y^{(i)}$  is the true target value and  $\hat{y}^{(i)}$  is the predicted value. MSE loss penalizes large errors more heavily than small errors, making it suitable for regression tasks where we want to minimize the average squared difference between predictions and

targets.

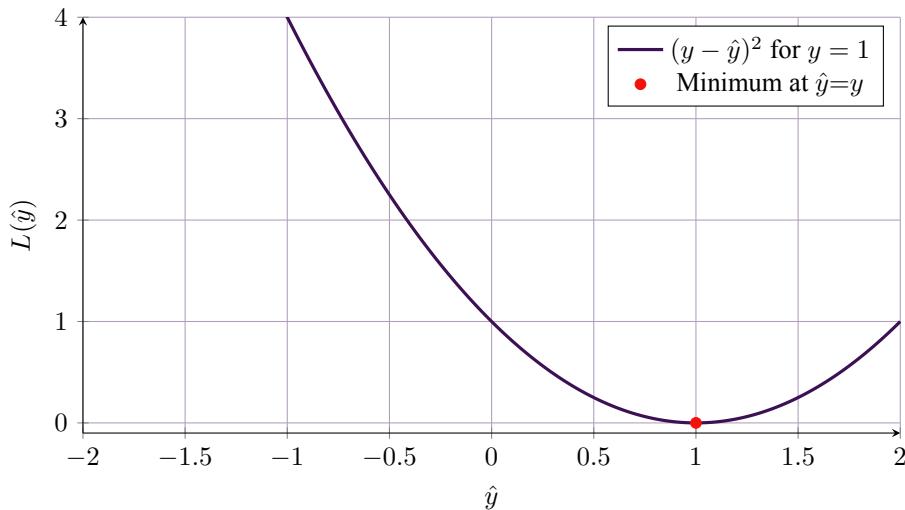


Figure 6.7: MSE loss in 1D: convex parabola at target  $y$ . Min at  $\hat{y}=y$ .

### 6.3.3 Sigmoid Output for Binary Classification

For binary classification:

$$\hat{y} = \sigma(\mathbf{W}^\top \mathbf{h} + b) \quad (6.18)$$

with binary cross-entropy loss:

$$L = -\frac{1}{n} \sum_{i=1}^n [y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})] \quad (6.19)$$

### 6.3.4 Softmax Output for Multiclass Classification

For  $K$  classes:

$$\hat{y}_k = \frac{\exp(z_k)}{\sum_{j=1}^K \exp(z_j)} \quad (6.20)$$

with categorical cross-entropy loss:

$$L = -\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K y_k^{(i)} \log \hat{y}_k^{(i)} \quad (6.21)$$

#### Remark 6.1: Softmax Loss in Deep Learning

The softmax function converts raw logits into a probability distribution over classes, ensuring all probabilities sum to 1. The categorical cross-entropy loss measures the difference between the predicted probability distribution and the true one-hot encoded target, providing strong gradients when predictions are wrong and encouraging the network to assign high probability to the correct class. This combination is the standard approach for multiclass classification in deep learning, used in image classification, natural language processing, and many other applications where we need to choose among multiple discrete categories.

## 6.4 Backpropagation ◆

Backpropagation is the fundamental algorithm for training neural networks that efficiently computes gradients by propagating errors backward through the network using the chain rule of calculus.

### 6.4.1 Intuition: Learning from Mistakes

Imagine you're learning to play basketball, where after each shot, you need to know how far off your shot was (the error), which part of your technique needs adjustment (which parameters to change), and how much you should adjust each part (how much to change each parameter). Backpropagation is like having a coach who watches your shot and tells you exactly what to adjust, such as "Your elbow was too high" (gradient for elbow angle), "You need to follow through more" (gradient for follow-through), and "Your timing was off" (gradient for release timing). The key insight is that we can efficiently compute how much each parameter contributed to the final error by working backwards through the network, where backpropagation efficiently computes gradients using the chain rule to determine the precise adjustments needed for each parameter.

### 6.4.2 The Chain Rule

The chain rule is fundamental for computing derivatives of composite functions. Let's start with simple examples to build intuition.

#### Simple Chain Rule: $f(g(x))$

For composition  $f(g(x))$ :

$$\frac{df}{dx} = \frac{df}{dg} \frac{dg}{dx} \quad (6.22)$$

#### Example 6.2

Let  $f(x) = (x^2 + 1)^3$ . We can think of this as  $f(g(x))$  where:

- $g(x) = x^2 + 1$  (inner function)
- $f(g) = g^3$  (outer function)

Using the chain rule:

$$\frac{df}{dx} = \frac{df}{dg} \cdot \frac{dg}{dx} \quad (6.23)$$

$$= 3g^2 \cdot (2x) \quad (6.24)$$

$$= 3(x^2 + 1)^2 \cdot 2x \quad (6.25)$$

$$= 6x(x^2 + 1)^2 \quad (6.26)$$

#### Multivariable Chain Rule: $f(g(x, y))$

For functions of multiple variables, we use partial derivatives:

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial g} \frac{\partial g}{\partial x} \quad (6.27)$$

**Example 6.3**

Let  $f(x, y) = (x^2 + y^2)^2$ . We can think of this as  $f(g(x, y))$  where:

- $g(x, y) = x^2 + y^2$  (inner function)
- $f(g) = g^2$  (outer function)

Using the chain rule for partial derivatives:

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial g} \cdot \frac{\partial g}{\partial x} \quad (6.28)$$

$$= 2g \cdot (2x) \quad (6.29)$$

$$= 2(x^2 + y^2) \cdot 2x \quad (6.30)$$

$$= 4x(x^2 + y^2) \quad (6.31)$$

Similarly:

$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial g} \cdot \frac{\partial g}{\partial y} \quad (6.32)$$

$$= 2g \cdot (2y) \quad (6.33)$$

$$= 2(x^2 + y^2) \cdot 2y \quad (6.34)$$

$$= 4y(x^2 + y^2) \quad (6.35)$$

**Vector Chain Rule**

For vectors, we use the Jacobian:

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \frac{\partial \mathbf{y}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{x}} \quad (6.36)$$

**Example 6.4: Vector Chain Rule**

Consider the composition  $\mathbf{y} = \mathbf{f}(\mathbf{g}(\mathbf{x}))$  where:

- $\mathbf{g}(\mathbf{x}) = \mathbf{x}^2 + \mathbf{1}$  (element-wise square plus ones)
- $\mathbf{f}(\mathbf{z}) = \mathbf{z}^3$  (element-wise cube)

For a 2D case with  $\mathbf{x} = [x_1, x_2]^T$ :

$$\mathbf{g}(\mathbf{x}) = [x_1^2 + 1, x_2^2 + 1]^T \quad (6.37)$$

$$\mathbf{y} = [(x_1^2 + 1)^3, (x_2^2 + 1)^3]^T \quad (6.38)$$

The Jacobian matrices are:

$$\frac{\partial \mathbf{g}}{\partial \mathbf{x}} = \begin{bmatrix} 2x_1 & 0 \\ 0 & 2x_2 \end{bmatrix} \quad (6.39)$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{g}} = \begin{bmatrix} 3(x_1^2 + 1)^2 & 0 \\ 0 & 3(x_2^2 + 1)^2 \end{bmatrix} \quad (6.40)$$

Using the vector chain rule:

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \frac{\partial \mathbf{y}}{\partial \mathbf{g}} \frac{\partial \mathbf{g}}{\partial \mathbf{x}} \quad (6.41)$$

$$= \begin{bmatrix} 3(x_1^2 + 1)^2 & 0 \\ 0 & 3(x_2^2 + 1)^2 \end{bmatrix} \begin{bmatrix} 2x_1 & 0 \\ 0 & 2x_2 \end{bmatrix} \quad (6.42)$$

$$= \begin{bmatrix} 6x_1(x_1^2 + 1)^2 & 0 \\ 0 & 6x_2(x_2^2 + 1)^2 \end{bmatrix} \quad (6.43)$$

**6.4.3 Backward Pass**

Starting from the loss  $L$ , we compute gradients layer by layer:

$$\delta^{(L)} = \nabla_{\mathbf{h}^{(L)}} L \quad (6.44)$$

$$\delta^{(l)} = (\mathbf{W}^{(l+1)})^\top \delta^{(l+1)} \odot \sigma'(\mathbf{z}^{(l)}) \quad (6.45)$$

where  $\odot$  denotes element-wise multiplication.

Parameter gradients:

$$\frac{\partial L}{\partial \mathbf{W}^{(l)}} = \delta^{(l)} (\mathbf{h}^{(l-1)})^\top \quad (6.46)$$

$$\frac{\partial L}{\partial \mathbf{b}^{(l)}} = \delta^{(l)} \quad (6.47)$$

**6.4.4 Computational Graph**

Modern frameworks use automatic differentiation on computational graphs, where forward mode is efficient when outputs greatly exceed inputs, and reverse mode (backprop) is efficient when inputs greatly exceed outputs, allowing for optimal gradient computation based on the specific network architecture and task requirements.

## 6.5 Design Choices ◆

Design choices in neural networks involve critical decisions about architecture, initialization, and training procedures that significantly impact the network's ability to learn and generalize to new data.

### 6.5.1 Intuition: Building the Right Network

Designing a neural network is like designing a building, where depth (layers) is like floors in a building where more floors can house more complex functions, but they're harder to build and maintain, width (neurons per layer) is like rooms per floor where more rooms give more space, but you need to fill them efficiently, initialization is like the foundation where if it's wrong, the whole building might collapse, and training is like the construction process where you need the right tools and techniques. The key is finding the right balance for your specific task and data, where each design choice affects the network's capacity, trainability, and performance.

### 6.5.2 Network Depth and Width

The choice of network depth and width involves important trade-offs, where deeper networks can learn more complex functions but can be harder to optimize due to vanishing or exploding gradients, though modern techniques enable very deep networks with 100+ layers. Wider networks have more capacity but there's a trade-off between width and depth, where very wide shallow networks compete with narrow deep networks in terms of representational power and computational efficiency, with the optimal choice depending on the specific task and available computational resources.

### 6.5.3 Weight Initialization

Poor initialization can prevent learning. Common strategies:

#### Definition 6.6: Xavier/Glorot Initialization

Xavier/Glorot initialization balances variance across layers by matching forward and backward signal scales for approximately linear activations (tanh/sigmoid in their linear regime). In the popular uniform variant:

$$W_{ij} \sim \mathcal{U}\left(-\sqrt{\frac{6}{n_{in}+n_{out}}}, \sqrt{\frac{6}{n_{in}+n_{out}}}\right) \quad (6.48)$$

and in the normal variant:

$$W_{ij} \sim \mathcal{N}\left(0, \frac{2}{n_{in} + n_{out}}\right) \quad (6.49)$$

The scaling arises by setting  $\text{Var}[\text{preact}]$  and  $\text{Var}[\text{grad}]$  approximately constant layer-to-layer under independence assumptions.

#### Definition 6.7: He Initialization

He initialization (for ReLU-family) uses a larger variance to compensate for half of activations being zeroed by  $\max(0, \cdot)$ . Common forms are

$$W_{ij} \sim \mathcal{N}\left(0, \frac{2}{n_{in}}\right) \quad (6.50)$$

or the uniform analogue

$$W_{ij} \sim \mathcal{U}\left(-\sqrt{\frac{6}{n_{in}}}, \sqrt{\frac{6}{n_{in}}}\right) \quad (6.51)$$

which maintains stable activation and gradient variances at initialization for ReLU/LeakyReLU networks.

### 6.5.4 Batch Training

Mini-batch gradient descent computes gradients on small batches of typically 32-256 examples, providing regularization through noise and enabling efficient parallel computation. This approach balances between stochastic gradient descent and full-batch gradient descent, offering a compromise between computational efficiency and gradient stability that is well-suited for modern deep learning applications.

#### Remark 6.2: Online Learning vs Mini-batch Training

Online learning processes one example at a time, making it memory-efficient but potentially noisy, while mini-batch training processes small groups of examples simultaneously, providing a good balance between computational efficiency and gradient stability. Mini-batch training is preferred in modern deep learning because it enables parallel processing on GPUs, provides more stable gradient estimates than online learning, and allows for efficient memory usage while maintaining good convergence properties.

## 6.6 Real World Applications

Deep feedforward networks serve as the foundation for many practical applications across industries. Here we explore how these networks solve real-world problems in accessible, less technical terms.

### 6.6.1 Medical Diagnosis Support

Feedforward networks help doctors make better decisions by analyzing patient data, where disease prediction networks analyze patient symptoms, medical history, and test results to predict the likelihood of diseases like diabetes or heart disease, learning patterns from thousands of past patient records to help identify at-risk individuals early. Treatment recommendation networks learn from successful treatment outcomes to suggest personalized treatment plans based on a patient's unique characteristics, improving recovery rates and reducing side effects. Drug dosage optimization networks help determine optimal medication dosages by considering factors like patient weight, age, kidney function, and drug interactions, reducing risks of under or over-medication while ensuring effective treatment outcomes.

### 6.6.2 Financial Fraud Detection

Banks and financial institutions use feedforward networks to protect customers from fraud, where credit card fraud detection networks analyze transaction patterns in real-time, flagging unusual purchases like expensive items bought far from home within milliseconds, happening seamlessly as you shop and protecting your account without interrupting legitimate purchases. Loan default prediction networks evaluate applicant information before approving loans to predict repayment likelihood, helping banks make fairer lending decisions while reducing financial risks. Insurance claim verification networks identify suspicious insurance claims by detecting patterns inconsistent with typical legitimate claims, saving companies billions while ensuring honest customers get quick payouts.

### 6.6.3 Product Recommendation Systems

Online platforms use feedforward networks to personalize your experience, where e-commerce recommendation networks analyze your browsing history, purchase patterns, and preferences to suggest products you're likely to enjoy, making shopping more efficient and helping you discover new items. Content recommendation networks used by streaming services suggest movies, shows, or music based on what you've watched or listened to before, where the network learns your taste profile and finds content matching your preferences. Targeted advertising networks help businesses show you relevant ads by understanding your interests and needs, benefiting both consumers who see useful products and businesses who reach interested customers.

### 6.6.4 Why These Applications Work

Feedforward networks excel at these tasks because they can learn complex patterns from historical data, make decisions quickly once trained, handle multiple input features simultaneously, and generalize to new, unseen situations. These applications demonstrate how deep learning moves from theory to practice, improving everyday life in ways both visible and behind-the-scenes, where the networks' ability to process large amounts of data and identify subtle patterns makes them invaluable tools for solving real-world problems across various industries.

## Key Takeaways

### Key Takeaways 6

- **Feedforward networks** compose layers of linear transformations and nonlinear activations to learn complex functions.
- **Activation functions** introduce nonlinearity; ReLU and variants balance expressiveness with gradient flow.
- **Backpropagation** efficiently computes gradients via the chain rule, enabling gradient-based optimisation.
- **Output layers and losses** must match the task: softmax/cross-entropy for classification, linear/MSE for regression.
- **Design choices** (depth, width, initialisation) profoundly affect training dynamics and generalisation.

## Exercises

### Easy

*Exercise 6.1 (Activation Functions).* Compare ReLU and sigmoid activation functions. List two advantages of ReLU over sigmoid for hidden layers in deep networks.

**Hint:**

*Consider gradient flow, computational efficiency, and the vanishing gradient problem.*

*Exercise 6.2 (Network Capacity).* A feedforward network has an input layer with 10 neurons, two hidden layers with 20 neurons each, and an output layer with 3 neurons. Calculate the total number of parameters (weights and biases).

**Hint:**

*For each layer transition, count weights and biases separately.*

*Exercise 6.3 (Output Layer Design).* For a binary classification task, what activation function and loss function would you use for the output layer? Justify your choice.

**Hint:**

*Think about probability outputs and the relationship between binary cross-entropy and sigmoid activation.*

*Exercise 6.4 (Backpropagation Basics).* Explain in simple terms why backpropagation is more efficient than computing gradients using finite differences for each parameter.

**Hint:**

*Consider the number of forward passes required and the chain rule of calculus.*

## Medium

*Exercise 6.5 (Gradient Computation).* For a simple network with one hidden layer:  $\mathbf{h} = \sigma(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1)$  and  $\mathbf{y} = \mathbf{W}_2 \mathbf{h} + \mathbf{b}_2$ , derive the gradient  $\frac{\partial L}{\partial \mathbf{W}_1}$  for mean squared error loss.

**Hint:**

Apply the chain rule:  $\frac{\partial L}{\partial \mathbf{W}_1} = \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{W}_1}$ .

*Exercise 6.6 (Universal Approximation).* The universal approximation theorem states that a feedforward network with a single hidden layer can approximate any continuous function. Discuss why we still use deep networks with multiple layers in practice.

**Hint:**

Consider efficiency of representation, number of neurons needed, and hierarchical feature learning.

## Hard

*Exercise 6.7 (Xavier Initialisation).* Derive the Xavier (Glorot) initialisation scheme for weights. Explain why it helps maintain variance of activations across layers.

**Hint:**

Start with the variance of layer outputs and the assumption that inputs and weights are independent.

*Exercise 6.8 (Residual Connections).* Analyse how residual connections (skip connections) help with gradient flow in very deep networks. Derive the gradient through a residual block.

**Hint:**

Consider  $\mathbf{y} = \mathbf{x} + F(\mathbf{x})$  and compute  $\frac{\partial L}{\partial \mathbf{x}}$  where  $F$  is a sub-network.

*Exercise 6.9 (Universal Approximation).* Explain the universal approximation theorem for neural networks. What are its practical limitations?

**Hint:**

The theorem states that a single hidden layer with sufficient neurons can approximate any continuous function, but doesn't guarantee efficient learning.

*Exercise 6.10 (Activation Function Properties).* Compare the properties of ReLU, Leaky ReLU, and ELU activation functions. When would you use each?

**Hint:**

Consider gradient flow, computational efficiency, and the "dying ReLU" problem.

*Exercise 6.11 (Network Depth vs Width).* Explain the trade-offs between making a network deeper versus wider. When might you prefer one approach?

**Hint:**

Deeper networks can learn more complex features but are harder to train; wider networks are easier to train but may need more parameters.

*Exercise 6.12 (Gradient Vanishing Exercise).* Explain why gradients can vanish in deep networks and how modern architectures address this issue.

**Hint:**

*Consider the chain rule and how gradients are multiplied through layers, especially with activation functions like sigmoid.*

*Exercise 6.13 (Batch Normalization Effects).* Explain how batch normalization affects training dynamics and why it can improve convergence.

**Hint:**

*Batch normalization reduces internal covariate shift and can act as a regularizer.*

*Exercise 6.14 (Dropout Regularization).* Compare dropout with other regularization techniques like L1/L2 regularization. When is dropout most effective?

**Hint:**

*Dropout prevents overfitting by randomly setting neurons to zero during training, creating an ensemble effect.*

*Exercise 6.15 (Weight Initialization Strategies).* Compare Xavier/Glorot initialization with He initialization. When would you use each?

**Hint:**

*Xavier assumes symmetric activation functions, while He initialization is designed for ReLU networks.*



# Chapter 7

## Regularization for Deep Learning

This chapter explores techniques to improve generalization and prevent overfitting in deep neural networks. Regularization helps models perform well on unseen data.

### Learning Objectives

1. Why regularization is needed and how it improves generalization
2. Norm penalties (L1, L2, Elastic Net) and when to use each
3. Data augmentation strategies for vision, text, and audio tasks
4. Early stopping and its interaction with optimization
5. Dropout and normalization layers and their effects
6. Advanced techniques: label smoothing, mixup, adversarial training, and gradient clipping

### 7.1 Parameter Norm Penalties ◆

**Parameter norm penalties** constrain model capacity by penalizing large weights.

#### 7.1.1 Intuition: Shrinking the Model's "Complexity"

Think of a model as a musical band with many instruments (parameters). If every instrument plays loudly (large weights), the result can be noisy and overfit to the training song. Norm penalties are like asking the band to lower the volume uniformly (L2) or mute many instruments entirely (L1) so the melody (true signal) stands out. This discourages memorization and encourages simpler patterns that generalize.

#### 7.1.2 L2 Regularization (Weight Decay)

Add squared L2 norm of weights to the loss:

$$\tilde{L}(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) + \frac{\lambda}{2} \|\boldsymbol{w}\|^2 \quad (7.1)$$

Gradient update becomes:

$$\boldsymbol{w} \leftarrow (1 - \alpha\lambda)\boldsymbol{w} - \alpha\nabla_{\boldsymbol{w}}L \quad (7.2)$$

The factor  $(1 - \alpha\lambda)$  causes "weight decay."

### 7.1.3 L1 Regularization

L1 regularization adds the L1 norm of weights to the loss function:

$$\tilde{L}(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) + \lambda \|\boldsymbol{w}\|_1 \quad (7.3)$$

L1 regularization promotes sparsity by driving many weights to become exactly zero, making it useful for feature selection where only the most important features are retained. The gradient of L1 regularization is  $\text{sign}(\boldsymbol{w})$ , which provides a constant magnitude update that encourages weights to move toward zero, effectively performing automatic feature selection by eliminating less important parameters.

### 7.1.4 Elastic Net

Combines L1 and L2:

$$\tilde{L}(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) + \lambda_1 \|\boldsymbol{w}\|_1 + \lambda_2 \|\boldsymbol{w}\|^2 \quad (7.4)$$

Why combine them? The L1 term promotes sparsity, selecting a compact set of features, while the L2 term stabilizes the solution by shrinking correlated coefficients together rather than arbitrarily picking one. In high-dimensional settings with collinearity, pure L1 can be unstable; the added L2 induces grouping and reduces variance, yielding models that are both interpretable (sparse) and numerically well-conditioned.

## 7.2 Dataset Augmentation ◆

**Data augmentation** artificially increases training set size by applying transformations that preserve labels.

### 7.2.1 Intuition: Seeing the Same Thing in Many Ways

Humans recognize an object despite different viewpoints, lighting, or small occlusions. Augmentation teaches models the same robustness by showing multiple, label-preserving variations of each example. This reduces overfitting by making spurious correlations less useful and forcing the model to focus on invariant structure.

### 7.2.2 Image Augmentation

Image augmentation employs various transformations to artificially expand the training dataset while preserving semantic content. Geometric transformations including rotation, translation, scaling, flipping, and cropping change the spatial properties of images while preserving the semantic content, where rotation and translation help models become invariant to object orientation and position, while scaling and cropping teach robustness to different object sizes and partial views. Color modifications such as brightness, contrast, and saturation adjustments simulate different lighting conditions and camera settings that occur in real-world scenarios, where by varying these color properties, models learn to focus on structural features rather than specific color characteristics, improving generalization across different environments. Noise augmentation including Gaussian noise and blur helps models become more robust to sensor imperfections and motion blur that occur in real images, where this regularization technique prevents overfitting to pixel-perfect training data and improves performance on noisy real-world inputs. Cutout and erasing techniques randomly remove rectangular regions from images, forcing the model to learn from partial information and encouraging the network to not rely on specific spatial locations, instead learning more distributed, robust feature representations. Mixup creates new training examples by linearly interpolating between pairs of images and their corresponding labels, encouraging smoother decision boundaries and reducing overconfident predictions, leading to better calibration and generalization.

Example: horizontal flip

$$\boldsymbol{x}_{\text{aug}} = \text{flip}(\boldsymbol{x}), \quad y_{\text{aug}} = y \quad (7.5)$$

### 7.2.3 Text Augmentation

Text augmentation techniques for natural language processing include synonym replacement, which replaces words with their synonyms while preserving the original meaning and label, helping models become more robust to vocabulary variations and reducing overfitting to specific word choices, improving generalization to unseen text variations. Random insertion and deletion operations randomly add or remove words from sentences, simulating natural language variations and typos, where this augmentation helps models become more robust to noisy text inputs and teaches them to focus on important content rather than exact word sequences. Back-translation translates text to another language and then back to the original language, creating paraphrased versions with the same meaning, generating diverse sentence structures while preserving semantic content and helping models learn more robust language representations. Paraphrasing rewrites sentences using different wording while maintaining the same meaning and label, exposing models to various ways of expressing the same concept and improving their ability to generalize to different writing styles and linguistic variations.

### 7.2.4 Audio Augmentation

Audio augmentation techniques for speech and audio processing include time stretching, which changes the duration of audio signals without affecting pitch, simulating different speaking rates and helping models become robust to variations in speech tempo while preserving the fundamental frequency characteristics and semantic content of the audio. Pitch shifting modifies the fundamental frequency of audio while keeping the duration constant, simulating different voice characteristics and helping models learn pitch-invariant features while improving generalization across speakers with different vocal ranges. Adding background noise introduces various types of noise to simulate real-world acoustic environments, helping models become robust to environmental factors like room acoustics, background conversations, and equipment noise, improving performance in noisy conditions. SpecAugment randomly masks frequency bands or time segments in spectrograms, forcing models to learn from partial information and encouraging the network to develop more robust acoustic representations that don't rely on specific frequency or temporal patterns.

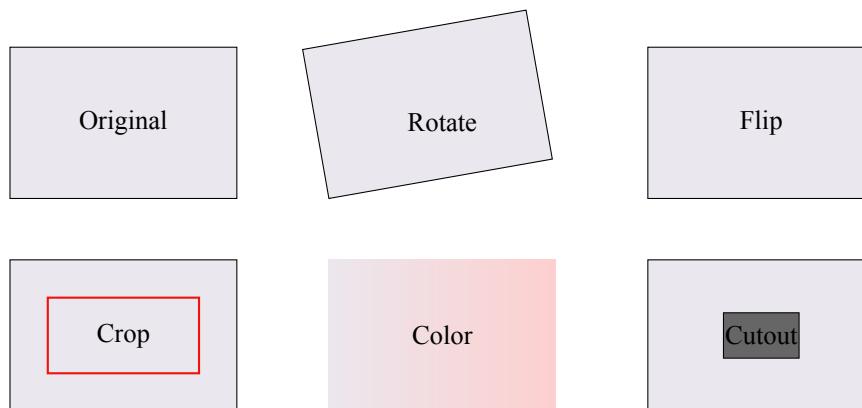


Figure 7.1: Illustration of common image augmentations. Variants preserve labels while encouraging invariances.

## 7.3 Early Stopping ♦

**Early stopping** monitors validation performance and stops training when it begins to degrade.

### 7.3.1 Intuition: Stop Before You Memorize

Imagine studying for an exam. Initially, practice improves your understanding (training and validation improve). If you keep cramming the exact same questions, you start memorizing answers that don't help with new questions (training improves, validation worsens). Early stopping is the principle of stopping at the point of best validation performance to avoid memorization.

### 7.3.2 Algorithm

The early stopping algorithm works by training the model and evaluating it on the validation set periodically, tracking the best validation performance throughout training. If there is no improvement for  $p$  epochs (patience), the training stops, and the algorithm returns the model with the best validation performance, ensuring that the network is saved at its optimal generalization point rather than continuing to overfit.

#### Remark 7.1: Early Stopping in Modern Frameworks

Modern deep learning frameworks like Hugging Face Transformers, TensorFlow, and PyTorch provide built-in early stopping capabilities that can be configured manually or systematically. These frameworks offer options to monitor validation metrics, set patience parameters, save checkpoints automatically, and restore the best model, making early stopping implementation straightforward and robust across different model architectures and training scenarios.

---

#### Algorithm 3 Early stopping meta-algorithm

---

```

Require:  $n \geq 1$  (number of steps between evaluations)
Require:  $p \geq 1$  (patience: number of worsened validations before stopping)
Require:  $\theta_0$  (initial parameters)
Ensure: Best parameters  $\theta^*$  and best step  $i^*$ 
 $\theta \leftarrow \theta_0, i \leftarrow 0, j \leftarrow 0, v \leftarrow \infty$ 
 $\theta^* \leftarrow \theta, i^* \leftarrow i$ 
while  $j < p$  do
    Update  $\theta$  by running the training algorithm for  $n$  steps
     $i \leftarrow i + n$ 
     $v' \leftarrow \text{ValidationSetError}(\theta)$ 
    if  $v' < v$  then
         $j \leftarrow 0$ 
         $\theta^* \leftarrow \theta, i^* \leftarrow i, v \leftarrow v'$ 
    else
         $j \leftarrow j + 1$ 
    end if
end while
return  $\theta^*, i^*$ 
```

---

### 7.3.3 Benefits

Early stopping is simple and effective, requiring only tracking validation performance and a patience parameter while being widely used in practice. It automatically determines training duration by finding a good stopping time without a pre-fixed epoch budget, often saving substantial compute resources. Early stopping provides implicit regularization by halting before convergence, which limits effective capacity by keeping weights smaller and preventing memorization, where in some regimes it mimics an L2 constraint under gradient descent. It is compatible with many settings, working with any loss function, architecture including MLPs, CNNs, and Transformers, and any optimizer. Early stopping reduces computational cost by stopping training as soon as overfitting begins, reducing energy and time requirements, while improving generalization stability by curbing validation variance late in training when overfitting spikes.

### 7.3.4 Considerations

Early stopping requires a reliable validation protocol with a proper validation set and evaluation cadence, where noisy metrics may trigger premature stops and should be addressed using smoothing or requiring monotone improvements. The patience parameter  $p$  and evaluation interval  $n$  interact with learning rate schedules, where too small  $p$  can stop before a scheduled learning rate drop helps. Checkpointing is crucial, as you should always restore the best model rather than the last one, keeping track of the weights at the best validation step. With warmup or long plateaus, consider larger patience or metric smoothing to avoid premature stopping. For tasks with multiple metrics like accuracy and calibration, pick the primary metric or create a

composite metric. In distributed training, ensure validation statistics are aggregated consistently across devices to avoid spurious decisions. Early stopping predates modern deep learning and was popular in classical neural nets and boosting as a strong regularizer, remaining a standard baseline technique.

### Example 7.1

**Example (vision):** Train a ResNet on CIFAR-10 with validation accuracy checked each epoch; use patience  $p = 20$ . Accuracy peaks at epoch 142; training halts at 162 without improvement, and the checkpoint from 142 is used for testing.

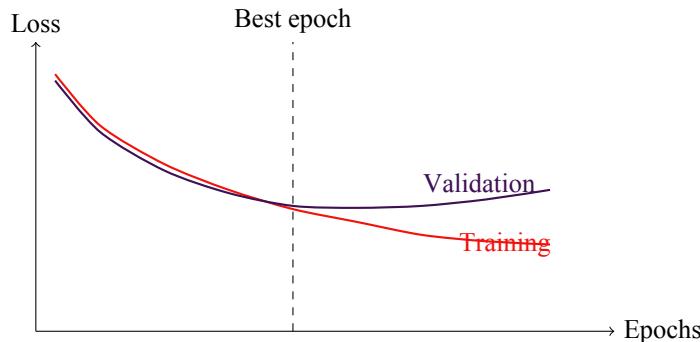


Figure 7.2: Early stopping: validation loss minimum before training loss; best checkpoint saved and restored.

## 7.4 Dropout ♦

**Dropout** randomly deactivates neurons during training, preventing co-adaptation.

### 7.4.1 Intuition: Training a Robust Ensemble

Dropout is like asking different subsets of a team to work on the same task on different days. No single member can rely on a particular colleague always being present, so each learns to be broadly useful. This results in a robust team (model) that performs well even when some members (neurons) are inactive.

### 7.4.2 Training with Dropout

Training with dropout involves sampling a binary mask  $\mathbf{m}$  with  $P(m_i = 1) = p$  for each layer at each training step, then applying the mask to the activations:  $\mathbf{h} = \mathbf{m} \odot \mathbf{h}$ . Mathematically, this is expressed as  $\mathbf{h}_{\text{dropout}} = \mathbf{m} \odot f(\mathbf{W}\mathbf{x} + \mathbf{b})$  where  $m_i \sim \text{Bernoulli}(p)$ , effectively randomly deactivating neurons with probability  $1 - p$  during training to prevent co-adaptation and improve generalization.

### 7.4.3 Inference

At test time, scale outputs by dropout probability:

$$\mathbf{h}_{\text{test}} = p \cdot f(\mathbf{W}\mathbf{x} + \mathbf{b}) \quad (7.6)$$

Or equivalently, scale weights during training by  $\frac{1}{p}$  (inverted dropout).

In practice, modern frameworks implement *inverted dropout*: during training, activations are scaled by  $\frac{1}{p}$  after masking so that the expected activation matches test-time activations, and no scaling is needed at inference [Sri+14; GBC16a]. For convolutional layers, use the same  $p$  per feature map to avoid distribution shift.

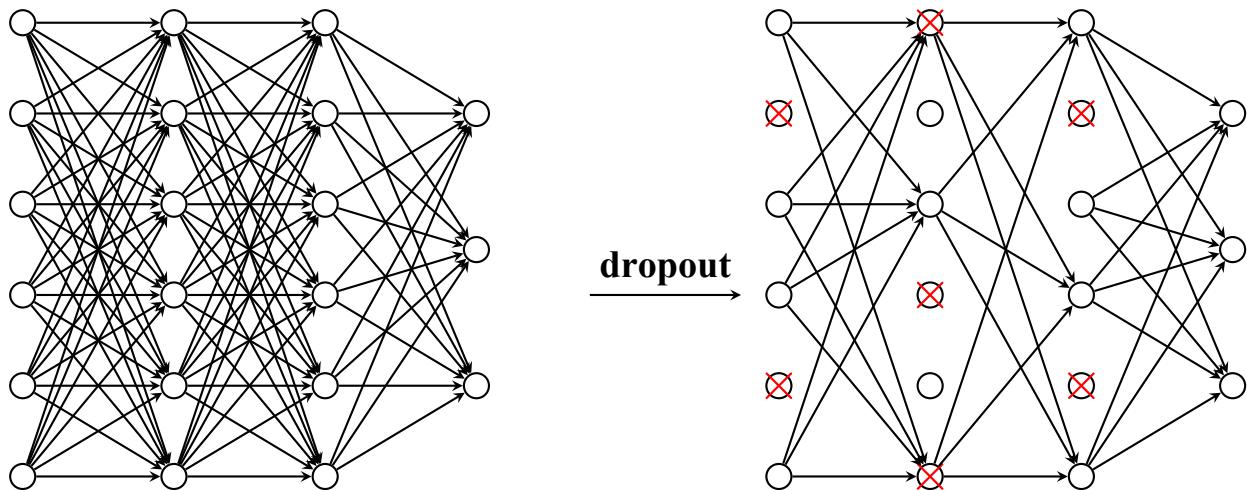


Figure 7.3: Dropout training: randomly deactivating neurons ( $\times$ ) creates subnetworks, forcing robustness.

#### 7.4.4 Interpretation

Dropout can be viewed as an implicit ensemble where sampling masks trains an ensemble of  $2^n$  subnetworks whose shared weights yield a form of model averaging. It also acts as noise injection where multiplicative Bernoulli noise on activations provides data-dependent regularization analogous to adding Gaussian noise for linear models. Additionally, dropout relates to approximate Bayesian inference where with appropriate priors, it connects to variational inference, and applying dropout at test time with multiple passes (MC dropout) estimates predictive uncertainty, making it useful for uncertainty quantification in safety-critical applications.

##### Example 7.2

**Example (uncertainty):** Run  $T = 20$  stochastic forward passes with dropout enabled at test time and average predictions to obtain mean and variance; high variance flags low-confidence inputs.

#### 7.4.5 Variants

DropConnect drops individual weights instead of activations, promoting sparsity at the parameter level and providing a different form of regularization. DropConnect is a variant of dropout that randomly sets individual weight connections to zero during training rather than entire neuron activations, creating a more fine-grained regularization approach that encourages the network to learn robust representations even when specific weight connections are missing. Spatial Dropout drops entire feature maps in CNNs to preserve spatial coherence and regularize channel reliance, which is particularly useful for convolutional layers. Variational Dropout uses the same dropout mask across time steps in RNNs to avoid injecting different noise per step that can harm temporal consistency, making it more suitable for sequential data. MC Dropout keeps dropout active at inference and averages predictions to quantify epistemic uncertainty, which is useful in safety-critical applications where uncertainty estimation is crucial. Concrete and Alpha Dropout provide continuous relaxations or distributions tailored for specific activations like SELU to maintain self-normalizing properties, offering more sophisticated regularization approaches for specific network architectures.

## 7.5 Batch Normalization ◆

**Batch normalization** normalizes layer inputs across the batch dimension.

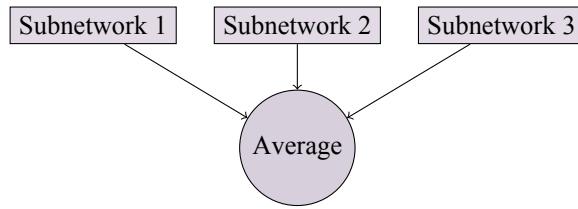


Figure 7.4: Dropout as implicit model averaging over many subnetworks.

Randomly dropped (training)

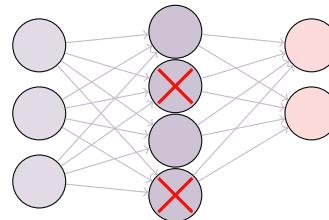


Figure 7.5: Dropout during training: randomly deactivating hidden units encourages redundancy.

### 7.5.1 Intuition: Keeping Scales Stable

Training can become unstable if the distribution of activations shifts as earlier layers update (internal covariate shift). Batch normalization re-centers and re-scales activations, keeping them in a predictable range so downstream layers see a more stable input distribution. This allows larger learning rates and speeds up training.

### 7.5.2 Algorithm

For mini-batch  $\mathcal{B}$  with activations  $x$ :

$$\mu_{\mathcal{B}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} x_i \quad (7.7)$$

$$\sigma_{\mathcal{B}}^2 = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} (x_i - \mu_{\mathcal{B}})^2 \quad (7.8)$$

$$\hat{x}_i = \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \quad (7.9)$$

$$y_i = \gamma \hat{x}_i + \beta \quad (7.10)$$

where  $\gamma$  and  $\beta$  are learnable parameters.

Implementation details: maintain running averages  $\mu_{\text{running}}, \sigma_{\text{running}}^2$  updated with momentum  $\rho$  per iteration; apply per-feature normalization for fully connected layers and per-channel per-spatial-location statistics for CNNs [IS15; GBC16a].

### 7.5.3 Benefits

Batch normalization stabilizes distributions by mitigating internal covariate shift and keeping activations in a stable range, enabling larger learning rates through better-conditioned optimization that allows faster training. It provides less sensitive initialization with a wider set of workable initializations, while the mini-batch noise in statistics acts as stochastic regularization that improves generalization. Batch normalization supports deeper networks by facilitating training of very deep architectures like ResNets, and improves gradient flow where normalized scales yield healthier signal-to-noise ratios in backpropagation, making it an essential component for modern deep learning architectures.

### 7.5.4 Inference

At test time, use running averages computed during training:

$$y = \gamma \frac{x - \mu_{\text{running}}}{\sqrt{\sigma_{\text{running}}^2 + \epsilon}} + \beta \quad (7.11)$$

Be careful when batch sizes are small at inference or differ from training: do not recompute batch statistics at test time; use the stored running averages. For distribution shift, consider recalibrating  $\mu_{\text{running}}$ ,  $\sigma_{\text{running}}^2$  with a small unlabeled buffer.

### 7.5.5 Variants

Layer Normalization normalizes across features for each sample and is effective in RNNs and Transformers where batch statistics are less stable. Group Normalization normalizes within groups of channels and is robust to small batch sizes common in detection and segmentation tasks. Instance Normalization normalizes each sample independently and is prominent in style transfer where contrast and style per instance varies. Batch Renormalization and Ghost BatchNorm adjust for mismatch between batch and population statistics or simulate small batches inside large ones for regularization, providing alternatives when standard batch normalization is not suitable.

#### Example 7.3

**Example (vision):** In a CNN with batch size 128, use per-channel BN after each convolution with momentum  $\rho = 0.9$  and  $\epsilon = 10^{-5}$ . During inference, freeze  $\gamma, \beta$  and use stored running statistics.

## 7.6 Other Regularization Techniques ◆

### 7.6.1 Intuition: Many Small Guards Against Overfitting

Beyond penalties and normalization, there are practical techniques that act like small guards during training. Each introduces a mild constraint or noise that nudges the model away from brittle solutions and encourages smoother decision boundaries.

### 7.6.2 Label Smoothing

Replace hard targets with smoothed distributions:

$$y'_k = (1 - \epsilon)y_k + \frac{\epsilon}{K} \quad (7.12)$$

Prevents overconfident predictions and improves calibration by discouraging saturated logits; commonly used in large-scale classification (e.g., ImageNet) and sequence models. Choose  $\epsilon$  in  $[0.05, 0.2]$  depending on class count  $K$  and desired calibration. It can also mitigate overfitting to annotator noise.

#### Example 7.4

**Example (ImageNet):** With  $K = 1000$  and  $\epsilon = 0.1$ , the target for the correct class becomes 0.9 while others receive 0.0001 each; top-1 accuracy and ECE often improve.

### 7.6.3 Gradient Clipping

Limit gradient magnitude to prevent exploding gradients:

**Definition 7.1: Clipping by Value**

Clipping by value limits each gradient component individually:

$$g \leftarrow \max(\min(g, \theta), -\theta) \quad (7.13)$$

where  $\theta$  is the clipping threshold. This method constrains each gradient element to lie within  $[-\theta, \theta]$ , providing component-wise control over gradient magnitudes.

**Definition 7.2: Clipping by Norm**

Clipping by norm rescales the entire gradient vector when its norm exceeds a threshold:

$$g \leftarrow \frac{g}{\max(1, \|g\|/\theta)} \quad (7.14)$$

where  $\theta$  is the clipping threshold. This method preserves the gradient direction while scaling its magnitude, making it preferred over value clipping as it maintains the relative importance of different parameters.

Clipping stabilizes training in RNNs and very deep nets by preventing exploding gradients, especially with large learning rates or noisy batches. Norm clipping with threshold  $\theta$  is preferred as it preserves direction while scaling magnitude. Excessive clipping can bias updates and slow convergence; tune  $\theta$  w.r.t. optimizer and batch size.

**Example 7.5**

**Example (NLP):** Train a GRU with global norm clip  $\theta = 1.0$ ; without clipping, gradients occasionally explode causing loss spikes.

$$g \leftarrow \frac{g}{\max(1, \|g\|/\theta)} \quad (7.15)$$

## 7.6.4 Stochastic Depth

Randomly skip residual blocks during training with survival probability  $p_l$  per layer  $l$ , while using the full network depth at test time. This shortens expected depth during training, improving gradient flow and reducing overfitting in very deep networks [Hua+16].

Let  $p_l$  decrease with depth (e.g., linearly from 1.0 to  $p_{\min}$ ). During training, with probability  $1 - p_l$  a residual block is bypassed; otherwise it is applied and its output is scaled to match test-time expectation.

**Example 7.6**

**Example (ResNets):** In a 110-layer ResNet, set  $p_l$  from 1.0 to 0.8 across depth; training converges faster and generalizes better on CIFAR-10.

## 7.6.5 Mixup

Train on convex combinations of examples:

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j \quad (7.16)$$

$$\tilde{y} = \lambda y_i + (1 - \lambda)y_j \quad (7.17)$$

where  $\lambda \sim \text{Beta}(\alpha, \alpha)$  and  $\text{Beta}(\alpha, \alpha)$  is the Beta distribution with shape parameters  $\alpha$  and  $\alpha$ , which controls the mixing strength and distribution of the interpolation weights.

Mixup encourages linear behavior between classes, reduces memorization of spurious correlations, and improves robustness to label noise [Zha+18]. Typical  $\alpha$  values are in  $[0.2, 1.0]$ . Variants include CutMix (patch-level mixing) [Yun+19] and Manifold Mixup (mix at hidden layers).

#### Example 7.7

**Example (vision):** With  $\alpha = 0.4$ , randomly pair images in a batch and form convex combinations; train using the mixed targets. Improves top-1 accuracy and calibration.

### 7.6.6 Adversarial Training

Add adversarially perturbed examples to training:

$$\mathbf{x}_{\text{adv}} = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} L(\mathbf{x}, y)) \quad (7.18)$$

This FGSM objective can be extended to multi-step PGD adversaries. Adversarial training improves worst-case robustness but often reduces clean accuracy and increases compute [Goo+14]. Robust features learned can transfer across tasks; careful tuning of  $\epsilon$ , steps, and randomness is crucial.

#### Example 7.8

**Example (robust CIFAR-10):** Use  $\epsilon = 8/255$ ,  $k = 7$  PGD steps, step size  $2/255$ ; train with a 1:1 mix of clean and adversarial samples.

#### Remark 7.2: Historical Context and Applications

Label smoothing and dropout popularized regularization at scale; gradient clipping stabilized early RNNs; stochastic depth enabled training very deep residual networks; mixup and CutMix improved data-efficient generalization; adversarial training established the modern paradigm for robustness. Applications span medical imaging, autonomous driving, speech recognition, and large-scale language models where calibration and robustness are critical [GBC16a; He+16; IS15].

## 7.7 Real World Applications

Regularization techniques are essential for making deep learning models work reliably in real-world scenarios where data is messy and models need to perform well on new, unseen examples.

### 7.7.1 Autonomous Vehicle Safety

Self-driving cars rely heavily on regularization to ensure safe operation, where robust object detection uses regularization techniques like dropout and data augmentation to help vehicles recognize pedestrians, cyclists, and other vehicles under diverse conditions including rain, fog, night driving, and unusual angles, preventing system failures when encountering weather or lighting conditions not heavily represented in training data. Generalization to new environments is crucial, as a self-driving car trained in sunny California needs to work safely in snowy Boston, where regularization prevents the model from memorizing specific training locations and instead learns general driving principles that transfer across different cities and climates. Regularization helps prevent overfitting to rare events by maintaining good performance on common scenarios like normal traffic while still being prepared for rare but critical situations such as emergency vehicles and unexpected obstacles.

### 7.7.2 Medical Imaging Analysis

Healthcare applications use regularization to make reliable diagnoses, where cancer detection from limited data benefits from regularization techniques like data augmentation and early stopping that allow models to learn effectively from hundreds rather than millions of examples, making them practical for clinical use since medical datasets are often small because annotating medical images requires expert radiologists. Consistent performance across hospitals is achieved through regularization, where different hospitals use different imaging equipment, and regularization ensures models trained at one hospital generalize to work at others despite variations in image quality, resolution, or equipment manufacturers. Regularization helps reduce false positives in medical diagnosis, where false alarms cause unnecessary anxiety and costly follow-up tests, and techniques like label smoothing help models be appropriately confident, reducing overconfident but incorrect predictions.

### 7.7.3 Natural Language Processing for Customer Service

Chatbots and virtual assistants benefit from regularization, where handling diverse customer queries is improved through regularization techniques like data augmentation including paraphrasing and synonym replacement that help chatbots understand intent even when people use unexpected wording, since customers phrase questions in countless ways. Regularization prevents memorization of training conversations, where without regularization, chatbots might memorize training examples and give nonsensical responses to new queries, but dropout and other techniques force the model to learn general conversation patterns rather than specific exchanges. Regularization helps models adapt to evolving language, where language changes constantly with new slang and terminology, and regularization helps models stay flexible and adapt to linguistic shifts without extensive retraining.

### 7.7.4 Key Benefits in Practice

Regularization provides crucial advantages in real applications, where it works with limited data since not every problem has millions of training examples, reduces maintenance costs by helping models generalize better and requiring less frequent retraining, increases reliability by ensuring systems work consistently even when deployed conditions differ from training, and enables deployment by making models trustworthy enough for safety-critical applications. These examples show that regularization isn't just a mathematical nicety—it's the difference between models that work only in labs and those that succeed in the real world, where the techniques discussed in this chapter are essential for bridging the gap between theoretical deep learning and practical applications.

## Key Takeaways

### Key Takeaways 7

- **Regularisation** constrains model capacity to improve generalisation and prevent overfitting to training data.
- **Norm penalties** ( $L_1$ ,  $L_2$ ) encourage simpler models;  $L_1$  induces sparsity whilst  $L_2$  shrinks weights uniformly.
- **Data augmentation** artificially expands training sets by applying semantics-preserving transformations.
- **Dropout** randomly drops units during training, forcing redundant representations and reducing co-adaptation.
- **Early stopping and batch normalisation** are simple yet powerful techniques for better training dynamics and generalisation.

## Exercises

### Easy

*Exercise 7.1 (L1 vs L2 Regularisation).* Explain the difference between L1 and L2 regularisation. Which one is more likely to produce sparse weights, and why?

**Hint:**

*Consider the shape of the L1 and L2 penalty terms and their gradients.*

*Exercise 7.2 (Data Augmentation Strategies).* List three common data augmentation techniques for image classification tasks and explain how each helps improve generalisation.

**Hint:**

*Think about geometric transformations, colour adjustments, and realistic variations.*

*Exercise 7.3 (Early Stopping).* Describe how early stopping works as a regularisation technique. What metric should you monitor, and when should you stop training?

**Hint:**

*Consider validation set performance and the risk of overfitting to the training set.*

*Exercise 7.4 (Dropout Interpretation).* During training, dropout randomly sets activations to zero with probability  $p$ . During inference, all neurons are active but their outputs are scaled. Explain why this scaling is necessary.

**Hint:**

*Think about the expected value of activations during training versus inference.*

### Medium

*Exercise 7.5 (Regularisation Trade-off).* Given a model with both L2 regularisation and dropout, discuss how you would tune the regularisation strength  $\lambda$  and dropout rate  $p$ . What signs would indicate too much or too little regularisation?

**Hint:**

*Monitor training and validation loss curves, and consider the bias-variance trade-off.*

*Exercise 7.6 (Batch Normalisation Effect).* Explain how batch normalisation acts as a regulariser. Discuss its interaction with dropout.

**Hint:**

*Consider the noise introduced by computing statistics on mini-batches and why dropout is often not needed with batch normalisation.*

### Hard

*Exercise 7.7 (Mixup Derivation).* Mixup trains on convex combinations of examples:  $\tilde{\mathbf{x}} = \lambda \mathbf{x}_i + (1 - \lambda) \mathbf{x}_j$  where  $\lambda \sim \text{Beta}(\alpha, \alpha)$ . Derive how this affects the decision boundary and explain why it improves generalisation.

**Hint:**

*Consider the effect on the loss surface and the implicit regularisation from interpolating between examples.*

*Exercise 7.8 (Adversarial Training).* Design an adversarial training procedure for a classification model. Explain how to generate adversarial examples using FGSM (Fast Gradient Sign Method) and why this improves robustness.

**Hint:**

*Adversarial examples are  $\mathbf{x}_{adv} = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} L)$ . Discuss the trade-off between clean and adversarial accuracy.*

*Exercise 7.9 (Early Stopping Strategy).* Explain how early stopping works as a regularisation technique. How do you determine the optimal stopping point?

**Hint:**

*Monitor validation loss and stop when it starts increasing, indicating overfitting.*

*Exercise 7.10 (Data Augmentation Effects).* Compare different data augmentation techniques for image classification. Which techniques are most effective for different types of images?

**Hint:**

*Consider geometric transformations, color jittering, and mixup techniques.*

*Exercise 7.11 (Weight Decay vs Dropout).* Compare the effects of weight decay and dropout regularisation. When would you use one over the other?

**Hint:**

*Weight decay penalises large weights globally, while dropout creates sparse activations locally.*

*Exercise 7.12 (Batch Normalization vs Layer Normalization).* Compare batch normalisation and layer normalisation. When is each more appropriate?

**Hint:**

*Batch normalisation depends on batch statistics, while layer normalisation is independent of batch size.*

*Exercise 7.13 (Regularisation in Convolutional Networks).* Explain how regularisation techniques differ when applied to convolutional layers versus fully connected layers.

**Hint:**

*Consider spatial structure preservation and parameter sharing in convolutional layers.*

*Exercise 7.14 (Ensemble Regularisation).* How can ensemble methods be viewed as a form of regularisation? Compare bagging and boosting approaches.

**Hint:**

*Ensembles reduce variance by averaging predictions from multiple models.*



# Chapter 8

# Optimization for Training Deep Models

This chapter covers optimization algorithms and strategies for training deep neural networks effectively. Modern optimizers go beyond basic gradient descent to accelerate convergence and improve performance.

## Learning Objectives

1. Gradient descent variants and appropriate batch size selection
2. Momentum-based methods including classical momentum and Nesterov accelerated gradient
3. Adaptive optimizers (AdaGrad, RMSProp, Adam) and hyperparameter tuning
4. Second-order approaches and when they are practical
5. Optimization challenges and their remedies
6. Optimization plan combining initializer, optimizer, and learning rate schedule

## 8.1 Gradient Descent Variants ◆

Gradient descent variants differ in how they compute gradients, balancing computational efficiency with convergence stability through different batch sizes and update strategies.

### 8.1.1 Intuition: Following the Steepest Downhill Direction

Imagine standing on a foggy hillside trying to reach the valley. You feel the slope under your feet and take a step downhill. **Batch gradient descent** measures the average slope using the whole landscape (dataset) before each step: accurate but slow to gauge. **Stochastic gradient descent (SGD)** feels the slope at a single point: fast, but noisy. **Mini-batch** is a compromise: feel the slope at a handful of nearby points to get a reliable yet efficient direction. This trade-off underlies most practical training regimes.mini-batch

Historical note: Early neural network training widely used batch gradient descent, but the rise of large datasets and GPUs made mini-batch SGD the de facto standard [GBC16a; Pri23].

### 8.1.2 Batch Gradient Descent

Batch gradient descent computes the gradient using the entire training set, providing the most accurate gradient estimate at each step. The mathematical formulation updates parameters by taking the average gradient across all training examples, ensuring each update is based on complete information about the loss landscape.

The update rule computes the gradient of the average loss over all training examples:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha \nabla_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^n L(\boldsymbol{\theta}, \mathbf{x}^{(i)}, y^{(i)}) \quad (8.1)$$

This approach provides deterministic and low-variance updates that are well-suited for convex optimization problems. However, each step requires processing all  $n$  examples, incurring high computational and memory costs for large datasets. In deep, non-convex landscapes, batch gradient descent remains stable but often responds too slowly to curvature changes, making it less practical for modern deep learning applications.

The mathematical foundation becomes clear when examining a convex quadratic example. For a loss function  $L(\boldsymbol{\theta}) = \frac{1}{2}a\boldsymbol{\theta}^2$  with gradient  $a\boldsymbol{\theta}$ , batch gradient descent with step size  $\alpha$  yields the update  $\boldsymbol{\theta}_{t+1} = (1 - \alpha a)\boldsymbol{\theta}_t$ . Convergence occurs when  $0 < \alpha < \frac{2}{a}$ , illustrating the critical interaction between learning rate and curvature. This relationship demonstrates why batch gradient descent works well for small datasets and convex objectives like linear or logistic regression, where precise convergence is desired.

Historical context: Full-batch methods trace back to classical numerical optimization. For massive datasets, stochastic approximations dating to [RM51] became essential; modern deep learning typically favors mini-batches [GBC16a; GBC16b; Zha+24b].

### 8.1.3 Stochastic Gradient Descent (SGD)

Stochastic gradient descent uses a single random example per update, providing very fast and streaming-friendly updates that can begin learning immediately. The mathematical formulation updates parameters using the gradient of a single training example, introducing controlled randomness that helps escape local minima and saddle points.

The update rule processes one example at a time:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \mathbf{x}^{(i)}, y^{(i)}) \quad (8.2)$$

The noisy gradients inherent in SGD add valuable exploration capabilities, helping the optimizer traverse saddle points and plateaus that might trap more deterministic methods. However, this high variance can hamper precise convergence, making it essential to use diminishing learning rates  $\alpha_t$  to stabilize the optimization process.

Practical implementation requires careful attention to learning rate scheduling. Using a decaying schedule such as  $\alpha_t = \alpha_0 / (1 + \lambda t)$  helps stabilize convergence, while switching to momentum or Adam methods later can provide fine-tuning capabilities. Additionally, shuffling examples every epoch prevents periodic bias and ensures the stochastic nature remains beneficial rather than introducing systematic patterns.[RM51; GBC16b; Zha+24b]

Applications: Early CNN training and many online/streaming scenarios employ SGD due to its simplicity and ability to handle large-scale data. In large vision models, SGD with momentum remains competitive [He+16].

### 8.1.4 Mini-Batch Gradient Descent

Mini-batch gradient descent balances the computational efficiency of stochastic methods with the stability of batch methods by processing small subsets of the training data. This approach provides an optimal compromise between gradient variance and computational cost, making it the de facto standard for modern deep learning.

The mathematical formulation processes a mini-batch of examples:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha \nabla_{\boldsymbol{\theta}} \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} L(\boldsymbol{\theta}, \mathbf{x}^{(i)}, y^{(i)}) \quad (8.3)$$

where  $\mathcal{B}$  is a mini-batch typically containing 32 – 1024 examples, depending on the model architecture and available hardware. This approach significantly reduces gradient variance compared to pure stochastic methods while maintaining computational efficiency through parallel processing on modern GPUs. The mini-batch size provides a crucial hyperparameter that balances

training speed with gradient stability, enabling efficient utilization of hardware resources while maintaining good convergence properties.

The choice of mini-batch size involves important trade-offs that affect both optimization dynamics and computational efficiency. Larger batches yield smoother gradient estimates but may require proportionally larger learning rates following the "linear scaling rule" heuristic. However, very large batches can hurt generalization unless paired with appropriate warmup strategies and regularization techniques, making it essential to carefully balance batch size with learning rate and other hyperparameters.[[GBC16b](#); [Zha+24b](#)]

Illustrative example (variance vs. batch size): For a fixed  $\alpha$ , increasing  $|\mathcal{B}|$  reduces the variance of the stochastic gradient approximately as  $\mathcal{O}(1/|\mathcal{B}|)$ , improving stability but diminishing returns once hardware is saturated.

## 8.2 Momentum-Based Methods ◆

One-liner: Momentum smooths gradients by accumulating an exponential moving average of past updates, accelerating descent along gentle directions while damping oscillations in steep ones.

### 8.2.1 Intuition: Rolling a Ball Down a Valley

Plain SGD can wobble like a marble on a bumpy path. **Momentum** acts like mass: it carries velocity so you keep moving in consistently good directions and smooth out small bumps. Imagine a heavy ball rolling down a hill—its mass (momentum) helps it maintain direction even when hitting small bumps, just like how momentum accumulates past gradients to maintain consistent movement in the loss landscape. The derivative of momentum with respect to time gives us the acceleration, but in optimization, we use the derivative of the loss with respect to parameters to determine the direction, while momentum provides the "inertia" to keep moving smoothly. **Nesterov acceleration** adds anticipation by peeking where the momentum will take you before correcting, often yielding crisper steps in curved valleys. Think of it like a skilled skier who looks ahead to anticipate the curve and adjusts their trajectory before reaching it, rather than just following their current momentum and then correcting afterward. This "look-ahead" approach helps the optimizer make more informed decisions by evaluating the gradient at the anticipated future position, leading to smoother and more efficient convergence.

Historical note: Momentum has deep roots in convex optimization and was popularized in early neural network training; Nesterov's variant provided stronger theoretical guarantees in convex settings and inspired practical variants in deep learning [[Pol64](#); [Nes83](#); [GBC16a](#); [Bis06](#)].

### 8.2.2 Momentum

Momentum accumulates gradients over time to maintain velocity in consistently good directions, effectively smoothing out oscillations and accelerating convergence in relevant directions. The mathematical formulation builds an exponentially weighted moving average of past gradients, creating a low-pass filter that suppresses high-frequency noise while preserving important directional information.

The update equations accumulate velocity and apply it to parameter updates:

$$\mathbf{v}_t = \beta \mathbf{v}_{t-1} - \alpha \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t) \quad (8.4)$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \mathbf{v}_t \quad (8.5)$$

where  $\beta \in [0, 1]$  is the momentum coefficient, typically set to 0.9, controlling how much past gradient information to retain. This approach accelerates convergence in relevant directions by maintaining velocity along consistent gradient directions while dampening oscillations that occur when gradients change rapidly. The momentum mechanism helps escape local minima and saddle points by providing sufficient inertia to overcome small gradient magnitudes that might otherwise trap the optimizer. The mathematical interpretation reveals that momentum implements an exponentially weighted moving average of past gradients, effectively creating a low-pass filter that suppresses high-frequency noise. In anisotropic valleys with different curvatures along different dimensions, momentum allows larger effective steps along shallow curvature directions while

reducing the characteristic zig-zag pattern across sharp directions. The choice of hyperparameters requires careful tuning, with  $\beta \in [0.8, 0.99]$  providing different levels of smoothing and memory, while the learning rate  $\alpha$  must be balanced to prevent divergence.[[Pol64](#); [GBC16b](#); [Zha+24b](#)]

Example (ravine function): For  $L(\theta_1, \theta_2) = 100\theta_1^2 + \theta_2^2$ , momentum reduces oscillations in the steep  $\theta_1$  direction and speeds travel along the gentle  $\theta_2$  direction.

### 8.2.3 Nesterov Accelerated Gradient (NAG)

Nesterov Accelerated Gradient implements a "look-ahead" version of momentum that evaluates the gradient at an anticipated future position, providing more informed updates that often yield better convergence properties. This anticipatory approach allows the optimizer to correct its course earlier, reducing overshoot in curved valleys and improving convergence rates.

The mathematical formulation computes gradients at the look-ahead position:

$$\mathbf{v}_t = \beta \mathbf{v}_{t-1} - \alpha \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t + \beta \mathbf{v}_{t-1}) \quad (8.6)$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \mathbf{v}_t \quad (8.7)$$

By computing the gradient at the look-ahead point  $\boldsymbol{\theta}_t + \beta \mathbf{v}_{t-1}$ , NAG corrects the course earlier than standard momentum, which reduces overshoot in curved valleys and can significantly improve convergence rates, particularly in convex optimization settings. This anticipatory mechanism allows the optimizer to make more informed decisions by evaluating the gradient at where the momentum will take it, rather than just following the current momentum and correcting afterward.[[Nes83](#); [GBC16b](#); [GBC16a](#)]

#### Remark 8.1: Nesterov Accelerated Gradient Practice Notes

Common defaults:  $\beta = 0.9$ , initial  $\alpha \in [10^{-3}, 10^{-1}]$  depending on scale. Widely used with SGD in large-scale vision models [[He+16](#)]. Start with  $\beta = 0.9$  and tune  $\alpha$  based on your loss scale; for well-normalized networks,  $\alpha = 0.01$  often works well. NAG typically requires fewer iterations than standard momentum to converge, making it particularly valuable for expensive training runs. The look-ahead gradient computation adds minimal computational overhead (one extra gradient evaluation per step) while often providing significant convergence improvements. Consider using NAG when training deep networks with many parameters, as the anticipation effect helps navigate complex loss landscapes more efficiently than standard momentum.

## 8.3 Adaptive Learning Rate Methods ◆

Adaptive learning rate methods automatically adjust step sizes per parameter based on historical gradient information, allowing faster progress on rarely-updated dimensions while stabilizing highly-volatile parameters.

### 8.3.1 Intuition: Per-Parameter Step Sizes

Different parameters learn at different speeds: some directions are steep, others are flat. **Adaptive methods** adjust the step size per parameter based on recent gradient information, allowing faster progress on rarely-updated or low-variance dimensions while stabilizing steps on highly-volatile ones.

Historical note: AdaGrad emerged for sparse exercises; RMSProp stabilized AdaGrad's decay; Adam blended momentum with RMSProp-style adaptation and became a widely used default in deep learning [[DHS11](#); [TH12](#); [KB14](#); [GBC16a](#)].

### 8.3.2 AdaGrad

Adapts learning rate per parameter based on historical gradients:

$$\mathbf{g}_t = \nabla_{\theta} L(\theta_t) \quad (8.8)$$

$$\mathbf{r}_t = \mathbf{r}_{t-1} + \mathbf{g}_t \odot \mathbf{g}_t \quad (8.9)$$

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{\mathbf{r}_t + \epsilon}} \odot \mathbf{g}_t \quad (8.10)$$

where  $\epsilon$  (e.g.,  $10^{-8}$ ) prevents division by zero. AdaGrad is well-suited to sparse features: infrequent parameters receive larger effective steps, accelerating learning in NLP and recommender settings [DHS11; GBC16b; Zha+24b]. A drawback is the ever-growing accumulator  $\mathbf{r}_t$ , which can shrink steps too aggressively over long runs.

### 8.3.3 RMSProp

RMSProp (Root Mean Square Propagation) addresses AdaGrad's aggressive decay using exponential moving average:

$$\mathbf{r}_t = \rho \mathbf{r}_{t-1} + (1 - \rho) \mathbf{g}_t \odot \mathbf{g}_t \quad (8.11)$$

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{\mathbf{r}_t + \epsilon}} \odot \mathbf{g}_t \quad (8.12)$$

Add a small  $\epsilon$  for numerical stability and tune decay  $\rho \in [0.9, 0.99]$ . RMSProp prevents the learning rate from decaying to zero as in AdaGrad, making it effective for non-stationary objectives typical in deep networks [TH12; GBC16b; Zha+24b].

### 8.3.4 Adam (Adaptive Moment Estimation)

Combines momentum and adaptive learning rates:

$$\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1) \mathbf{g}_t \quad (8.13)$$

$$\mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1 - \beta_2) \mathbf{g}_t \odot \mathbf{g}_t \quad (8.14)$$

$$\hat{\mathbf{m}}_t = \frac{\mathbf{m}_t}{1 - \beta_1^t} \quad (8.15)$$

$$\hat{\mathbf{v}}_t = \frac{\mathbf{v}_t}{1 - \beta_2^t} \quad (8.16)$$

$$\theta_{t+1} = \theta_t - \frac{\alpha \hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t} + \epsilon} \quad (8.17)$$

Default hyperparameters:  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ ,  $\epsilon = 10^{-8}$ ,  $\alpha = 0.001$  [KB14]. Adam often converges quickly and is robust to poorly scaled gradients. For best generalization in some vision tasks, SGD with momentum can still outperform Adam; consider switching optimizers during fine-tuning [GBC16a; Zha+24b; He+16].

### 8.3.5 Learning Rate Schedules

Learning rate schedules systematically adjust the learning rate during training to balance exploration and exploitation, often improving convergence speed and final performance. Different schedules provide various strategies for managing the learning rate over time, each suited to different optimization scenarios and model architectures.

Step decay reduces the learning rate by a factor at predetermined intervals, providing a simple yet effective approach for many applications:

$$\alpha_t = \alpha_0 \cdot \gamma^{\lfloor t/s \rfloor} \quad (8.18)$$

Exponential decay provides smooth reduction of the learning rate over time, often useful for fine-tuning and convergence:

$$\alpha_t = \alpha_0 e^{-\lambda t} \quad (8.19)$$

Cosine annealing creates a smooth, periodic learning rate schedule that can help escape local minima:

$$\alpha_t = \alpha_{\min} + \frac{1}{2}(\alpha_{\max} - \alpha_{\min}) \left( 1 + \cos \left( \frac{t}{T} \pi \right) \right) \quad (8.20)$$

Additional scheduling strategies include warmup techniques that start from a small learning rate and increase linearly over the first  $T_w$  steps to reduce early instabilities in large-batch training. The one-cycle policy increases then anneals the learning rate, often paired with momentum decay, to speed convergence and improve generalization. Practical implementation typically combines cosine decay with warmup for transformer-like models, step decay for CNNs trained with SGD, and exponential decay for simple baselines.[[GBC16b](#); [Zha+24b](#)]

## 8.4 Second-Order Methods ◆

Second-order methods utilize curvature information from the Hessian matrix to make more informed optimization steps, providing faster convergence than first-order methods but requiring significant computational resources for large neural networks.

### 8.4.1 Intuition: Curvature-Aware Steps

First-order methods follow the slope; second-order methods also consider the *curvature* of the landscape to choose better-scaled steps. If the valley is sharply curved in one direction and flat in another, curvature-aware steps shorten strides along the sharp direction and lengthen them along the flat one.

Historical note: Classical optimization popularized Newton and quasi-Newton methods; in deep learning, memory and compute constraints motivated approximations like L-BFGS and natural gradient [[LN89](#); [Ama98](#); [GBC16a](#); [Bis06](#)].

### 8.4.2 Newton's Method

Newton's method uses second-order Taylor expansion to make curvature-aware optimization steps, providing theoretically optimal convergence rates for quadratic functions. The mathematical formulation incorporates the Hessian matrix to rescale gradients according to local curvature, yielding steps that are invariant to parameter scaling.

The update rule incorporates second-order information:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mathbf{H}^{-1} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t) \quad (8.21)$$

where  $\mathbf{H}$  is the Hessian matrix containing second-order derivatives. The Hessian rescales the gradient by local curvature, yielding steps that naturally adapt to the geometry of the loss landscape. In quadratic bowls, this approach takes longer steps along shallow directions and shorter steps along steep directions, providing optimal convergence properties.

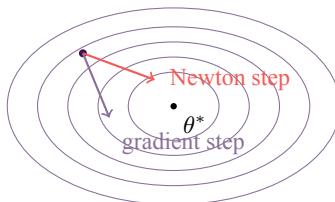


Figure 8.1: Newton's method rescales gradient by curvature: longer steps in shallow, shorter in steep.

Despite its theoretical advantages, Newton's method faces significant computational challenges that limit its practical application in deep learning. Computing the Hessian matrix requires  $O(n^2)$  operations in the number of parameters, while inverting the Hessian demands  $O(n^3)$  operations, making the approach computationally infeasible for large neural networks with millions of parameters. These computational requirements grow quadratically and cubically with the number of parameters, respectively, creating prohibitive memory and time costs for modern deep learning architectures.

### 8.4.3 Quasi-Newton Methods

Quasi-Newton methods approximate the Hessian inverse using iterative updates that avoid the computational burden of computing and inverting the full Hessian matrix. These approaches maintain low-rank approximations of the Hessian inverse, providing a practical compromise between computational efficiency and second-order optimization benefits.

The L-BFGS (Limited-memory Broyden-Fletcher-Goldfarb-Shanno) algorithm represents the most widely used quasi-Newton method in deep learning, maintaining a low-rank approximation of the Hessian inverse that is significantly more efficient than full Newton's method. While still computationally expensive for very large models, L-BFGS remains valuable for smaller networks or specific applications where second-order information provides substantial benefits. The limited-memory approach stores only recent gradient differences, making it feasible for moderate-sized neural networks while preserving much of the convergence advantages of second-order methods.

Historical note: Quasi-Newton methods, notably BFGS and its limited-memory variant L-BFGS [LN89], performed well on moderate-sized networks and remain valuable for fine-tuning smaller models or optimizing differentiable components inside larger systems [GBC16a; Bis06].

### 8.4.4 Natural Gradient

Uses Fisher information matrix instead of Hessian:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha \mathbf{F}^{-1} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t) \quad (8.22)$$

Provides parameter updates invariant to reparameterization. In probabilistic models,  $\mathbf{F}$  is the Fisher information, defining a Riemannian metric on the parameter manifold; stepping along  $\mathbf{F}^{-1} \nabla L$  follows the steepest descent in information geometry [Ama98]. Approximations (e.g., K-FAC) make natural gradient practical in deep nets by exploiting layer structure (see Figure 8.2).

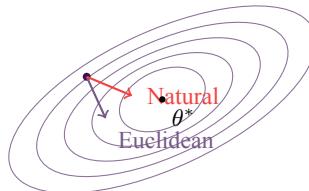


Figure 8.2: Natural gradient accounts for local geometry (Fisher metric), directing updates orthogonally.

## 8.5 Optimization Challenges ◆

Deep neural network optimization faces unique challenges including vanishing and exploding gradients, saddle points, and plateaus that require specialized techniques and careful hyperparameter tuning to overcome.

### 8.5.1 Intuition: Why Training Gets Stuck

Deep networks combine nonlinearity and depth, creating landscapes with flat plateaus, narrow valleys, and saddle points. Noise (SGD), momentum, and schedules act like navigational aids to keep moving and avoid getting stuck.

### 8.5.2 Vanishing and Exploding Gradients

In deep networks, gradients can become exponentially small or large during backpropagation, creating fundamental challenges for training. Vanishing gradients occur when gradients become exponentially small as they propagate backward through the network, particularly common with sigmoid and tanh activation functions that have small derivatives. This phenomenon severely limits the ability of early layers to learn effectively, as their gradients become too small to drive meaningful parameter updates.

Exploding gradients represent the opposite problem, where gradients grow exponentially large during backpropagation, particularly common in recurrent neural networks due to the repeated application of the same weight matrices. These large gradients can cause parameter updates to become unstable, leading to training divergence or numerical overflow.

Modern deep learning addresses these challenges through several key techniques. ReLU activation functions, batch normalization, and residual connections help mitigate vanishing gradients by providing alternative pathways for gradient flow. Gradient clipping provides a direct solution to exploding gradients by limiting the magnitude of gradients during backpropagation. The mathematical formulation of gradient clipping rescales gradients when they exceed a threshold:

$$\mathbf{g} \leftarrow \frac{\mathbf{g}}{\max(1, \|\mathbf{g}\|/\theta)} \quad (8.23)$$

where  $\theta$  represents the clipping threshold. This approach prevents individual parameter updates from becoming too large while preserving the relative direction of the gradient vector.

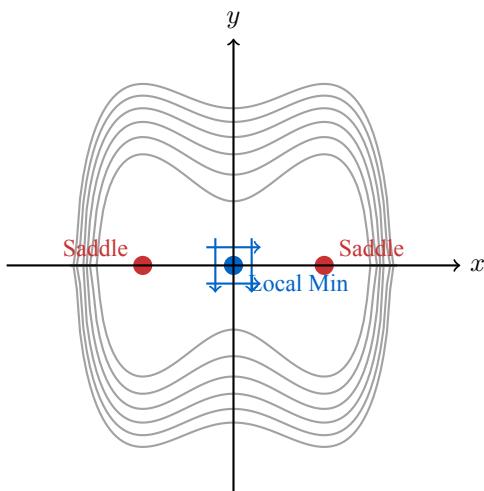
### 8.5.3 Local Minima and Saddle Points

In high-dimensional optimization landscapes, saddle points are significantly more common than local minima, making them the primary challenge for gradient-based optimization methods. Saddle points represent critical points where the gradient is zero but the curvature is mixed, containing both positive and negative eigenvalues in the Hessian matrix. These points create deceptive flat regions where gradient descent can become trapped, as the zero gradient provides no directional information for escape. The mathematical characterization of saddle points reveals their deceptive nature. At a saddle point, the gradient vanishes completely, but unlike local minima where all eigenvalues of the Hessian are positive, saddle points have mixed curvature with both positive and negative eigenvalues. This mixed curvature means that while some directions lead downhill (negative eigenvalues), others lead uphill (positive eigenvalues), creating a complex optimization landscape where simple gradient descent can become trapped.

Momentum and noise provide essential mechanisms for escaping saddle points in high-dimensional spaces. Momentum accumulates velocity from past gradients, providing sufficient inertia to overcome the zero gradient at saddle points and continue optimization. Stochastic noise from mini-batch sampling adds random perturbations that help the optimizer escape these deceptive flat regions by providing small random forces that can push the optimization trajectory away from saddle points.

#### Example: Critical Points in Two Dimensions

Consider the function  $f(x, y) = x^4 - 2x^2 + y^2$  to illustrate different types of critical points. This function provides a clear example of how different critical points create distinct optimization challenges in the loss landscape.



Critical Points in Optimization Landscape

Figure 8.3: Critical points in  $f(x, y) = x^4 - 2x^2 + y^2$ : minimum at  $(0, 0)$ , saddle points at  $(\pm 1, 0)$ .

The function  $f(x, y) = x^4 - 2x^2 + y^2$  demonstrates three distinct types of critical points that commonly appear in optimization landscapes. At the local minimum  $(0, 0)$ , the function value  $f(0, 0) = 0$  represents the lowest point in the neighborhood, with all nearby points having higher function values. The gradient vanishes at this point, and the Hessian matrix has positive eigenvalues, indicating concave-up curvature in all directions.

Saddle points at  $(\pm 1, 0)$  represent the most challenging critical points for optimization, where  $f(\pm 1, 0) = -1$  creates deceptive flat regions. These points have zero gradients but mixed curvature, with the Hessian containing both positive and negative eigenvalues. This mixed curvature means that while some directions lead downhill, others lead uphill, creating a complex landscape where gradient descent can become trapped.

In high-dimensional optimization, saddle points are much more common than local minima, making them the primary challenge for gradient-based methods. The visualization clearly shows how saddle points create deceptive flat regions where optimization can stall, while local minima provide clear convergence targets.

### 8.5.4 Plateaus

Flat regions with small gradients slow convergence. Adaptive methods and learning rate schedules help navigate plateaus.

### 8.5.5 Practical Optimization Strategy

Developing an effective optimization strategy requires understanding the interplay between optimizer choice, learning rate scheduling, and problem-specific considerations. The recommended approach begins with Adam for rapid initial progress, using learning rates in the range  $\{10^{-3}, 3 \cdot 10^{-4}, 10^{-4}\}$  to establish a strong foundation for training. This adaptive method provides robust performance across diverse architectures and datasets while requiring minimal hyperparameter tuning.

Learning rate scheduling plays a crucial role in optimization success, with different strategies suited to different model architectures. Cosine decay with warmup proves particularly effective for transformer-like models, providing smooth transitions from exploration to exploitation phases. For convolutional networks trained with SGD and momentum, step decay schedules often yield superior results by allowing the optimizer to make large initial progress before fine-tuning with smaller learning rates. When validation accuracy saturates, switching from Adam to SGD with Nesterov momentum can improve generalization by providing different optimization dynamics. This transition requires careful tuning of both learning rate and momentum parameters, but often yields better final performance on vision tasks. Gradient clipping becomes essential in recurrent models and unstable training scenarios, preventing parameter updates from becoming too large while preserving optimization direction. Continuous monitoring of training and validation metrics provides essential feedback for optimization strategy adjustment. Early stopping prevents overfitting by halting training when validation performance plateaus, while careful attention to learning rate schedules ensures optimal convergence behavior throughout the training process.

Applications and heuristics:

- Vision: SGD+momentum or Nesterov often yields state-of-the-art with careful schedules and augmentations [He+16].
- NLP/Transformers: Adam/AdamW with warmup+cosine is a strong default; clip global norm in seq2seq models.
- Reinforcement learning: Adam with small  $\alpha$  stabilizes non-stationary objectives.

Common failure modes:

- Divergence at start: reduce  $\alpha$ , add warmup, or increase  $\epsilon$  for Adam.
- Plateau: try larger batch with warmup, use cosine schedule, or add momentum.
- Overfitting: increase regularization (weight decay, dropout), add data augmentation.

## 8.6 Key Takeaways ◆

### Key Takeaways 8

- **Mini-batch SGD is the workhorse:** balances speed and stability; pair with momentum and schedules.
- **Momentum and Nesterov reduce oscillations:** accelerate along consistent directions and dampen noise.
- **Adaptive methods ease tuning:** Adam often works out-of-the-box; consider switching to SGD+momentum late for best generalisation.
- **Curvature matters:** second-order ideas inspire preconditioning; full Hessians are usually impractical in deep nets.
- **Schedules are critical:** cosine or step decay often yield large gains; warmup stabilises early training.
- **Mitigate pathologies:** use proper initialisation, normalisation, and gradient clipping to handle vanishing/exploding gradients and plateaus.

## Exercises

### Easy

*Exercise 8.1 (Batch Size Selection).* Explain the trade-offs between using a large batch size versus a small batch size for training. Consider computation time, memory usage, and convergence properties.

**Hint:**

*Think about GPU utilization, gradient noise, and generalization gap.*

*Exercise 8.2 (Momentum Intuition).* Describe how momentum helps accelerate optimization. Use the analogy of a ball rolling down a hill to explain the concept.

**Hint:**

*Consider how previous gradients influence the current update and help overcome small local variations.*

*Exercise 8.3 (Learning Rate Scheduling).* List three common learning rate scheduling strategies and explain when each is most appropriate.

**Hint:**

*Consider step decay, exponential decay, cosine annealing, and cyclical learning rates.*

*Exercise 8.4 (Adam Hyperparameters).* Adam optimizer has hyperparameters  $\beta_1$  (typically 0.9) and  $\beta_2$  (typically 0.999). Explain the role of each parameter.

**Hint:**

*$\beta_1$  controls momentum (first moment),  $\beta_2$  controls adaptive learning rates (second moment).*

## Medium

*Exercise 8.5 (Optimizer Comparison).* Compare SGD with momentum, RMSProp, and Adam on a simple optimization problem. Discuss their convergence behavior and when to prefer one over another.

**Hint:**

*Consider sparse gradients, non-stationary objectives, and computational cost.*

*Exercise 8.6 (Gradient Clipping).* Explain why gradient clipping is important for training recurrent neural networks. Derive the gradient clipping formula and discuss the choice of threshold.

**Hint:**

*Consider exploding gradients and the norm  $\|\nabla L\|$ . Clip by value or by norm.*

## Hard

*Exercise 8.7 (Natural Gradient Descent).* Derive the natural gradient update rule and explain why it is invariant to reparameterisations. Discuss the computational challenges of using natural gradient in deep learning.

**Hint:**

*Consider the Fisher information matrix  $\mathbf{F}$  and the update  $\theta \leftarrow \theta - \eta \mathbf{F}^{-1} \nabla L$ .*

*Exercise 8.8 (Learning Rate Warmup).* Analyse why learning rate warmup is beneficial when training with large batch sizes. Provide theoretical justification based on the optimization landscape.

**Hint:**

*Consider the stability of gradients in early training and the sharpness of the loss landscape.*

*Exercise 8.9 (Optimizer Comparison).* Compare the convergence properties of SGD, Adam, and RMSprop on a non-convex optimization problem.

**Hint:**

*Consider the adaptive learning rates and momentum effects of each optimizer.*

*Exercise 8.10 (Gradient Clipping).* Explain when and why gradient clipping is necessary. How does it affect the optimization dynamics?

**Hint:**

*Consider the exploding gradient problem and the relationship between gradient norms and learning stability.*

*Exercise 8.11 (Learning Rate Scheduling).* Design a learning rate schedule for training a deep network. Compare step decay, exponential decay, and cosine annealing.

**Hint:**

*Consider the trade-off between exploration and exploitation in different phases of training.*

*Exercise 8.12 (Second-Order Methods).* Explain why second-order optimization methods are rarely used in deep learning despite their theoretical advantages.

**Hint:**

*Consider computational complexity, memory requirements, and the stochastic nature of deep learning.*

*Exercise 8.13 (Optimization Landscape).* Analyse the relationship between the optimization landscape and the choice of optimizer for deep networks.

**Hint:**

*Consider saddle points, local minima, and the role of noise in optimization.*

*Exercise 8.14 (Batch Size Effects).* Investigate how batch size affects optimization dynamics and generalization in deep learning.

**Hint:**

*Consider the relationship between batch size, gradient noise, and the implicit regularization effect.*

# Chapter 9

## Convolutional Networks

This chapter introduces convolutional neural networks (CNNs), which are particularly effective for processing grid-structured data like images. We build intuition first, then progressively add mathematical detail and modern algorithms, with historical context and key takeaways throughout [GBC16a; Pri23].

### Learning Objectives

1. Convolution, pooling, and receptive fields for translation equivariance and parameter sharing
2. Output shapes and parameter counts for CNN layer configurations
3. Core CNN algorithms: convolution, pooling, and residual connections
4. Classic CNN architectures and their design trade-offs
5. CNN applications in vision: classification, detection, and segmentation
6. Historical milestones that motivated CNN advances

### 9.1 The Convolution Operation

#### Intuition

Convolution extracts local patterns by sliding small filters across the input, producing feature maps that respond strongly where patterns occur. Parameter sharing means the same filter detects the same pattern anywhere, yielding translation equivariance and dramatic parameter efficiency [GBC16a; Pri23].

#### Example 9.1: The Detective's Magnifying Glass

Imagine a detective examining a crime scene photograph with a magnifying glass, systematically scanning every area to look for specific patterns like fingerprints or footprints. The magnifying glass (convolution kernel) is the same tool used everywhere, but it can detect the same type of evidence whether it appears in the top-left corner or bottom-right corner of the photo. This is exactly how convolution works in neural networks—the same learned pattern detector scans across the entire image, finding similar features wherever they appear.

**Example 9.2: Edge Detection**

Consider a simple edge detection filter that looks for vertical edges in an image. This filter might have values like  $[-1, 0, 1]$  for a horizontal line of pixels. When this filter slides across an image, it produces high responses where there are strong vertical edges (like the side of a building) and low responses in flat areas (like the sky). The same filter works identically whether the edge is in the center of the image or near the edges, demonstrating the translation equivariance property that makes CNNs so powerful for image processing.

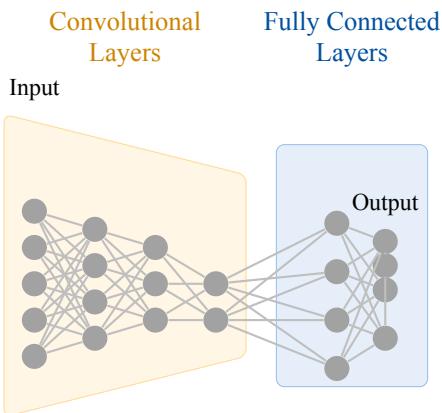


Figure 9.1: Deep CNN: conv layers (orange) maintain spatial structure, FC layers (blue) process global features.

### 9.1.1 Definition

The **convolution** operation applies a filter (kernel) across an input:

**Definition 9.1: Discrete 2D Convolution**

For discrete 2D convolution:

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(i - m, j - n)K(m, n) \quad (9.1)$$

where  $I$  is the input and  $K$  is the kernel.

where  $I$  is the input and  $K$  is the kernel. In deep learning libraries, the implemented operation is often cross-correlation (no kernel flip).

**Definition 9.2: Cross-Correlation**

In practice, we often use cross-correlation:

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(i + m, j + n)K(m, n) \quad (9.2)$$

where the kernel is not flipped, making it the standard operation in deep learning libraries.

### 9.1.2 Properties

The convolution operation possesses several fundamental properties that make it particularly well-suited for processing grid-structured data like images. Parameter sharing represents one of the most important characteristics, where the same kernel is applied across all spatial locations, dramatically reducing the number of parameters compared to fully connected layers acting

on flattened images. This parameter sharing yields translation equivariance, meaning that if the input shifts spatially, the feature map shifts in exactly the same way. Formally, letting  $T_\delta$  denote a spatial shift, we have  $(T_\delta I) * K = T_\delta(I * K)$  under appropriate boundary conditions, making CNNs robust to object translations in images.

Local connectivity ensures that each output depends only on a local input region, known as the receptive field, which exploits the spatial locality inherent in natural images where nearby pixels are statistically dependent. This local processing enables compositionality, where stacking layers progressively grows the effective receptive field, allowing the network to detect increasingly complex patterns that build from simple edges to corners to object parts. The linear nature of convolution operations, when combined with pointwise nonlinearities like ReLU, creates the foundation for modeling complex functions while maintaining computational efficiency.

Padding and boundary effects play crucial roles in preserving spatial information and mitigating edge shrinkage that occurs when kernels extend beyond image boundaries. Without proper padding, repeated convolutions rapidly reduce feature map size and bias learned features toward the center of the image. Stride and downsampling provide efficient ways to reduce spatial resolution, though aggressive striding early in the network can remove important fine details, leading many modern designs to delay downsampling to later stages. Dilation, or atrous convolutions, offers an alternative approach by inserting holes between kernel elements to increase the receptive field without increasing parameters, proving particularly useful for dense prediction tasks like semantic segmentation.

The combination of convolution with pooling or global aggregation introduces partial translation invariance in the representation, while multi-channel mixing through kernels of shape  $k \times k \times C_{\text{in}}$  enables learning both spatial and cross-channel interactions. The strategic use of  $1 \times 1$  convolutions allows mixing channels without spatial coupling, serving as bottlenecks and dimension reduction techniques in modern architectures like Inception and ResNet.[[GBC16a](#); [Pri23](#); [KSH12](#); [He+16](#)]

### 9.1.3 Multi-Channel Convolution

For input with  $C_{\text{in}}$  channels and  $C_{\text{out}}$  output channels:

$$S_{c_{\text{out}}}(i, j) = \sum_{c_{\text{in}}=1}^{C_{\text{in}}} (I_{c_{\text{in}}} * K_{c_{\text{out}}, c_{\text{in}}})(i, j) + b_{c_{\text{out}}} \quad (9.3)$$

Multi-channel convolution is the fundamental operation that enables CNNs to process complex inputs like RGB images (3 channels) and produce rich feature representations. The notation  $S_{c_{\text{out}}}(i, j)$  represents the output feature map value at spatial position  $(i, j)$  for output channel  $c_{\text{out}}$ . This operation allows the network to learn both spatial patterns (through kernel sliding) and cross-channel interactions (by mixing information from different input channels). Each output channel  $S_{c_{\text{out}}}(i, j)$  captures how strongly a particular learned pattern is detected at location  $(i, j)$ , where higher values indicate stronger pattern matches. This enables the network to detect complex features that span across multiple input channels while maintaining spatial locality and translation equivariance.

### 9.1.4 Hyperparameters

The design of convolutional layers involves several critical hyperparameters that significantly impact both computational efficiency and learning performance. Kernel size represents one of the most fundamental choices, with  $3 \times 3$  and  $5 \times 5$  kernels being the most common selections in practice. Modern architectures often prefer stacked  $3 \times 3$  convolutions over single  $5 \times 5$  kernels, as demonstrated in VGG networks, because they provide the same receptive field with fewer parameters and better gradient flow.[[GBC16a](#)]

Stride controls the step size for sliding the kernel across the input, directly affecting the spatial resolution of the output feature maps. The relationship between input size, kernel size, and stride determines the output dimensions through the formula Output size =  $\lfloor \frac{n-k}{s} \rfloor + 1$ , where  $n$  is the input size,  $k$  is the kernel size, and  $s$  is the stride. Padding strategies provide different approaches to handling boundary effects, with "valid" padding using no padding, "same" padding preserving spatial size, and "full" padding providing maximum coverage. For "same" padding with stride 1, the required padding is calculated as  $p = \lfloor \frac{k-1}{2} \rfloor$ , ensuring that the output maintains the same spatial dimensions as the input.

## 9.2 Pooling

**Pooling** reduces spatial dimensions and provides translation invariance.

### Intuition

Pooling summarizes nearby activations so that small translations of the input do not significantly change the summary. Max pooling keeps the strongest response, while average pooling smooths responses. It provides a degree of translation *invariance* complementary to convolution's translation *equivariance*. Modern designs sometimes prefer strided convolutions to make downsampling learnable [GBC16a; He+16].

#### 9.2.1 Max Pooling

Takes maximum value in each pooling region:

$$\text{MaxPool}(i, j) = \max_{m, n \in \mathcal{R}_{ij}} I(m, n) \quad (9.4)$$

Common:  $2 \times 2$  max pooling with stride 2 (halves spatial dimensions).

Max pooling takes the maximum value within each pooling region, preserving only the strongest activation while discarding weaker responses. This operation provides translation invariance by keeping the strongest response in each region, making the output robust to small spatial shifts in the input. Max pooling is particularly effective for detecting the presence of features (like edges or textures) rather than their exact location, making it useful for building hierarchical representations in CNNs. The operation reduces spatial dimensions while maintaining the most important information, helping the network focus on the most salient features.

#### Example 9.3

Max Pooling Example For input of size  $H \times W = 32 \times 32$  and a  $2 \times 2$  window with stride 2, the output is  $16 \times 16$ . Channels are pooled independently.

#### 9.2.2 Average Pooling

Computes average:

$$\text{AvgPool}(i, j) = \frac{1}{|\mathcal{R}_{ij}|} \sum_{m, n \in \mathcal{R}_{ij}} I(m, n) \quad (9.5)$$

Average pooling computes the mean value across each pooling region, providing a smooth summary of local activations rather than preserving only the strongest response like max pooling. This operation reduces spatial dimensions while providing translation invariance by averaging nearby activations, making the output less sensitive to small spatial shifts in the input. Unlike max pooling which preserves the strongest features, average pooling creates a smoother representation that can help reduce noise and provide more stable feature maps. It's particularly useful when you want to preserve information about the overall activation level in a region rather than just the peak response.

#### 9.2.3 Global Pooling

Global pooling operations extend the concept of local pooling to cover entire spatial dimensions, providing powerful mechanisms for reducing feature map complexity while preserving essential information. Global Average Pooling (GAP) computes the average value across all spatial locations for each channel, while Global Max Pooling identifies the maximum value across all spatial locations. These operations prove particularly useful for reducing the parameter count before fully connected layers and for connecting convolutional backbones to classification heads, such as using GAP before softmax layers. The strategic use of global average pooling can replace large fully connected layers by averaging each feature map to a single scalar value, dramatically reducing overfitting and parameter count while maintaining classification performance. This approach

has become particularly popular in modern architectures where the final feature maps are globally pooled before classification, eliminating the need for expensive fully connected layers and providing better generalization properties.[[GBC16a](#)]

### 9.2.4 Alternative: Strided Convolutions

Instead of relying on non-learned pooling operators, strided convolutions with stride  $s > 1$  perform learned downsampling that can adapt to the specific requirements of the task. For kernel size  $k$ , stride  $s$ , and padding  $p$ , the output spatial dimensions are calculated as  $H' = \left\lfloor \frac{H-k+2p}{s} \right\rfloor + 1$  and  $W' = \left\lfloor \frac{W-k+2p}{s} \right\rfloor + 1$ , providing precise control over the downsampling process.

The primary advantages of strided convolutions include their learnable nature and ability to combine feature extraction with downsampling in a single operation, making them particularly effective for stage transitions in modern architectures like ResNet. However, these operations may introduce aliasing artifacts if high-frequency content is not properly low-pass filtered before sub-sampling, leading to the development of anti-aliasing variants that apply blurring before strided operations to preserve signal quality.[[He+16](#)]

**Example.** A  $3 \times 3$  convolution with stride 2 and padding 1 keeps spatial size roughly halved (e.g.,  $32 \rightarrow 16$ ) while learning filters.

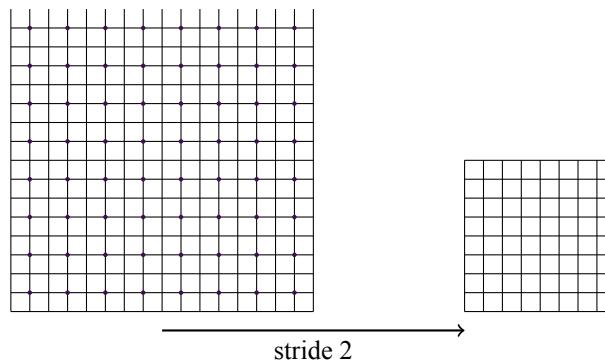


Figure 9.2: Downsampling via stride 2: fewer spatial samples after a strided convolution compared to pooling.

## 9.3 CNN Architectures ◆

### Historical Context

CNNs evolved from early biologically inspired work to practical systems. **LeNet-5** established the template for digit recognition [[LeC+89](#)]. **AlexNet** showed large-scale training with ReLU, dropout, and data augmentation could dominate ImageNet [[KSH12](#)]. **VGG** emphasized simplicity via small filters, while **Inception** exploited multi-scale processing with  $1 \times 1$  dimension reduction. **ResNet** enabled very deep networks via residual connections [[He+16](#)]. Efficiency-driven families like **MobileNet** and **EfficientNet** target edge devices and compound scaling.

### 9.3.1 LeNet-5 (1998)

LeNet-5 demonstrated the viability of CNNs for handwritten digit recognition on the MNIST dataset, establishing the foundational architecture that would influence all subsequent convolutional networks. The network combined convolution, subsampling through pooling, and small fully connected layers in a carefully designed topology:

$\text{Conv}(6@5 \times 5) \rightarrow \text{Pool}(2 \times 2) \rightarrow \text{Conv}(16@5 \times 5) \rightarrow \text{Pool}(2 \times 2) \rightarrow \text{FC}(120) \rightarrow \text{FC}(84) \rightarrow \text{Softmax}(10)$ .

The architecture utilized sigmoid and tanh activation functions, though later pedagogical treatments often retrofit ReLU activations to demonstrate modern practices. LeNet-5's key properties included local receptive fields that captured spatial patterns, parameter sharing that dramatically reduced the number of learnable parameters, and early evidence of translation invariance through pooling operations. The network's impact extended far beyond digit recognition, establishing the core

conv-pool pattern and demonstrating end-to-end learning for computer vision tasks, fundamentally changing how researchers approached image classification problems.[[LeC+89](#); [GBC16a](#)]

#### Remark 9.1: LeNet-5's Historical Importance

LeNet-5 was the first successful deep convolutional network that demonstrated the practical viability of CNNs for real-world applications, establishing the fundamental conv-pool-FC architecture pattern that became the foundation for all subsequent CNN designs. Its success on MNIST digit recognition proved that deep learning could solve complex pattern recognition tasks, directly inspiring the development of modern CNN architectures and marking the beginning of the deep learning revolution in computer vision.

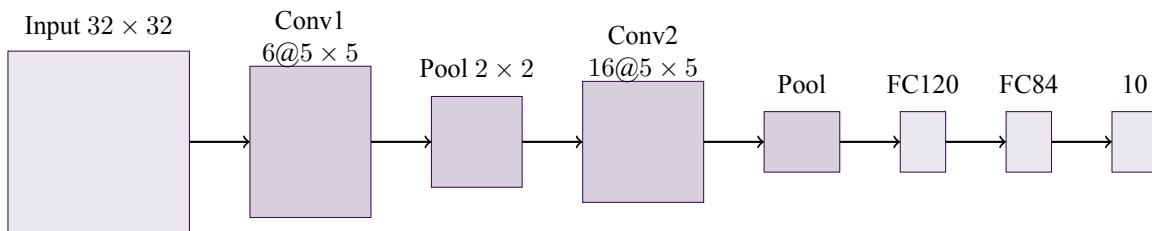


Figure 9.3: LeNet-5 architecture: alternating conv and pooling, followed by small fully connected layers.

### 9.3.2 AlexNet (2012)

AlexNet achieved a decisive victory at ILSVRC 2012, winning by a large margin and catalyzing the deep learning revolution in computer vision. The network's design featured 5 convolutional layers followed by 3 fully connected layers, incorporating local response normalization (LRN) and overlapping pooling to improve feature extraction. The optimization strategy proved equally important, with ReLU activations enabling significantly faster training compared to traditional sigmoid and tanh functions, while heavy data augmentation and dropout in the fully connected layers effectively reduced overfitting.

The systems engineering aspects of AlexNet were equally groundbreaking, as the network was trained on 2 GPUs using model parallelism to handle the computational demands. The architecture used large kernels in early layers combined with stride for rapid downsampling, creating an efficient pipeline for processing high-resolution images. AlexNet's impact extended far beyond its immediate performance gains, establishing large-scale supervised pretraining on ImageNet as the standard approach for computer vision research and demonstrating the practical viability of deep learning for real-world applications.[[KSH12](#)]

### 9.3.3 VGG Networks (2014)

VGG networks emphasized architectural depth through a simple yet effective recipe: stacks of  $3 \times 3$  convolutions with stride 1 combined with  $2 \times 2$  max pooling for downsampling. The uniform block design represented a key innovation, where replacing large kernels with multiple  $3 \times 3$  layers increased nonlinearity and receptive field while using fewer parameters than equivalent single large kernels. This approach demonstrated that depth could be achieved through careful architectural choices rather than simply adding more layers.

The VGG-16 and VGG-19 models achieved strong accuracy on ImageNet, though they came with significant computational costs due to very large parameter counts in the dense layers. The trade-offs inherent in VGG networks included strong classification accuracy balanced against heavy memory and computation requirements, making them particularly suitable as feature extractors for transfer learning applications. Despite their computational demands, VGG networks established the importance of systematic architectural design and demonstrated that simple, uniform building blocks could achieve competitive performance when properly scaled.[[GBC16a](#)]

### Remark 9.2: VGG Networks' Impact on Deep Learning

VGG networks revolutionized CNN design by proving that depth could be achieved through systematic use of small  $3 \times 3$  filters, establishing the principle that multiple small convolutions are more effective than single large filters. Their uniform architecture design and strong performance made VGG networks the standard feature extractor for transfer learning, directly influencing the development of modern CNN architectures and demonstrating that systematic architectural choices could achieve state-of-the-art performance.

### 9.3.4 ResNet (2015)

ResNet introduced identity skip connections to learn residual functions, fundamentally changing how deep networks could be trained and optimized. The core equation  $y = \mathcal{F}(x, \{\mathbf{W}_i\}) + x$  represents a simple yet powerful idea: instead of learning the direct mapping from input to output, the network learns the residual function  $\mathcal{F}$  that represents the difference between input and desired output, where  $\mathcal{F}$  is typically a small stack of convolutions with normalization and activation.

This architectural innovation enabled unprecedented depth, allowing stable training of 50, 101, and 152-layer models that would have been impossible to train with traditional architectures. The gradient flow benefits were equally important, as the Jacobian includes an identity term that mitigates vanishing gradients by providing direct paths for gradient propagation. ResNet blocks come in two main variants: basic blocks with two  $3 \times 3$  convolutions and bottleneck blocks with  $1 \times 1$ - $3 \times 3$ - $1 \times 1$  convolutions, both incorporating projection shortcuts to handle dimension changes between layers.[He+16]

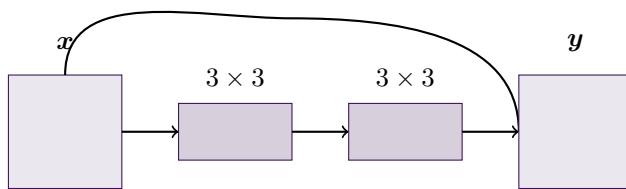


Figure 9.4: ResNet basic residual block with identity skip connection.

### 9.3.5 Inception/GoogLeNet (2014)

GoogLeNet popularized Inception modules that revolutionized CNN design through parallel multi-scale processing and strategic use of  $1 \times 1$  bottlenecks for computational efficiency. The Inception architecture featured parallel branches with  $1 \times 1$ ,  $3 \times 3$ , and  $5 \times 5$  convolutions alongside a pooled branch, all followed by channel-wise concatenation to combine multi-scale features. This design philosophy recognized that different scales of features are important for visual recognition and that processing them in parallel could capture more diverse patterns than sequential processing.

The efficiency gains came primarily from  $1 \times 1$  convolutions that reduced channel dimensions before applying larger kernels, dramatically cutting floating-point operations while preserving network capacity. This approach proved particularly effective for reducing computational costs in deeper layers where channel counts are high. GoogLeNet's impact extended beyond its immediate performance, achieving competitive accuracy with significantly fewer parameters than VGG networks while influencing the design of later hybrid networks that incorporated multi-scale processing and efficient channel mixing techniques.

### 9.3.6 MobileNet and EfficientNet

**MobileNet.** Prioritizes efficiency for edge devices using depthwise separable convolutions (depthwise  $k \times k$  followed by pointwise  $1 \times 1$ ), drastically reducing FLOPs and parameters while maintaining accuracy.

**EfficientNet.** Introduces compound scaling to jointly scale depth, width, and resolution with a principled coefficient, yielding strong accuracy/efficiency trade-offs. Variants (B0–B7) demonstrate near-optimal Pareto fronts.

For introductory treatment of these families, see *D2L (modern CNNs)* and *Deep Learning (convolutional networks)*.

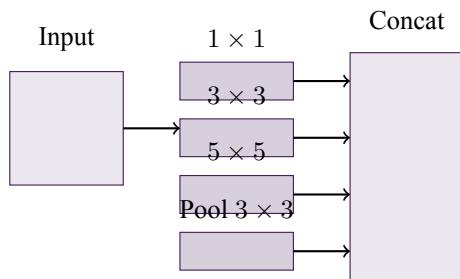


Figure 9.5: Inception module: parallel multi-scale branches with  $1 \times 1$  bottlenecks.

## 9.4 Applications of CNNs ◆

### Intuition

Backbones of stacked convolutions extract spatially local features that become increasingly abstract with depth. Task-specific heads (classification, detection, segmentation) transform backbone features into outputs appropriate to the problem [GBC16a; Pri23].

#### 9.4.1 Image Classification

Image classification represents the foundational task of assigning a single label to an entire image, serving as the basis for most computer vision applications. The standard approach combines a convolutional backbone that extracts hierarchical features with a task-specific head that produces the final classification. The backbone processes the input through multiple convolutional layers with downsampling via pooling or strided convolutions, progressively building more abstract and semantically meaningful representations. The head typically uses global average pooling followed by a small fully connected layer or  $1 \times 1$  convolution with softmax activation to produce class probabilities.

This architecture has proven remarkably effective across diverse datasets including CIFAR-10/100 and ImageNet (ILSVRC), establishing the foundation for transfer learning where ImageNet pretraining commonly improves performance on downstream tasks. The hierarchical feature extraction enables the network to learn from low-level edges and textures to high-level object parts and complete objects, making it particularly suitable for natural image classification where spatial structure and local patterns provide strong discriminative signals.

#### 9.4.2 Object Detection

Object detection extends image classification to simultaneously localize and classify multiple objects within a single image, requiring the network to predict bounding boxes and class labels for each detected object. This task presents unique challenges as it combines spatial localization with classification, demanding architectures that can handle variable numbers of objects at different scales and positions.

Region-based approaches, also known as two-stage methods, first generate region proposals and then classify each region. R-CNN pioneered this approach by applying CNN features to region proposals, though it was computationally expensive. Fast R-CNN and Faster R-CNN improved efficiency by integrating feature extraction, with Faster R-CNN learning region proposals through a Region Proposal Network (RPN). Mask R-CNN extended this framework with an additional instance segmentation branch for pixel-level object boundaries.

Single-shot approaches like YOLO and SSD perform detection in a single pass, making dense predictions at multiple scales for real-time performance. YOLO processes the entire image at once, predicting bounding boxes and class probabilities directly, while SSD uses default boxes across different feature map scales for efficient multi-scale detection. These methods employ specialized heads and losses including classification losses (cross-entropy or focal loss), box regression losses (smooth- $\ell_1$  or IoU losses), and non-maximum suppression (NMS) at inference to eliminate duplicate detections.

### 9.4.3 Semantic Segmentation

Semantic segmentation represents the most fine-grained computer vision task, requiring the assignment of a class label to each individual pixel in the image. This pixel-level classification demands architectures that can maintain spatial resolution while providing rich semantic understanding, making it particularly challenging compared to classification or detection tasks.

Fully Convolutional Networks (FCNs) revolutionized semantic segmentation by replacing dense layers with  $1 \times 1$  convolutions and upsampling through deconvolution to restore input resolution. U-Net introduced the encoder-decoder architecture with skip connections that preserve fine-grained spatial details for precise localization, becoming particularly popular in medical imaging applications where pixel-level accuracy is critical. Atrous or dilated convolutions provide an alternative approach by enlarging the receptive field without losing spatial resolution, enabling the network to capture both local details and global context simultaneously.

The training and evaluation of semantic segmentation models requires specialized losses and metrics that account for the pixel-level nature of the task. Pixel-wise cross-entropy loss provides the foundation for training, while Dice and IoU losses offer better handling of class imbalance. Mean Intersection over Union (mIoU) serves as the standard evaluation metric, measuring the overlap between predicted and ground truth segmentations across all classes to provide a comprehensive assessment of segmentation quality.[RFB15]

## 9.5 Core CNN Algorithms ◆

We introduce algorithms progressively, starting from basic cross-correlation to residual learning.

### 9.5.1 Cross-Correlation and Convolution

Given input  $I \in \mathbb{R}^{H \times W \times C_{\text{in}}}$  and kernel  $K \in \mathbb{R}^{k \times k \times C_{\text{in}} \times C_{\text{out}}}$ , the output feature map  $S \in \mathbb{R}^{H' \times W' \times C_{\text{out}}}$  under stride  $s$  and padding  $p$  is computed by cross-correlation:

$$S(i, j, c_{\text{out}}) = \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} \sum_{c_{\text{in}}=0}^{C_{\text{in}}-1} I(i+m, j+n, c_{\text{in}}) \cdot K(m, n, c_{\text{in}}, c_{\text{out}}) + b_{c_{\text{out}}} \quad (9.6)$$

where  $b_{c_{\text{out}}}$  is the bias term for output channel  $c_{\text{out}}$ . Libraries often refer to this as "convolution" [GBC16a].

### 9.5.2 Backpropagation Through Convolution

Backpropagation through convolution follows the chain rule, where gradients flow backward through the cross-correlation operation to update both input and kernel parameters. For loss  $\mathcal{L}$  and pre-activation output  $S = I * K$ , the gradients are computed as  $\frac{\partial \mathcal{L}}{\partial K} = I * \frac{\partial \mathcal{L}}{\partial S}$  and  $\frac{\partial \mathcal{L}}{\partial I} = \frac{\partial \mathcal{L}}{\partial S} * K^{\text{rot}}$ , where  $*$  denotes cross-correlation,  $*$  denotes convolution, and  $K^{\text{rot}}$  is the kernel rotated by  $180^\circ$ . The mathematical derivation shows that gradient computation mirrors the forward pass but with rotated kernels, enabling efficient parameter updates through cross-correlation operations.[GBC16a]

#### Example 9.4

Backpropagation Through Convolution - Shape-aware Example For  $I \in \mathbb{R}^{32 \times 32 \times 64}$  and  $K \in \mathbb{R}^{3 \times 3 \times 64 \times 128}$  with stride 1 and same padding,  $S \in \mathbb{R}^{32 \times 32 \times 128}$ . The gradient w.r.t.  $K$  accumulates over spatial locations and batch.

### 9.5.3 Pooling Backpropagation

Pooling is a downsampling operation that reduces spatial dimensions by aggregating local regions, typically using max or average operations to preserve important features while reducing computational complexity.

Pooling backpropagation requires careful handling of gradient routing to maintain the mathematical properties of the forward pass. For max pooling, the upstream gradient is routed exclusively to the maximal input in each pooling region, since only the maximum value contributed to the output. For average pooling, the gradient is evenly divided among all elements in each

region, reflecting the equal contribution of each input to the average. This gradient routing ensures that the backward pass accurately reflects the forward computation, enabling proper parameter updates during training.

#### 9.5.4 Residual Connections

A residual connection is a skip connection that adds the input directly to the output of a layer, allowing the network to learn residual functions rather than direct mappings, which helps with gradient flow and enables training of very deep networks. Residual connections introduce identity skip paths that fundamentally change gradient flow in deep networks. In a residual block with input  $\mathbf{x}$  and residual mapping  $\mathcal{F}$ , the output becomes  $\mathbf{y} = \mathcal{F}(\mathbf{x}) + \mathbf{x}$ , where the identity connection provides a direct path for information and gradients. The mathematical derivation of backpropagation yields  $\frac{\partial \mathcal{L}}{\partial \mathbf{x}} = \frac{\partial \mathcal{L}}{\partial \mathbf{y}} (\frac{\partial \mathcal{F}}{\partial \mathbf{x}} + \mathbf{I})$ , where the identity matrix  $\mathbf{I}$  ensures that gradients can flow directly through the skip connection, stabilizing training and enabling very deep networks.[He+16]

#### 9.5.5 Progressive Complexity: Depthwise Separable Convolutions

Depthwise separable convolution factors standard convolution into depthwise (per-channel) and pointwise ( $1 \times 1$ ) operations, reducing FLOPs and parameters (used by MobileNet). This keeps representational power while improving efficiency.

##### Remark 9.3: Parameter Comparison

For  $C_{\text{in}} = C_{\text{out}} = c$  and kernel  $k \times k$ :

$$\text{standard} = k^2 c^2, \quad (9.7)$$

$$\text{depthwise separable} = k^2 c + c^2, \quad \text{saving} \approx 1 - \frac{k^2 c + c^2}{k^2 c^2}. \quad (9.8)$$

This shows that depthwise separable convolutions achieve significant parameter reduction while maintaining representational power, making them particularly effective for mobile and edge computing applications.

#### 9.5.6 Normalization and Activation

Batch normalization and ReLU-family activations play crucial roles in modern CNN training by improving optimization dynamics and generalization performance. Batch normalization addresses covariate shift by normalizing activations across the batch dimension, smoothing the loss landscape and enabling faster convergence with higher learning rates. ReLU and its variants provide sparse activations that reduce computational complexity while maintaining gradient flow, with techniques like Leaky ReLU and ELU addressing the "dying ReLU" problem. These components work together to create more stable training dynamics, allowing networks to learn more effectively from complex visual data.[IS15]

#### 9.5.7 Other Useful Variants

Modern CNN architectures incorporate several additional techniques that address specific challenges in deep learning. Group and Layer Normalization provide alternatives to batch normalization when batch sizes are small or when batch statistics are unreliable, offering different approaches to activation normalization that can improve training stability. Dilated convolutions expand the receptive field without requiring pooling operations, making them particularly effective in segmentation tasks where spatial resolution must be preserved. Anti-aliased downsampling techniques, including blur pooling and low-pass filtering before strided operations, reduce aliasing artifacts in feature maps, leading to more robust representations that better preserve spatial information during downsampling operations.

## 9.6 Real World Applications ●

Convolutional neural networks have revolutionized how computers understand images and videos, transforming abstract computer vision research into practical tools that improve healthcare, safety, and daily convenience. Their applications touch nearly every aspect of modern visual technology, from medical diagnosis to autonomous vehicles and everyday smartphone features.

### 9.6.1 Medical Image Analysis

CNNs help doctors diagnose diseases more accurately and quickly, revolutionizing medical imaging across multiple specialties. Cancer detection in radiology represents one of the most impactful applications, where CNNs analyze X-rays, CT scans, and MRIs to detect tumors often invisible to the human eye. Mammography systems using CNNs can spot breast cancer earlier than traditional methods, potentially saving thousands of lives annually by learning to recognize subtle patterns that indicate malignancy.

Diabetic retinopathy screening demonstrates how CNNs can democratize healthcare access, examining photos of patients' eyes to detect diabetes-related damage before vision loss occurs. This allows automated screening in remote areas without specialist ophthalmologists, making eye care accessible to millions more people worldwide. Skin cancer classification through smartphone apps represents another accessible application, allowing people to photograph suspicious moles for instant preliminary assessment, encouraging early medical consultation when something looks concerning.

### 9.6.2 Autonomous Driving

Self-driving cars rely on CNNs to understand their surroundings, requiring real-time processing of complex visual information under varied conditions. Object detection and tracking systems process camera feeds to identify pedestrians, other vehicles, traffic signs, and lane markings in real-time, working perfectly under challenging conditions including rain, snow, nighttime, and construction zones where lives depend on accurate perception.

Depth estimation capabilities enable CNNs to analyze images and determine how far away objects are, helping vehicles make safe decisions about braking, turning, and merging. This works effectively even with regular cameras, though it's enhanced when combined with other sensors. Semantic segmentation provides pixel-level understanding by labeling every pixel in the camera view as road, sidewalk, vehicle, sky, etc., giving the vehicle complete environmental understanding that enables precise navigation.

## Key Takeaways

### Key Takeaways 9

- **Convolution and pooling** exploit spatial structure through parameter sharing and local receptive fields, achieving translation equivariance.
- **Classic architectures** progressively deepened networks (LeNet → AlexNet → VGG → ResNet) via innovations like batch normalisation and residual connections.
- **ResNets solve vanishing gradients** by adding skip connections, enabling training of very deep networks.
- **Modern CNNs balance efficiency and accuracy** through depthwise separable convolutions (MobileNet) and compound scaling (EfficientNet).
- **CNNs excel in vision tasks:** classification, object detection, and semantic segmentation, with task-specific heads and losses.

## Exercises

### Easy

*Exercise 9.1 (Receptive Field Calculation).* A CNN has two convolutional layers with  $3 \times 3$  kernels (no padding, stride 1). Calculate the receptive field of a neuron in the second layer.

**Hint:**

*Each layer expands the receptive field. For the second layer, consider how many input pixels affect it.*

*Exercise 9.2 (Parameter Counting).* Calculate the number of parameters in a convolutional layer with 64 input channels, 128 output channels, and  $3 \times 3$  kernels (including bias).

**Hint:**

*Each output channel has a  $3 \times 3$  kernel for each input channel, plus one bias term.*

*Exercise 9.3 (Pooling Operations).* Explain the difference between max pooling and average pooling. When would you prefer one over the other?

**Hint:**

*Consider feature prominence, spatial information retention, and gradient flow.*

*Exercise 9.4 (Translation Equivariance).* Explain what translation equivariance means in the context of CNNs and why it is a desirable property for image processing.

**Hint:**

*If the input is shifted, how does the output change? Consider the relationship  $f(T(x)) = T(f(x))$ .*

### Medium

*Exercise 9.5 (Output Shape Calculation).* Given an input image of size  $224 \times 224 \times 3$ , apply the following operations and calculate the output shape at each step:

1. Conv2D: 64 filters,  $7 \times 7$  kernel, stride 2, padding 3
2. MaxPool2D:  $3 \times 3$ , stride 2
3. Conv2D: 128 filters,  $3 \times 3$  kernel, stride 1, padding 1

**Hint:**

*Use the formula:  $\text{output\_size} = \lfloor \frac{\text{input\_size} + 2 \times \text{padding} - \text{kernel\_size}}{\text{stride}} \rfloor + 1$ .*

*Exercise 9.6 (ResNet Skip Connections).* Explain why residual connections (skip connections) help train very deep networks. Discuss the gradient flow through skip connections.

**Hint:**

*Consider the identity mapping  $y = x + F(x)$  and compute  $\frac{\partial y}{\partial x}$ .*

**Hard**

*Exercise 9.7 (Dilated Convolutions).* Derive the receptive field for a stack of dilated convolutions with dilation rates [1, 2, 4, 8]. Compare computational cost with standard convolutions achieving the same receptive field.

**Hint:**

*Dilated convolution with rate  $r$  introduces  $(r - 1)$  gaps between kernel elements. Track receptive field growth layer by layer.*

*Exercise 9.8 (Depthwise Separable Convolutions).* Analyse the computational savings of depthwise separable convolutions (as used in MobileNets) compared to standard convolutions. Derive the reduction factor for a layer with  $C_{in}$  input channels,  $C_{out}$  output channels, and  $K \times K$  kernel size.

**Hint:**

*Depthwise separable splits into depthwise ( $C_{in}$  groups) and pointwise ( $1 \times 1$ ) convolutions. Compare FLOPs.*

*Exercise 9.9 (Convolutional Layer Design).* Design a convolutional layer architecture for a specific computer vision task. Justify your choices of kernel size, stride, and padding.

**Hint:**

*Consider the trade-off between computational efficiency and feature extraction capability.*

*Exercise 9.10 (Pooling Operations).* Compare different pooling operations (max, average, L2) and their effects on feature maps.

**Hint:**

*Consider the impact on spatial information, gradient flow, and computational efficiency.*

*Exercise 9.11 (Feature Map Visualization).* Explain how to visualize and interpret feature maps in different layers of a CNN.

**Hint:**

*Consider activation maximization, gradient-based methods, and occlusion analysis.*

*Exercise 9.12 (CNN Architecture Search).* Design an automated method for finding optimal CNN architectures for a given dataset.

**Hint:**

*Consider neural architecture search (NAS), evolutionary algorithms, and reinforcement learning approaches.*

*Exercise 9.13 (Transfer Learning Strategies).* Compare different transfer learning approaches for CNNs and when to use each.

**Hint:**

*Consider feature extraction, fine-tuning, and progressive unfreezing strategies.*

*Exercise 9.14 (CNN Interpretability).* Explain methods for understanding and interpreting CNN decisions, including attention mechanisms.

**Hint:**

*Consider gradient-based attribution methods, attention maps, and adversarial analysis.*

# Chapter 10

# Sequence Modeling: Recurrent and Recursive Nets

This chapter covers architectures designed for sequential and temporal data, including recurrent neural networks (RNNs) and their variants.

## Learning Objectives

1. Why sequence models are needed and data modalities requiring temporal context
2. Comparison of vanilla RNNs, LSTMs, and GRUs with their gating mechanisms
3. Backpropagation through time (BPTT) and truncated BPTT with gradient clipping
4. Sequence-to-sequence models with attention and alignment mechanisms
5. Advanced decoding: bidirectional RNNs, teacher forcing, and beam search
6. Common failure modes and mitigation strategies

## 10.1 Recurrent Neural Networks ◆

### Intuition

An RNN carries a running summary of the past—like a notepad you update after reading each word. This hidden state lets the model use prior context to influence current predictions. However, keeping reliable notes over long spans is hard: small errors can compound, and gradients may shrink or grow too much [GBC16a].

#### 10.1.1 Motivation

Sequential data exhibits *temporal dependencies* and *order-sensitive* structure that cannot be modeled well by i.i.d. assumptions or fixed-size context windows alone [GBC16a; Pri23; Bis06]. Consider the diverse nature of sequential data: time series forecasting energy load, financial returns, or physiological signals like ECG readings; natural language where the meaning of a word depends entirely on its context and sentences must conform to syntax and discourse structure; speech and audio where phonemes combine to form words and coarticulation effects span multiple frames; video where actions unfold over time and temporal cues disambiguate similar frames; and control and reinforcement learning where actions influence future observations, requiring persistent memory.

### Example 10.1: Temporal Dependencies in Social Media

Consider a social media post sequence: "Just finished my final exam" → "Feeling relieved" → "Time to celebrate!" The meaning of "celebrate" depends entirely on the temporal context of the previous posts about completing an exam. A traditional model without temporal awareness might interpret "celebrate" in isolation, but an RNN can maintain the context that this celebration is specifically about academic achievement, demonstrating how temporal dependencies are crucial for understanding sequential social media content where each post builds upon previous context.

Classic feedforward networks assume fixed-size inputs and lack a persistent state, making them ill-suited for long-range dependencies. Recurrent architectures introduce a hidden state that acts as a compact, learned memory and enables conditioning on arbitrary-length histories. Historically, recurrent ideas trace back to early neural sequence models and dynamical systems; practical training matured with BPTT [RHW86] and later with gated units to mitigate vanishing/exploding gradients [HS97]. For further background see the RNN overview in [GBC16a] and educational treatments in [Zha+24c; Wik25b; GBC16c].

### 10.1.2 Why Sequences Matter

The unique value of sequence modeling lies in its ability to understand how earlier elements affect later ones through context awareness, enabling the model to capture the rich dependencies that exist in temporal data. Unlike traditional approaches that treat each input independently, sequence models can work with inputs of any length, automatically adapting to the complexity and duration of the input sequence. This temporal pattern recognition capability allows the model to capture how things change over time, whether it's the evolution of language meaning, the progression of musical notes, or the development of market trends. Most importantly, this temporal understanding enables natural interaction with machines, allowing human-like communication where the system can maintain context across extended conversations and respond appropriately to the full history of the interaction.

### 10.1.3 Basic RNN Architecture

An RNN maintains a hidden state  $\mathbf{h}_t$  that evolves over time:

$$\mathbf{h}_t = \sigma(\mathbf{W}_{hh}\mathbf{h}_{t-1} + \mathbf{W}_{xh}\mathbf{x}_t + \mathbf{b}_h) \quad (10.1)$$

$$\mathbf{y}_t = \mathbf{W}_{hy}\mathbf{h}_t + \mathbf{b}_y \quad (10.2)$$

where  $\mathbf{x}_t$  is input at time  $t$ , and  $\sigma$  is typically tanh.

**Visual aid.** The following unrolled diagram shows shared parameters across time:

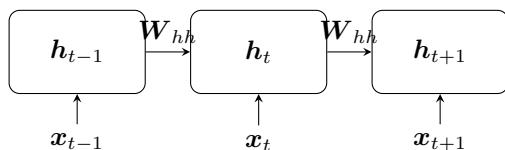


Figure 10.1: Unrolled RNN with shared parameters across time steps.

### 10.1.4 Unfolding in Time

RNNs can be "unrolled" into a deep feedforward computation graph over time with *shared parameters*. This perspective clarifies how gradients flow backward through temporal connections and why depth-in-time can cause vanishing/exploding gradients [GBC16a].

$$\mathbf{h}_t = f(\mathbf{h}_{t-1}, \mathbf{x}_t; \theta) \quad (10.3)$$

Unfolding reveals repeated applications of the same transition function across steps. The equation  $\mathbf{h}_t = f(\mathbf{h}_{t-1}, \mathbf{x}_t; \theta)$  represents the core RNN computation where the current hidden state  $\mathbf{h}_t$  depends on both the previous hidden state  $\mathbf{h}_{t-1}$  and the current input  $\mathbf{x}_t$ , with the same parameters  $\theta$  being reused at each time step. This recursive structure allows the network to maintain memory of past information while processing new inputs, enabling the modeling of temporal dependencies across arbitrary sequence lengths.

We annotate inputs, hidden states, and outputs to emphasize sharing and the temporal chain rule during BPTT.

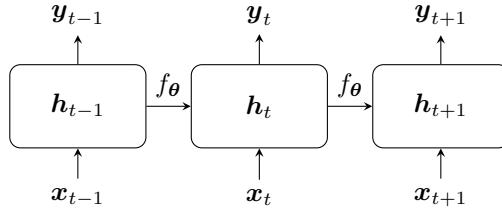


Figure 10.2: Unrolling an RNN across time: the same parameters  $\theta$  are reused at each step.

This view connects RNNs to dynamic Bayesian networks and emphasizes that training complexity scales with the unroll length. See [Wik25b; GBC16a; Zha+24c].

### 10.1.5 Types of Sequences

Modern sequence modeling encompasses various architectural patterns and training strategies that address different sequence-to-sequence mapping requirements. Teacher forcing is a training technique where the model receives the ground truth previous output as input during training, accelerating convergence but potentially causing exposure bias during inference. Bidirectionality allows models to process sequences in both forward and backward directions, capturing context from both past and future information simultaneously. Attention mechanisms enable models to focus on relevant parts of the input sequence dynamically, while beam search provides a decoding strategy that maintains multiple candidate sequences during generation to find more optimal outputs.

Type	Description	Examples
One-to-one	Fixed-size input to fixed-size output with temporal structure ignored or not present	Static image classification
One-to-many	Single input to sequence output	Image captioning
Many-to-one	Sequence input to single output	Sentiment classification; keyword spotting
Many-to-many (synchronous)	Sequence labeling with aligned input/output lengths	Part-of-speech tagging; frame labeling
Many-to-many (asynchronous)	Sequence transduction with potentially different lengths. Attention helps bridge length mismatch by learning soft alignments [BCB14]	Machine translation; speech recognition

Table 10.1: Types of sequence models and their characteristics.

Design choices (teacher forcing, bidirectionality, attention, beam search) depend on whether future context is available and whether output timing must be causal. See [Zha+24c; Wik25b] for further taxonomy.

See [GBC16a; Pri23; Bis06; Wik25b; GBC16c; Zha+24c] for introductions to sequence modeling and RNNs.

## 10.2 Backpropagation Through Time (BPTT) ♦

### Intuition

BPTT treats the unrolled RNN as a deep network across time and applies backpropagation through each time slice. Gradients flow backward along temporal edges, accumulating effects from future steps. Truncation limits how far signals propagate to balance cost and dependency length [GBC16a].

#### Remark 10.1: Historical Context of BPTT

The backpropagation algorithm [RHW86] enabled efficient training of deep networks; applying it to unrolled RNNs became known as BPTT. Awareness of vanishing/exploding gradients led to clipping and gated architectures [GBC16a; HS97].

### 10.2.1 BPTT Algorithm

Gradients are computed by unrolling the network and applying backpropagation through the temporal graph. Let  $L = \sum_{t=1}^T L_t$  and  $\mathbf{h}_t = f(\mathbf{h}_{t-1}, \mathbf{x}_t; \theta)$ ,  $\mathbf{y}_t = g(\mathbf{h}_t; \theta_y)$ . The total derivative w.r.t. hidden states satisfies the recurrence:

$$\frac{\partial L}{\partial \mathbf{h}_t} = \frac{\partial L}{\partial \mathbf{y}_t} \frac{\partial \mathbf{y}_t}{\partial \mathbf{h}_t} + \frac{\partial L}{\partial \mathbf{h}_{t+1}} \frac{\partial \mathbf{h}_{t+1}}{\partial \mathbf{h}_t} \quad (10.4)$$

For any parameter block  $\mathbf{W} \in \theta$  appearing at each time step:

$$\frac{\partial L}{\partial \mathbf{W}} = \sum_{t=1}^T \frac{\partial L_t}{\partial \mathbf{W}} \quad (10.5)$$

---

#### Algorithm 4 Backpropagation Through Time (BPTT)

```

1: Input: Sequence  $\{\mathbf{x}_1, \dots, \mathbf{x}_T\}$ , parameters  $\theta$ 
2: Output: Gradients  $\frac{\partial L}{\partial \theta}$ 
3:
4: ▷ Forward pass
5: for  $t = 1$  to  $T$  do
6:   Compute  $\mathbf{h}_t = f(\mathbf{h}_{t-1}, \mathbf{x}_t; \theta)$ 
7:   Compute  $\mathbf{y}_t = g(\mathbf{h}_t; \theta_y)$ 
8:   Compute  $L_t = \text{loss}(\mathbf{y}_t, \mathbf{y}_t^{\text{target}})$ 
9: end for
10:
11: ▷ Initialize temporal gradients
12:  $\frac{\partial L}{\partial \mathbf{h}_{T+1}} \leftarrow \mathbf{0}$ 
13:
14: ▷ Backward pass
15: for  $t = T$  downto 1 do
16:    $\delta_t \leftarrow \frac{\partial L}{\partial \mathbf{y}_t} \frac{\partial \mathbf{y}_t}{\partial \mathbf{h}_t} + \frac{\partial L}{\partial \mathbf{h}_{t+1}} \frac{\partial \mathbf{h}_{t+1}}{\partial \mathbf{h}_t}$ 
17:    $\frac{\partial L}{\partial \mathbf{W}} += \delta_t \frac{\partial \mathbf{h}_t}{\partial \mathbf{W}}$  ▷ Accumulate gradients for all parameters at time  $t$ 
18: end for
19:
20: ▷ Apply gradient clipping if needed and update parameters
21: Apply gradient clipping if  $\|\frac{\partial L}{\partial \theta}\| >$  threshold
22: Update parameters:  $\theta \leftarrow \theta - \alpha \frac{\partial L}{\partial \theta}$ 

```

---

This view aligns with the computational-graph treatment in [GBC16a] and standard expositions [GBC16c; Zha+24c].

### 10.2.2 Vanishing and Exploding Gradients

Gradients can vanish or explode exponentially due to the chain rule applied across time steps. The gradient of the loss with respect to an earlier hidden state  $\mathbf{h}_k$  involves a product of Jacobian matrices:

$$\frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_k} = \prod_{i=k+1}^t \frac{\partial \mathbf{h}_i}{\partial \mathbf{h}_{i-1}} = \prod_{i=k+1}^t \mathbf{W}^\top \text{diag}(\sigma'(\mathbf{z}_i)) \quad (10.6)$$

The behavior depends on the eigenvalues of the weight matrix  $\mathbf{W}$ :

- **Eigenvalues  $< 1$ :** Each multiplication by  $\mathbf{W}^\top$  shrinks the gradient magnitude, causing exponential decay as the product accumulates over time steps. This leads to vanishing gradients where early time steps receive negligible gradient signals.
- **Eigenvalues  $> 1$ :** Each multiplication amplifies the gradient, causing exponential growth that can lead to numerical instability and training divergence.

**Gradient clipping** prevents exploding gradients by scaling down gradients when their norm exceeds a threshold, maintaining training stability while preserving gradient direction. **Careful initialization** techniques like Xavier/He initialization set initial weights to have appropriate variance, reducing the likelihood of extreme eigenvalues that cause gradient problems. **ReLU activation** helps because its derivative is either 0 or 1, avoiding the multiplicative shrinking effect of sigmoid/tanh derivatives that compound vanishing gradients. **LSTM/GRU architectures** introduce gating mechanisms that create direct paths for gradient flow, bypassing the problematic multiplicative chains and enabling learning of long-range dependencies.

### 10.2.3 Truncated BPTT

For very long sequences, truncate gradient computation by limiting backpropagation to a sliding window of  $k$  steps [GBC16a]. This approach processes inputs in segments of length  $k$ , possibly overlapping with stride  $s$ , allowing the model to handle arbitrarily long sequences while maintaining computational efficiency. The key insight is that gradients are backpropagated only within each segment to reduce memory and computational time, making it feasible to train on very long sequences that would otherwise be computationally prohibitive. The hidden state is carried forward between segments but treated as a constant during the truncated backward step, creating a balance between maintaining temporal continuity and computational tractability.

**Trade-offs:** Lower memory and latency versus potentially missing very long-range dependencies. Increasing  $k$  improves dependency length coverage but increases cost. Hybrids with dilated RNNs or attention can mitigate the trade-off.

## 10.3 Long Short-Term Memory (LSTM) ◆

### Intuition

The LSTM adds a highway for information (the cell state) that can pass signals forward with minimal modification. Gates act like valves to forget unhelpful information, write new content, and reveal outputs, which preserves gradients over long spans [HS97; GBC16a].

LSTM (Long Short-Term Memory) is a type of recurrent neural network architecture designed to solve the vanishing gradient problem by using gating mechanisms to control information flow and maintain long-term memory through a dedicated cell state.

#### Remark 10.2: Historical Context of LSTM

Introduced in the 1990s to address vanishing gradients [HS97], LSTMs unlocked practical sequence learning across speech, language, and time-series tasks before attention-based Transformers became dominant [Vas+17].

**Algorithm 5** Truncated Backpropagation Through Time (Truncated BPTT)

---

```

1: Input: Sequence  $\{x_1, \dots, x_T\}$ , parameters  $\theta$ , chunk size  $k$ , stride  $s$ 
2: Output: Gradients  $\frac{\partial L}{\partial \theta}$ 
3:
4: Initialize  $\mathbf{h}_0$  ▷ Initial hidden state
5:
6: for  $t = 1$  to  $T$  step  $s$  do
7:   ▷ Take chunk  $[t, t + k - 1]$ 
8:      $t_{\text{end}} \leftarrow \min(t + k - 1, T)$ 
9:
10:    ▷ Forward pass over the chunk
11:    for  $i = t$  to  $t_{\text{end}}$  do
12:      Compute  $\mathbf{h}_i = f(\mathbf{h}_{i-1}, \mathbf{x}_i; \theta)$ 
13:      Compute  $\mathbf{y}_i = g(\mathbf{h}_i; \theta_y)$ 
14:      Compute  $L_i = \text{loss}(\mathbf{y}_i, \mathbf{y}_i^{\text{target}})$ 
15:    end for
16:
17:    ▷ Backpropagate losses only within the chunk
18:    Initialize  $\frac{\partial L}{\partial \mathbf{h}_{t_{\text{end}}+1}} \leftarrow \mathbf{0}$ 
19:    for  $i = t_{\text{end}}$  downto  $t$  do
20:       $\delta_i \leftarrow \frac{\partial L}{\partial \mathbf{y}_i} \frac{\partial \mathbf{y}_i}{\partial \mathbf{h}_i} + \frac{\partial L}{\partial \mathbf{h}_{i+1}} \frac{\partial \mathbf{h}_{i+1}}{\partial \mathbf{h}_i}$ 
21:       $\frac{\partial L}{\partial \mathbf{W}} += \delta_i \frac{\partial \mathbf{h}_i}{\partial \mathbf{W}}$  ▷ Accumulate gradients
22:    end for
23:
24:    ▷ Optionally detach hidden state to bound gradient length
25:     $\mathbf{h}_t \leftarrow \text{detach}(\mathbf{h}_{t_{\text{end}}})$  ▷ Prevent gradients from flowing beyond chunk
26: end for

```

---

### 10.3.1 Architecture

**Remark 10.3: Gates in LSTM**

Gates in LSTM are learnable mechanisms that control information flow by deciding what information to forget, what new information to store, and what to output, enabling the network to selectively maintain or discard information over long time spans.

**Remark 10.4: Recurrence and Hidden State - The Memory**

The hidden state in LSTM serves as the network's working memory, carrying forward processed information from previous time steps while the cell state acts as long-term memory that can store information across extended sequences, with gates controlling how information flows between these memory components.

LSTM uses **gating mechanisms** to control information flow and maintain a persistent cell state that supports long-range credit assignment [HS97; GBC16a]. Gating mechanisms are essential because they allow the network to learn when to remember, forget, or output information, solving the fundamental problem of how to maintain useful information over long sequences while discarding irrelevant details.

$$\mathbf{f}_t = \sigma(\mathbf{W}_f[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f) \quad (\text{forget gate}) \quad (10.7)$$

$$\mathbf{i}_t = \sigma(\mathbf{W}_i[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i) \quad (\text{input gate}) \quad (10.8)$$

$$\tilde{\mathbf{c}}_t = \tanh(\mathbf{W}_c[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_c) \quad (\text{candidate}) \quad (10.9)$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t \quad (\text{cell state}) \quad (10.10)$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_o[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o) \quad (\text{output gate}) \quad (10.11)$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \quad (\text{hidden state}) \quad (10.12)$$

### 10.3.2 Key Ideas

The LSTM's key innovations address fundamental limitations of traditional RNNs and feedforward networks. Unlike standard RNNs where information must pass through repeated nonlinear transformations that cause gradient decay, the LSTM introduces a dedicated **cell state**  $\mathbf{c}_t$  that acts as a highway for information flow with minimal transformation, enabling gradients to flow directly across long time spans. The **gating mechanisms** (forget, input, and output gates) provide selective control over information flow, allowing the network to learn when to remember, forget, and output information—a capability that was impossible with fixed-weight feedforward networks or vanilla RNNs. This selective memory management solves the vanishing gradient problem by creating direct paths for gradient flow while maintaining the ability to learn complex temporal dependencies that exceed the capacity of traditional sequence models.

**Cell state**  $\mathbf{c}_t$ : Long-term memory

- Information flows with minimal transformation
- Gates control what to remember/forget

**Forget gate**  $\mathbf{f}_t$ : Decides what to discard from cell state

**Input gate**  $\mathbf{i}_t$ : Decides what new information to store

**Output gate**  $\mathbf{o}_t$ : Decides what to output

### 10.3.3 Advantages

LSTMs address the vanishing gradient problem through their innovative cell state and gating mechanisms, where the cell state acts as a memory highway allowing gradients to flow relatively unimpeded across many time steps, preventing them from shrinking to zero. Unlike vanilla RNNs where gradients must pass through repeated nonlinear transformations, LSTMs provide direct paths for gradient flow, creating a more stable training environment. The architecture's ability to learn long-term dependencies stems from the forget and input gates that explicitly control what information is retained or discarded from the cell state, allowing LSTMs to store relevant information for extended periods and access it when needed, effectively capturing dependencies across hundreds of time steps without degradation. This selective memory mechanism, combined with the linear-like flow through the cell state path involving simple additions and multiplications by gate activations, has made LSTMs a de facto standard for various sequential data processing tasks, with robust performance across diverse applications from natural language processing to speech recognition.

## 10.4 Gated Recurrent Units (GRU) ◆

### Intuition

Gated Recurrent Unit (GRU): A simplification of the LSTM that uses only two gates (Reset and Update) and merges the hidden state and cell state, making it faster to train with fewer parameters while achieving comparable performance in many tasks. GRU simplifies LSTM by merging cell and hidden state and combining gates, often matching performance with fewer parameters—useful when data or compute is limited [Cho+14; GBC16a].

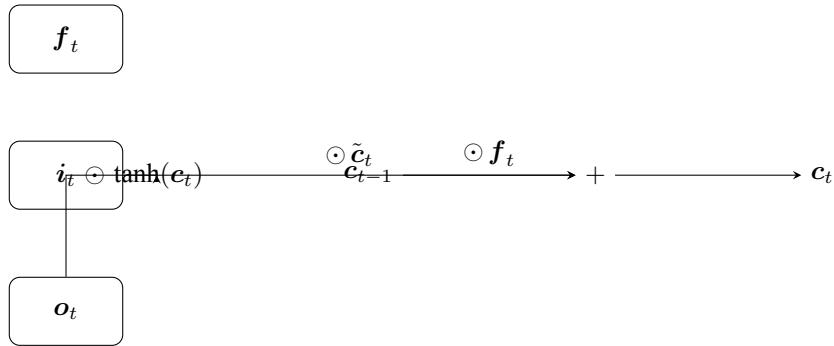


Figure 10.3: Visual aid: A compact LSTM cell diagram.

**Remark 10.5: Historical Context of GRU**

Proposed in the mid-2010s, GRU offered a practical alternative to LSTM with competitive empirical results and simpler implementation [Cho+14]. GRU became important in modern deep learning as it provided a more efficient alternative to LSTM for many sequence modeling tasks, particularly valuable for resource-constrained applications and when computational efficiency is crucial.

### 10.4.1 Architecture

GRU simplifies LSTM with fewer parameters and merges the cell and hidden state into a single vector, often yielding comparable performance with less computation [Cho+14; GBC16a]. The mathematical formulation involves two key gates: the update gate  $z_t$  controls how much of the previous hidden state to retain, while the reset gate  $r_t$  determines how much of the previous hidden state to forget when computing the candidate hidden state. The candidate hidden state  $\tilde{h}_t$  is computed using the reset gate to selectively incorporate information from the previous state, and the final hidden state  $h_t$  is a weighted combination of the previous state and the candidate, where the weights are determined by the update gate.

**Algorithm 6** GRU Forward Pass

```

1: Input:  $x_t$  (input at time  $t$ ),  $h_{t-1}$  (previous hidden state)
2: Output:  $h_t$  (current hidden state)
3:
4:
5:  $z_t \leftarrow \sigma(\mathbf{W}_z[h_{t-1}, x_t])$                                 ▷ Compute gates
6:  $r_t \leftarrow \sigma(\mathbf{W}_r[h_{t-1}, x_t])$                                 ▷ Update gate
7:
8:  $\tilde{h}_t \leftarrow \tanh(\mathbf{W}[r_t \odot h_{t-1}, x_t])$                       ▷ Reset gate
9:
10:
11:  $h_t \leftarrow (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$                   ▷ Compute candidate hidden state
12:
13:
14: return  $h_t$                                                        ▷ Compute final hidden state

```

$$z_t = \sigma(\mathbf{W}_z[h_{t-1}, x_t]) \quad (\text{update gate}) \quad (10.13)$$

$$r_t = \sigma(\mathbf{W}_r[h_{t-1}, x_t]) \quad (\text{reset gate}) \quad (10.14)$$

$$\tilde{h}_t = \tanh(\mathbf{W}[r_t \odot h_{t-1}, x_t]) \quad (\text{candidate}) \quad (10.15)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (10.16)$$

### 10.4.2 Architecture (visual)

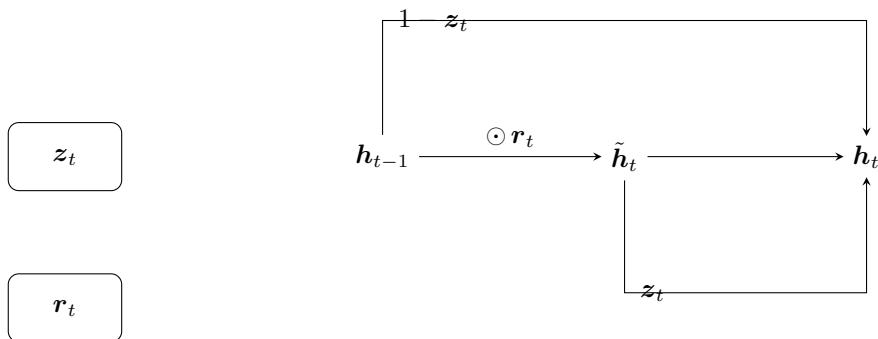


Figure 10.4: Illustrative GRU flow with update and reset gates.

### 10.4.3 Comparison with LSTM

GRU and LSTM represent two different approaches to solving the vanishing gradient problem in recurrent networks, each with distinct advantages depending on the application requirements. GRU uses a single hidden state  $h_t$  with two gates (update and reset), making it computationally more efficient and easier to implement, while LSTM maintains separate hidden  $h_t$  and cell  $c_t$  states with three gates (input, forget, and output), providing more expressive power for complex sequence modeling tasks. The choice between them often depends on the specific requirements: GRU is typically preferred for smaller datasets or when computational efficiency is crucial, while LSTM often performs better on very long sequences or when the additional model capacity is beneficial for capturing complex temporal dependencies.

## 10.5 Sequence-to-Sequence Models ◆

### Intuition

Encoder – decoder models compress a source sequence into a representation and then generate a target sequence step-by-step. Attention augments this by letting the decoder look back at encoder states as needed, creating a dynamic context per output token [Cho+14; BCB14].

#### Remark 10.6: Sequence-to-Sequence Models

A Seq2Seq (Sequence-to-Sequence) model is a Encoder-Decoder model, a specific type of neural network architecture designed to transform one sequence of data (the input) into another sequence of data (the output). It is the backbone of many key applications in Natural Language Processing (NLP).

### 10.5.1 Encoder-Decoder Architecture

The encoder-decoder architecture revolutionized sequence-to-sequence learning by solving a fundamental challenge: how to transform variable-length input sequences into variable-length output sequences of potentially different lengths. Traditional approaches struggled with this asymmetry, as they required fixed input-output dimensions or relied on hand-crafted features that couldn't capture the complex relationships between source and target sequences. The key insight behind this architecture lies in its elegant separation of concerns: the encoder compresses the entire input sequence into a rich, fixed-size representation that captures all essential information, while the decoder uses this representation to generate the output sequence step-by-step, maintaining the temporal dependencies crucial for coherent generation. This design was revolutionary compared to previous architectures because it eliminated the need for explicit alignment between input and output positions, allowing the model to

learn implicit correspondences through end-to-end training. Unlike rule-based systems or traditional statistical methods that required extensive linguistic knowledge, the encoder-decoder framework could automatically discover complex mappings between any two sequence domains, making it the foundation for modern neural machine translation and countless other sequence transduction tasks.

For sequence transduction tasks like machine translation [Cho+14; BCB14]:

**Encoder:** Processes input sequence into representation (fixed or per-step states)

$$\mathbf{c} = f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T) \quad (10.17)$$

**Decoder:** Generates output sequence from representation

A minimal encoder – decoder with context vector.

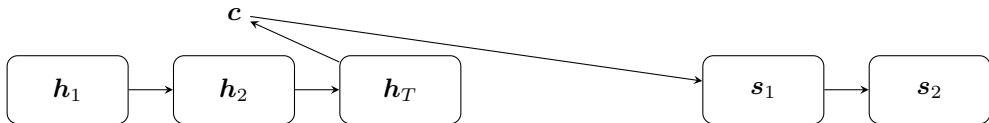


Figure 10.5: Encoder – decoder with fixed context vector  $\mathbf{c}$ . Attention uses step-dependent  $c_t$ .

$$\mathbf{y}_t = g(\mathbf{y}_{t-1}, \mathbf{c}, \mathbf{s}_{t-1}) \quad (10.18)$$

### 10.5.2 Attention Mechanism: Attention Is All You Need

Attention represents one of the most crucial breakthroughs in deep learning, so fundamental that the seminal paper "Attention Is All You Need" [Vas+17] demonstrated that attention mechanisms alone could replace entire architectural components like recurrent layers. This paradigm shift occurred because attention solves the fundamental information bottleneck problem in sequence-to-sequence models, where standard approaches compress entire input sequences into a single fixed vector  $\mathbf{c}$ , inevitably losing critical information and context. The attention mechanism elegantly addresses this limitation by allowing the decoder to dynamically focus on different parts of the input sequence at each generation step, creating a flexible and context-aware representation that adapts to the specific requirements of each output token.

**Attention** allows the decoder to focus on relevant input parts by computing a content-based weighted average of encoder states [BCB14]. At each decoding step  $t$ :

$$e_{ti} = a(\mathbf{s}_{t-1}, \mathbf{h}_i) \quad (\text{alignment scores}) \quad (10.19)$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_j \exp(e_{tj})} \quad (\text{attention weights}) \quad (10.20)$$

$$\mathbf{c}_t = \sum_i \alpha_{ti} \mathbf{h}_i \quad (\text{context vector}) \quad (10.21)$$

Common scoring functions  $a(\cdot)$  include additive (Bahdanau) attention using a small MLP, and multiplicative/dot-product attention which is parameter-efficient and forms the basis for scaled dot-product attention in Transformers [Vas+17]. Attention weights  $\alpha_{ti}$  are interpretable as soft alignments between target position  $t$  and source position  $i$  (see [Wik25a]). For a broader overview, consult standard references [GBC16c; Zha+24a].

**Training and inference.** Attention is trained end-to-end with the seq2seq objective. During inference, attention enables the model to retrieve the most relevant encoder features for each generated token, improving long-input performance and handling reordering. Variants include multi-head attention, local/monotonic attention for streaming, and coverage terms to reduce repetition.

**Dynamic context for each output:** Unlike traditional approaches that use a single, static context vector for all output positions, attention creates a unique, dynamically computed context vector for each decoding step. This means that when generating the

word "cat" in a translation, the model can focus on the relevant source words like "gato" or "chat," while when generating "dog," it shifts its attention to "perro" or "chien." This adaptive context selection allows the model to maintain fine-grained relationships between source and target elements, dramatically improving translation quality and coherence across different linguistic structures.

**Better for long sequences:** Traditional encoder-decoder models suffer from severe performance degradation on long sequences because the fixed-size context vector becomes an information bottleneck, unable to preserve all the nuanced details from lengthy inputs. Attention mechanisms solve this by providing direct access to all encoder states, allowing the model to selectively retrieve relevant information regardless of sequence length. This capability is particularly crucial for tasks like document summarization or translating lengthy articles, where the model must maintain awareness of information from the beginning of the document even when generating the final sentences.

**Interpretable (visualize attention weights):** One of the most remarkable aspects of attention mechanisms is their inherent interpretability, as the attention weights  $\alpha_{ti}$  provide a clear, visual representation of which input positions the model considers most relevant for each output token. Researchers can visualize these attention patterns as heatmaps, revealing fascinating insights about how the model learns linguistic alignments, syntactic structures, and semantic relationships. This interpretability has proven invaluable for debugging model behavior, understanding translation errors, and even discovering novel linguistic patterns that the model has learned autonomously, making attention not just a powerful computational tool but also a window into the model's decision-making process.

### 10.5.3 Applications

**Machine translation (NMT):** Modern neural machine translation systems power real-time communication across language barriers, enabling instant translation of web pages, documents, and conversations in applications like Google Translate and Microsoft Translator. These systems handle complex linguistic phenomena such as idiomatic expressions, cultural references, and context-dependent meanings that traditional rule-based systems struggled with, making global communication seamless for billions of users worldwide.

**Text summarization (extractive and abstractive):** News organizations and content platforms rely on sequence-to-sequence models to automatically generate concise summaries of lengthy articles, research papers, and legal documents, helping readers quickly grasp key information without reading entire texts. Financial institutions use these systems to summarize market reports and regulatory documents, while healthcare organizations employ them to distill complex medical literature into actionable insights for practitioners.

**Question answering and dialogue systems:** Virtual assistants like Siri, Alexa, and Google Assistant leverage sequence-to-sequence architectures to understand user queries and generate natural, contextually appropriate responses across diverse topics and conversation styles. Customer service chatbots powered by these models can handle complex inquiries, maintain conversation context across multiple turns, and provide personalized assistance while reducing human workload and improving response times.

**Image captioning (CNN/ViT encoder, RNN/Transformer decoder):** Social media platforms and accessibility tools use image captioning to automatically generate descriptive text for photos, helping visually impaired users understand visual content and improving content discoverability through search. Medical imaging systems employ these models to generate detailed reports from X-rays and MRI scans, while autonomous vehicles use them to describe road conditions and potential hazards for safety systems.

**Speech recognition and speech translation:** Real-time meeting transcription services like Otter.ai and Zoom's live transcription feature use sequence-to-sequence models to convert spoken language into accurate text, enabling accessibility and note-taking for millions of users. Simultaneous interpretation systems at international conferences and diplomatic meetings leverage these technologies to provide real-time translation between speakers of different languages, breaking down communication barriers in global settings.

**OCR and handwriting recognition:** Banking and financial institutions use OCR systems to automatically process handwritten checks, forms, and documents, dramatically reducing manual data entry and processing time while minimizing human errors. Educational platforms employ handwriting recognition to digitize student notes and assignments, enabling digital archiving, search, and analysis of handwritten content across academic institutions.

**Code generation and program repair:** Software development platforms like GitHub Copilot and Tabnine use sequence-to-sequence models to suggest code completions, generate functions from natural language descriptions, and automatically fix bugs in existing codebases. These systems help developers write code faster, catch potential errors early, and learn new programming patterns, while automated program repair tools can identify and fix security vulnerabilities and performance issues in large-scale software projects.

## 10.6 Advanced Topics ◆

### Intuition

Variants extend context (bidirectional), depth (stacked layers), supervision signals (teacher forcing), and search quality (beam search). Beam search maintains multiple candidate sequences during generation, exploring promising paths rather than committing to a single greedy choice, which often leads to higher-quality outputs. These architectural and algorithmic choices create fundamental trade-offs that practitioners must navigate carefully. Training stability often comes at the cost of inference complexity, while improved context modeling may increase computational requirements. The optimal configuration depends heavily on the specific task requirements, available computational resources, and latency constraints. For example, real-time applications may prioritize speed over quality, while offline processing can afford more sophisticated search strategies. Understanding these trade-offs is crucial for designing effective RNN-based systems that meet both performance and practical deployment requirements.

### Historical Context

Bidirectional RNNs (BiRNN) emerged post-2000 as a key innovation to improve context use for labeling tasks like Part-of-Speech tagging. Instead of processing sequences unidirectionally, BiRNNs use two independent RNNs (often LSTMs): one running forward and one backward, with their outputs concatenated. This structure allows any element  $x_t$  to benefit from both past and future context, resulting in much richer local context for classification. Teacher forcing stabilized decoder training but highlighted exposure bias; beam search became standard for autoregressive decoding in translation and speech.

#### 10.6.1 Bidirectional RNNs

The term "bidirectional" refers to processing the input sequence in both temporal directions—forward and backward—unlike standard RNNs that only process left-to-right. Mathematically, this means maintaining two separate hidden state sequences:

$$\vec{h}_t = f(\mathbf{x}_t, \vec{h}_{t-1}) \quad (10.22)$$

$$\overleftarrow{h}_t = f(\mathbf{x}_t, \overleftarrow{h}_{t+1}) \quad (10.23)$$

$$\mathbf{h}_t = [\vec{h}_t; \overleftarrow{h}_t] \quad (10.24)$$

The forward arrow  $\vec{h}_t$  processes the sequence from left to right (time  $t - 1$  to  $t$ ), while the backward arrow  $\overleftarrow{h}_t$  processes from right to left (time  $t + 1$  to  $t$ ). The final representation  $\mathbf{h}_t$  concatenates both directions, giving each position access to both past and future context.

#### Example 10.2: Bidirectional Reading Metaphor

Imagine reading a sentence twice: first normally from left to right to understand the flow, then reading it backward from right to left to catch any nuances you might have missed. A bidirectional RNN does exactly this—it "reads" the sequence in both directions simultaneously, allowing each word to be understood in the full context of what comes before and after it.

Useful when future context is available.

#### Example 10.3: Use Cases and Caveats

Effective for tagging, chunking, and ASR with full utterances, but not suitable for strictly causal, low-latency streaming since backward states require future tokens. Alternatives include limited-lookahead or online approximations.

### 10.6.2 Deep RNNs

Stack multiple RNN layers:

$$\mathbf{h}_t^{(l)} = f(\mathbf{h}_t^{(l-1)}, \mathbf{h}_{t-1}^{(l)}) \quad (10.25)$$

The term "deep" refers to the vertical stacking of multiple recurrent layers, where each layer  $l$  processes the hidden states from the previous layer  $l - 1$  at each time step. This creates a hierarchical representation where lower layers capture local patterns and dependencies, while higher layers learn more complex, long-range temporal relationships. Unlike feedforward networks where depth refers to the number of layers, in deep RNNs, depth combines both the number of layers and the temporal dimension, creating a two-dimensional computational graph. Each layer can specialize in different aspects of the sequence modeling task, with early layers often focusing on local features and later layers integrating information across longer time horizons.

**Visual aid.** Stacked recurrent layers over time.

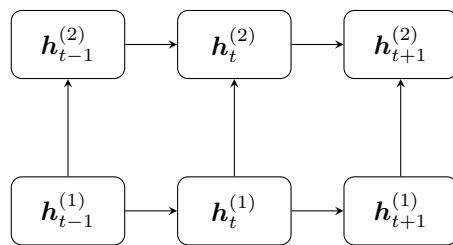


Figure 10.6: Deep RNN: multiple recurrent layers stacked over time.

Practical tips: residual connections, layer normalization, and dropout between layers help optimization and generalization.

### 10.6.3 Teacher Forcing

Teacher forcing is a training strategy where the model receives the ground truth previous output as input during training instead of using its own predictions, which accelerates convergence but can cause exposure bias during inference.

During training, teacher forcing uses ground truth as decoder input instead of the model's own predictions, providing several key advantages for sequence learning. By providing the correct previous token during training, the model receives consistent, high-quality input signals that guide it toward the target distribution more directly, eliminating the compounding effect of early prediction errors that would otherwise propagate through the sequence. This approach allows the model to focus on learning the mapping from context to next token rather than recovering from its own mistakes, creating a more predictable gradient landscape where the model can learn the underlying sequence patterns without being derailed by cascading errors. However, teacher forcing may cause exposure bias at test time—a train – test mismatch where the model never learns to recover from its own errors, since during training it only sees ground truth inputs but at test time must generate sequences using its own predictions, which may contain errors that compound over time. Mitigations include scheduled sampling, where the model gradually transitions from using ground truth to its own predictions during training, and sequence-level training that optimizes end-to-end sequence quality rather than individual token predictions.

### 10.6.4 Beam Search

Beam search is a decoding algorithm that maintains multiple candidate sequences during generation, exploring the most promising paths rather than committing to a single greedy choice, which often leads to higher-quality outputs.

For inference, beam search maintains top- $k$  hypotheses during sequence generation, providing significant advantages over greedy decoding approaches. Unlike greedy decoding that always selects the single most probable token at each step, which can lead to suboptimal global sequences since locally optimal choices don't guarantee globally optimal solutions, beam search explores multiple promising paths simultaneously, allowing it to recover from early mistakes and find sequences with higher overall probability. This is particularly important in sequence generation tasks where the best next word might not be the most probable one when considering the full sequence context. The approach creates a fundamental trade-off between quality and speed: larger beam sizes explore more hypotheses and generally produce higher-quality outputs, but require exponentially more computation as the beam width increases, with computational cost growing as  $O(k \times V)$  where  $k$  is the beam size and  $V$  is the vocabulary size. Common beam sizes of 5 – 10 provide a good compromise for most applications, while length normalization and coverage penalties are often used in neural machine translation to prevent bias toward shorter sequences and encourage the model to attend to all parts of the input sequence.

## 10.7 Real World Applications

Recurrent and recursive networks excel at understanding sequences—whether words in sentences, notes in music, or events over time. These capabilities enable numerous practical applications.

### 10.7.1 Machine Translation

Sequence models have revolutionized global communication by breaking down language barriers through sophisticated translation services that understand context and nuance. Google Translate and similar services now translate between over 100 languages, helping billions of people access information and communicate across language barriers, with the models understanding context to translate "bank" correctly as either a financial institution or riverbank depending on the surrounding words. Real-time conversation translation apps enable tourists to have natural conversations with locals, facilitate business meetings across languages, and support international collaboration by processing speech patterns and converting them to another language while preserving meaning and tone. Document translation systems allow businesses to automatically translate contracts, user manuals, and websites, making multilingual business operations feasible and affordable while maintaining the quality that requires human review for critical applications.

### 10.7.2 Voice Assistants and Speech Recognition

Sequence models have transformed human-computer interaction by making voice-based communication natural and intuitive across diverse applications. Smartphone assistants like Siri, Google Assistant, and Alexa use sequence models to understand voice commands despite accents, background noise, and casual phrasing, processing sound waves sequentially to recognize words even when spoken quickly or unclearly. Automated transcription services now transcribe meetings, podcasts, and lectures automatically, making content searchable and accessible while handling multiple speakers, technical terminology, and varying audio quality—tasks that once required hours of human effort. Voice-to-text applications provide crucial accessibility tools for people with mobility or vision impairments, enabling them to interact with devices, write documents, and access information independently, with these tools becoming more accurate and responsive through advances in sequence modeling.

### 10.7.3 Predictive Text and Content Generation

Sequence models have revolutionized writing and communication by providing intelligent assistance that understands context and user patterns. Email clients now predict what users will type next through smart compose features, suggesting complete sentences based on writing patterns and message context, saving time and reducing typing effort especially on mobile devices where typing is slower. Development tools like GitHub Copilot leverage sequence models trained on billions of lines of code to suggest code as developers type, understanding programming context and patterns to help write software faster with fewer bugs. Social media platforms employ sequence models for content moderation, detecting toxic comments, spam, and harmful content by understanding context, slang, and subtle linguistic patterns that indicate problematic content, enabling automated content filtering at scale.

### 10.7.4 Financial Forecasting and Analysis

Sequence models have transformed financial analysis by understanding temporal patterns in markets and financial behavior, enabling more sophisticated risk assessment and decision-making. While markets remain notoriously difficult to predict, sequence models analyze historical price patterns, trading volumes, and news sentiment to identify trends and inform trading decisions, providing quantitative traders with powerful tools for market analysis. Banks employ sequence models for fraud detection by analyzing transaction sequences to identify unusual patterns that might indicate stolen cards or fraudulent activity, where the temporal aspect is crucial since legitimate behavior follows certain patterns over time. Credit risk assessment has been revolutionized as lenders analyze sequences of financial behaviors including payment histories, spending patterns, and income changes to assess creditworthiness more accurately than traditional snapshot-based approaches, enabling more precise lending decisions and risk management.

These applications demonstrate how sequence models transform our ability to process language, speech, and time-series data at scale.

## Key Takeaways

### Key Takeaways 10

- **RNNs process sequential data** by maintaining hidden states that capture temporal dependencies.
- **LSTMs and GRUs** mitigate vanishing gradients via gating mechanisms that control information flow.
- **Backpropagation through time (BPTT)** computes gradients by unrolling the recurrent computation graph.
- **Attention mechanisms** allow models to focus on relevant parts of the input sequence, improving alignment in seq2seq tasks.
- **Practical challenges** include gradient clipping, teacher forcing, and exposure bias in autoregressive generation.

## Exercises

### Easy

*Exercise 10.1 (RNN vs Feedforward).* Explain why standard feedforward networks are not suitable for sequence modeling tasks. What key capability do RNNs provide?

**Hint:**

*Consider variable-length inputs and the need to maintain temporal context.*

*Exercise 10.2 (LSTM Gates).* Name the three gates in an LSTM cell and briefly describe the role of each.

**Hint:**

*Think about what information needs to be forgotten, what new information to store, and what to output.*

*Exercise 10.3 (Vanishing Gradients).* Explain why vanilla RNNs suffer from the vanishing gradient problem when processing long sequences.

**Hint:**

*Consider repeated matrix multiplication during backpropagation through time.*

*Exercise 10.4 (Sequence-to-Sequence Tasks).* Give three examples of sequence-to-sequence tasks and explain what makes them challenging.

**Hint:**

*Consider machine translation, speech recognition, and video captioning.*

## Medium

*Exercise 10.5 (BPTT Implementation).* Describe how truncated backpropagation through time (BPTT) works. What are the trade-offs compared to full BPTT?

**Hint:**

*Consider memory requirements, gradient approximation quality, and the effective temporal window.*

*Exercise 10.6 (Attention Mechanism).* Explain the intuition behind attention mechanisms in sequence-to-sequence models. How does attention address the bottleneck of fixed-size context vectors?

**Hint:**

*Consider how different parts of the input sequence should influence different parts of the output.*

## Hard

*Exercise 10.7 (GRU vs LSTM).* Compare GRU (Gated Recurrent Unit) and LSTM architectures mathematically. Derive their update equations and analyse computational complexity.

**Hint:**

*Count the number of parameters and operations per cell. GRU has fewer gates.*

*Exercise 10.8 (Bidirectional RNN Gradient).* Derive the gradient flow in a bidirectional RNN. Explain why bidirectional RNNs cannot be used for online prediction tasks.

**Hint:**

*Consider that backward pass requires seeing the entire future sequence.*

*Exercise 10.9 (Advanced Topic 1).* Explain a key concept from this chapter and its practical applications.

**Hint:**

*Consider the theoretical foundations and real-world implications.*

*Exercise 10.10 (Advanced Topic 2).* Analyse the relationship between different techniques covered in this chapter.

**Hint:**

*Look for connections and trade-offs between methods.*

*Exercise 10.11 (Advanced Topic 3).* Design an experiment to test a hypothesis related to this chapter's content.

**Hint:**

*Consider experimental design, metrics, and potential confounding factors.*

*Exercise 10.12 (Advanced Topic 4).* Compare different approaches to solving a problem from this chapter.

**Hint:**

*Consider computational complexity, accuracy, and practical considerations.*

*Exercise 10.13 (Advanced Topic 5).* Derive a mathematical relationship or prove a theorem from this chapter.

**Hint:**

*Start with the definitions and work through the logical steps.*

*Exercise 10.14 (Advanced Topic 6).* Implement a practical solution to a problem discussed in this chapter.

**Hint:**

*Consider the implementation details and potential challenges.*

*Exercise 10.15 (Advanced Topic 7).* Evaluate the limitations and potential improvements of techniques from this chapter.

**Hint:**

*Consider both theoretical limitations and practical constraints.*



# Chapter 11

## Practical Methodology

This chapter provides practical guidelines for successfully applying deep learning to real-world problems.

### ⌚ Learning Objectives

1. Deep learning project scoping: objectives, constraints, and success metrics
2. Data pipelines: dataset splitting, leakage management, and reproducibility
3. Architecture selection and baseline comparison strategies
4. Hyperparameter tuning and learning rate scheduling
5. Debugging techniques using loss curves, ablations, and sanity checks
6. Model deployment: monitoring drift, distribution shift, and documentation

## Intuition

Practical deep learning succeeds when we reduce uncertainty early and iterate quickly. Start with *simple, auditable baselines* to validate data and objectives, then progressively add complexity only when it measurably helps. Prefer experiments that answer the biggest unknowns first (e.g., data quality vs. model capacity). Treat metrics, validation splits, and ablations as your instrumentation layer; they convert intuition into evidence. See also Goodfellow, Bengio, and Courville [GBC16a] for methodology patterns.

### 11.1 Performance Metrics ◆

Performance metrics provide quantitative measures to evaluate model effectiveness, guide model selection, and ensure models meet business requirements across different machine learning tasks.

#### 11.1.1 Classification Metrics

Robust model evaluation depends on selecting metrics aligned with task requirements and operational costs . Accuracy alone can be misleading under class imbalance ; prefer precision/recall, AUC, PR-AUC, calibration, and cost-sensitive metrics when appropriate Goodfellow, Bengio, and Courville [GBC16a] and Prince [Pri23].

**Definition 11.1: Confusion Matrix**

For binary classification with positive/negative classes, define true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). The confusion matrix summarizes counts:

	Predicted +	Predicted -
Actual +	TP	FN
Actual -	FP	TN

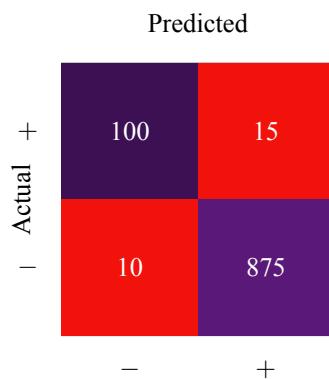


Figure 11.1: Confusion matrix heatmap (example counts). High diagonal values indicate good performance.

**Definition 11.2: Accuracy**

accuracy measures overall correctness but can obscure minority-class performance:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \quad (11.1)$$

**Definition 11.3: Precision and Recall**

precision and recall quantify quality on the positive class:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (11.2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (11.3)$$

**Definition 11.4: F1 Score**

The harmonic mean balances precision and recall:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (11.4)$$

**Definition 11.5: ROC and AUC**

The ROC curve plots TPR vs. FPR as the decision threshold varies; AUC summarizes ranking quality and is threshold-independent.

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}. \quad (11.5)$$

**Definition 11.6: Precision-Recall (PR) Curve**

Under heavy class imbalance, the PR curve and average precision (AP) are often more informative than ROC Prince [Pri23].

**Definition 11.7: Calibration**

A calibrated classifier's predicted probabilities match observed frequencies. Use reliability diagrams and expected calibration error (ECE). Calibration matters in risk-sensitive applications Goodfellow, Bengio, and Courville [GBC16a].

**ROC and PR curves.** The following figures illustrate key classification metrics:

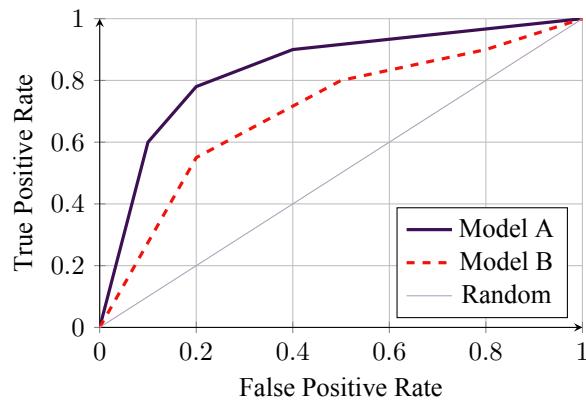


Figure 11.2: ROC curves for two models. Higher AUC indicates better ranking quality.

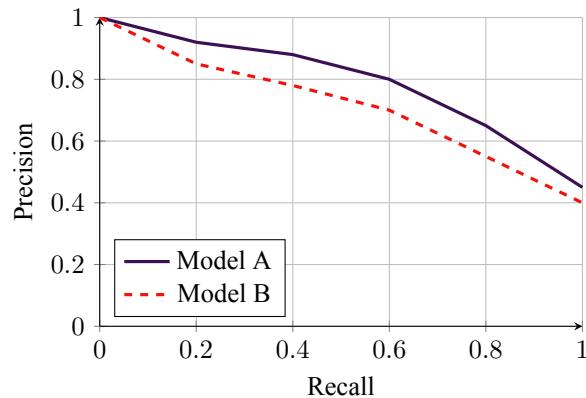


Figure 11.3: Precision–recall curves emphasize performance on the positive class under imbalance.

**Calibration diagram.** Shows how well predicted probabilities match observed frequencies:

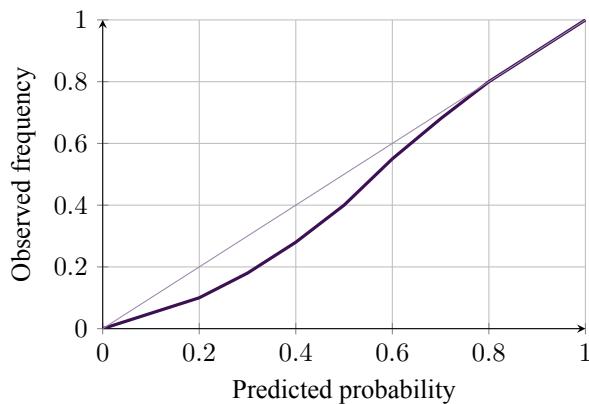


Figure 11.4: Reliability diagram illustrating calibration. The diagonal is perfect calibration.

### 11.1.2 Regression Metrics

Choose metrics that reflect business loss and robustness to outliers . Mean squared error (MSE) penalizes large errors more heavily than mean absolute error (MAE). Root mean squared error (RMSE) is in the original units. Coefficient of determination  $R^2$  measures variance explained.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad \text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad \text{RMSE} = \sqrt{\text{MSE}}. \quad (11.6)$$

#### Remark 11.1: Regression Metrics for Robustness

For heavy-tailed noise, consider Huber loss and quantile losses for pinball objectives. Huber loss is a robust loss function that combines the benefits of squared loss (for small errors) and absolute loss (for large errors), making it less sensitive to outliers than mean squared error while still being differentiable everywhere.

#### Definition 11.8: Huber Loss

Huber loss is defined as:

$$L_\delta(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2 & \text{if } |y - \hat{y}| \leq \delta \\ \delta|y - \hat{y}| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases} \quad (11.7)$$

where  $\delta$  is a threshold parameter that determines the transition point between squared and absolute loss.

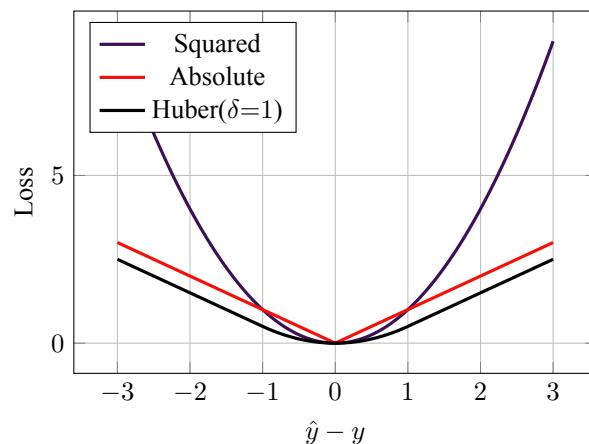


Figure 11.5: Comparison of squared, absolute, and Huber losses.

### 11.1.3 NLP and Sequence Metrics

Sequence generation quality is commonly measured by **BLEU** and **ROUGE** (n-gram overlap), while language models use *perplexity* (negative log-likelihood in exponential form) Goodfellow, Bengio, and Courville [GBC16a] and Zhang et al. [Zha+24c]:

$$\text{PPL} = \exp\left(-\frac{1}{N} \sum_{i=1}^N \log P(x_i)\right). \quad (11.8)$$

#### Remark 11.2: Mean Average Precision (mAP)

Mean Average Precision (mAP) measures how well a system ranks relevant items by computing the average precision across different recall levels, making it crucial for information retrieval where ranking quality matters more than binary classification.

#### Remark 11.3: Normalized Discounted Cumulative Gain (nDCG)

Normalized Discounted Cumulative Gain (nDCG) evaluates ranking quality by considering both relevance and position, giving higher weight to items ranked earlier in the list, which reflects real-world user behavior where top results are most important.

#### Remark 11.4: Recall@k

Recall@k measures the proportion of relevant items found in the top-k results, providing a practical metric for applications where users only examine the first few results.

For retrieval and ranking, report mean average precision (mAP), normalized discounted cumulative gain (nDCG), and recall@k. These metrics are essential because they capture the nuanced performance of ranking systems where the order of results significantly impacts user experience, unlike traditional classification metrics that treat all predictions equally regardless of their position in a ranked list.

### 11.1.4 Worked examples

#### Example 11.1: Imbalanced Disease Detection

In a 1% prevalence setting, a classifier with 99% accuracy can be worthless. Reporting PR-AUC and calibration surfaces early detection quality and absolute risk estimates valued by clinicians Ronneberger, Fischer, and Brox [RFB15].

#### Example 11.2: Threshold Selection

Optimize thresholds against a cost matrix or utility function (e.g., false negative cost  $\gg$  false positive). Plot utility vs. threshold to choose operating points.

#### Example 11.3: Macro vs. Micro Averaging

For multi-class, macro-averaged F1 treats classes equally; micro-averaged F1 weights by support. Choose based on fairness vs. prevalence alignment Prince [Pri23].

## 11.2 Baseline Models and Debugging ◆

Baseline models and systematic debugging strategies are essential for establishing performance floors, identifying implementation issues, and ensuring that complex models provide genuine improvements over simple approaches.

### 11.2.1 Establishing Baselines

#### Definition 11.9: Baseline Model

A baseline model is a simple, well-understood reference system that provides a performance floor for comparison against more complex approaches.

Strong baselines de-risk projects by validating data quality, metrics, and feasibility Goodfellow, Bengio, and Courville [GBC16a] and Prince [Pri23]. They serve as sanity checks to ensure that sophisticated models actually improve upon simple solutions rather than introducing unnecessary complexity. Baselines help identify whether poor performance stems from model limitations or fundamental issues with data quality, preprocessing, or evaluation methodology. By establishing multiple baselines across different complexity levels, practitioners can quantify the marginal value of each architectural choice and avoid over-engineering solutions. These reference points become immutable benchmarks that prevent performance regression and provide confidence that improvements are genuine rather than artifacts of experimental variance.

Start with simple baselines:

1. **Random baseline:** Random predictions
2. **Simple heuristics:** Rule-based systems
3. **Classical ML:** Logistic regression, random forests
4. **Simple neural networks:** Small architectures

Compare deep learning improvements against these baselines.

#### Example 11.4: Data Leakage Prevention

Use data leakage checks (e.g., time-based splits, patient-level splits) and ensure identical preprocessing across baselines. Data leakage occurs when information from the future or test set inadvertently influences model training, leading to overly optimistic performance estimates that don't generalize to real-world scenarios.

### 11.2.2 Debugging Strategy

#### Remark 11.5: Debugging Deep Learning vs Traditional Software

Debugging deep learning models is fundamentally different from traditional software debugging because DL models fail silently with no clear error messages, have non-deterministic behavior due to random initialization and data ordering, and require understanding of both code and mathematical concepts like gradients and optimization landscapes. Unlike traditional software where bugs cause immediate crashes, DL models can appear to work while producing incorrect results, making systematic debugging essential for identifying whether failures stem from implementation bugs, data quality issues, hyperparameter choices, or fundamental model limitations.

Deep learning models often fail silently or produce unexpected results due to the complexity of neural architectures and the non-convex optimization landscape. Systematic debugging is essential because model failures can stem from multiple sources: implementation bugs, data quality issues, hyperparameter choices, or fundamental limitations of the approach. Without proper debugging methodology, practitioners may waste significant time pursuing ineffective solutions or miss critical insights about their data and model behavior.

#### Step 1: Overfit a small dataset

- Take 10-100 examples
- Turn off regularization
- If can't overfit, suspect implementation, data, or optimization bugs

## Step 2: Check intermediate outputs

- Visualize activations
- Check gradient magnitudes
- Verify loss decreases on training set
- Plot learning-rate vs. loss; test different seeds

## Step 3: Diagnose underfitting vs. overfitting

- **Underfitting:** Poor train performance → increase capacity
- **Overfitting:** Good train, poor validation → add regularization

### 11.2.3 Common Issues

#### Remark 11.6: Most Common Deep Learning Issues

The three most common issues in deep learning are:

- **Vanishing/Exploding Gradients:** Gradients become too small or large during backpropagation, preventing effective learning
- **Dead ReLUs:** Neurons never activate due to negative weights, reducing model capacity
- **Loss Not Decreasing:** Training fails due to incorrect learning rates, gradient bugs, or data preprocessing errors

Vanishing and exploding gradients represent fundamental challenges in deep network training that can prevent effective learning. Batch normalization stabilizes training by normalizing inputs to each layer, preventing gradients from becoming too small or large during backpropagation—for example, in a 50-layer network without batch norm, gradients might vanish to near-zero values by layer 20, but with batch norm they remain stable throughout the entire network. Gradient clipping prevents exploding gradients by capping gradient magnitudes at a threshold (e.g., 1.0 or 5.0), which is particularly important in RNNs where gradients can grow exponentially over long sequences, causing training instability and preventing convergence. Proper weight initialization (Xavier/He initialization) ensures gradients start at reasonable magnitudes rather than vanishing or exploding from the first forward pass, with He initialization for ReLU networks setting weights to  $\sqrt{2/n}$  where  $n$  is the input size, preventing the "dead neuron" problem where all activations become zero. Residual connections provide direct paths for gradient flow, allowing information to bypass layers where gradients might vanish, as seen in ResNet architectures where the skip connection  $y = F(x) + x$  ensures that even if  $F(x)$  becomes zero, the gradient can still flow through the identity connection.

Dead ReLUs occur when neurons never activate because their weights become too negative, often due to aggressive learning rates, where reducing the learning rate from 0.01 to 0.001 can prevent neurons from being "killed" during early training, allowing them to recover and contribute to learning. Unlike standard ReLU which outputs zero for negative inputs, Leaky ReLU allows small negative values (e.g., 0.01x) and ELU provides smooth negative outputs, preventing the "dying ReLU" problem where neurons become permanently inactive, as seen in networks where 30-50% of neurons might never fire after initialization. Loss not decreasing can stem from multiple sources that require systematic investigation. Learning rates that are too high cause the optimizer to overshoot the minimum and oscillate around it, while rates that are too low make training painfully slow—a learning rate of 0.1 might cause loss to bounce between 0.5 and 0.7, while 0.0001 might show no improvement for 100 epochs. Gradient computation bugs can cause the optimizer to move in wrong directions or not move at all, with common issues including incorrect backpropagation implementations, wrong loss function derivatives, or gradient accumulation errors that result in gradients being zero or pointing away from the minimum. Incorrect data preprocessing can make learning impossible by normalizing inputs to the wrong scale or introducing data leakage, such as normalizing images to [0,1] when the model expects [-1,1], or accidentally including future information in time series data that prevents the model from learning meaningful patterns. Misaligned labels or incorrect class indexing can cause the model to learn the wrong mappings, with a common

mistake being using 1-based indexing for labels when the model expects 0-based indexing, causing the model to predict class 0 when it should predict class 1, resulting in consistently wrong predictions.

### 11.2.4 Ablation and sanity checks

Perform *ablation studies* to quantify the contribution of each component (augmentation, architecture blocks, regularizers). Use *label shuffling* to verify the pipeline cannot learn when labels are randomized. Train with *frozen features* to isolate head capacity.

#### Remark 11.7: Ablation Studies and Sanity Checks

- **Ablation studies:** Systematically remove components to measure their individual contribution to model performance
- **Label shuffling:** Randomize labels to verify the model cannot learn when ground truth is meaningless
- **Frozen features:** Fix feature extractor weights to test if the classification head has sufficient capacity

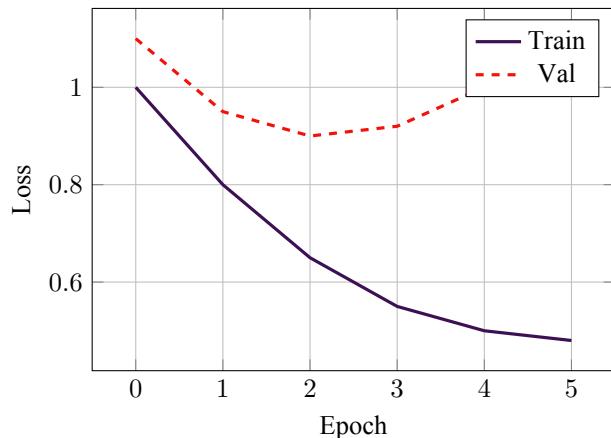


Figure 11.6: Typical overfitting: training loss decreases while validation loss bottoms out and rises.

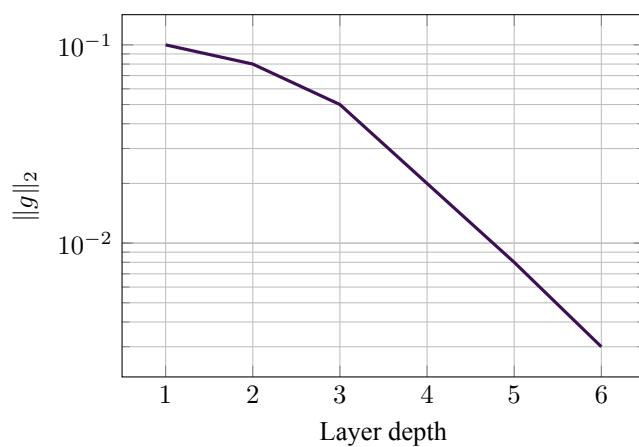


Figure 11.7: Gradient norms vanishing with depth; motivates normalization and residual connections.

#### Remark 11.8: Historical Notes and References

Debugging by overfitting a tiny subset and systematic ablations has roots in classical ML practice and was emphasized in early deep learning methodology Goodfellow, Bengio, and Courville [GBC16a]. Modern best practices are also surveyed in open textbooks Prince [Pri23] and Zhang et al. [Zha+24b].

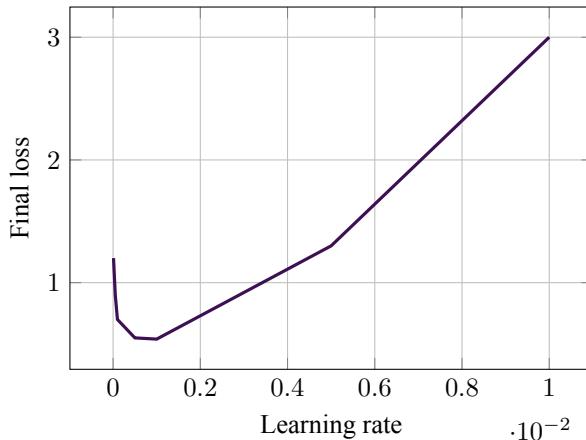


Figure 11.8: Learning-rate sweep to identify a stable training regime.

## 11.3 Hyperparameter Tuning ◆

**Hyperparameter tuning** is the process of selecting optimal configuration settings that control how a machine learning model learns, rather than the parameters the model learns itself. Unlike classical ML algorithms like linear regression or decision trees that have few, well-understood hyperparameters, deep learning models have dozens of hyperparameters that interact in complex ways, making tuning much more challenging and critical for success.

### Example 11.5: Musical Instrument Metaphor

Think of hyperparameter tuning like tuning a musical instrument—classical ML is like tuning a simple guitar with just a few strings, where each adjustment has a clear, predictable effect. Deep learning is like tuning a complex orchestra with dozens of instruments, where changing one instrument’s tuning affects how all the others sound together, and the perfect harmony requires careful coordination of many interdependent settings.

### 11.3.1 Key Hyperparameters (Priority Order)

Effective tuning prioritizes learning rate, regularization, and capacity before fine details. This systematic approach prevents wasting time on minor optimizations when fundamental issues remain unresolved—for example, tuning dropout rates while using a learning rate that’s 10x too high will yield poor results regardless of regularization choices. Treat the validation set as your instrumentation layer and control randomness via fixed seeds Goodfellow, Bengio, and Courville [GBC16a], Prince [Pri23], and Zhang et al. [Zha+24b].

Priority	Hyperparameter	Key Considerations
1	Learning rate	Most critical; consider warmup and cosine decay
2	Network architecture	Depth/width, normalization, residuals
3	Batch size	Affects noise scale and generalization
4	Regularization	Weight decay, dropout, label smoothing
5	Optimizer parameters	Momentum, $\beta$ values in Adam

Table 11.1: Priority order for hyperparameter tuning in deep learning.

### 11.3.2 Search Strategies

Manual search involves human-guided exploration of hyperparameter space based on domain knowledge and intuition, beginning with hyperparameter values that have worked well in similar problems or are recommended in literature, such as

starting with a learning rate of 0.001 for Adam optimizer or using 0.5 dropout rate for fully connected layers, as these are commonly successful starting points across many deep learning tasks. The approach systematically modifies hyperparameters based on validation performance, typically changing one parameter at a time to understand its individual effect—if validation loss plateaus, try increasing learning rate; if overfitting occurs, increase regularization strength or reduce model capacity. While manual search requires significant human effort and can take days or weeks, it provides deep understanding of how different hyperparameters affect model behavior, building intuition that proves valuable for future projects and helping identify the most promising regions of hyperparameter space.

Grid search systematically evaluates all combinations of hyperparameters from predefined discrete sets, defining a grid of possible values for each hyperparameter (e.g., learning rates [0.001, 0.01, 0.1] and batch sizes [32, 64, 128]) and training a model for every possible combination, ensuring comprehensive coverage of the specified search space without missing any potential configurations. While grid search guarantees finding the best combination within the defined grid, computational cost grows exponentially with the number of hyperparameters—for 3 hyperparameters with 5 values each, you need 125 training runs, making it computationally prohibitive for large models or extensive search spaces. Grid search works well when you have few hyperparameters to tune, as the search space remains manageable, but with more than 3-4 hyperparameters, the curse of dimensionality makes grid search impractical, and most combinations will likely be suboptimal anyway.

Random search samples hyperparameters from probability distributions rather than evaluating all combinations, randomly sampling hyperparameter values from appropriate distributions (e.g., learning rate from log-uniform distribution, dropout from uniform distribution) to explore the search space more efficiently by avoiding the rigid structure of grid search. Random search often finds good hyperparameters with fewer trials than grid search because it doesn't waste time on systematically poor regions of the search space, with studies showing that random search can achieve similar performance to grid search with 10-100x fewer evaluations, making it much more practical for expensive training procedures. As the number of hyperparameters increases, random search becomes increasingly advantageous over grid search, since in high-dimensional spaces, most of the volume lies near the boundaries, and random sampling naturally explores these regions more effectively than the structured approach of grid search.

Bayesian optimization uses probabilistic models to guide hyperparameter search intelligently, building a probabilistic model (typically Gaussian Process) that predicts the performance of untested hyperparameter configurations based on previous evaluations, capturing both the expected performance and uncertainty to allow informed decisions about where to search next. Instead of random sampling, Bayesian optimization uses acquisition functions (like Expected Improvement or Upper Confidence Bound) to select the most promising hyperparameter configurations to evaluate next, balancing exploration of uncertain regions with exploitation of areas likely to contain good solutions. Bayesian optimization typically requires far fewer evaluations than random or grid search to find good hyperparameters, especially when each evaluation is expensive, making it particularly valuable for neural architecture search or when training large models, where each hyperparameter trial might take hours or days to complete.

### 11.3.3 Best Practices

Effective hyperparameter tuning requires systematic approaches that balance thoroughness with computational efficiency while maintaining scientific rigor, helping practitioners avoid common pitfalls and maximize the value of their tuning efforts. Use logarithmic scale for learning rate and sweep  $[10^{-5}, 10^{-1}]$  because learning rates span several orders of magnitude, and linear spacing would miss critical regions where small changes have dramatic effects—for example, the difference between 0.001 and 0.01 can mean the difference between convergence and divergence, while the difference between 0.1 and 0.11 is usually negligible, with logarithmic sampling ensuring equal attention to each order of magnitude and capturing the full range of potentially useful learning rates.

Vary batch size and adjust learning rate proportionally since larger batch sizes provide more stable gradients but require higher learning rates to maintain the same effective step size, where doubling batch size from 32 to 64 requires increasing learning rate by approximately 2x to maintain similar convergence dynamics, a relationship that stems from the fact that larger batches reduce gradient noise, allowing for more aggressive updates without destabilizing training. Track results with a consistent random seed and multiple repeats because deep learning results can vary significantly due to random initialization and data shuffling, making single runs unreliable for hyperparameter comparison—use fixed seeds for reproducibility and run multiple trials (3-5) to

estimate the variance and ensure that performance differences are statistically significant rather than due to random chance. Early-stop poor runs and allocate budget adaptively by monitoring training progress and terminating clearly failing experiments early, where if a configuration shows no improvement after 20% of the planned training time, stop it and redirect computational resources to more promising candidates, as this adaptive allocation can reduce total tuning time by 50-70% while focusing effort on the most promising regions of hyperparameter space. Use a fixed validation protocol to avoid leakage by establishing a single, immutable validation split before beginning any hyperparameter tuning to prevent data leakage and overfitting to the validation set, since changing validation splits during tuning can lead to overly optimistic estimates and poor generalization, with the validation set remaining completely untouched until the final evaluation and all hyperparameter decisions based on this consistent benchmark.

Retrain with best setting on train+val and report on held-out test by identifying the best hyperparameters, then retraining the model using both training and validation data to maximize the information available for learning, and reporting final performance on a completely held-out test set that was never used for any hyperparameter decisions. This two-stage approach ensures that the final model uses all available training data while maintaining an unbiased estimate of true generalization performance.

### 11.3.4 Historical notes

The scaling of deep learning models and search spaces necessitated a move beyond rudimentary optimization methods, driving the evolution from simple grid search to more sophisticated hyperparameter tuning strategies. The challenge of efficiently finding optimal hyperparameters in high-dimensional spaces led to the rise of Random Search and Bayesian Optimization, offering superior coverage and effectiveness compared to exhaustive grid searches that became computationally prohibitive. For the stability of model weights during the immense computation required for large language models and Transformers, learning-rate schedules became a standard component of modern training pipelines. Specifically, practices like Warmup ensure that training begins with a small learning rate to stabilize early-stage gradients before gradually increasing it, while techniques such as Step Decay or Cosine Annealing then regulate the descent toward the optimum in later phases. These scheduling mechanisms are essential for preventing gradient explosion and achieving reliable convergence across large batches and deep network architectures. The historical progression from manual tuning to automated optimization reflects the broader trend toward systematic, data-driven approaches in deep learning methodology Goodfellow, Bengio, and Courville [GBC16a], Prince [Pri23], and Zhang et al. [Zha+24b].

## 11.4 Data Preparation and Preprocessing ◆

### 11.4.1 Data Splitting

#### Remark 11.9: Data Splitting in Deep Learning

Data splitting is not new to deep learning—it is already widely used in classical ML. However, it becomes even more critical in deep learning because neural networks have high capacity and can easily overfit to small datasets, making proper train/validation/test splits essential for reliable evaluation and preventing data leakage that could lead to overly optimistic performance estimates.

**Train/Validation/Test split:** This three-way split ensures unbiased model evaluation by keeping the test set completely isolated until final assessment, while the validation set guides hyperparameter tuning without contaminating the final performance estimate. The validation set acts as a proxy for the test set during development, allowing you to make informed decisions about model architecture and hyperparameters without peeking at the true test performance.

**Cross-validation:** For small datasets where a single train/validation split might not provide reliable estimates, k-fold cross-validation uses all available data for both training and validation by rotating which subset serves as the validation set. This approach maximizes the use of limited data while providing more robust performance estimates, especially crucial when you have fewer than 1000 examples and need to make the most of every data point.

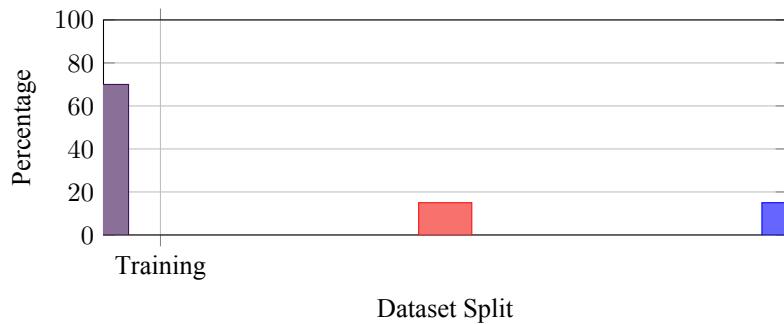


Figure 11.9: Train/Validation/Test split with typical proportions: 70% training, 15% validation, 15% test.

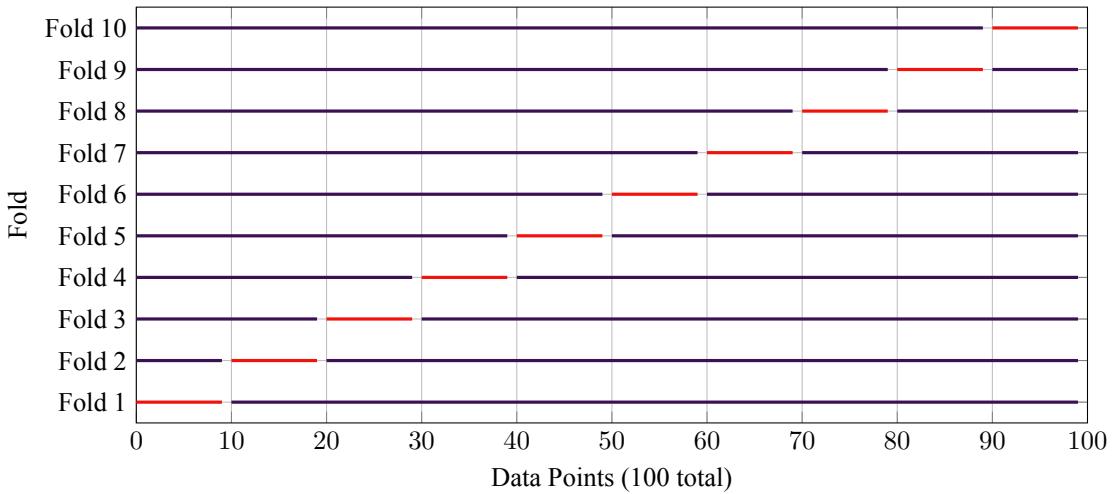


Figure 11.10: 10-fold cross-validation: red=validation (10 each), purple=training (90 each).

- **k-fold cross-validation:** Divides data into k equal folds, using each fold as validation set once
- **Stratified splits for imbalanced data:** Ensures each fold maintains the same class distribution as the original dataset

### 11.4.2 Normalization

Normalization is essential because neural networks are sensitive to the scale of input features, and features with vastly different ranges can cause training instability and poor convergence. When one feature ranges from 0 to 1 while another spans 0 to 1000, the larger-scale feature dominates the learning process, causing the network to ignore the smaller-scale feature entirely. This scale imbalance leads to slow convergence, as the optimizer struggles to find appropriate learning rates that work for both features simultaneously.

#### Definition 11.10: Min-Max Scaling

This method rescales features to a fixed range, typically [0,1], by subtracting the minimum value and dividing by the range. Min-max scaling preserves the original distribution shape and is particularly useful when you know the expected range of your data or when you need features to have the same scale for distance-based algorithms.

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (11.9)$$

**Definition 11.11: Standardization (Z-score)**

This approach transforms features to have zero mean and unit variance, making them follow a standard normal distribution. Standardization is more robust to outliers than min-max scaling and is the preferred method for most deep learning applications, as it centers the data around zero and gives equal importance to all features regardless of their original scale.

$$x' = \frac{x - \mu}{\sigma} \quad (11.10)$$

Always compute statistics on training set only! Using validation or test set statistics would create data leakage, as the model would have access to information from future data during training, leading to overly optimistic performance estimates that don't generalize to truly unseen data.

### 11.4.3 Handling Imbalanced Data

**Example 11.6: Imbalanced Data in Social Media**

Consider a social media platform with 1 million posts where 950,000 are normal content and only 50,000 are spam (5%)

Imbalanced data occurs when one or more classes have significantly fewer examples than others, creating a skewed class distribution that can severely bias model training toward the majority class, which is problematic because standard machine learning algorithms assume balanced class distributions and will naturally favor the majority class, leading to poor performance on minority classes that are often the most important to identify correctly. Oversampling duplicates minority class examples to balance the dataset artificially, increasing the representation of minority classes by creating exact copies of existing examples, which helps the model see more minority class instances during training, though simple duplication can lead to overfitting since the model sees identical examples multiple times, potentially memorizing specific patterns rather than learning generalizable features.

Undersampling removes majority class examples to create a more balanced dataset by randomly discarding instances from the overrepresented class, reducing computational cost and training time while forcing the model to pay more attention to minority classes, though the main drawback is the loss of potentially valuable information from the majority class, which can hurt overall model performance if the discarded examples contain important patterns. SMOTE (Synthetic Minority Oversampling Technique) creates new synthetic examples for minority classes by interpolating between existing minority class instances in feature space, generating realistic synthetic data points by finding k-nearest neighbors of minority examples and creating new instances along the line segments connecting them, providing more diverse training examples than simple duplication while maintaining the statistical properties of the original minority class distribution.

Class weights penalize errors on minority class more heavily during training by assigning higher loss weights to minority class misclassifications, adjusting the loss function to make the model more sensitive to minority class errors and effectively forcing it to prioritize learning the underrepresented classes, with weights typically set inversely proportional to class frequency so a class with 10% representation gets 10x higher weight than a class with 100% representation. Focal loss focuses on hard examples by down-weighting easy examples and up-weighting difficult-to-classify instances, particularly useful for extreme class imbalance, as this loss function automatically adapts to the difficulty of each example, reducing the contribution of well-classified majority class examples while emphasizing misclassified minority class instances, making it especially effective for object detection and segmentation tasks where background pixels vastly outnumber foreground objects.

### 11.4.4 Data Augmentation

**Data augmentation** is a crucial strategy in deep learning to artificially increase the size and diversity of a training dataset, which is vital for achieving generalization and mitigating overfitting, especially when working with limited real-world samples. By generating additional examples through domain-specific transformations, such as flips, crops, or color jitter for images, the model learns to recognize the core object or pattern regardless of minor variations. In the NLP space, techniques like back-translation (translating text to another language and back) introduce crucial syntactic and vocabulary variance that

stabilizes large language models. The primary challenge lies in calibrating the strength of these augmentations, as excessive or unrealistic noise, such as extreme time stretching for audio or radical color shifts for images, can distort the underlying signal and cause a debilitating distribution shift that undermines model performance. Ultimately, intelligent data augmentation expands the effective manifold of the training data without the cost of collecting new samples.

For images: flips, crops, color jitter; for text: back-translation; for audio: time stretch, noise. Calibrate augmentation strength to avoid distribution shift Prince [Pri23].

### 11.4.5 Visual aids

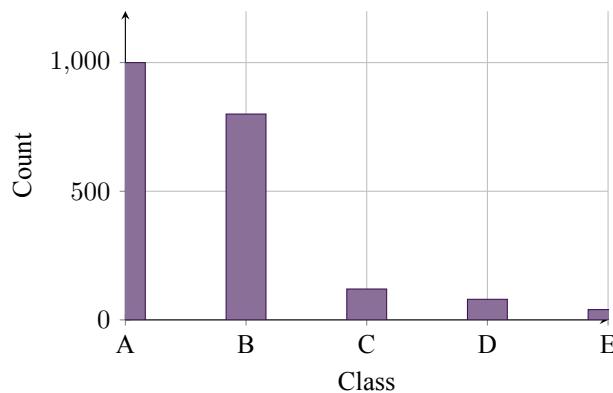


Figure 11.11: Imbalanced dataset example motivating class weights or resampling.

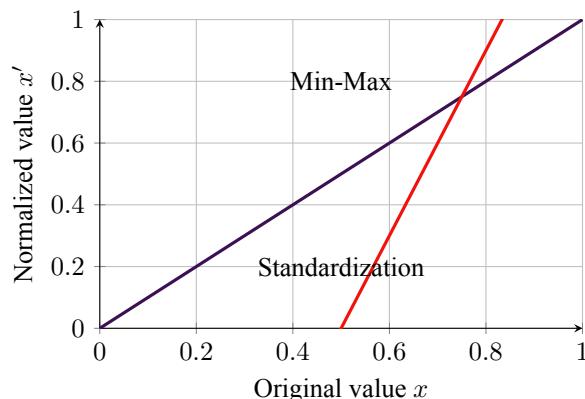


Figure 11.12: Min-max scaling (purple) vs. standardization (red) schematic.

#### Remark 11.10: Transfer Learning

Transfer learning is like using a pre-trained chef's knowledge to cook a new dish—instead of learning cooking from scratch, you start with someone who already knows how to handle ingredients and techniques, then adapt their skills to your specific recipe. In deep learning, this means using a model trained on a large dataset (like ImageNet) as a starting point for your specific task, dramatically reducing the data and time needed to achieve good performance.

### 11.4.6 Historical notes

Careful dataset design (train/val/test segregation, leakage prevention) has long underpinned reliable evaluation in ML and remains essential at scale in deep learning Bishop [Bis06] and Goodfellow, Bengio, and Courville [GBC16a]. The evolution from classical ML to deep learning fundamentally transformed data preprocessing requirements, as traditional methods like linear regression and decision trees were relatively robust to feature scaling, while neural networks require careful normalization

to prevent gradient instability. The introduction of batch normalization by Ioffe and Szegedy in 2015 marked a pivotal moment, as it automated the normalization process during training, eliminating the need for manual feature scaling in many cases. Unlike classical methods where data splitting was primarily about preventing overfitting, deep learning's data augmentation techniques (pioneered in computer vision by Krizhevsky et al. in 2012) became essential for generalization, as neural networks' high capacity made them prone to memorizing training data. The rise of transfer learning and pre-trained models further complicated data preparation, as practitioners now needed to understand how to adapt datasets for models trained on different distributions, a challenge that classical ML rarely faced. Modern frameworks like TensorFlow and PyTorch have democratized these sophisticated preprocessing techniques, making advanced data preparation accessible to practitioners who previously relied on simpler methods like one-hot encoding for categorical variables or basic standardization for continuous features.

## 11.5 Production Considerations ◆

Production environments present fundamentally different challenges compared to localhost, development, or staging environments where models are initially developed and tested. While development focuses on model accuracy and training efficiency, production must handle real-world constraints like user traffic spikes, hardware limitations, and the unpredictable nature of live data streams. Unlike controlled testing environments with curated datasets, production systems face distribution shifts, adversarial inputs, and edge cases that can cause models to fail catastrophically if not properly monitored and managed. The transition from prototype to production requires careful consideration of scalability, reliability, and maintainability—factors that are often overlooked during initial development but become critical when serving millions of users or processing real-time data streams.

### 11.5.1 Model Deployment

#### Remark 11.11: P95 and P99 Percentiles

P95 and P99 percentiles represent the response time thresholds where 95

**Model deployment** is the process of making trained models available to serve predictions in production environments, transforming research prototypes into reliable services that can handle real-world traffic and constraints. Unlike development environments where models run on single machines with unlimited resources, production deployment requires careful orchestration of infrastructure, monitoring, and continuous improvement to ensure models perform reliably at scale. Production environments often have strict resource constraints that require careful optimization of model size and computational requirements. Mobile devices with limited memory or edge servers with computational budgets necessitate model compression techniques like pruning, which removes unnecessary weights, and quantization, which reduces precision from 32-bit to 8-bit floating point numbers. These techniques can reduce model size by 70-90% while maintaining acceptable accuracy, making them essential for deploying large models like BERT or GPT variants on resource-constrained devices. This compression enables real-time inference without requiring expensive cloud infrastructure, making advanced AI capabilities accessible on devices with limited computational resources.

Production systems need robust infrastructure to safely deploy new models without disrupting existing services, requiring sophisticated deployment strategies that balance innovation with reliability. A/B testing allows comparing new model versions against current ones using a small percentage of traffic, providing statistical validation of improvements while minimizing risk to the overall system. Canary deployments gradually roll out changes to detect issues early, enabling quick rollbacks if problems arise and ensuring that model improvements are validated with real user data before full deployment. This infrastructure prevents catastrophic failures by providing multiple safety nets and validation mechanisms that protect both users and business operations.

Real-time applications like recommendation systems or fraud detection require sub-100ms response times, making latency optimization crucial for user experience and business success. While average latency might be acceptable for many applications, tail latency measured by p95 and p99 percentiles can cause significant user frustration when 5% of requests take 10x longer than expected. Optimizing for tail performance involves techniques like request queuing, intelligent caching strategies, and model

optimization to ensure consistent response times across all user interactions. This focus on tail performance is particularly important for applications where user experience directly impacts business metrics like conversion rates and user retention. Production systems must choose between batch processing, which predicts on large datasets periodically, and online inference, which provides real-time predictions for individual requests. Batch processing is more computationally efficient and allows for complex feature engineering that might be too expensive for real-time systems, but online inference provides immediate results that are essential for time-sensitive applications. The choice between these approaches depends heavily on feature freshness requirements—recommendation systems might tolerate 1-hour-old features that can be pre-computed, while fraud detection needs real-time transaction data to be effective in preventing fraudulent activities as they occur.

### 11.5.2 Monitoring

**Monitoring** is the continuous observation and measurement of model performance, system health, and data quality in production environments to detect issues before they impact users. Unlike development environments where you can manually inspect results, production monitoring requires automated systems that can detect subtle changes in model behavior, data distribution, or system performance that might indicate degradation or failure.

Models trained on historical data often fail when the input distribution changes, such as when user behavior shifts or new data sources are introduced, making distribution shift monitoring essential for maintaining model performance. Covariate shift occurs when input features change, such as new user demographics entering the system, while label shift happens when the relationship between inputs and outputs changes, such as economic conditions affecting fraud patterns. Monitoring these shifts using statistical tests like Kolmogorov-Smirnov or population stability index helps detect when models need retraining before performance degrades significantly, enabling proactive responses to changing data conditions that could otherwise lead to silent failures.

While accuracy might remain stable over time, model calibration—the reliability of probability estimates—can drift significantly, leading to overconfident or underconfident predictions that can have serious consequences. This is particularly critical in applications like medical diagnosis or financial risk assessment where probability estimates directly impact decision-making processes and outcomes. Monitoring calibration drift involves tracking metrics like expected calibration error and reliability diagrams to ensure that a model’s confidence scores remain trustworthy as data distributions evolve, preventing situations where models appear to perform well but provide misleading confidence estimates.

Production systems must maintain consistent performance under varying load conditions, requiring careful monitoring of response times, request throughput, and resource utilization to ensure optimal user experience. Autoscaling systems automatically adjust computational resources based on demand, but they need proper configuration to prevent over-provisioning that wastes resources or under-provisioning that causes timeouts and service degradation. Monitoring these metrics helps optimize cost-performance trade-offs and ensures that user experience remains consistent during traffic spikes or system updates, balancing operational efficiency with service reliability.

Automated monitoring can detect performance degradation, but understanding the root cause often requires human expertise to analyze failure patterns and edge cases that aren’t apparent from aggregate metrics. Human-in-the-loop systems combine automated error detection with expert review to identify systematic issues, data quality problems, or model limitations that could lead to broader system failures. This approach is essential for complex applications where errors can have significant consequences, such as autonomous vehicles or medical diagnosis systems, where understanding the specific nature of failures is crucial for developing effective solutions and preventing similar issues in the future.

### 11.5.3 Iterative Improvement

**Iterative improvement** is the continuous cycle of deploying models, monitoring their performance, collecting feedback, and refining them based on real-world usage patterns and user behavior. Unlike one-time model development, production systems require ongoing optimization to maintain performance as data distributions change, user preferences evolve, and new edge cases emerge that weren’t present during initial training.

The first deployment establishes a baseline model in production, typically using a conservative approach with extensive monitoring and gradual rollout to minimize risk and ensure system stability. This initial model serves as the foundation for all

future improvements and provides the first real-world performance data that reveals how the model behaves with actual users and data. The deployment process includes setting up comprehensive monitoring infrastructure, establishing performance baselines that will be used to measure future improvements, and creating rollback procedures that can quickly restore the system to a known good state in case of unexpected issues.

Continuous monitoring tracks model performance across multiple dimensions, including accuracy metrics, user satisfaction indicators, system health metrics, and business outcomes that directly impact the organization's success. This monitoring reveals how the model behaves with real users and data, identifying areas for improvement that weren't apparent during development and highlighting edge cases that require attention. The monitoring data provides valuable insights into model limitations, unexpected failure modes, and opportunities for enhancement that guide future development efforts and help prioritize which improvements will have the greatest impact.

Production systems generate valuable data that can be used to improve models, including user interactions, feedback signals, and edge cases that weren't present in the original training set. This data collection includes both explicit feedback such as user ratings and corrections, and implicit signals like user behavior patterns and engagement metrics that provide rich information about model performance and user needs. The collected data becomes the foundation for retraining and improving models with more representative and comprehensive datasets that better reflect the real-world conditions the system will encounter.

Using the collected production data, models are retrained with updated datasets that include real-world examples and edge cases that weren't available during initial development. This retraining process may involve architectural changes to better handle the new data patterns, hyperparameter tuning to optimize performance on the updated dataset, or incorporating new features that weren't available during initial development. The improved models are thoroughly tested against the original baseline to ensure they provide meaningful improvements before deployment, using both offline metrics and small-scale online testing to validate the changes.

Before full deployment, improved models are tested against the current production model using controlled experiments with a subset of users to ensure that improvements are real and not due to random variation. A/B testing provides statistical validation of improvements while also allowing for gradual rollout to minimize risk and detect any unexpected issues before they impact the entire user base. This testing process ensures that model improvements translate to better user experience and business outcomes before committing to full deployment, providing confidence that the changes will have the intended positive impact on the system's overall performance.

#### 11.5.4 Model Drift

Model drift occurs when the statistical properties of input data change over time, causing model performance to degrade as the model was trained on data that no longer represents the current distribution.

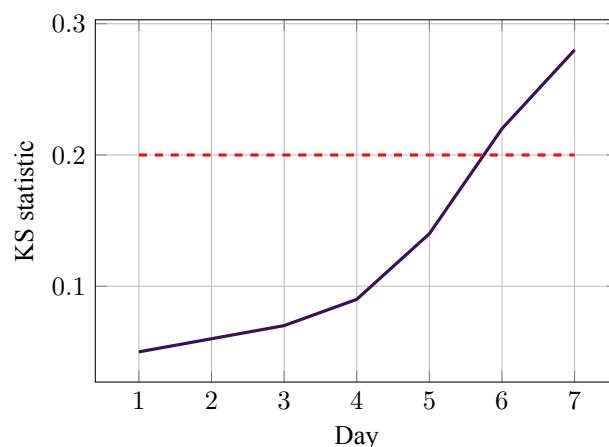


Figure 11.13: Simple drift monitor: KS statistic over time with an alert threshold. The purple line shows the Kolmogorov-Smirnov statistic measuring distribution shift between training and production data, while the red dashed line represents the alert threshold (0.2) that triggers when drift becomes significant.

**Remark 11.12: Model Monitoring**

Model monitoring is about both model development and operation. Tools such as Hugging Face, PyTorch, and TensorFlow provide dashboards to monitor model performance during tuning, enabling practitioners to track training progress, identify issues early, and optimize hyperparameters effectively.

**11.5.5 Applications and context**

Production considerations vary significantly across different application domains, each with unique requirements and constraints that influence deployment strategies. Mobile applications prioritize model compression and low-latency serving, as edge devices have limited computational resources and users expect instant responses for vision and speech applications. Recommender systems require real-time inference with sub-100ms latency to maintain user engagement, making optimization techniques like quantization and caching essential for delivering personalized content at scale. Healthcare and finance applications demand the highest levels of model reliability and calibration, as prediction errors can have life-threatening or financially catastrophic consequences. These domains require extensive post-deployment monitoring, human-in-the-loop validation, and rigorous testing protocols to ensure models maintain their performance over time. The stakes are so high that even small calibration drift or distribution shifts can lead to incorrect diagnoses or fraudulent transactions going undetected.

**Remark 11.13: Agile Methodology and MLOps**

Agile methodology and MLOps pipelines can significantly help with deploying models by enabling rapid iteration, continuous integration, and automated testing that catches issues early in the development cycle. These practices ensure that model improvements are systematically validated, deployed, and monitored, reducing the risk of production failures while accelerating the delivery of valuable AI capabilities to users.

The choice of production strategies depends heavily on the specific application requirements: real-time systems prioritize latency optimization, safety-critical applications emphasize monitoring and validation, while resource-constrained environments focus on compression and efficient serving architectures Ronneberger, Fischer, and Brox [RFB15] and Prince [Pri23].

**11.6 Real World Applications**

Practical methodology—the systematic approach to designing, training, and deploying deep learning systems—is what separates successful real-world projects from academic experiments.

**11.6.1 Healthcare Diagnostic System Deployment**

Bringing AI from lab to clinic requires a fundamentally different approach than academic research, as the stakes involve human lives and regulatory compliance. Companies developing AI diagnostic tools must follow rigorous methodologies that begin with careful dataset collection from diverse hospitals, ensuring representation across different demographics, equipment types, and clinical settings. A stroke detection system, for example, must work reliably across different scanners, patient populations, and hospital settings before doctors trust it with patient care, requiring systematic validation on held-out test sets and extensive clinical trials that can take years to complete.

Real medical data presents unique challenges that academic datasets rarely capture—images contain artifacts from equipment malfunctions, labels include errors from overworked radiologists, and rare diseases are severely underrepresented in training sets. Practical methodology addresses these issues through comprehensive data cleaning procedures, sophisticated techniques for handling class imbalance, and the establishment of confidence thresholds that determine when the system should defer to human experts rather than risk misdiagnosis.

Once deployed, medical AI systems require ongoing validation that goes far beyond typical model monitoring. This includes establishing comprehensive monitoring dashboards that track not just accuracy metrics but also clinical outcomes, detecting

distribution shift when patient populations change due to seasonal diseases or demographic shifts, and developing protocols for updating models without disrupting critical clinical workflows. The methodology must account for the fact that a single misdiagnosis can have life-threatening consequences, making continuous improvement both essential and extremely delicate.

### 11.6.2 Recommendation System Development

Building and maintaining large-scale personalization systems requires methodologies that balance user satisfaction with business objectives while handling the complexity of serving millions of users simultaneously. When Netflix develops new recommendation algorithms, they don't just optimize offline metrics—practical methodology involves carefully designed A/B tests with real users that balance multiple competing objectives including user engagement, content diversity, discovery of new material, and business goals like subscription retention. This approach requires understanding long-term effects beyond immediate clicks, as a recommendation that increases short-term engagement might reduce long-term satisfaction if it creates filter bubbles.

The cold start problem presents a fundamental challenge in recommendation systems, where new users have no interaction history and new items have no ratings to guide recommendations. Practical methodology addresses this through strategic initialization techniques that leverage demographic information and content features, hybrid approaches that combine collaborative filtering with content-based methods, and active learning strategies that quickly gather useful information through carefully designed user interactions. These techniques must work seamlessly with existing recommendation pipelines while providing meaningful value to users from their first interaction.

Serving recommendations to millions of users simultaneously requires careful system design that goes far beyond model accuracy. The methodology includes choosing appropriate model architectures that balance accuracy with inference speed, implementing sophisticated caching strategies that can handle the scale of global user bases, and developing gradual rollout procedures that can detect problems early before they impact large user segments. This infrastructure must be robust enough to handle traffic spikes during popular content releases while maintaining sub-100ms response times that users expect from modern applications.

### 11.6.3 Autonomous Vehicle Development

The most safety-critical deep learning application requires methodologies that prioritize safety above all other considerations, as any failure can result in catastrophic consequences. Self-driving cars must handle rare but critical scenarios like a child running into the street, situations that are difficult to encounter in real-world testing but must be thoroughly validated before deployment. Companies use systematic methodologies that combine extensive real-world data collection from diverse driving conditions, photorealistic simulation of dangerous scenarios that would be too risky to test in reality, and extensive closed-track testing with professional drivers before any public road trials.

When test vehicles make mistakes, teams follow rigorous procedures that go far beyond typical software debugging to understand root causes, reproduce issues in simulation environments, develop fixes, and validate improvements through comprehensive testing protocols. This includes systematic logging of all sensor data and decision-making processes for later analysis, creating detailed incident reports that can inform future development, and implementing changes that are validated through multiple testing stages before being approved for deployment.

Models progress through increasingly realistic testing environments that mirror the complexity of real-world deployment: simulation environments that can test millions of scenarios, closed tracks with controlled conditions, controlled public roads with safety drivers, and finally broader deployment with extensive monitoring. Each stage has specific success criteria and methodologies for objective evaluation that must be met before progression to the next level, ensuring that safety is never compromised in the pursuit of technological advancement.

## 11.6.4 Key Methodological Principles

### Remark 11.14: Bringing Ideas to Products

Bringing an idea to a product is not easy. You need to learn software engineering, develop a model using AI coding agents with Agile methodology, and follow DevOps and MLOps practices. This requires combining technical skills with project management, understanding business requirements, and implementing robust deployment pipelines that can scale to serve real users.

What makes real-world projects succeed requires a systematic approach that balances technical excellence with practical constraints and business objectives. These principles have emerged from countless real-world deployments and represent the accumulated wisdom of practitioners who have learned to navigate the complex landscape of production deep learning systems.

**Start simple:** Baseline models first, then increase complexity as needed. This principle prevents teams from over-engineering solutions before understanding the fundamental requirements and constraints of their specific problem. A simple linear model or basic neural network often provides surprising insights into the data and problem structure, revealing which features matter most and where the real challenges lie. Only after establishing a working baseline should teams consider more sophisticated architectures, as complexity without understanding leads to systems that are difficult to debug, maintain, and improve.

**Measure what matters:** Align metrics with actual business or user goals rather than academic benchmarks that may not reflect real-world value. A model that achieves 99% accuracy on a test set but fails to improve user engagement or business outcomes is ultimately useless, regardless of its technical sophistication. This requires close collaboration with domain experts and stakeholders to understand what success looks like in practice, then designing evaluation frameworks that capture these nuanced objectives rather than relying solely on standard machine learning metrics.

**Understand your data:** Invest time in data exploration and cleaning before building any models, as the quality and characteristics of your data will fundamentally determine what's possible. Real-world datasets contain surprises, biases, and patterns that aren't apparent from high-level summaries, requiring systematic exploration to understand missing values, outliers, and distributional properties. This upfront investment pays dividends throughout the project lifecycle, as models built on well-understood data are more robust, interpretable, and maintainable than those built on poorly characterized datasets.

**Iterate systematically:** Change one thing at a time to understand impact, avoiding the temptation to make multiple changes simultaneously that can obscure the effects of individual modifications. This disciplined approach enables teams to build causal understanding of what drives performance improvements, making it easier to debug issues and replicate successes. Systematic iteration also provides a clear audit trail of decisions and their consequences, which is essential for maintaining and improving systems over time.

**Plan for production:** Consider deployment constraints from the beginning rather than treating them as afterthoughts that can be addressed later. This includes understanding latency requirements, resource constraints, security considerations, and integration challenges that will ultimately determine whether a model can be successfully deployed. Early consideration of these constraints often leads to architectural decisions that make deployment much easier, while ignoring them can result in models that perform well in development but are impossible to deploy in practice.

**Monitor continuously:** Real-world conditions change constantly, requiring models to adapt to new data distributions, user behaviors, and environmental factors that weren't present during training. This monitoring goes beyond simple accuracy metrics to include system health, user satisfaction, business outcomes, and early warning signs of degradation. Continuous monitoring enables proactive responses to changing conditions, preventing catastrophic failures and ensuring that models remain effective as the world around them evolves.

These examples show that methodology—the "how" of deep learning—is just as important as the "what" when building systems that work reliably in practice.

## Key Takeaways

### Key Takeaways 11

- **Start simple:** Establish a reliable baseline to validate data, metrics, and training code.
- **Measure relentlessly:** Use clear validation splits, confidence intervals, and learning curves.
- **Ablate to learn:** Prefer small, controlled changes to isolate causal effects.
- **Prioritise data:** Label quality, coverage, and augmentation often beat model complexity.
- **Tune methodically:** Track hyperparameters, seeds, and environments for reproducibility.
- **Deploy with monitoring:** Watch for drift, performance decay, and fairness regressions; plan periodic re-training.

## Exercises

### Easy

*Exercise 11.1 (Define Success Metrics).* You are building a classifier for defect detection. Propose suitable metrics beyond accuracy and justify validation splits.

**Hint:**

*Consider precision/recall, AUROC vs. AUPRC under class imbalance, and stratified splits.*

*Exercise 11.2 (Sanity Checks).* List three quick sanity checks to run before large-scale training and explain expected outcomes.

**Hint:**

*Overfit a tiny subset; randomise labels; train with shuffled pixels.*

*Exercise 11.3 (Baseline First).* Explain why a strong non-deep baseline can accelerate iteration on a deep model.

**Hint:**

*Separates data/metric issues from model capacity; provides performance floor.*

*Exercise 11.4 (Data Leakage).* Define data leakage and give two concrete examples.

**Hint:**

*Temporal leakage; using normalised stats computed on the full dataset.*

### Medium

*Exercise 11.5 (Hyperparameter Search Budget).* Given budget for 30 runs, propose an allocation between exploration (random search) and exploitation (local search). Defend your choice.

**Hint:**

*Start broad (e.g., 20 random), then refine top configurations (e.g., 10 local).*

*Exercise 11.6 (Early Stopping vs. Schedules).* Compare early stopping with cosine decay schedules under limited training budget.

**Hint:**

*Consider variance, bias, and checkpoint selection.*

## Hard

*Exercise 11.7 (Confidence Intervals).* Derive a 95% Wilson interval for a classifier with  $n$  samples and accuracy  $\hat{p}$ .

**Hint:**

*Use  $\frac{\hat{p} + z^2/(2n) \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1+z^2/n}$  with  $z \approx 1.96$ .*

*Exercise 11.8 (Causal Confounding).* Your model uses a spurious feature. Propose an experimental protocol to detect and mitigate it.

**Hint:**

*Counterfactual augmentation, environment splitting, invariant risk minimisation.*

*Exercise 11.9 (Advanced Topic 1).* Explain a key concept from this chapter and its practical applications.

**Hint:**

*Consider the theoretical foundations and real-world implications.*

*Exercise 11.10 (Advanced Topic 2).* Analyse the relationship between different techniques covered in this chapter.

**Hint:**

*Look for connections and trade-offs between methods.*

*Exercise 11.11 (Advanced Topic 3).* Design an experiment to test a hypothesis related to this chapter's content.

**Hint:**

*Consider experimental design, metrics, and potential confounding factors.*

*Exercise 11.12 (Advanced Topic 4).* Compare different approaches to solving a problem from this chapter.

**Hint:**

*Consider computational complexity, accuracy, and practical considerations.*

*Exercise 11.13 (Advanced Topic 5).* Derive a mathematical relationship or prove a theorem from this chapter.

**Hint:**

*Start with the definitions and work through the logical steps.*

*Exercise 11.14 (Advanced Topic 6).* Implement a practical solution to a problem discussed in this chapter.

**Hint:**

*Consider the implementation details and potential challenges.*

*Exercise 11.15 (Advanced Topic 7).* Evaluate the limitations and potential improvements of techniques from this chapter.

**Hint:**

*Consider both theoretical limitations and practical constraints.*



# Chapter 12

## Applications

This chapter showcases deep learning applications across various domains, demonstrating the breadth and impact of the field.

### Learning Objectives

1. Deep learning opportunities in vision, language, speech, and tabular domains
2. Model families and data representations for different application areas
3. Evaluation protocols and metrics for different task types
4. Common failure modes and deployment considerations

## Intuition

Applications differ not only in architectures but in *data assumptions and metrics*, requiring practitioners to think beyond model selection to understand the fundamental characteristics of each domain. A robust recipe for approaching any application is: clarify the task and loss function, understand the data distribution and constraints, then choose the simplest model that can plausibly meet the requirements before scaling up.

This systematic approach prevents the common pitfall of over-engineering solutions before understanding the problem's core requirements. For example, a computer vision task might require real-time inference on mobile devices, constraining the choice of architectures regardless of theoretical performance, while a natural language processing application might prioritize interpretability over raw accuracy for regulatory compliance. The key insight is that successful applications balance technical sophistication with practical constraints, often achieving better results through careful problem formulation and data understanding than through complex model architectures.

The applications in this chapter demonstrate how different domains require different approaches: computer vision emphasizes spatial relationships and translation invariance, natural language processing focuses on sequential patterns and semantic understanding, while tabular data applications often require careful feature engineering and handling of missing values. Each domain has its own evaluation metrics, failure modes, and deployment considerations that must be understood before selecting appropriate model families and training strategies.

Real-world success depends on recognizing that the "best" model is not always the most complex or highest-performing on benchmark datasets, but rather the one that best fits the specific constraints, requirements, and context of the intended application. This chapter provides the framework for making these critical decisions systematically across diverse application domains.

## 12.1 Computer Vision Applications ●

Computer vision has revolutionized how machines interpret visual information, enabling applications from autonomous vehicles to medical diagnosis that were impossible just a decade ago. These systems can now match or exceed human performance in many visual tasks, transforming industries and creating new possibilities for human-computer interaction.

### 12.1.1 Image Classification

Image classification assigns labels to images, serving as the foundation for most computer vision applications and catalyzing the widespread adoption of deep CNNs through benchmarks like ImageNet Krizhevsky, Sutskever, and Hinton [KSH12], He et al. [He+16], Goodfellow, Bengio, and Courville [GBC16a], and Prince [Pri23]. This technology powers everything from smartphone camera apps that automatically tag photos to industrial quality control systems that inspect products on assembly lines.

The ImageNet dataset with its 1000-class object recognition challenge became the standard benchmark that drove CNN development, enabling systems to distinguish between hundreds of object categories with human-level accuracy. This capability powers modern search engines that can find images by content, social media platforms that automatically suggest tags, and e-commerce sites that can identify products from photos uploaded by users. The dataset's scale and diversity made it possible to train models that generalize well to real-world scenarios, establishing transfer learning as a standard practice in computer vision. Fine-grained classification tackles the challenge of distinguishing between visually similar categories, such as identifying specific bird species from photographs or recognizing different car models on the road. These applications require models to focus on subtle distinguishing features rather than obvious differences, making them valuable for wildlife conservation efforts where researchers need to identify species from camera trap images, or for automotive applications where distinguishing between similar vehicle models is crucial for traffic analysis and autonomous driving systems.

Medical imaging applications use image classification to assist radiologists in diagnosing diseases from X-rays, CT scans, and MRI images, often detecting conditions that might be missed by human observers. These systems can identify pneumonia in chest X-rays, detect diabetic retinopathy in eye scans, and classify skin lesions for early cancer detection, providing second opinions that improve diagnostic accuracy and help address the shortage of specialist radiologists in many regions. The technology is particularly valuable in telemedicine applications where expert radiologists are not locally available.

Modern image classification architectures typically use a CNN backbone such as ResNet or EfficientNet combined with a classification head that outputs probability distributions over the target classes. Transfer learning from pretrained weights has become standard practice, allowing practitioners to leverage models trained on large datasets like ImageNet and fine-tune them for specific applications with relatively small amounts of domain-specific data. This approach dramatically reduces training time and computational requirements while often achieving better performance than training from scratch, making advanced computer vision capabilities accessible to organizations with limited resources.

### 12.1.2 Object Detection

Object detection locates and classifies multiple objects within images, representing a significant advancement over simple classification by providing spatial information about where objects are located. Real-time variants like YOLO emphasize speed for applications requiring immediate responses, while two-stage models like Faster R-CNN prioritize accuracy for applications where precision is more important than speed.

Autonomous driving systems rely heavily on object detection to identify pedestrians, vehicles, traffic signs, and other road users in real-time, enabling self-driving cars to navigate safely through complex traffic environments. These systems must process video streams at high frame rates while maintaining accuracy, as any missed detection could lead to catastrophic accidents. The technology also powers advanced driver assistance systems (ADAS) that warn drivers about potential hazards, automatically apply brakes in emergency situations, and help with parking by detecting obstacles around the vehicle.

Surveillance and security applications use object detection for person detection and tracking in crowded environments like airports, shopping malls, and public transportation hubs. These systems can identify suspicious behavior patterns, track individuals across multiple camera feeds, and provide real-time alerts to security personnel. The technology is also used in smart cities to monitor traffic flow, detect accidents, and optimize traffic light timing based on real-time vehicle and pedestrian counts.

Retail applications leverage object detection for product recognition and inventory management, enabling cashierless stores where customers can simply pick up items and walk out without traditional checkout processes. These systems can identify thousands of different products in real-time, track customer behavior for analytics, and help with loss prevention by detecting shoplifting attempts. The technology also powers recommendation systems that can suggest products based on what customers are looking at in physical stores.

Modern object detection methods include YOLO for real-time applications, Faster R-CNN for high-accuracy scenarios, and RetinaNet for handling class imbalance in dense detection scenarios. The focal loss function was specifically designed to address the class imbalance problem in dense detectors, where background pixels vastly outnumber object pixels, making it easier to train models that can detect small or rare objects effectively. These methods have enabled object detection to become practical for real-world applications across diverse industries.

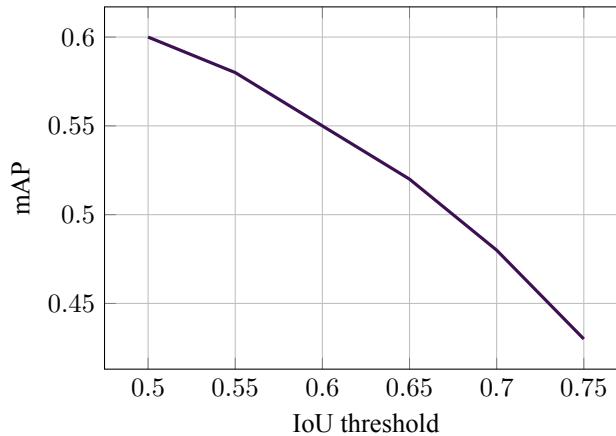


Figure 12.1: Detector performance (mAP) vs. IoU threshold schematic.

### 12.1.3 Semantic Segmentation

Semantic segmentation represents the most detailed level of computer vision analysis, classifying every pixel in an image to create dense, pixel-level understanding of visual scenes. This technology enables applications that require precise spatial understanding, from autonomous vehicles that need to distinguish between drivable road surfaces and sidewalks to medical imaging systems that must identify exact boundaries of tumors and organs.

Autonomous driving applications use semantic segmentation to create detailed maps of the driving environment, distinguishing between road surfaces, sidewalks, vehicles, pedestrians, and other objects with pixel-level precision. This information is crucial for path planning algorithms that must navigate safely through complex urban environments, avoiding obstacles while staying within designated lanes. The technology also enables advanced features like automatic lane-keeping, adaptive cruise control, and collision avoidance systems that can respond to subtle changes in the driving environment.

Medical imaging applications leverage semantic segmentation for precise tumor segmentation and organ delineation, enabling radiologists to identify exact boundaries of cancerous tissues and healthy organs with unprecedented accuracy. These systems can assist in surgical planning by providing detailed 3D reconstructions of patient anatomy, help with radiation therapy planning by identifying healthy tissues that must be protected, and enable automated measurement of tumor size and progression over time. The technology is particularly valuable for early cancer detection where precise boundary identification can mean the difference between successful treatment and disease progression.

Satellite imagery analysis uses semantic segmentation for land use classification, creating detailed maps of urban development, agricultural areas, forests, and water bodies that are essential for environmental monitoring and urban planning. These systems can track deforestation, monitor crop health, assess damage from natural disasters, and provide data for climate change research. The technology enables automated analysis of vast areas that would be impossible to survey manually, providing valuable insights for government agencies, environmental organizations, and commercial applications like precision agriculture.

Modern semantic segmentation architectures include U-Net with its encoder-decoder structure and skip connections that preserve fine-grained details, DeepLab with its atrous convolutions for multi-scale feature extraction, and Mask R-CNN that

combines object detection with instance segmentation. These architectures have been specifically designed to handle the challenges of dense prediction tasks, including the need to preserve spatial resolution, handle multi-scale objects, and maintain computational efficiency for real-world deployment. The success of these methods has enabled semantic segmentation to become practical for applications requiring pixel-level understanding across diverse domains.

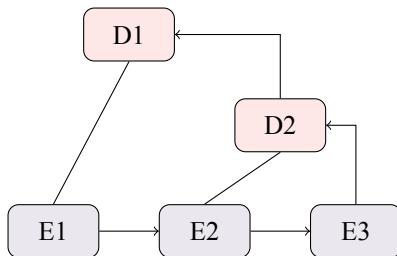


Figure 12.2: U-Net style encoder-decoder with skip connections.

#### 12.1.4 Face Recognition

Face recognition technology has become one of the most widely deployed computer vision applications, enabling secure and convenient identification of individuals across diverse scenarios. This technology combines face detection, feature extraction, and similarity matching to create robust systems that can identify people with high accuracy while maintaining user privacy and security.

Security and access control applications use face recognition to replace traditional key cards and passwords with biometric authentication that is both more secure and more convenient. These systems are deployed in corporate offices, government buildings, and residential complexes to control access to restricted areas, automatically unlock doors for authorized personnel, and maintain security logs of who accessed which areas and when. The technology is also used in border control and immigration systems to verify traveler identities and detect individuals on watch lists, significantly improving security while reducing wait times at airports and border crossings.

Photo organization applications leverage face recognition to automatically tag and organize personal photo collections, making it easy to find pictures of specific people across years of memories. Social media platforms use this technology to suggest tags for photos, helping users identify friends and family members in group pictures, while cloud storage services can automatically create albums based on the people present in photos. The technology also powers photo editing applications that can automatically enhance portraits, remove red-eye, and apply filters based on facial features.

Payment authentication systems use face recognition as a biometric security measure for mobile payments, online banking, and cryptocurrency transactions, providing a more secure alternative to passwords and PINs. These systems must balance security with usability, requiring extremely low false acceptance rates to prevent unauthorized access while maintaining high user convenience. The technology is also used in retail environments for cashierless checkout systems, where customers can make purchases simply by looking at a camera, eliminating the need for physical payment methods.

Modern face recognition systems typically use a three-stage approach: face detection to locate faces in images, feature extraction using deep learning models like FaceNet or ArcFace to create compact embeddings that capture facial characteristics, and similarity matching to compare embeddings and determine identity. These systems are evaluated using metrics like ROC curves and precision-recall curves, with particular attention to false acceptance rates (FAR) and false rejection rates (FRR) at specific operating points. The choice of operating point depends on the application requirements, with security-critical applications prioritizing low false acceptance rates while convenience-focused applications may tolerate slightly higher error rates for better user experience.

#### 12.1.5 Image Generation and Manipulation

Image generation and manipulation represent the creative frontier of computer vision, enabling applications that can create, enhance, and modify visual content in ways that were previously impossible. These technologies have applications ranging from

entertainment and social media to professional content creation and scientific visualization, though they also raise important ethical considerations about authenticity and misuse.

Style transfer applications allow users to apply artistic styles from famous paintings to their own photographs, creating unique digital artwork that combines the content of one image with the style of another. This technology powers popular mobile apps that let users transform their photos into paintings in the style of Van Gogh, Picasso, or other artists, making high-quality artistic effects accessible to everyone. Professional photographers and graphic designers use these tools to create unique visual effects for marketing materials, social media content, and artistic projects, while researchers use the technology to study the relationship between artistic style and visual perception.

Super-resolution technology enhances image quality by increasing resolution and removing artifacts, making it possible to create high-quality images from low-resolution sources. This technology is used in medical imaging to improve the quality of scans for better diagnosis, in satellite imagery to create detailed maps from lower-resolution satellite data, and in entertainment to upscale classic films and television shows to modern high-definition standards. The technology is also valuable for forensic applications where investigators need to enhance surveillance footage to identify suspects or read license plates from distant cameras.

Inpainting technology fills missing or damaged regions in images, automatically reconstructing plausible content based on surrounding areas. This capability is used in photo editing software to remove unwanted objects from images, restore damaged historical photographs, and fill in missing areas in panoramic photos. Professional photographers use these tools to clean up images by removing distracting elements, while archivists use the technology to restore damaged historical documents and photographs. The technology is also valuable for creating seamless image composites and removing watermarks or other unwanted elements from images.

Deepfake technology, while raising significant ethical concerns, has legitimate applications in entertainment, education, and accessibility. In entertainment, it can be used to create realistic visual effects for movies and video games, or to allow actors to perform in multiple languages without dubbing. Educational applications include creating historical reenactments or language learning content with native speakers. However, the technology's potential for creating convincing fake videos has led to concerns about misinformation, privacy violations, and the erosion of trust in visual media, highlighting the need for robust detection methods and ethical guidelines for its use.

### 12.1.6 Historical context and references

Modern computer vision breakthroughs stem from the revolutionary ImageNet-scale training that popularized CNNs Krizhevsky, Sutskever, and Hinton [KSH12], demonstrating that deep learning could achieve superhuman performance on complex visual recognition tasks. The introduction of deeper residual networks He et al. [He+16] solved the vanishing gradient problem that had limited network depth, enabling the training of much deeper architectures that could learn more complex visual representations. Specialized architectures for segmentation like U-Net Ronneberger, Fischer, and Brox [RFB15] introduced encoder-decoder structures with skip connections that became the foundation for dense prediction tasks, enabling pixel-level understanding of images.

The ImageNet competition catalyzed a decade of rapid progress in computer vision, with each year bringing new architectural innovations that pushed the boundaries of what was possible. The transition from hand-crafted features to learned representations marked a paradigm shift that extended far beyond image classification, influencing object detection, semantic segmentation, and even natural language processing. These advances have enabled real-world applications that were previously impossible, from autonomous vehicles that can navigate complex urban environments to medical imaging systems that can detect diseases with superhuman accuracy.

The success of computer vision has also driven advances in related fields, with techniques developed for image understanding influencing natural language processing, robotics, and even scientific discovery. The availability of large-scale datasets, powerful computing resources, and open-source frameworks has democratized access to these technologies, enabling researchers and practitioners worldwide to contribute to the field's continued advancement. See Goodfellow, Bengio, and Courville [GBC16a] and Prince [Pri23] for broader context on the theoretical foundations and practical applications of these transformative technologies.

## 12.2 Natural Language Processing ●

Natural Language Processing represents one of the most widely adopted applications of deep learning, transforming how humans interact with computers through text and speech. From search engines that understand natural language queries to virtual assistants that can engage in meaningful conversations, NLP has become an integral part of modern digital experiences. The field has experienced revolutionary advances with the introduction of transformer architectures and large language models, enabling applications that were previously impossible, such as real-time translation between hundreds of languages and AI systems that can generate human-like text. These technologies power everything from customer service chatbots and content recommendation systems to advanced research tools and educational platforms, making NLP one of the most impactful areas of artificial intelligence in terms of daily user interaction and business value.

### 12.2.1 Text Classification

Text classification serves as the foundation for many NLP applications, automatically categorizing documents and messages to enable intelligent routing, filtering, and analysis. This technology has become essential for managing the vast amounts of textual data generated daily across digital platforms, from social media posts and customer feedback to legal documents and scientific papers.

Sentiment analysis applications analyze the emotional tone and opinion expressed in text, enabling businesses to monitor customer satisfaction, track brand perception, and respond to public sentiment in real-time. Social media platforms use this technology to detect hate speech and harmful content, while e-commerce sites analyze product reviews to provide sentiment-based recommendations and identify trending items. Financial institutions employ sentiment analysis to gauge market sentiment from news articles and social media, helping traders make informed investment decisions. The technology is also used in political campaigns to understand public opinion and in healthcare to monitor patient satisfaction and mental health indicators from text communications.

Spam detection systems protect users from unwanted and potentially harmful messages by automatically identifying and filtering spam emails, phishing attempts, and malicious content. Email providers use sophisticated NLP models to achieve near-perfect spam detection rates while minimizing false positives that might block legitimate messages. Social media platforms employ similar technology to detect and remove spam accounts, fake reviews, and malicious content that could harm users or manipulate public opinion. The technology has evolved to handle increasingly sophisticated spam techniques, including AI-generated content and context-aware attacks that traditional rule-based systems cannot detect.

Topic classification systems automatically organize large collections of documents by subject matter, enabling efficient information retrieval and content management. News organizations use this technology to automatically categorize articles by topic, making it easier for readers to find relevant content and for editors to manage their content libraries. Academic institutions employ topic classification to organize research papers and help researchers discover relevant studies in their fields. Legal firms use the technology to categorize case documents and contracts, while government agencies apply it to process and organize public records and regulatory documents.

Modern text classification systems typically use pretrained transformer models like BERT, RoBERTa, and DistilBERT, which can be fine-tuned for specific tasks with relatively small amounts of domain-specific data. These models achieve state-of-the-art performance across diverse classification tasks while being computationally efficient enough for real-time applications.

Evaluation metrics include accuracy and F1 scores for general applications, while risk-sensitive domains like healthcare and finance require additional attention to model calibration to ensure that confidence scores accurately reflect prediction reliability. The success of these models has made advanced text classification capabilities accessible to organizations of all sizes, democratizing access to sophisticated NLP technology.

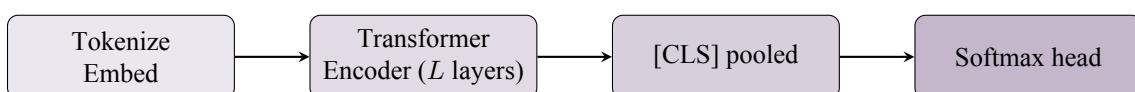


Figure 12.3: Transformer fine-tuning for text classification.

## 12.2.2 Machine Translation

Machine translation represents one of the most transformative applications of NLP, breaking down language barriers and enabling global communication on an unprecedented scale. The technology has evolved from rule-based systems to statistical methods and finally to neural approaches that can translate between hundreds of language pairs with remarkable accuracy. Modern translation systems can handle complex linguistic phenomena, cultural nuances, and domain-specific terminology, making them essential tools for international business, education, and diplomacy. The introduction of attention mechanisms and transformer architectures revolutionized the field, enabling translations that often rival human quality while being fast enough for real-time applications.

Google Translate and DeepL represent the most widely used commercial translation services, processing billions of translation requests daily across diverse applications. Google Translate supports over 100 languages and is integrated into search results, email, and mobile apps, enabling users to understand content in foreign languages instantly. DeepL has gained recognition for its high-quality translations, particularly for European languages, and is widely used by professionals who require accurate translations for business documents and academic papers. These services power everything from tourist apps that translate signs and menus to e-commerce platforms that automatically translate product descriptions for international markets.

Sequence-to-sequence models with attention mechanisms revolutionized machine translation by enabling the model to focus on relevant parts of the source sentence when generating each word of the translation. This approach solved the bottleneck problem of earlier encoder-decoder architectures, allowing the model to handle long sentences and complex linguistic structures effectively. The attention mechanism enables the model to learn alignment between source and target languages, making translations more accurate and contextually appropriate. These models are particularly effective for languages with different word orders and grammatical structures, enabling high-quality translation between linguistically distant language pairs.

Transformer models have become the standard architecture for machine translation, offering superior performance and efficiency compared to earlier recurrent neural network approaches. These models can be trained on massive parallel corpora, learning to translate between multiple language pairs simultaneously while sharing knowledge across languages. The self-attention mechanism allows the model to capture long-range dependencies and complex linguistic patterns that are essential for accurate translation. Modern transformer-based translation systems can handle specialized domains like legal, medical, and technical texts, making them valuable tools for professional translators and international organizations.

Modern machine translation architectures use encoder-decoder transformers with subword tokenization to handle the vocabulary size and morphological complexity of different languages. The encoder processes the source language text to create contextual representations, while the decoder generates the target language text using cross-attention to focus on relevant source information. Subword tokenization enables the model to handle rare words and morphological variations effectively, improving translation quality for languages with rich morphology. Evaluation uses metrics like BLEU and chrf for automated assessment, combined with human evaluation to measure translation quality, fluency, and cultural appropriateness. These systems have achieved near-human performance on many language pairs, making real-time translation practical for applications like video conferencing, live streaming, and international communication.

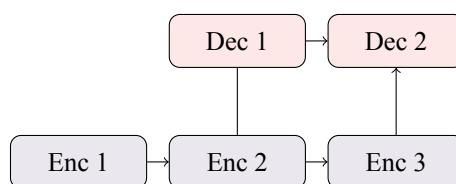


Figure 12.4: Encoder – decoder Transformer schematic with cross-attention.

## 12.2.3 Question Answering

Question answering systems represent one of the most challenging and valuable applications of NLP, enabling users to obtain specific information by asking natural language questions. These systems must understand the semantic meaning of questions, locate relevant information from vast knowledge sources, and generate accurate, coherent answers that directly address the user's query. The technology has evolved from simple keyword matching to sophisticated neural models that can reason about

complex relationships and provide nuanced answers to diverse question types. Modern QA systems are essential components of search engines, virtual assistants, and educational platforms, transforming how people access and interact with information. Extractive question answering systems find the answer span within a given text passage, making them particularly effective for applications where the answer exists verbatim in the source material. The SQuAD dataset has become the standard benchmark for this task, enabling systems to answer questions like "What is the capital of France?" by locating the relevant text span. These systems are widely used in customer service applications where agents need to quickly find answers from knowledge bases, in legal research where lawyers search through case documents, and in educational platforms where students can ask questions about course materials. The technology is also valuable for fact-checking applications that need to verify claims against reliable sources and for research tools that help scientists find relevant information in academic papers.

Open-domain question answering systems can answer questions by searching through large corpora of text, including the entire internet, making them incredibly powerful tools for information retrieval. These systems combine information retrieval techniques with reading comprehension models to first find relevant documents and then extract or generate answers from them. They power modern search engines that can provide direct answers to questions rather than just returning lists of relevant web pages. The technology is used in virtual assistants like Siri and Alexa to answer general knowledge questions, in educational platforms to help students with homework and research, and in professional tools that help researchers and analysts find information quickly. These systems are particularly valuable for applications that require up-to-date information, as they can access the latest content from the web.

Visual question answering systems combine computer vision and natural language processing to answer questions about images, enabling applications that can understand and describe visual content. These systems can answer questions like "What color is the car?" or "How many people are in the image?" by analyzing both the visual content and the linguistic structure of the question. The technology is used in accessibility applications that help visually impaired users understand images, in educational tools that can answer questions about diagrams and illustrations, and in content moderation systems that can understand the context of images to detect inappropriate content. These systems are also valuable for medical imaging applications where doctors can ask questions about X-rays or scans to get AI-assisted diagnostic insights.

#### 12.2.4 Language Models and Text Generation

Large language models represent the most transformative development in NLP, enabling AI systems to generate human-like text across diverse applications and domains. These models have revolutionized how we interact with computers, enabling natural language interfaces that can understand context, maintain conversation flow, and provide helpful responses to complex queries. The technology has achieved remarkable capabilities in text generation, from creative writing and technical documentation to code generation and educational content creation. However, the power of these models also raises important considerations about safety, accuracy, and responsible deployment, requiring careful evaluation and human oversight to ensure they provide reliable and beneficial assistance.

GPT models have become the most widely used large language models for general text generation, with ChatGPT/OpenAI holding approximately 80

Code generation applications like GitHub Copilot, Gemini, Revo, and Codex have transformed software development by providing AI-powered coding assistance that can understand natural language descriptions and generate corresponding code. These tools can suggest function implementations, debug code, write tests, and even generate entire applications from high-level specifications. Developers use these systems to accelerate their workflow, learn new programming languages, and explore different approaches to solving problems. The technology is particularly valuable for educational applications where students can learn programming concepts through interactive AI assistance, and for professional development where engineers can quickly prototype ideas and explore new technologies. These systems have become essential tools for modern software development, enabling faster iteration and more creative problem-solving.

Chatbots and conversational AI systems have become ubiquitous in customer service, education, and entertainment applications, providing 24/7 assistance that can handle a wide range of user queries. These systems power virtual assistants that can help with scheduling, information retrieval, and task automation, making them valuable tools for personal productivity and professional efficiency. Educational chatbots can provide personalized tutoring and answer student questions, while entertainment chatbots can engage users in creative conversations and interactive storytelling. The technology is also used in mental health applications

to provide supportive conversations and in language learning platforms to practice conversational skills with AI partners. Content creation applications leverage large language models to generate articles, reports, marketing copy, and creative content across diverse industries. News organizations use these systems to generate initial drafts of articles and summaries, while marketing agencies employ them to create compelling copy for advertisements and social media campaigns. The technology is used in legal applications to draft contracts and legal documents, in academic settings to generate research proposals and grant applications, and in creative industries to develop story ideas and character descriptions. These systems can adapt their writing style to different audiences and purposes, making them valuable tools for content creators who need to produce large volumes of high-quality text efficiently.

### 12.2.5 Named Entity Recognition

Named Entity Recognition represents a fundamental task in NLP that identifies and classifies specific entities within text, enabling systems to understand the key information and relationships present in documents. This technology is essential for information extraction, knowledge graph construction, and semantic understanding of text, making it a critical component of many advanced NLP applications. The task involves identifying entities such as people, organizations, locations, dates, and technical terms, then classifying them into predefined categories. Modern NER systems use sophisticated neural architectures with BIO tagging schemes and conditional random field (CRF) heads to achieve high accuracy across diverse domains and languages.

Person, organization, and location name recognition enables systems to identify the key actors and places mentioned in text, making it valuable for applications that need to understand the who, what, and where of information. News organizations use this technology to automatically tag articles with relevant people, companies, and locations, making it easier for readers to find related content and for editors to organize their archives. Social media platforms employ NER to identify mentions of celebrities, brands, and places, enabling targeted advertising and content recommendation. The technology is also used in legal applications to identify parties involved in cases, in financial services to track company mentions and market sentiment, and in intelligence applications to extract information about individuals and organizations from large text corpora.

Date, quantity, and technical term recognition enables systems to identify temporal information, numerical data, and domain-specific terminology that is crucial for understanding the context and significance of information. Financial institutions use this technology to extract dates, monetary amounts, and financial terms from documents to automate data entry and analysis. Medical applications employ NER to identify drug names, dosages, and medical conditions from patient records, enabling automated coding and analysis. The technology is used in scientific literature analysis to identify chemical compounds, measurements, and technical concepts, helping researchers discover relevant studies and track scientific progress. Legal applications use NER to identify case numbers, dates, and legal terminology from court documents, enabling automated case analysis and precedent identification.

Information extraction applications leverage NER to automatically extract structured information from unstructured text, enabling the creation of knowledge bases and databases from large text collections. These systems can identify relationships between entities, extract facts and events, and organize information in ways that make it easily searchable and analyzable. The technology is used in business intelligence applications to extract information about competitors, markets, and trends from news articles and reports. Academic institutions employ NER to extract information from research papers, enabling automated literature reviews and knowledge discovery. Government agencies use the technology to process and organize public records, legal documents, and regulatory filings, making information more accessible to citizens and researchers. The combination of NER with other NLP techniques enables the creation of sophisticated information extraction systems that can understand complex relationships and extract nuanced information from text.

### 12.2.6 Historical context and references

The transformer architecture introduced by Vaswani et al. Vaswani et al. [Vas+17] revolutionized NLP by replacing recurrent neural networks with self-attention mechanisms, enabling parallel processing and capturing long-range dependencies more effectively. This breakthrough led to the development of BERT Devlin et al. [Dev+18], which demonstrated the power of bidirectional context understanding and established the paradigm of pretraining followed by fine-tuning that became standard practice across NLP applications. The introduction of GPT models Radford et al. [Rad+19] showed that autoregressive language

modeling could achieve remarkable performance on diverse tasks, paving the way for the large language models that dominate the field today.

The success of these models has transformed NLP from a specialized field requiring domain expertise to a widely accessible technology that powers everyday applications. The availability of pretrained models has democratized access to state-of-the-art NLP capabilities, enabling researchers and practitioners to achieve high performance on specific tasks with minimal training data. This shift has accelerated innovation across industries, from healthcare and finance to education and entertainment, as organizations can now integrate sophisticated language understanding into their products and services.

The impact of these advances extends far beyond academic research, with transformer-based models now powering search engines, translation services, virtual assistants, and content generation systems used by billions of people worldwide. The field continues to evolve rapidly, with new architectures, training methods, and applications emerging regularly, making NLP one of the most dynamic and impactful areas of artificial intelligence. See Goodfellow, Bengio, and Courville [GBC16a], Prince [Pri23], and Zhang et al. [Zha+24a] for broader context and comprehensive tutorials on the theoretical foundations and practical applications of these transformative technologies.

## 12.3 Speech Recognition and Synthesis

Speech recognition and synthesis represent one of the most transformative applications of deep learning, enabling natural human-computer interaction through voice interfaces. These technologies have revolutionized how we interact with devices, from smartphones and smart speakers to automotive systems and accessibility tools. The field has evolved from rule-based systems to end-to-end neural architectures that can understand and generate human speech with remarkable accuracy and naturalness. Modern speech systems power virtual assistants, real-time translation services, and assistive technologies that improve accessibility for millions of users worldwide. The integration of self-supervised learning and transformer architectures has further enhanced the robustness and multilingual capabilities of these systems, making them essential components of modern AI applications.

### 12.3.1 Automatic Speech Recognition (ASR)

Automatic Speech Recognition (ASR) converts spoken language into text, serving as the foundation for voice-controlled applications and accessibility tools. Modern ASR systems use deep neural networks to process audio signals, extracting meaningful text from speech with high accuracy across diverse languages and accents. The technology has evolved from simple command recognition to sophisticated systems that can handle continuous speech, multiple speakers, and noisy environments. Self-supervised pretraining techniques like wav2vec have significantly reduced the need for large amounts of labeled data, making ASR more accessible and cost-effective. These systems are now integrated into everyday applications, from virtual assistants and transcription services to real-time translation and voice commands in smart devices.

Virtual assistants like Siri, Alexa, and Google Assistant have become ubiquitous in modern life, providing hands-free access to information, entertainment, and smart home control. These systems use advanced ASR to understand natural language commands, enabling users to set reminders, play music, control lighting, and access weather information through simple voice interactions. The technology has transformed how we interact with technology, making it more accessible to users with mobility limitations and providing convenience in situations where hands-free operation is essential. These assistants continue to evolve, incorporating more sophisticated language understanding and context awareness to provide increasingly natural and helpful interactions.

Transcription services have revolutionized documentation and accessibility, converting spoken content into text for meetings, lectures, interviews, and media content. Professional transcription services use advanced ASR to provide accurate, real-time transcription for legal proceedings, medical consultations, and educational content. The technology has made information more accessible to deaf and hard-of-hearing individuals, enabling them to participate fully in conversations and access audio content. Real-time transcription services are now integrated into video conferencing platforms, providing live captions for remote meetings and webinars, enhancing accessibility and comprehension for all participants.

Voice commands have transformed user interfaces across multiple domains, enabling hands-free control of devices and applications. In automotive systems, voice commands allow drivers to control navigation, make phone calls, and adjust settings

without taking their hands off the wheel, improving safety and convenience. Smart home systems use voice commands to control lighting, temperature, security systems, and entertainment devices, creating seamless and intuitive living environments. The technology has also found applications in healthcare, where voice commands enable hands-free operation of medical equipment, reducing the risk of contamination and improving workflow efficiency in sterile environments.

The architectures used in modern ASR systems have evolved significantly, with CTC-based models providing efficient training for sequence-to-sequence tasks by allowing flexible alignment between audio and text. Listen-Attend-Spell architectures use attention mechanisms to focus on relevant parts of the audio signal, improving accuracy for long sequences and complex utterances. Transducer models combine the benefits of CTC and attention-based approaches, providing both efficiency and accuracy for real-time applications. Self-supervised pretraining with models like wav2vec has revolutionized the field by learning rich audio representations from unlabeled data, reducing the need for expensive labeled datasets and improving performance on low-resource languages and domains.

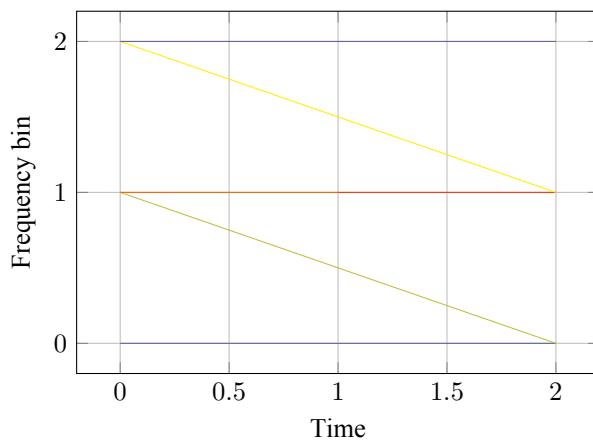


Figure 12.5: Schematic spectrogram input for ASR.

### 12.3.2 Text-to-Speech (TTS)

Text-to-Speech (TTS) technology converts written text into natural-sounding speech, enabling applications that require audio output from text content. Modern TTS systems use neural vocoders and advanced synthesis techniques to generate speech that is virtually indistinguishable from human speech in terms of naturalness and expressiveness. The technology has evolved from robotic-sounding concatenative synthesis to sophisticated neural approaches that can capture the nuances of human speech, including emotion, intonation, and prosody. TTS systems are now integrated into a wide range of applications, from accessibility tools and virtual assistants to entertainment and educational content, making information more accessible and engaging for users with visual impairments or learning preferences.

WaveNet and Tacotron represent breakthrough architectures in neural TTS, using deep learning to generate high-quality speech from text input. WaveNet uses dilated convolutions to model the temporal dependencies in speech, producing more natural-sounding audio with better prosody and expressiveness. Tacotron combines sequence-to-sequence learning with attention mechanisms to generate mel-spectrograms from text, which are then converted to audio using neural vocoders. These architectures have enabled the development of voice synthesis systems that can mimic specific speakers, create custom voices for applications, and generate speech in multiple languages with native-like pronunciation. The technology has found applications in audiobook production, virtual assistants, and personalized voice interfaces.

Voice cloning technology has emerged as a powerful application of TTS, enabling the creation of synthetic voices that closely mimic specific individuals. This technology has applications in entertainment, where it can be used to create voiceovers for characters or restore the voices of actors who have passed away. In accessibility, voice cloning can help individuals with speech impairments create personalized synthetic voices that sound more natural and expressive than traditional text-to-speech systems. The technology has also found applications in language learning, where learners can practice pronunciation with native speaker voices, and in customer service, where synthetic voices can provide consistent and professional interactions. However, the technology also raises important ethical considerations regarding consent, privacy, and the potential for misuse in creating

deepfake audio content.

Accessibility tools powered by TTS technology have transformed the lives of individuals with visual impairments, learning disabilities, and other conditions that affect reading ability. Screen readers use TTS to convert digital text into speech, enabling blind and visually impaired users to access websites, documents, and applications independently. Educational TTS systems can read textbooks, articles, and other learning materials aloud, supporting students with dyslexia and other reading difficulties. The technology has also enabled the creation of audiobooks and spoken content for entertainment and education, making information more accessible to diverse audiences. These tools continue to evolve, incorporating more natural-sounding voices, better pronunciation of technical terms, and support for multiple languages and accents.

### 12.3.3 Speaker Recognition

Speaker recognition technology identifies and verifies individuals based on their unique vocal characteristics, providing secure and convenient authentication methods. The technology analyzes various acoustic features of speech, including pitch, formants, and spectral characteristics, to create unique voiceprints for each individual. Modern speaker recognition systems use deep learning approaches to extract robust features that are resistant to noise, aging, and other factors that might affect voice quality. The technology has applications in security, authentication, and personalization, enabling hands-free access to devices and services. Speaker recognition is also used in forensic applications, where it can help identify individuals from audio recordings in criminal investigations.

Voice biometrics use speaker recognition technology to provide secure authentication for banking, healthcare, and other sensitive applications. Banks and financial institutions use voice biometrics to verify customer identity during phone calls, reducing fraud and improving security while maintaining convenience. Healthcare systems use voice authentication to secure access to patient records and medical devices, ensuring that only authorized personnel can access sensitive information. The technology has also found applications in smart home security, where it can recognize authorized users and grant access to restricted areas or functions. Voice biometrics offer several advantages over traditional authentication methods, including hands-free operation, resistance to spoofing attacks, and the ability to work across different devices and platforms.

Speaker diarization technology identifies who spoke when in multi-speaker conversations, providing valuable insights for meeting analysis, transcription, and content organization. The technology uses machine learning algorithms to segment audio recordings by speaker, enabling automatic identification of different participants in conversations. This capability is essential for accurate transcription of meetings, interviews, and other multi-speaker events, where it's important to know which person said what. Speaker diarization is also used in call center analytics, where it can help identify customer service representatives and customers in recorded calls for quality assurance and training purposes. The technology has applications in media production, where it can automatically identify and label different speakers in podcasts, interviews, and other audio content.

### 12.3.4 Historical context and references

The development of speech recognition and synthesis has been marked by several key breakthroughs that have transformed the field from rule-based systems to sophisticated neural architectures. The introduction of Hidden Markov Models (HMMs) in the 1970s provided the first statistical approach to speech recognition, enabling systems to learn from data rather than relying on hand-crafted rules. The development of Mel-Frequency Cepstral Coefficients (MFCCs) in the 1980s provided robust feature extraction methods that became the standard for speech processing systems. The introduction of neural networks in the 1990s marked a significant shift, with systems like TIMIT demonstrating the potential of deep learning for speech recognition tasks. The 2000s saw the development of more sophisticated neural architectures, including Long Short-Term Memory (LSTM) networks and attention mechanisms that improved the ability to handle long sequences and complex speech patterns. The introduction of Connectionist Temporal Classification (CTC) in 2006 provided an efficient method for training sequence-to-sequence models without requiring frame-level alignments, making it easier to train ASR systems on large datasets. The development of WaveNet in 2016 marked a breakthrough in neural TTS, demonstrating that deep learning could generate high-quality speech that was virtually indistinguishable from human speech. The introduction of transformer architectures in 2017 further revolutionized the field, enabling more efficient processing of sequential data and better capture of long-range dependencies in speech.

The 2020s have seen the integration of self-supervised learning and large-scale pretraining, with models like wav2vec and Whisper demonstrating the power of learning from unlabeled audio data. These developments have made speech technology more accessible and cost-effective, enabling the development of systems that can work across multiple languages and domains with minimal labeled data. The field continues to evolve rapidly, with new architectures, training methods, and applications emerging regularly, making speech recognition and synthesis one of the most dynamic and impactful areas of artificial intelligence. See Prince [Pri23] for broader context and comprehensive tutorials on the theoretical foundations and practical applications of these transformative technologies.

## 12.4 Healthcare and Medical Imaging ●

Healthcare and medical imaging represent one of the most impactful applications of deep learning, where AI technologies are directly improving patient outcomes and advancing medical research. These applications have transformed diagnostic accuracy, treatment planning, and drug discovery processes, enabling healthcare professionals to make more informed decisions based on comprehensive data analysis. Deep learning models can analyze medical images with superhuman accuracy, detecting subtle patterns and anomalies that might be missed by human observers, while also processing vast amounts of genomic and clinical data to identify disease patterns and treatment responses. The integration of AI in healthcare has led to significant improvements in early disease detection, personalized medicine, and clinical workflow efficiency, ultimately saving lives and reducing healthcare costs. However, these applications also require careful consideration of regulatory compliance, data privacy, and model interpretability to ensure safe and effective deployment in clinical settings.

### 12.4.1 Medical Image Analysis

Medical image analysis has revolutionized diagnostic medicine by enabling automated detection and analysis of diseases from various imaging modalities. Deep learning models can process medical images with remarkable accuracy, identifying patterns and anomalies that may be difficult for human observers to detect consistently. These systems are particularly valuable for screening applications, where they can analyze large volumes of images to identify patients who may require further examination or treatment. The technology has been successfully deployed across multiple medical specialties, from radiology and pathology to ophthalmology and dermatology, improving diagnostic accuracy and reducing the time required for image interpretation. Modern medical imaging AI systems often incorporate explainability features to help clinicians understand the reasoning behind AI recommendations, fostering trust and enabling effective human-AI collaboration in clinical decision-making.

Cancer detection in mammograms and CT scans has been transformed by deep learning algorithms that can identify malignant lesions with accuracy comparable to or exceeding that of experienced radiologists. These systems are particularly valuable for breast cancer screening programs, where they can help radiologists identify suspicious areas in mammograms and reduce the number of missed cancers. In CT scans, AI models can detect lung nodules, liver lesions, and other abnormalities that may indicate cancer, enabling earlier diagnosis and treatment. The technology has been integrated into clinical workflows at major medical centers, where it assists radiologists in prioritizing cases and identifying patients who may need immediate attention. These systems continue to evolve, with newer models incorporating multi-modal data and longitudinal imaging to provide more comprehensive cancer detection and monitoring capabilities.

Diabetic retinopathy detection from retinal images has become a standard application of AI in ophthalmology, enabling automated screening for this common complication of diabetes. AI systems can analyze retinal photographs to identify signs of diabetic retinopathy, including microaneurysms, hemorrhages, and other pathological changes that may indicate the need for treatment. This technology has been particularly valuable in resource-limited settings, where access to specialist ophthalmologists may be limited, enabling primary care providers to screen patients for diabetic retinopathy and refer those who need specialist care. The technology has been deployed in screening programs worldwide, helping to identify patients at risk of vision loss and enabling timely intervention to prevent blindness. These systems have demonstrated high accuracy in clinical trials and are now being integrated into routine diabetes care protocols.

Pneumonia detection from chest X-rays has been significantly improved by deep learning models that can identify signs of pneumonia and other respiratory conditions with high accuracy. These systems are particularly valuable in emergency departments and intensive care units, where rapid diagnosis of respiratory conditions is critical for patient outcomes. AI models

can detect various types of pneumonia, including bacterial, viral, and fungal pneumonia, and can also identify other respiratory conditions such as tuberculosis and COVID-19. The technology has been deployed in hospitals worldwide, where it assists radiologists in prioritizing cases and identifying patients who may need immediate treatment. These systems have been particularly valuable during the COVID-19 pandemic, where they have helped healthcare providers quickly identify patients with pneumonia and other respiratory complications.

Tumor boundary delineation is a critical application of deep learning in medical imaging, where AI models can accurately identify and segment tumor boundaries in various imaging modalities. This technology is essential for treatment planning, where precise tumor boundaries are needed to determine the optimal radiation therapy dose and surgical approach. AI models can segment tumors in brain, lung, liver, and other organs with high accuracy, providing clinicians with detailed information about tumor size, shape, and location. The technology has been integrated into radiation therapy planning systems, where it helps oncologists design treatment plans that maximize tumor coverage while minimizing damage to healthy tissue. These systems continue to evolve, with newer models incorporating multi-modal imaging and longitudinal data to provide more comprehensive tumor analysis and monitoring.

Organ segmentation for surgical planning has been transformed by deep learning algorithms that can accurately identify and segment various organs and anatomical structures in medical images. This technology is essential for surgical planning, where precise knowledge of organ boundaries and relationships is critical for successful outcomes. AI models can segment organs in CT, MRI, and other imaging modalities, providing surgeons with detailed 3D models of patient anatomy. The technology has been integrated into surgical planning systems, where it helps surgeons visualize complex anatomical relationships and plan optimal surgical approaches. These systems have been particularly valuable in minimally invasive surgery, where precise knowledge of anatomy is essential for successful outcomes.

#### **12.4.2 Drug Discovery**

Drug discovery has been revolutionized by deep learning approaches that can predict molecular properties, identify potential drug candidates, and optimize drug-target interactions. These technologies have significantly accelerated the drug development process, reducing the time and cost required to bring new treatments to market. AI models can analyze vast databases of molecular structures and biological data to identify compounds with desired properties, enabling more efficient drug discovery and development. The technology has been particularly valuable for rare diseases and conditions where traditional drug discovery methods may be less effective. These systems continue to evolve, with newer models incorporating multi-modal data and advanced molecular representations to provide more comprehensive drug discovery capabilities.

Predicting molecular properties using deep learning has become a standard approach in drug discovery, where AI models can predict various molecular characteristics such as solubility, toxicity, and binding affinity. These predictions are essential for identifying promising drug candidates and optimizing molecular structures for desired properties. AI models can analyze molecular structures and predict properties with high accuracy, enabling researchers to focus on the most promising compounds for further development. The technology has been integrated into drug discovery pipelines at major pharmaceutical companies, where it helps researchers identify and optimize drug candidates more efficiently. These systems continue to evolve, with newer models incorporating advanced molecular representations and multi-property optimization to provide more comprehensive drug discovery capabilities.

Protein structure prediction, exemplified by AlphaFold, has revolutionized our understanding of protein biology and drug discovery by providing accurate predictions of protein structures from amino acid sequences. This technology has been particularly valuable for understanding protein function and identifying potential drug targets, enabling more effective drug discovery and development. AlphaFold and similar systems have been used to predict structures for millions of proteins, providing researchers with unprecedented access to structural information. The technology has been integrated into drug discovery pipelines, where it helps researchers understand protein function and identify potential drug targets. These systems continue to evolve, with newer models incorporating advanced molecular representations and multi-scale modeling to provide more comprehensive protein structure prediction capabilities.

Drug-target interaction prediction has been significantly improved by deep learning models that can predict how drugs interact with various biological targets. These predictions are essential for understanding drug mechanisms and identifying potential side effects, enabling more effective drug development and clinical use. AI models can analyze drug and target structures to predict

interaction strength and specificity, helping researchers optimize drug design and identify potential drug combinations. The technology has been integrated into drug discovery pipelines, where it helps researchers identify and optimize drug-target interactions more efficiently. These systems continue to evolve, with newer models incorporating advanced molecular representations and multi-target optimization to provide more comprehensive drug discovery capabilities.

### 12.4.3 Clinical Decision Support

Clinical decision support systems powered by deep learning have transformed healthcare by providing clinicians with evidence-based recommendations and risk assessments. These systems can analyze vast amounts of patient data to identify patterns and trends that may not be apparent to human observers, enabling more informed clinical decision-making. The technology has been integrated into electronic health records and clinical workflows, where it assists clinicians in diagnosis, treatment planning, and patient monitoring. These systems have been particularly valuable in complex cases where multiple factors must be considered, enabling clinicians to make more informed decisions based on comprehensive data analysis. The technology continues to evolve, with newer models incorporating multi-modal data and advanced reasoning capabilities to provide more comprehensive clinical decision support.

Diagnosis assistance has been significantly improved by AI systems that can analyze patient data to provide diagnostic recommendations and identify potential conditions. These systems can process symptoms, medical history, laboratory results, and imaging data to suggest possible diagnoses and recommend further testing. The technology has been integrated into clinical workflows at major medical centers, where it assists clinicians in complex diagnostic cases and helps identify patients who may need immediate attention. These systems have been particularly valuable in emergency departments and intensive care units, where rapid diagnosis is critical for patient outcomes. The technology continues to evolve, with newer models incorporating advanced reasoning capabilities and multi-modal data analysis to provide more comprehensive diagnostic support.

Treatment recommendation systems have been transformed by AI models that can analyze patient data to suggest optimal treatment strategies and monitor treatment responses. These systems can consider patient characteristics, disease severity, treatment history, and other factors to recommend personalized treatment approaches. The technology has been integrated into clinical workflows, where it assists clinicians in treatment planning and helps identify patients who may need treatment adjustments. These systems have been particularly valuable in oncology and other complex medical specialties, where treatment decisions must consider multiple factors and patient characteristics. The technology continues to evolve, with newer models incorporating advanced reasoning capabilities and multi-modal data analysis to provide more comprehensive treatment recommendations.

Risk prediction for readmission and mortality has been significantly improved by AI models that can analyze patient data to identify those at highest risk for adverse outcomes. These systems can process clinical data, laboratory results, and other information to predict patient risk and recommend interventions to improve outcomes. The technology has been integrated into clinical workflows, where it helps clinicians identify high-risk patients and implement preventive measures. These systems have been particularly valuable in intensive care units and other high-risk settings, where early identification of at-risk patients can significantly improve outcomes. The technology continues to evolve, with newer models incorporating advanced risk assessment capabilities and multi-modal data analysis to provide more comprehensive risk prediction.

### 12.4.4 Genomics

Genomics has been revolutionized by deep learning approaches that can analyze DNA sequences, identify genetic variants, and predict gene expression patterns. These technologies have enabled personalized medicine approaches that consider individual genetic characteristics when making treatment decisions. AI models can analyze genomic data to identify disease-associated variants, predict drug responses, and recommend personalized treatment strategies. The technology has been integrated into clinical workflows, where it helps clinicians make more informed decisions based on genetic information. These systems continue to evolve, with newer models incorporating advanced genomic representations and multi-scale analysis to provide more comprehensive genomic insights.

DNA sequence analysis has been transformed by deep learning models that can identify genetic variants, predict functional effects, and analyze sequence patterns. These systems can process vast amounts of genomic data to identify variants associated with disease and drug response, enabling more personalized medicine approaches. The technology has been integrated into

clinical genomics workflows, where it helps researchers and clinicians identify clinically relevant variants and understand their functional effects. These systems have been particularly valuable for rare disease diagnosis and treatment, where genetic information can provide crucial insights into disease mechanisms and treatment options. The technology continues to evolve, with newer models incorporating advanced genomic representations and multi-scale analysis to provide more comprehensive sequence analysis capabilities.

Variant calling has been significantly improved by AI models that can identify genetic variants with high accuracy and distinguish between pathogenic and benign variants. These systems can process genomic data to identify single nucleotide variants, insertions, deletions, and other genetic changes that may be associated with disease. The technology has been integrated into clinical genomics workflows, where it helps researchers and clinicians identify clinically relevant variants and understand their functional effects. These systems have been particularly valuable for rare disease diagnosis and treatment, where genetic information can provide crucial insights into disease mechanisms and treatment options. The technology continues to evolve, with newer models incorporating advanced genomic representations and multi-scale analysis to provide more comprehensive variant calling capabilities.

Gene expression prediction has been transformed by deep learning models that can predict gene expression levels from various molecular data types. These predictions are essential for understanding gene function and identifying potential drug targets, enabling more effective drug discovery and development. AI models can analyze genomic data to predict gene expression patterns and identify genes that may be associated with disease or drug response. The technology has been integrated into drug discovery pipelines, where it helps researchers identify potential drug targets and understand gene function. These systems continue to evolve, with newer models incorporating advanced genomic representations and multi-scale analysis to provide more comprehensive gene expression prediction capabilities.

#### 12.4.5 Historical context and references

The development of deep learning in healthcare has been marked by several key breakthroughs that have transformed medical imaging, drug discovery, and clinical decision support. The introduction of CNNs in medical imaging, particularly U-Net for biomedical image segmentation, revolutionized the field by enabling accurate and automated analysis of medical images. The development of transfer learning approaches allowed medical AI systems to leverage pretrained models and achieve high performance with limited medical data. The introduction of attention mechanisms and transformer architectures further improved the ability to process complex medical data and identify subtle patterns in medical images and clinical data.

The 2010s saw significant advances in medical AI, with systems like DeepMind's AlphaFold revolutionizing protein structure prediction and enabling new approaches to drug discovery. The development of federated learning approaches allowed medical AI systems to learn from distributed data while maintaining patient privacy, addressing one of the major challenges in medical AI deployment. The introduction of explainable AI techniques has been crucial for medical applications, where clinicians need to understand and trust AI recommendations. The development of regulatory frameworks for medical AI has been essential for ensuring the safety and effectiveness of these systems in clinical practice.

The 2020s have seen the integration of multi-modal AI approaches that can process various types of medical data, from images and genomic sequences to clinical notes and laboratory results. The development of large-scale medical AI models has enabled more comprehensive analysis of patient data and improved diagnostic accuracy. The integration of AI into clinical workflows has been accelerated by the COVID-19 pandemic, where AI systems have been used for diagnosis, treatment planning, and drug discovery. The field continues to evolve rapidly, with new architectures, training methods, and applications emerging regularly, making healthcare AI one of the most dynamic and impactful areas of artificial intelligence. See Ronneberger, Fischer, and Brox [RFB15] and Prince [Pri23] for broader context and comprehensive tutorials on the theoretical foundations and practical applications of these transformative technologies.

## 12.5 Reinforcement Learning Applications

Reinforcement learning represents one of the most exciting frontiers of artificial intelligence, where agents learn to make optimal decisions through trial and error in complex environments. Unlike supervised learning, which relies on labeled examples, reinforcement learning agents learn by interacting with their environment, receiving rewards or penalties for their

actions, and gradually improving their decision-making strategies. This paradigm has enabled breakthrough achievements in game playing, robotics, autonomous systems, and resource optimization, demonstrating the power of learning from experience rather than explicit instruction. The technology has found applications in diverse domains, from entertainment and gaming to critical infrastructure and safety systems, where adaptive decision-making is essential for success. The integration of deep learning with reinforcement learning has created powerful systems that can master complex tasks and environments, opening new possibilities for artificial intelligence applications.

### 12.5.1 Game Playing

Game playing has been revolutionized by reinforcement learning, where AI agents have achieved superhuman performance in complex strategic games through self-play and deep learning. These systems learn optimal strategies by playing millions of games against themselves, gradually improving their decision-making through trial and error. The success of these systems has demonstrated the power of reinforcement learning to master complex, multi-step decision problems that require long-term planning and strategic thinking. These achievements have not only advanced the field of artificial intelligence but have also provided insights into human cognition and decision-making processes.

AlphaGo's victory over world champion Lee Sedol in 2016 marked a historic milestone in artificial intelligence, demonstrating that AI could master the ancient game of Go, which was considered too complex for computers. The system used deep neural networks combined with Monte Carlo tree search to evaluate board positions and plan moves, learning from millions of games played against itself. AlphaGo's success has inspired new approaches to strategic decision-making in various domains, from business strategy to military planning. The technology has also been applied to other complex games and strategic problems, demonstrating the versatility of reinforcement learning approaches for mastering challenging decision environments.

AlphaZero represents a breakthrough in general game-playing AI, demonstrating that a single algorithm could master multiple complex games without domain-specific knowledge. The system learned to play chess, shogi, and Go at superhuman levels by playing against itself, discovering novel strategies and playing styles that surprised human experts. AlphaZero's success has inspired new approaches to multi-domain learning and transfer learning, where knowledge gained in one domain can be applied to related problems. The technology has been applied to various strategic games and decision problems, demonstrating the power of general-purpose reinforcement learning algorithms for mastering complex environments.

OpenAI Five's success in Dota 2 demonstrated that reinforcement learning could master complex real-time strategy games that require teamwork, coordination, and long-term planning. The system learned to play the game at a professional level by training on thousands of games, developing sophisticated strategies and coordination patterns. OpenAI Five's achievements have inspired new approaches to multi-agent reinforcement learning and team coordination, where multiple AI agents must work together to achieve common goals. The technology has been applied to various multi-agent systems and collaborative AI applications, demonstrating the potential of reinforcement learning for complex, multi-agent environments.

AlphaStar's mastery of StarCraft II demonstrated that reinforcement learning could handle complex real-time strategy games with imperfect information and continuous action spaces. The system learned to play the game at a grandmaster level by training on millions of games, developing sophisticated strategies and micro-management skills. AlphaStar's success has inspired new approaches to real-time decision-making and resource management, where agents must make rapid decisions under uncertainty. The technology has been applied to various real-time systems and resource optimization problems, demonstrating the power of reinforcement learning for complex, dynamic environments.

### 12.5.2 Robotics

Robotics has been transformed by reinforcement learning, where robots learn to perform complex manipulation and navigation tasks through trial and error in real-world environments. These systems can adapt to changing conditions and learn new skills without explicit programming, making them more versatile and capable than traditional robotic systems. The technology has enabled robots to perform complex tasks in unstructured environments, from manufacturing and assembly to household chores and caregiving. These systems continue to evolve, with newer approaches incorporating simulation-to-real transfer and multi-modal learning to improve performance and generalization.

Manipulation tasks in robotics have been significantly improved by reinforcement learning, where robots learn to grasp, manipulate, and assemble objects through trial and error. These systems can adapt to different object shapes, sizes, and

materials, learning optimal grasping strategies and manipulation techniques. The technology has been applied to manufacturing and assembly tasks, where robots must handle various products and components with high precision and reliability. These systems have been particularly valuable in flexible manufacturing environments, where robots must adapt to changing product requirements and production schedules. The technology continues to evolve, with newer approaches incorporating tactile feedback and multi-modal sensing to improve manipulation capabilities.

Navigation in robotics has been revolutionized by reinforcement learning, where robots learn to move autonomously through complex environments while avoiding obstacles and reaching their destinations. These systems can adapt to different environments and conditions, learning optimal navigation strategies and path-planning techniques. The technology has been applied to various robotic applications, from autonomous vehicles and drones to service robots and mobile platforms. These systems have been particularly valuable in dynamic environments, where robots must adapt to changing conditions and unexpected obstacles. The technology continues to evolve, with newer approaches incorporating advanced perception and planning capabilities to improve navigation performance.

Locomotion in robotics has been transformed by reinforcement learning, where robots learn to walk, run, and move in various ways through trial and error. These systems can adapt to different terrains and conditions, learning optimal locomotion strategies and movement patterns. The technology has been applied to various robotic applications, from humanoid robots and exoskeletons to quadruped robots and mobile platforms. These systems have been particularly valuable in challenging environments, where robots must adapt to rough terrain and changing conditions. The technology continues to evolve, with newer approaches incorporating advanced control and sensing capabilities to improve locomotion performance.

### 12.5.3 Autonomous Vehicles

Autonomous vehicles have been revolutionized by reinforcement learning, where AI systems learn to make complex driving decisions through trial and error in simulated and real-world environments. These systems can adapt to different driving conditions and scenarios, learning optimal driving strategies and decision-making techniques. The technology has been applied to various autonomous vehicle applications, from passenger cars and trucks to drones and autonomous ships. These systems have been particularly valuable in complex driving scenarios, where vehicles must make rapid decisions under uncertainty and changing conditions. The technology continues to evolve, with newer approaches incorporating advanced perception and planning capabilities to improve driving performance.

Path planning and decision making in autonomous vehicles have been significantly improved by reinforcement learning, where AI systems learn to plan optimal routes and make driving decisions through trial and error. These systems can adapt to different traffic conditions and scenarios, learning optimal driving strategies and decision-making techniques. The technology has been applied to various autonomous vehicle applications, from highway driving and city navigation to parking and maneuvering. These systems have been particularly valuable in complex driving scenarios, where vehicles must make rapid decisions under uncertainty and changing conditions. The technology continues to evolve, with newer approaches incorporating advanced perception and planning capabilities to improve driving performance.

The combination of reinforcement learning with computer vision has created powerful autonomous vehicle systems that can perceive their environment and make driving decisions based on visual information. These systems can identify and track other vehicles, pedestrians, and obstacles, enabling safe and efficient navigation through complex environments. The technology has been applied to various autonomous vehicle applications, from passenger cars and trucks to drones and autonomous ships.

These systems have been particularly valuable in complex driving scenarios, where vehicles must make rapid decisions based on visual information and changing conditions. The technology continues to evolve, with newer approaches incorporating advanced perception and planning capabilities to improve driving performance.

Safety-critical systems in autonomous vehicles have been transformed by reinforcement learning, where AI systems learn to make safe driving decisions through trial and error in controlled environments. These systems can adapt to different safety requirements and scenarios, learning optimal safety strategies and decision-making techniques. The technology has been applied to various autonomous vehicle applications, from passenger cars and trucks to drones and autonomous ships. These systems have been particularly valuable in complex driving scenarios, where vehicles must make rapid decisions under uncertainty and changing conditions. The technology continues to evolve, with newer approaches incorporating advanced perception and planning capabilities to improve driving performance.

#### 12.5.4 Recommendation Systems

Recommendation systems have been revolutionized by reinforcement learning, where AI systems learn to make personalized recommendations through trial and error in user interaction environments. These systems can adapt to different user preferences and behaviors, learning optimal recommendation strategies and decision-making techniques. The technology has been applied to various recommendation applications, from content and product recommendations to personalized services and experiences. These systems have been particularly valuable in dynamic environments, where user preferences and behaviors change over time. The technology continues to evolve, with newer approaches incorporating advanced user modeling and recommendation capabilities to improve performance.

Netflix and YouTube content recommendations have been transformed by reinforcement learning, where AI systems learn to recommend personalized content based on user behavior and preferences. These systems can adapt to different user tastes and viewing habits, learning optimal recommendation strategies and content selection techniques. The technology has been applied to various content recommendation applications, from movies and TV shows to music and podcasts. These systems have been particularly valuable in dynamic environments, where user preferences and content availability change over time. The technology continues to evolve, with newer approaches incorporating advanced user modeling and content analysis capabilities to improve recommendation performance.

E-commerce product suggestions have been significantly improved by reinforcement learning, where AI systems learn to recommend personalized products based on user behavior and preferences. These systems can adapt to different shopping patterns and preferences, learning optimal recommendation strategies and product selection techniques. The technology has been applied to various e-commerce applications, from online shopping and retail to marketplace and auction platforms. These systems have been particularly valuable in dynamic environments, where user preferences and product availability change over time. The technology continues to evolve, with newer approaches incorporating advanced user modeling and product analysis capabilities to improve recommendation performance.

Personalized news feeds have been transformed by reinforcement learning, where AI systems learn to recommend personalized news content based on user behavior and preferences. These systems can adapt to different reading habits and interests, learning optimal recommendation strategies and content selection techniques. The technology has been applied to various news and media applications, from social media and news aggregators to personalized content platforms. These systems have been particularly valuable in dynamic environments, where user interests and content availability change over time. The technology continues to evolve, with newer approaches incorporating advanced user modeling and content analysis capabilities to improve recommendation performance.

#### 12.5.5 Resource Management

Resource management has been revolutionized by reinforcement learning, where AI systems learn to optimize resource allocation and utilization through trial and error in complex environments. These systems can adapt to different resource constraints and requirements, learning optimal management strategies and decision-making techniques. The technology has been applied to various resource management applications, from data center optimization and energy management to supply chain and logistics optimization. These systems have been particularly valuable in dynamic environments, where resource availability and demand change over time. The technology continues to evolve, with newer approaches incorporating advanced optimization and planning capabilities to improve resource management performance.

Data center cooling optimization has been transformed by reinforcement learning, where AI systems learn to optimize cooling systems and energy consumption through trial and error in data center environments. These systems can adapt to different cooling requirements and conditions, learning optimal cooling strategies and energy management techniques. The technology has been applied to various data center applications, from server cooling and energy management to facility optimization and maintenance. These systems have been particularly valuable in dynamic environments, where cooling requirements and energy costs change over time. The technology continues to evolve, with newer approaches incorporating advanced optimization and planning capabilities to improve cooling performance.

Traffic light control has been significantly improved by reinforcement learning, where AI systems learn to optimize traffic flow and reduce congestion through trial and error in traffic environments. These systems can adapt to different traffic patterns and conditions, learning optimal control strategies and traffic management techniques. The technology has been applied to various

traffic management applications, from intersection control and signal optimization to traffic flow management and congestion reduction. These systems have been particularly valuable in dynamic environments, where traffic patterns and conditions change over time. The technology continues to evolve, with newer approaches incorporating advanced optimization and planning capabilities to improve traffic control performance.

Energy grid optimization has been transformed by reinforcement learning, where AI systems learn to optimize energy distribution and consumption through trial and error in power grid environments. These systems can adapt to different energy demands and supply conditions, learning optimal distribution strategies and energy management techniques. The technology has been applied to various energy management applications, from power grid optimization and renewable energy integration to energy storage and demand response. These systems have been particularly valuable in dynamic environments, where energy demand and supply change over time. The technology continues to evolve, with newer approaches incorporating advanced optimization and planning capabilities to improve energy management performance.

### 12.5.6 Historical context and references

The development of reinforcement learning has been marked by several key breakthroughs that have transformed the field from theoretical concepts to practical applications. The introduction of Q-learning in the 1980s provided the first practical algorithm for learning optimal policies in Markov decision processes, enabling agents to learn from experience without requiring a model of the environment. The development of policy gradient methods in the 1990s provided more efficient approaches to learning continuous control policies, enabling applications in robotics and autonomous systems. The introduction of deep reinforcement learning in the 2010s combined the power of deep neural networks with reinforcement learning, enabling agents to learn from high-dimensional sensory input and master complex environments.

The 2010s saw significant advances in reinforcement learning, with systems like DeepMind's AlphaGo demonstrating the power of deep reinforcement learning combined with tree search for mastering complex strategic games. The development of actor-critic methods and advanced policy gradient algorithms provided more stable and efficient approaches to learning complex policies. The introduction of experience replay and target networks improved the stability and efficiency of deep Q-learning, enabling applications in various domains. The development of multi-agent reinforcement learning opened new possibilities for collaborative AI and competitive environments.

The 2020s have seen the integration of reinforcement learning with other AI technologies, creating powerful systems that can learn from multiple modalities and adapt to complex environments. The development of transfer learning and meta-learning approaches has enabled agents to learn new tasks more efficiently by leveraging knowledge from related tasks. The integration of reinforcement learning with computer vision and natural language processing has created powerful systems that can understand and interact with complex environments. The field continues to evolve rapidly, with new algorithms, architectures, and applications emerging regularly, making reinforcement learning one of the most dynamic and impactful areas of artificial intelligence. See Silver et al. [Sil+16] and Prince [Pri23] for broader context and comprehensive tutorials on the theoretical foundations and practical applications of these transformative technologies.

## 12.6 Other Applications ●

Beyond the major application domains of computer vision, natural language processing, speech recognition, healthcare, and reinforcement learning, deep learning has found transformative applications across numerous other sectors. These applications demonstrate the versatility and adaptability of deep learning technologies to diverse problems and industries, from financial services and scientific research to agriculture and manufacturing. The technology has enabled new approaches to complex problems that were previously intractable with traditional methods, opening up possibilities for innovation and efficiency gains across multiple sectors. These applications often require domain-specific adaptations and considerations, highlighting the importance of understanding both the technical capabilities of deep learning and the unique requirements of each application domain. The continued expansion of deep learning into new areas demonstrates its potential to transform industries and create new opportunities for technological advancement.

## 12.6.1 Finance

Finance has been revolutionized by deep learning applications that can analyze vast amounts of financial data to make predictions, detect patterns, and optimize trading strategies. These systems can process complex market data, news sentiment, and economic indicators to identify trading opportunities and manage risk more effectively than traditional methods. The technology has been particularly valuable for high-frequency trading, where rapid decision-making is essential for success. These systems continue to evolve, with newer approaches incorporating advanced risk management and regulatory compliance capabilities to improve performance and reliability.

Algorithmic trading has been transformed by deep learning systems that can analyze market data and execute trades with superhuman speed and accuracy. These systems can identify patterns in price movements, volume changes, and market sentiment to make profitable trading decisions. The technology has been applied to various financial markets, from stocks and bonds to commodities and currencies, enabling traders to capitalize on market opportunities more effectively. These systems have been particularly valuable for institutional investors and hedge funds, where they can process vast amounts of data and execute complex trading strategies. The technology continues to evolve, with newer approaches incorporating advanced risk management and regulatory compliance capabilities to improve trading performance.

Fraud detection has been significantly improved by deep learning models that can identify suspicious transactions and activities with high accuracy. These systems can analyze transaction patterns, user behavior, and other factors to detect fraudulent activities in real-time. The technology has been applied to various financial services, from credit cards and banking to insurance and investment platforms, helping to protect customers and reduce financial losses. These systems have been particularly valuable for online transactions and digital payments, where fraud risks are higher and traditional detection methods may be less effective. The technology continues to evolve, with newer approaches incorporating advanced behavioral analysis and anomaly detection capabilities to improve fraud prevention.

Credit risk assessment has been revolutionized by deep learning approaches that can analyze borrower characteristics and predict default probabilities with high accuracy. These systems can process various types of data, from credit scores and income statements to social media activity and spending patterns, to assess creditworthiness. The technology has been applied to various lending applications, from personal loans and mortgages to business credit and investment decisions, enabling more accurate risk assessment and pricing. These systems have been particularly valuable for alternative lending and fintech applications, where traditional credit assessment methods may be less effective. The technology continues to evolve, with newer approaches incorporating advanced behavioral analysis and alternative data sources to improve credit assessment accuracy.

Market prediction has been transformed by deep learning models that can forecast price movements and market trends with improved accuracy. These systems can analyze various factors, from economic indicators and news sentiment to technical analysis and market microstructure, to predict future market behavior. The technology has been applied to various financial markets, from stocks and bonds to commodities and currencies, enabling investors to make more informed decisions. These systems have been particularly valuable for portfolio management and risk assessment, where accurate market predictions are essential for success. The technology continues to evolve, with newer approaches incorporating advanced time series analysis and multi-modal data processing to improve prediction accuracy.

## 12.6.2 Scientific Research

Scientific research has been transformed by deep learning applications that can analyze complex data and identify patterns that may be difficult for human researchers to detect. These systems can process vast amounts of experimental data, simulations, and observations to discover new insights and accelerate the pace of scientific discovery. The technology has been applied to various scientific disciplines, from physics and chemistry to biology and environmental science, enabling researchers to tackle complex problems more effectively. These systems continue to evolve, with newer approaches incorporating advanced data analysis and pattern recognition capabilities to improve scientific research outcomes.

Physics research has been revolutionized by deep learning applications that can analyze particle physics data and detect gravitational waves with unprecedented sensitivity. These systems can process complex experimental data to identify particle interactions, classify events, and detect rare phenomena that may be missed by traditional analysis methods. The technology has been applied to various physics experiments, from particle accelerators and detectors to gravitational wave observatories and neutrino experiments, enabling researchers to discover new particles and phenomena. These systems have been particularly

valuable for high-energy physics experiments, where vast amounts of data must be processed to identify rare events and interactions. The technology continues to evolve, with newer approaches incorporating advanced signal processing and pattern recognition capabilities to improve physics research outcomes.

Climate science has been transformed by deep learning models that can analyze climate data and predict weather patterns with improved accuracy. These systems can process various types of data, from satellite imagery and weather stations to ocean currents and atmospheric conditions, to forecast weather and climate changes. The technology has been applied to various climate applications, from weather forecasting and climate modeling to extreme weather prediction and climate change analysis, enabling researchers to better understand and predict climate behavior. These systems have been particularly valuable for long-term climate projections and extreme weather events, where accurate predictions are essential for planning and preparedness. The technology continues to evolve, with newer approaches incorporating advanced data assimilation and multi-scale modeling to improve climate prediction accuracy.

Astronomy has been revolutionized by deep learning applications that can analyze astronomical data and identify celestial objects with superhuman accuracy. These systems can process images from telescopes and observatories to classify galaxies, detect exoplanets, and identify other astronomical phenomena. The technology has been applied to various astronomical surveys, from galaxy classification and exoplanet detection to supernova identification and gravitational lensing analysis, enabling researchers to discover new celestial objects and phenomena. These systems have been particularly valuable for large-scale astronomical surveys, where vast amounts of data must be processed to identify rare and interesting objects. The technology continues to evolve, with newer approaches incorporating advanced image processing and pattern recognition capabilities to improve astronomical research outcomes.

### 12.6.3 Agriculture

Agriculture has been transformed by deep learning applications that can analyze crop data and optimize farming practices to improve yields and reduce environmental impact. These systems can process various types of data, from satellite imagery and weather data to soil conditions and crop health, to provide farmers with actionable insights and recommendations. The technology has been applied to various agricultural applications, from crop monitoring and disease detection to yield prediction and precision agriculture, enabling farmers to make more informed decisions and improve productivity. These systems have been particularly valuable for large-scale farming operations, where data-driven insights can significantly impact profitability and sustainability. The technology continues to evolve, with newer approaches incorporating advanced sensor data and machine learning capabilities to improve agricultural outcomes.

Crop disease detection has been revolutionized by deep learning models that can identify plant diseases and pests with high accuracy from images and sensor data. These systems can analyze various types of data, from drone imagery and satellite photos to ground-based sensors and weather data, to detect diseases early and recommend treatment strategies. The technology has been applied to various crops, from wheat and corn to fruits and vegetables, enabling farmers to prevent disease outbreaks and minimize crop losses. These systems have been particularly valuable for organic farming and sustainable agriculture, where early disease detection is essential for maintaining crop health without chemical treatments. The technology continues to evolve, with newer approaches incorporating advanced image processing and sensor fusion to improve disease detection accuracy.

Yield prediction has been significantly improved by deep learning approaches that can forecast crop yields with high accuracy based on various factors. These systems can analyze weather data, soil conditions, planting patterns, and historical yields to predict future harvests and optimize farming decisions. The technology has been applied to various crops and farming systems, from traditional agriculture to precision farming and vertical farming, enabling farmers to plan their operations more effectively. These systems have been particularly valuable for commodity trading and food security planning, where accurate yield predictions are essential for market stability and food supply management. The technology continues to evolve, with newer approaches incorporating advanced weather modeling and soil analysis to improve yield prediction accuracy.

Precision agriculture has been transformed by deep learning systems that can optimize farming practices at the individual plant or field level. These systems can analyze various types of data, from soil sensors and weather stations to drone imagery and satellite data, to provide personalized recommendations for each part of a field. The technology has been applied to various farming operations, from crop planting and fertilization to irrigation and harvesting, enabling farmers to maximize yields while minimizing resource use. These systems have been particularly valuable for sustainable farming and environmental conservation,

where precision agriculture can reduce water usage, fertilizer application, and environmental impact. The technology continues to evolve, with newer approaches incorporating advanced robotics and automation to improve precision agriculture outcomes.

#### 12.6.4 Manufacturing

Manufacturing has been revolutionized by deep learning applications that can optimize production processes, detect defects, and predict maintenance needs with unprecedented accuracy. These systems can analyze various types of data, from production sensors and quality control measurements to supply chain information and customer feedback, to improve manufacturing efficiency and product quality. The technology has been applied to various manufacturing processes, from automotive and electronics to pharmaceuticals and food production, enabling companies to reduce costs and improve competitiveness. These systems have been particularly valuable for high-volume manufacturing operations, where small improvements in efficiency can have significant impact on profitability. The technology continues to evolve, with newer approaches incorporating advanced robotics and automation to improve manufacturing outcomes.

Quality control and defect detection have been transformed by deep learning models that can identify product defects and quality issues with superhuman accuracy. These systems can analyze various types of data, from visual inspections and sensor measurements to production parameters and test results, to detect defects early and prevent quality issues. The technology has been applied to various manufacturing processes, from automotive assembly and electronics production to pharmaceutical manufacturing and food processing, enabling companies to maintain high quality standards and reduce waste. These systems have been particularly valuable for automated production lines, where real-time quality control is essential for maintaining product consistency and customer satisfaction. The technology continues to evolve, with newer approaches incorporating advanced computer vision and sensor fusion to improve quality control accuracy.

Predictive maintenance has been significantly improved by deep learning approaches that can predict equipment failures and maintenance needs with high accuracy. These systems can analyze various types of data, from vibration sensors and temperature readings to production parameters and historical maintenance records, to forecast when equipment will need maintenance or replacement. The technology has been applied to various manufacturing equipment, from motors and pumps to conveyor belts and robotic systems, enabling companies to prevent unplanned downtime and reduce maintenance costs. These systems have been particularly valuable for critical manufacturing processes, where equipment failures can cause significant production losses and safety risks. The technology continues to evolve, with newer approaches incorporating advanced sensor data and machine learning capabilities to improve predictive maintenance accuracy.

Supply chain optimization has been revolutionized by deep learning systems that can optimize supply chain operations and predict demand patterns with improved accuracy. These systems can analyze various types of data, from sales forecasts and inventory levels to supplier performance and transportation costs, to optimize supply chain decisions and reduce costs. The technology has been applied to various supply chain applications, from inventory management and demand forecasting to supplier selection and logistics optimization, enabling companies to improve efficiency and reduce supply chain risks. These systems have been particularly valuable for global supply chains, where complex logistics and demand variability can significantly impact costs and service levels. The technology continues to evolve, with newer approaches incorporating advanced optimization algorithms and real-time data processing to improve supply chain performance.

#### 12.6.5 References

For domain-specific overviews, see Prince [Pri23] (applications survey). The applications discussed in this section represent just a sample of the diverse ways in which deep learning is transforming industries and creating new opportunities for technological advancement. These applications demonstrate the versatility and adaptability of deep learning technologies to diverse problems and industries, from financial services and scientific research to agriculture and manufacturing. The continued expansion of deep learning into new areas demonstrates its potential to transform industries and create new opportunities for technological advancement. As the field continues to evolve, we can expect to see even more innovative applications and breakthroughs that will further demonstrate the power and potential of deep learning technologies.

## 12.7 Ethical Considerations ●

Deep learning applications raise important ethical concerns that must be carefully considered and addressed throughout the development and deployment process. These concerns span multiple dimensions, from technical challenges like bias and transparency to broader societal impacts like job displacement and environmental sustainability. As deep learning systems become more powerful and widely deployed, the ethical implications of their use become increasingly significant, requiring proactive measures to ensure responsible development and deployment. The field has recognized the importance of ethical considerations, with researchers, practitioners, and organizations developing frameworks, guidelines, and best practices to address these challenges. Responsible AI development requires addressing these challenges through comprehensive documentation, human oversight, continuous monitoring, and careful consideration of environmental costs during model selection.

Bias and fairness represent critical ethical concerns in deep learning applications, where models may perpetuate or amplify societal biases present in training data. These biases can manifest in various ways, from discriminatory hiring decisions in automated screening systems to unfair loan approvals in financial applications, potentially reinforcing existing inequalities and creating new forms of discrimination. The challenge is particularly acute because bias can be subtle and difficult to detect, especially in complex deep learning models that process high-dimensional data and make decisions through non-linear transformations. Addressing bias requires careful attention to data collection and preprocessing, diverse and representative training datasets, and ongoing monitoring of model performance across different demographic groups. The field has developed various techniques for bias detection and mitigation, including fairness constraints, adversarial debiasing, and post-processing methods, though achieving true fairness remains an ongoing challenge that requires both technical and social solutions.

Privacy concerns have become increasingly prominent as deep learning systems process vast amounts of personal data, from medical records and financial information to social media posts and location data. These systems can extract sensitive information from seemingly innocuous data, creating risks for individual privacy and data protection. The challenge is compounded by the fact that deep learning models can memorize training data and potentially leak sensitive information through their outputs or intermediate representations. Privacy-preserving techniques such as differential privacy, federated learning, and secure multi-party computation have been developed to address these concerns, but implementing them effectively while maintaining model performance remains challenging. Organizations must carefully balance the benefits of data-driven insights with the need to protect individual privacy, often requiring legal compliance with regulations like GDPR and CCPA while implementing technical safeguards.

Transparency and interpretability represent fundamental challenges in deep learning, where the "black box" nature of complex models makes it difficult to understand how decisions are made. This lack of transparency can be problematic in high-stakes applications like healthcare, finance, and criminal justice, where understanding the reasoning behind decisions is crucial for trust, accountability, and regulatory compliance. The challenge is particularly acute for deep learning models, which often have millions of parameters and make decisions through complex, non-linear transformations that are difficult to interpret. Various techniques have been developed to improve model interpretability, including attention mechanisms, gradient-based methods, and surrogate models, but achieving full transparency while maintaining model performance remains an ongoing challenge. The field continues to evolve, with new approaches to explainable AI and interpretable machine learning being developed to address these concerns.

Security vulnerabilities in deep learning systems represent a significant ethical concern, where models can be manipulated through adversarial attacks and other malicious techniques. These attacks can cause models to make incorrect predictions or decisions, potentially leading to safety risks, financial losses, or other harmful outcomes. The challenge is particularly concerning for safety-critical applications like autonomous vehicles, medical diagnosis, and financial systems, where adversarial attacks could have serious consequences. Adversarial training, robust optimization, and other defensive techniques have been developed to improve model security, but achieving robust security while maintaining model performance remains challenging. The field continues to evolve, with new approaches to adversarial robustness and secure machine learning being developed to address these concerns.

Job displacement represents a significant societal concern as deep learning systems automate tasks previously performed by humans, potentially leading to unemployment and economic disruption. This displacement can affect workers across various industries, from manufacturing and transportation to healthcare and finance, creating challenges for individuals, communities,

and society as a whole. The challenge is particularly acute because automation can affect both routine and non-routine tasks, potentially displacing workers who may lack the skills or resources to transition to new roles. Addressing job displacement requires proactive measures, including education and training programs, social safety nets, and policies that support workers through transitions. The field has recognized the importance of responsible automation, with researchers and practitioners developing frameworks and guidelines to ensure that AI deployment benefits society as a whole.

Environmental impact has become an increasingly important ethical consideration as deep learning models become larger and more computationally intensive, requiring significant energy consumption for training and inference. The environmental cost of large models can be substantial, with some models requiring energy equivalent to that used by small cities, raising concerns about sustainability and climate change. The challenge is particularly acute for large language models and other resource-intensive applications, where the environmental cost of training and deployment can be significant. Addressing environmental impact requires careful consideration of model efficiency, renewable energy sources, and the trade-offs between model performance and environmental cost. The field has developed various techniques for efficient model design, including model compression, quantization, and knowledge distillation, though achieving optimal efficiency while maintaining performance remains an ongoing challenge.

## Key Takeaways

### Key Takeaways 12

- **Domain-specific applications:** Deep learning has transformed computer vision, NLP, speech recognition, health-care, and reinforcement learning, with each domain requiring specialized architectures and approaches tailored to specific data types and constraints.
- **Real-world impact:** Applications span from entertainment and gaming to critical infrastructure and safety systems, demonstrating the versatility of deep learning across diverse industries including finance, scientific research, agriculture, and manufacturing.
- **Systematic approach:** Successful applications require understanding the fundamental characteristics of each domain, including data assumptions, evaluation metrics, and deployment considerations, before selecting appropriate model families and training strategies.
- **Ethical responsibility:** As deep learning systems become more powerful and widely deployed, addressing bias, privacy, transparency, security, job displacement, and environmental impact becomes crucial for responsible AI development.
- **Practical constraints:** Real-world success depends on balancing technical sophistication with practical constraints, often achieving better results through careful problem formulation and data understanding than through complex model architectures.

## Exercises

### Easy

*Exercise 12.1 (Vision Metrics).* Choose between accuracy, AUROC, and mAP for object detection. Justify.

#### Hint:

*Class imbalance and localisation vs. classification.*

*Exercise 12.2 (NLP Tokenisation).* Explain implications of BPE vs. WordPiece for rare words.

**Hint:**

*Subword frequency, OOV handling, sequence length.*

*Exercise 12.3 (Tabular Baselines).* Why can tree ensembles outperform deep nets on small tabular datasets?

**Hint:**

*Inductive bias, feature interactions, sample efficiency.*

*Exercise 12.4 (Data Privacy).* List two privacy risks when deploying medical models and mitigations.

**Hint:**

*Re-identification, membership inference; anonymisation, DP.*

## Medium

*Exercise 12.5 (Dataset Shift).* Design a test to detect covariate shift between training and production.

**Hint:**

*Train a domain classifier; compare feature distributions.*

*Exercise 12.6 (Active Learning).* Propose an acquisition strategy for labelling a limited budget.

**Hint:**

*Uncertainty sampling, diversity, class balance.*

## Hard

*Exercise 12.7 (Cost-aware Serving).* Formalise an objective trading accuracy vs. compute cost and latency.

**Hint:**

*Multi-objective optimisation; constrained maximisation.*

*Exercise 12.8 (Ethical Deployment).* Specify post-deployment monitoring and escalation for high-stakes tasks.

**Hint:**

*Thresholds, audits, rollback, human oversight.*

*Exercise 12.9 (Advanced Topic 1).* Explain a key concept from this chapter and its practical applications.

**Hint:**

*Consider the theoretical foundations and real-world implications.*

*Exercise 12.10 (Advanced Topic 2).* Analyse the relationship between different techniques covered in this chapter.

**Hint:**

*Look for connections and trade-offs between methods.*

Exercise 12.11 (Advanced Topic 3). Design an experiment to test a hypothesis related to this chapter's content.

**Hint:**

*Consider experimental design, metrics, and potential confounding factors.*

Exercise 12.12 (Advanced Topic 4). Compare different approaches to solving a problem from this chapter.

**Hint:**

*Consider computational complexity, accuracy, and practical considerations.*

Exercise 12.13 (Advanced Topic 5). Derive a mathematical relationship or prove a theorem from this chapter.

**Hint:**

*Start with the definitions and work through the logical steps.*

Exercise 12.14 (Advanced Topic 6). Implement a practical solution to a problem discussed in this chapter.

**Hint:**

*Consider the implementation details and potential challenges.*

Exercise 12.15 (Advanced Topic 7). Evaluate the limitations and potential improvements of techniques from this chapter.

**Hint:**

*Consider both theoretical limitations and practical constraints.*



---

## **Part III**

# **Deep Learning Research**



# Chapter 13

## Linear Factor Models

This chapter introduces probabilistic models with linear structure, which form the foundation for many unsupervised learning methods.

### Learning Objectives

1. Probabilistic PCA and factor analysis relationships to PCA
2. EM updates for latent variable models with linear-Gaussian structure
3. Identifiability and rotation issues in factor models
4. Linear latent models and modern representation learning

## Intuition

Linear factor models posit a small set of hidden sources generating observed data through linear mixing plus noise. The learning problem is to recover those hidden coordinates that best explain variance while respecting uncertainty.

### 13.1 Probabilistic PCA ◆

Probabilistic PCA extends classical PCA by providing a probabilistic framework that handles uncertainty and enables principled treatment of noise and missing data in dimensionality reduction.

#### 13.1.1 Principal Component Analysis Review

PCA finds orthogonal directions of maximum variance through the transformation  $\mathbf{z} = \mathbf{W}^\top(\mathbf{x} - \boldsymbol{\mu})$  where  $\mathbf{W}$  contains principal components (eigenvectors of covariance matrix). In deep learning, this formula is fundamental for dimensionality reduction in preprocessing steps, where for example, when training a neural network on high-dimensional image data, PCA can reduce the input dimensionality from thousands of pixels to a smaller set of principal components, making training more efficient while preserving the most important variance in the data.

### 13.1.2 Probabilistic Formulation

**Remark 13.1: Key Differences between PCA and Probabilistic PCA**

The key difference is that classical PCA is deterministic and finds orthogonal directions of maximum variance, while Probabilistic PCA provides a probabilistic framework that models uncertainty through a generative process with noise, enabling principled treatment of missing data and uncertainty quantification.

The probabilistic formulation models observations through a generative process:

$$z \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (13.1)$$

$$\mathbf{x} | z \sim \mathcal{N}(\mathbf{W}z + \boldsymbol{\mu}, \sigma^2 \mathbf{I}) \quad (13.2)$$

**Explanations.**

- $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  means the latent variable has a *standard normal* prior: zero mean and identity covariance (independent unit-variance components).
- $\mathbf{x}|z \sim \mathcal{N}(\mathbf{W}z + \boldsymbol{\mu}, \sigma^2 \mathbf{I})$  is a *Gaussian likelihood*: data is a linear mapping of  $z$  via loadings  $\mathbf{W}$  plus mean  $\boldsymbol{\mu}$ , with isotropic noise variance  $\sigma^2$ .

Marginalizing the latent variable yields

$$\mathbf{x} \sim \mathcal{N}\left(\boldsymbol{\mu}, \underbrace{\mathbf{W}\mathbf{W}^\top}_{\text{signal}} + \underbrace{\sigma^2 \mathbf{I}}_{\text{noise}}\right). \quad (13.3)$$

Here  $\mathbf{W}\mathbf{W}^\top + \sigma^2 \mathbf{I}$  is the *covariance matrix* of the observed data:  $\mathbf{W}\mathbf{W}^\top$  captures variance explained by latent factors and  $\sigma^2 \mathbf{I}$  is residual (isotropic) noise.

**Definition 13.1: Probabilistic PCA**

Probabilistic PCA is a probabilistic extension of classical PCA that models observed data through a generative process where latent variables  $z$  follow a standard normal distribution and observed data  $\mathbf{x}$  is generated as a linear transformation of the latent variables plus Gaussian noise, enabling principled treatment of uncertainty and missing data.

### 13.1.3 Learning

**Remark 13.2: EM Algorithm**

The EM (Expectation-Maximization) algorithm is an iterative optimization method for maximum likelihood estimation in latent variable models, defined as  $\theta^{(t+1)} = \arg \max_\theta \mathbb{E}_{p(z|\mathbf{x}, \theta^{(t)})} [\log p(\mathbf{x}, z|\theta)]$  where the E-step computes the posterior expectation and the M-step maximizes the expected log-likelihood.

Learning in probabilistic PCA maximizes the likelihood using the EM algorithm, where the E-step computes the posterior distribution  $p(z|\mathbf{x})$  to estimate the latent variables given the observed data, and the M-step updates the parameters  $\mathbf{W}$ ,  $\boldsymbol{\mu}$ , and  $\sigma^2$  to maximize the expected log-likelihood. This iterative process allows the model to learn the optimal linear transformation and noise parameters that best explain the observed data, where the EM algorithm is particularly useful because it handles the latent variables naturally and provides a principled way to estimate parameters in the presence of missing data. As the noise variance  $\sigma^2 \rightarrow 0$ , the probabilistic PCA recovers standard PCA, showing that classical PCA is a special case of the probabilistic formulation when we assume no noise in the data generation process.

### 13.1.4 Historical context and references

The probabilistic formulation of PCA was developed to connect classical PCA with latent variable models and enable principled handling of noise and missing data, representing a significant advancement in dimensionality reduction techniques. This

probabilistic approach has been widely adopted in machine learning and deep learning applications, where it provides a foundation for understanding how linear transformations can capture the essential structure in high-dimensional data. Previous work has been establishing the theoretical foundations and practical applications of probabilistic PCA, while Goodfellow and colleagues have shown how these concepts extend to modern deep learning architectures. Real-world applications include image compression, noise reduction in signal processing, and preprocessing for neural network training, where the probabilistic framework allows for better handling of uncertainty and missing data compared to classical PCA.

## 13.2 Factor Analysis ◆

Factor analysis extends probabilistic PCA by allowing different noise variances for each observed dimension, providing more flexibility in modeling measurement error and enabling better handling of heterogeneous data where different variables may have different levels of reliability. The model assumes that observed data  $\mathbf{x}|\mathbf{z} \sim \mathcal{N}(\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \boldsymbol{\Psi})$  is generated from latent factors through linear transformation plus diagonal noise covariance  $\boldsymbol{\Psi}$ , where each observed dimension has its own noise variance, allowing for more realistic modeling of real-world data where different measurements may have different levels of precision. In deep learning, factor analysis is particularly useful for understanding the underlying structure of high-dimensional data and for preprocessing steps where we need to account for different levels of noise in different features.

### 13.2.1 Learning via EM

EM alternates between inferring posteriors over factors and updating loadings  $\mathbf{W}$  and noise  $\boldsymbol{\Psi}$ . Diagonal noise permits modeling idiosyncratic measurement error per dimension.

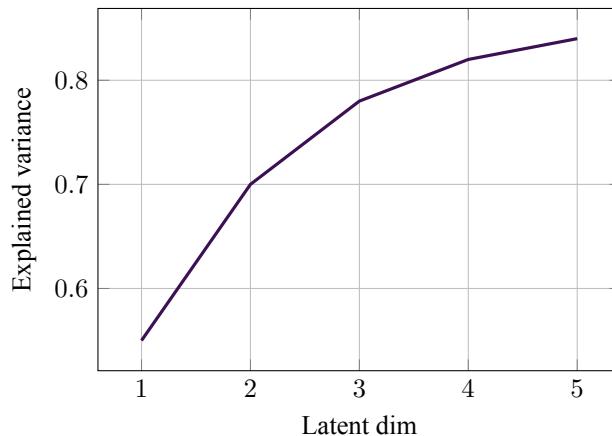


Figure 13.1: Explained variance as a function of latent dimensionality (illustrative).

## 13.3 Independent Component Analysis ◆

Independent Component Analysis (ICA) is a powerful technique for separating mixed signals into their independent source components, particularly useful when the sources are non-Gaussian and statistically independent.

### 13.3.1 Objective

The objective of ICA is to find independent sources from linear mixtures through the equation  $\mathbf{x} = \mathbf{As}$  where  $\mathbf{s}$  contains independent sources and  $\mathbf{A}$  is the mixing matrix. This formulation assumes that observed data is a linear combination of independent source signals, where the goal is to recover the original sources by finding the unmixing matrix that separates the mixed signals. ICA is particularly powerful because it can separate sources even when they are mixed in unknown proportions, making it useful for blind source separation problems where we don't know the mixing process beforehand. The key insight is

that independent sources have different statistical properties than their mixtures, allowing ICA to identify and separate them based on these differences.

### 13.3.2 Non-Gaussianity

ICA exploits the fact that independent signals are typically non-Gaussian, where the non-Gaussianity of source signals provides the key information needed to separate them from their mixtures. This is because Gaussian signals have the property that their mixtures are also Gaussian, making it impossible to separate them based on statistical independence alone, but non-Gaussian signals retain their distinctive statistical properties even when mixed. The non-Gaussianity assumption is crucial for ICA's success, as it allows the algorithm to identify which directions in the data space correspond to the original source signals rather than arbitrary linear combinations. This principle is widely applied in blind source separation problems like the cocktail party problem, where multiple speakers' voices are mixed together and need to be separated, in signal processing applications where noise needs to be separated from the signal of interest, and in feature extraction where we want to identify the most informative and independent features in high-dimensional data.

## 13.4 Sparse Coding ◆

Sparse coding learns an overcomplete dictionary where data has sparse representation through the optimization problem

$$\min_{\mathbf{D}, \mathbf{z}} \underbrace{\|\mathbf{x} - \mathbf{D}\mathbf{z}\|^2}_{\text{reconstruction error}} + \lambda \underbrace{\|\mathbf{z}\|_1}_{\text{sparsity penalty}}. \quad (13.4)$$

**Explanation.** The objective trades off fidelity to the data (small reconstruction error) with sparsity of the code  $\mathbf{z}$  (few nonzeros), controlled by  $\lambda > 0$ . This encourages parts-based, interpretable representations.

### 13.4.1 Optimization and interpretation

The  $\ell_1$  penalty promotes sparsity, yielding parts-based representations and robust denoising. The  $\ell_1$  penalty promotes sparsity by encouraging many coefficients to become exactly zero, leading to sparse solutions where only a few features are selected. Alternating minimization over dictionary  $\mathbf{D}$  and codes  $\mathbf{z}$  is common; convolutional variants are used in images Goodfellow, Bengio, and Courville [GBC16a].

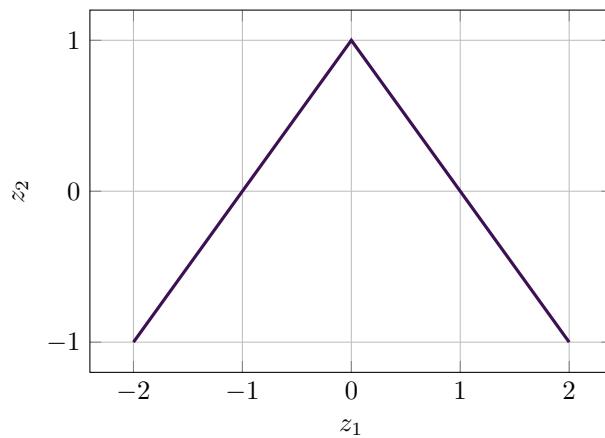


Figure 13.2: Schematic illustrating  $\ell_1$ -induced sparsity geometry.

## 13.5 Real World Applications

Linear factor models, including PCA, ICA, and sparse coding, provide interpretable representations of complex data. These techniques underpin many practical systems for data compression, signal processing, and feature extraction.

### 13.5.1 Facial Recognition Systems

Facial recognition systems benefit significantly from linear factor models, where eigenfaces for face recognition represent one of the earliest successful face recognition systems that used PCA to represent faces efficiently, with each face expressed as a weighted combination of "eigenfaces" (principal components) that reduces storage requirements dramatically by storing just a few dozen numbers per person while maintaining recognition accuracy instead of storing full images. Factor models provide robustness to lighting and expression variations by capturing the most important variations in face appearance (identity) while being less sensitive to less important variations like lighting and expression, making recognition work under different conditions without requiring massive datasets of each person. The compressed representations from factor models enable privacy-preserving face recognition without storing actual face images, providing better privacy protection where the low-dimensional codes contain enough information for matching but can't easily be reversed to reconstruct recognizable faces.

### 13.5.2 Audio Signal Processing

Audio signal processing applications leverage linear factor models for extracting meaning from sound, where music analysis and recommendation systems like Spotify use factor models to decompose songs into latent features including mood, genre, tempo, and instrumentation, with these compact representations enabling efficient similarity search across millions of songs so that when you like a song, the system finds others with similar factor patterns. Modern hearing aids use sparse coding to separate speech from background noise, where the factor model learns to represent speech efficiently with few active components while requiring many more components for noise, enabling selective amplification of speech while suppressing noise. Source separation applications isolate individual instruments in music recordings or separate overlapping speakers using independent component analysis (ICA), enabling remixing old recordings, improving audio quality, and creating karaoke tracks from normal songs.

### 13.5.3 Anomaly Detection in Systems

Anomaly detection systems use linear factor models to find unusual patterns in complex data, where network intrusion detection systems use factor models to represent normal network traffic patterns compactly, with unusual activities like potential attacks not fitting well into this low-dimensional representation and triggering alerts, enabling detection of novel attacks without explicitly programming rules for every possible threat. Manufacturing quality control applications use factor models to analyze sensor data from equipment, where normal operations cluster in low-dimensional space and deviations indicate problems like tool wear, calibration drift, or defects, enabling early detection that prevents defective products and costly equipment damage. Healthcare monitoring systems use wearable devices to compress continuous health metrics including heart rate, activity, and sleep patterns into factor representations, allowing doctors to spot concerning trends without examining raw data streams and enabling anomaly detection that alerts patients to unusual patterns warranting attention.

### 13.5.4 Practical Advantages

Linear factor models remain valuable because they provide interpretability where components often correspond to meaningful concepts, efficiency by dramatically reducing data storage and transmission costs, generalization by capturing essential patterns while ignoring noise, and serve as foundations for more complex deep learning systems. These applications demonstrate that relatively simple factor models continue to provide practical value, either standalone or as components within larger deep learning systems, where their interpretability makes them particularly valuable for understanding data structure and for applications where transparency is important, while their efficiency makes them suitable for resource-constrained environments and real-time applications.

## Key Takeaways

### Key Takeaways 13

- **Linear-Gaussian latent models** provide probabilistic PCA and factor analysis with uncertainty estimates.
- **EM algorithm** alternates inference of latents and parameter updates, exploiting conjugacy.
- **Identifiability** requires fixing rotations/scales; solutions are not unique without conventions.
- **Bridge to deep models:** Linear factors motivate nonlinear representation learning and VAEs.

## Exercises

### Easy

*Exercise 13.1 (PPCA vs PCA).* Contrast PCA and probabilistic PCA in assumptions and outputs.

**Hint:**

*Deterministic vs. probabilistic view; noise model; likelihood.*

*Exercise 13.2 (Dimensionality Choice).* List two heuristics to select latent dimensionality.

**Hint:**

*Explained variance, information criteria.*

*Exercise 13.3 (Gaussian Conditionals).* Recall the conditional of a joint Gaussian and its role in E-steps.

**Hint:**

*Use block-partitioned mean and covariance formulas.*

*Exercise 13.4 (Rotation Ambiguity).* Explain why factor loadings are identifiable only up to rotation.

**Hint:**

*Orthogonal transforms preserve latent covariance.*

### Medium

*Exercise 13.5 (PPCA Likelihood).* Derive the log-likelihood of PPCA and the M-step for  $\sigma^2$ .

**Hint:**

*Marginalise latents; differentiate w.r.t. variance.*

*Exercise 13.6 (EM for FA).* Write the E and M steps for Factor Analysis with diagonal noise.

**Hint:**

*Use expected sufficient statistics of latents.*

**Hard**

*Exercise 13.7 (Equivalence of PPCA Solution).* Show that the MLE loading matrix spans the top- $k$  eigenvectors of the sample covariance.

**Hint:**

*Use spectral decomposition of covariance.*

*Exercise 13.8 (Nonlinear Extension).* Sketch how to generalise linear factor models to VAEs.

**Hint:**

*Replace linear-Gaussian with neural encoder/decoder.*

*Exercise 13.9 (Advanced Topic 1).* Explain a key concept from this chapter and its practical applications.

**Hint:**

*Consider the theoretical foundations and real-world implications.*

*Exercise 13.10 (Advanced Topic 2).* Analyse the relationship between different techniques covered in this chapter.

**Hint:**

*Look for connections and trade-offs between methods.*

*Exercise 13.11 (Advanced Topic 3).* Design an experiment to test a hypothesis related to this chapter's content.

**Hint:**

*Consider experimental design, metrics, and potential confounding factors.*

*Exercise 13.12 (Advanced Topic 4).* Compare different approaches to solving a problem from this chapter.

**Hint:**

*Consider computational complexity, accuracy, and practical considerations.*

*Exercise 13.13 (Advanced Topic 5).* Derive a mathematical relationship or prove a theorem from this chapter.

**Hint:**

*Start with the definitions and work through the logical steps.*

*Exercise 13.14 (Advanced Topic 6).* Implement a practical solution to a problem discussed in this chapter.

**Hint:**

*Consider the implementation details and potential challenges.*

*Exercise 13.15 (Advanced Topic 7).* Evaluate the limitations and potential improvements of techniques from this chapter.

**Hint:**

*Consider both theoretical limitations and practical constraints.*



# Chapter 14

## Autoencoders

This chapter explores autoencoders, neural networks designed for unsupervised learning through data reconstruction.

### Learning Objectives

1. Autoencoder framework and common variants (denoising, sparse, contractive)
2. Role of bottlenecks and regularization in learning useful representations
3. Training objectives and reconstruction vs. downstream utility evaluation
4. Connection between autoencoders and generative models

## Intuition

Autoencoders are like skilled artists who can recreate a masterpiece from just a few key brushstrokes - they learn to compress complex data into essential features while discarding unnecessary details. The compression process involves reducing the dimensionality of input data while preserving the most important information needed for accurate reconstruction, where the goal is to learn a compact representation that captures the underlying structure and patterns in the data. This compression enables the model to identify and retain only the salient features that are necessary for understanding the data, while discarding noise and redundancies that don't contribute to the essential structure, ultimately surfacing meaningful patterns that can transfer to other tasks and applications.

### 14.1 Undercomplete Autoencoders ◆

Undercomplete autoencoders are neural networks with a bottleneck architecture that forces the model to learn compressed representations by constraining the latent space to have fewer dimensions than the input space.

#### 14.1.1 Architecture

An autoencoder consists of an encoder  $h = f(x)$  that maps input to latent representation and a decoder  $\hat{x} = g(h)$  that reconstructs from latent code, where the encoder compresses the input data into a lower-dimensional representation and the decoder attempts to reconstruct the original input from this compressed representation. Autoencoders differ from PCA in that they can learn non-linear transformations and capture complex patterns in the data, while PCA is limited to linear transformations, making autoencoders more powerful for handling non-linear data structures. Shannon's Information Theory is relevant to autoencoders because it provides the theoretical foundation for understanding how much information can be compressed without loss, where the bottleneck constraint forces the model to learn the most efficient representation that

preserves the essential information needed for reconstruction. The architecture typically consists of a symmetric network where the encoder gradually reduces dimensionality through hidden layers, creating a bottleneck at the latent layer, and the decoder mirrors this structure to reconstruct the original input, with the mathematical formulation being  $\mathbf{h} = f(\mathbf{x}) = \sigma(\mathbf{W}_e \mathbf{x} + \mathbf{b}_e)$  for the encoder and  $\hat{\mathbf{x}} = g(\mathbf{h}) = \sigma(\mathbf{W}_d \mathbf{h} + \mathbf{b}_d)$  for the decoder, where  $\sigma$  is the activation function and the weights are learned through training.

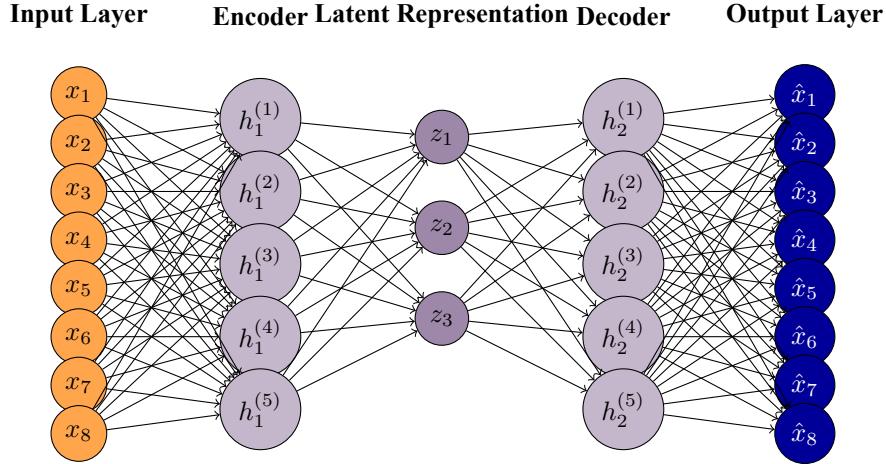


Figure 14.1: Undercomplete autoencoder architecture with bottleneck at latent layer.

### 14.1.2 Training Objective

The training objective is to minimize reconstruction error, where the loss function  $L = \|\mathbf{x} - g(f(\mathbf{x}))\|^2$  measures the difference between the original input and its reconstruction, or more generally  $L = -\log p(\mathbf{x}|g(f(\mathbf{x})))$  for probabilistic reconstruction, where the model learns to encode and decode data by minimizing the reconstruction error between input and output.

### 14.1.3 Undercomplete Constraint

The undercomplete constraint requires that  $\dim(\mathbf{h}) < \dim(\mathbf{x})$ , forcing the autoencoder to learn compressed representations by constraining the latent space to have fewer dimensions than the input space, where this bottleneck forces the model to learn the most important features and discard redundant information, acting as dimensionality reduction similar to PCA but with non-linear transformations.

## 14.2 Regularized Autoencoders ◆

Regularized autoencoders extend basic autoencoders by adding various forms of regularization to encourage learning of more useful and robust representations through constraints on the latent space or training process.

### 14.2.1 Sparse Autoencoders

Sparse autoencoders add a sparsity penalty on hidden activations through the loss function  $L = \|\mathbf{x} - \hat{\mathbf{x}}\|^2 + \lambda \sum_j |h_j|$ , where the L1 penalty encourages learning of sparse, interpretable features by forcing most hidden units to be inactive for any given input, promoting the discovery of meaningful and independent features.

### 14.2.2 Denoising Autoencoders (DAE)

Denoising autoencoders train to reconstruct clean input from corrupted versions by first corrupting the input  $\tilde{\mathbf{x}} \sim q(\tilde{\mathbf{x}}|\mathbf{x})$ , then encoding the corrupted input  $\mathbf{h} = f(\tilde{\mathbf{x}})$ , decoding and reconstructing  $\hat{\mathbf{x}} = g(\mathbf{h})$ , and minimizing the loss  $L = \|\mathbf{x} - \hat{\mathbf{x}}\|^2$ . This approach learns robust representations by forcing the model to recover the original signal from noisy inputs, where corruption

types include additive Gaussian noise, masking that randomly sets inputs to zero, and salt-and-pepper noise, with the model learning to identify and remove these corruptions while preserving the essential structure of the data.

### 14.2.3 Contractive Autoencoders (CAE)

Contractive autoencoders add a penalty on the Jacobian of the encoder through the loss function  $L = \|\mathbf{x} - \hat{\mathbf{x}}\|^2 + \lambda \left\| \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \right\|_F^2$ , where the Frobenius norm penalty encourages locally contractive mappings that are robust to small perturbations by penalizing large gradients in the encoder function.

## 14.3 Variational Autoencoders ◆

Variational Autoencoders (VAEs) are probabilistic generative models that learn to encode data into a latent space and generate new samples by combining the encoder-decoder architecture with variational inference techniques.

### 14.3.1 Probabilistic Framework

VAE is a generative model with the marginal likelihood  $p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$ , where the prior  $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$  is a standard normal distribution and the likelihood  $p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_\theta(\mathbf{z}), \boldsymbol{\sigma}_\theta^2(\mathbf{z})\mathbf{I})$  is parameterized by neural networks that output the mean and variance of the reconstruction distribution.

### 14.3.2 Evidence Lower Bound (ELBO)

Since we cannot directly maximize  $\log p(\mathbf{x})$ , we instead maximize the ELBO  $\mathcal{L} = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})] - D_{KL}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))$ , where the first term encourages accurate reconstruction and the second term regularizes the encoder distribution  $q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\sigma}_\phi^2(\mathbf{x})\mathbf{I})$  to match the prior.

### 14.3.3 Reparameterization Trick

The reparameterization trick enables backpropagation through stochastic nodes by writing sampling as a deterministic function of parameters and an auxiliary noise variable. For a Gaussian encoder

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_\phi(\mathbf{x}), \text{diag}(\boldsymbol{\sigma}_\phi^2(\mathbf{x}))), \quad (14.1)$$

sample via

$$\mathbf{z} = \boldsymbol{\mu}_\phi(\mathbf{x}) + \boldsymbol{\sigma}_\phi(\mathbf{x}) \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (14.2)$$

This yields a low-variance gradient estimator of the ELBO because the expectation over  $q_\phi(\mathbf{z}|\mathbf{x})$  becomes an expectation over  $\boldsymbol{\epsilon}$  independent of  $\phi$ :

$$\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[f(\mathbf{z})] = \nabla_\phi \mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}[f(\boldsymbol{\mu}_\phi(\mathbf{x}) + \boldsymbol{\sigma}_\phi(\mathbf{x}) \odot \boldsymbol{\epsilon})] \quad (14.3)$$

$$= \mathbb{E}_{\boldsymbol{\epsilon}}[\nabla_\phi f(\boldsymbol{\mu}_\phi(\mathbf{x}) + \boldsymbol{\sigma}_\phi(\mathbf{x}) \odot \boldsymbol{\epsilon})]. \quad (14.4)$$

Thus gradients flow through  $\boldsymbol{\mu}_\phi$  and  $\boldsymbol{\sigma}_\phi$  via the deterministic mapping while preserving stochasticity through  $\boldsymbol{\epsilon}$ .

### 14.3.4 Generation

Generation in VAEs involves sampling from the prior  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  and decoding to generate new data, where the decoder network learns to map latent samples to realistic data samples by training on the reconstruction task. This process enables the generation of new samples that follow the learned data distribution, where the quality of generated samples depends on how well the model has learned to capture the underlying data structure and the effectiveness of the latent space representation. The generation process is particularly powerful because it allows for controlled sampling and interpolation in the latent space,

enabling the creation of new data points that maintain the statistical properties of the training data while potentially exploring new combinations of learned features.

---

**Algorithm 7** Variational Autoencoder Training Algorithm
 

---

- 1: **Input:** Dataset  $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N$ , learning rate  $\alpha$
- 2: Initialize encoder parameters  $\phi$  and decoder parameters  $\theta$
- 3: **for** epoch = 1 to max\_epochs **do**
- 4:   **for** batch  $\mathcal{B} \subset \mathcal{D}$  **do**
- 5:     Compute encoder outputs:  $\mu_\phi(\mathbf{x}), \sigma_\phi(\mathbf{x})$
- 6:     Sample  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
- 7:     Reparameterize:  $\mathbf{z} = \mu_\phi(\mathbf{x}) + \sigma_\phi(\mathbf{x}) \odot \epsilon$
- 8:     Compute decoder outputs:  $\mu_\theta(\mathbf{z}), \sigma_\theta(\mathbf{z})$
- 9:     Compute reconstruction loss:  $\mathcal{L}_{rec} = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})]$
- 10:    Compute KL divergence:  $\mathcal{L}_{KL} = D_{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$
- 11:    Total loss:  $\mathcal{L} = \mathcal{L}_{rec} - \mathcal{L}_{KL}$
- 12:    Update parameters:  $\phi \leftarrow \phi - \alpha \nabla_\phi \mathcal{L}, \theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}$
- 13:   **end for**
- 14: **end for**
- 15: **Return** trained encoder  $q(\mathbf{z}|\mathbf{x})$  and decoder  $p(\mathbf{x}|\mathbf{z})$

---

### 14.3.5 Notes and references

VAEs provide a principled probabilistic framework for representation learning and generation, representing a significant milestone in generative modeling that combines the power of neural networks with variational inference techniques. The work by Kingma and Welling in 2013 introduced the reparameterization trick that made VAEs trainable, while subsequent research has extended VAEs to various domains including image generation, text modeling, and molecular design. These models have achieved remarkable success in applications ranging from image compression and denoising to drug discovery and creative applications, demonstrating their versatility and practical impact in modern machine learning systems.

## 14.4 Applications of Autoencoders ◆

Autoencoders find widespread applications in dimensionality reduction, anomaly detection, denoising, and generative modeling, where their ability to learn compressed representations makes them valuable for various machine learning tasks.

### 14.4.1 Dimensionality Reduction

Autoencoders learn compact representations for visualization tasks similar to t-SNE and UMAP, where they can reduce high-dimensional data to lower dimensions while preserving important structure and relationships. They serve as effective preprocessing tools for downstream tasks by extracting meaningful features that can improve the performance of subsequent machine learning models. The learned representations are particularly valuable for feature extraction in domains where the original data is high-dimensional and contains redundant information, enabling more efficient processing and analysis.

### 14.4.2 Anomaly Detection

Autoencoders excel at anomaly detection because high reconstruction error indicates anomalies, where the model learns to reconstruct normal patterns well but struggles with unusual or anomalous data. This principle is applied in fraud detection systems that identify suspicious transactions by measuring reconstruction error, in manufacturing quality control where defects produce high reconstruction errors, and in network intrusion detection where unusual network patterns are flagged based on their reconstruction difficulty.

### 14.4.3 Denoising

Denoising autoencoders (DAEs) remove noise from various types of data by learning to reconstruct clean signals from corrupted inputs, where they are particularly effective for image denoising by learning to identify and remove various types of noise while preserving important image features. They also excel at audio signal processing where they can separate speech from background noise, and in sensor data applications where they can filter out measurement noise while preserving the underlying signal patterns.

### 14.4.4 References

Autoencoders have evolved significantly since their introduction, with key milestones including the development of denoising autoencoders that improved robustness, sparse autoencoders that learned interpretable features, and variational autoencoders that enabled generative modeling. The work by Goodfellow and colleagues has been particularly influential in establishing the theoretical foundations and practical applications of autoencoders, while Prince's contributions have advanced our understanding of their connections to other machine learning techniques. These models have achieved remarkable success in applications ranging from image compression and denoising to drug discovery and creative applications, demonstrating their versatility and practical impact in modern machine learning systems.

## 14.5 Real World Applications

Autoencoders learn to compress and reconstruct data, finding compact representations that capture essential information. This capability enables numerous practical applications in compression, denoising, and anomaly detection.

### 14.5.1 Image and Video Compression

Autoencoders enable efficient storage and transmission of visual data through next-generation image compression that achieves better quality at the same file size or smaller files at the same quality compared to traditional formats like JPEG, where learned compression algorithms matter significantly for websites, cloud storage, and mobile apps where bandwidth and storage costs are substantial. Video streaming optimization applications by Netflix and YouTube experiment with autoencoder-based video compression to stream higher quality video at lower bitrates, reducing buffering, saving bandwidth costs, and enabling HD streaming in areas with limited internet connectivity by learning to preserve perceptually important details humans notice while discarding subtle information we don't. Satellite imagery compression applications handle the terabytes of imagery generated daily by Earth observation satellites, where autoencoder compression reduces transmission bandwidth from space to ground stations, allowing more frequent imagery updates or higher resolution within bandwidth constraints and improving applications from weather forecasting to agriculture monitoring.

### 14.5.2 Denoising and Enhancement

Autoencoders improve signal quality in degraded data through medical image enhancement where denoising autoencoders improve quality of MRI and CT scans, reducing radiation exposure needed for diagnostic-quality images or enabling faster scanning by learning the manifold of healthy tissue appearance and removing noise while preserving medically relevant details like tumor boundaries. Old photo restoration applications use autoencoders to remove scratches, stains, and aging artifacts from old photographs, where the models learn the structure of clean images and infer what damaged regions likely looked like originally, helping preserve family histories and restore historical photographs. Audio enhancement applications clean up audio recordings by removing background noise, hum, or compression artifacts, improving voice clarity in phone calls, enhancing podcast quality, and helping restore old audio recordings, where unlike simple filtering, autoencoders understand speech structure and preserve natural sound.

### 14.5.3 Anomaly Detection

Autoencoders identify unusual patterns in complex systems through credit card fraud detection where they learn to represent normal spending patterns compactly, with fraudulent transactions often not fitting these patterns well and resulting in poor reconstruction, where high reconstruction error flags potential fraud for investigation and catches novel fraud schemes without requiring examples of every possible type of fraud. Industrial equipment monitoring applications use autoencoders to monitor vibration patterns, temperatures, and other sensor data from machinery, where normal operation reconstructs well but unusual patterns indicating bearing wear, misalignment, or impending failure show high reconstruction error, triggering maintenance before catastrophic breakdowns. Cybersecurity threat detection systems use autoencoders trained on normal traffic patterns, where malware, intrusions, and data exfiltration create unusual patterns that reconstruct poorly, alerting security teams and detecting zero-day attacks and insider threats that evade signature-based detection.

### 14.5.4 Why Autoencoders Excel

Autoencoders excel in practical applications because they enable unsupervised learning without requiring labeled examples, just normal data, where they capture essential information compactly through dimensionality reduction and learn underlying structure despite corrupted inputs through noise robustness. Their reconstruction ability allows them to generate clean versions of corrupted data, making them particularly valuable for applications where data quality is important but labels are scarce or expensive to obtain.

### 14.5.5 Autoencoders Compared to other NN Algorithms

Feature	Autoencoder (AE)	Supervised NN (e.g., MLP, CNN)	Generative Adversarial Network (GAN)
Primary Goal	Data Compression, Feature Learning, Anomaly Detection	Classification or Regression (Mapping input to a specific output label)	Data Generation (Creating new, realistic samples)
Learning Type	Unsupervised (Learns from data structure itself)	Supervised (Requires labelled data)	Unsupervised (Learns via a competitive game)
Input/Output	Input ( $x$ ) = Output ( $x'$ ) (It learns the identity function under constraints)	Input ( $x$ ) → Output ( $y$ ) (Target is a label)	Input (Random Noise) → Output (Realistic Data Sample)
Loss Function	Reconstruction Error (e.g., Mean Squared Error)	Classification Loss (e.g., Cross-Entropy) or Regression Loss (e.g., MSE)	Adversarial Loss (between Generator and Discriminator)
Key Use Case	Dimensionality reduction, Denoising, Anomaly detection (poor reconstruction = anomaly)	Image recognition, Natural Language Processing, Stock prediction	Image synthesis, Deepfakes, Generating realistic data

Table 14.1: Comparison of Autoencoders with other neural network algorithms.

These applications show how autoencoders bridge classical compression and modern deep learning, providing practical solutions for data efficiency, quality enhancement, and anomaly detection.

## Key Takeaways

### Key Takeaways 14

- **Bottlenecks and noise** force representations to capture structure, not memorisation.
- **Regularised variants** (denoising, sparse, contractive) improve robustness and usefulness.
- **Utility beyond reconstruction:** learned codes transfer to downstream tasks.

## Exercises

### Easy

*Exercise 14.1 (Undercomplete AE).* Explain why undercomplete AEs can avoid trivial identity mapping.

**Hint:**

*Bottleneck limits capacity.*

*Exercise 14.2 (Denoising Noise).* How does noise type affect learned features?

**Hint:**

*Gaussian vs. masking vs. salt-and-pepper.*

*Exercise 14.3 (Sparse Codes).* List benefits of sparsity in latent codes.

**Hint:**

*Interpretability, robustness, compression.*

*Exercise 14.4 (Contractive Penalty).* What does a Jacobian penalty encourage?

**Hint:**

*Local invariance.*

### Medium

*Exercise 14.5 (Loss Choices).* Compare MSE vs. cross-entropy for images.

**Hint:**

*Data scale/likelihood assumptions.*

*Exercise 14.6 (Regularisation Trade-offs).* Contrast denoising vs. contractive penalties.

**Hint:**

*Noise robustness vs. local smoothing.*

## Hard

*Exercise 14.7 (Jacobian Penalty).* Derive gradient of contractive loss w.r.t. encoder parameters.

**Hint:**

*Chain rule through Jacobian norm.*

*Exercise 14.8 (Generative Link).* Explain links between AEs and VAEs/flows.

**Hint:**

*Likelihood vs. reconstruction objectives.*

*Exercise 14.9 (Advanced Topic 1).* Explain a key concept from this chapter and its practical applications.

**Hint:**

*Consider the theoretical foundations and real-world implications.*

*Exercise 14.10 (Advanced Topic 2).* Analyse the relationship between different techniques covered in this chapter.

**Hint:**

*Look for connections and trade-offs between methods.*

*Exercise 14.11 (Advanced Topic 3).* Design an experiment to test a hypothesis related to this chapter's content.

**Hint:**

*Consider experimental design, metrics, and potential confounding factors.*

*Exercise 14.12 (Advanced Topic 4).* Compare different approaches to solving a problem from this chapter.

**Hint:**

*Consider computational complexity, accuracy, and practical considerations.*

*Exercise 14.13 (Advanced Topic 5).* Derive a mathematical relationship or prove a theorem from this chapter.

**Hint:**

*Start with the definitions and work through the logical steps.*

*Exercise 14.14 (Advanced Topic 6).* Implement a practical solution to a problem discussed in this chapter.

**Hint:**

*Consider the implementation details and potential challenges.*

*Exercise 14.15 (Advanced Topic 7).* Evaluate the limitations and potential improvements of techniques from this chapter.

**Hint:**

*Consider both theoretical limitations and practical constraints.*

# Chapter 15

# Representation Learning

This chapter discusses the central challenge of deep learning: learning meaningful representations from data.

## Learning Objectives

1. Representations and desirable properties (invariance, disentanglement, sparsity)
2. Supervised, self-supervised, and contrastive learning objectives
3. Representation evaluation via linear probes and transfer learning
4. Information-theoretic perspectives in empirical practice

## Intuition

Good representations separate task-relevant factors from nuisance variability, where learning signals that compare positive and negative pairs, or predict masked content, shape geometry in embedding spaces to reflect semantics. For example, in facial recognition systems, a good representation would capture identity-related features like facial structure and bone geometry while being invariant to lighting conditions, pose variations, and facial expressions. A good representation is like a highly efficient, specialized library catalogue designed for a specific purpose, where it organizes and indexes information in a way that makes the most relevant details immediately accessible while filtering out irrelevant noise and redundancy.

## Representation Function

Representation learning can be mathematically defined as learning a function that maps raw input data from a high-dimensional space to a lower-dimensional, more informative space. Mathematically, representation learning aims to find a mapping function  $\mathbf{f}$  that transforms the input data  $\mathbf{x} \in \mathbb{R}^D$  into a representation  $\mathbf{z} \in \mathbb{R}^d$ , where typically the input dimension  $D$  is much greater than the representation dimension  $d$  ( $D \gg d$ ). The key is that  $\mathbf{z}$  is designed to capture the salient semantic factors of  $\mathbf{x}$  while discarding noise and redundancy, where the function  $\mathbf{f}$  is learned by defining an objective function that encourages  $\mathbf{z}$  to possess desirable properties like invariance, disentanglement, and sparsity.

### 15.1 What Makes a Good Representation? ◆

In the context of deep learning, a representation is a learned transformation that maps high-dimensional raw data into a lower-dimensional space that captures the essential semantic information needed for downstream tasks, where the goal is to extract meaningful features that are useful for classification, generation, or other machine learning objectives.

### 15.1.1 Desirable Properties

Disentanglement refers to the property where different factors of variation are separated in the representation space, where changes in one dimension affect only one factor, making the representation easier to interpret and manipulate by allowing independent control over different aspects of the data. Invariance ensures that the representation remains unchanged under irrelevant transformations, such as translation and rotation invariance for objects or speaker invariance for speech content, where the model learns to focus on task-relevant features while ignoring nuisance variability. Smoothness means that similar inputs have similar representations, enabling generalization to unseen data and supporting interpolation between known examples, where the representation space maintains a coherent structure that reflects the underlying data manifold. Sparsity refers to the property where few features are active for each input, providing computational efficiency by reducing the number of computations needed and improving interpretability by focusing on the most important features for each specific input.

### 15.1.2 Manifold Hypothesis

The manifold hypothesis states that natural data lies on low-dimensional manifolds embedded in high-dimensional space, where despite the high dimensionality of raw data, the underlying structure can be captured by a much smaller number of parameters. Deep learning learns to discover the manifold structure by identifying the low-dimensional subspace that contains the essential information, where the model maps data to meaningful coordinates on the manifold that correspond to the underlying factors of variation in the data, enabling efficient representation and manipulation of complex high-dimensional data.

### 15.1.3 Notes and references

Desirable properties are discussed in modern deep learning texts, where disentanglement and invariance connect to inductive biases and data augmentation, representing key milestones in understanding how to design effective representation learning systems. The work by Goodfellow and colleagues has been particularly influential in establishing the theoretical foundations for representation learning, while Prince's contributions have advanced our understanding of how these properties emerge in practice. These models have achieved remarkable success in applications ranging from computer vision and natural language processing to scientific discovery and creative applications, demonstrating their versatility and practical impact in modern machine learning systems.

## 15.2 Transfer Learning and Domain Adaptation ◆

Transfer learning and domain adaptation enable the reuse of learned representations across different tasks and domains, where knowledge from a source task is leveraged to improve performance on a target task, even when the target task has limited labeled data or operates in a different domain.

### 15.2.1 Transfer Learning

Transfer learning leverages knowledge from a source task to improve performance on a target task, where the learned representations from the source task are reused to accelerate learning on the target task. Feature extraction involves pre-training on a large dataset like ImageNet, freezing the convolutional layers to preserve the learned features, and training only the final classification layers on the target task, where this approach is particularly effective when the target task has limited labeled data. Fine-tuning starts with a pre-trained model and continues training on the target task with a lower learning rate to avoid destroying the learned features, where optionally freezing early layers can help preserve the low-level features while allowing the higher-level features to adapt to the target task.

### 15.2.2 Domain Adaptation

Domain adaptation addresses the challenge when training and test distributions differ, where the model needs to adapt to the target domain while maintaining performance on the source domain. Domain-adversarial training learns domain-invariant features by using an adversarial training procedure that encourages the feature extractor to learn representations that are

indistinguishable between source and target domains, where this approach helps the model generalize across different data distributions. Self-training uses confident predictions on the target domain to generate pseudo-labels for unlabeled target data, where the model is iteratively trained on these pseudo-labels to improve its performance on the target domain. Multi-task learning involves joint training on both domains, where the model learns to perform well on both the source and target tasks simultaneously, enabling better transfer of knowledge between domains.

### 15.2.3 Few-Shot Learning

Few-shot learning addresses the challenge of learning from few examples per class, where the model needs to quickly adapt to new tasks with limited labeled data. Meta-learning approaches like MAML (Model-Agnostic Meta-Learning) learn to learn quickly by training the model to adapt to new tasks with just a few gradient steps, where the meta-learner learns initialization parameters that enable fast adaptation to new tasks. Prototypical networks learn a metric space where examples from the same class are close together and examples from different classes are far apart, where classification is performed by computing distances to class prototypes in this learned metric space. Matching networks use attention-based comparison to find the most similar examples in the support set for each query example, where the model learns to attend to relevant examples and make predictions based on these similarities.

## 15.3 Self-Supervised Learning ◆

Self-supervised learning learns representations without manual labels by solving pretext tasks that can be automatically generated from the data itself, where the model learns useful representations by predicting parts of the input from other parts or by solving auxiliary tasks that don't require human annotation.

### 15.3.1 Pretext Tasks

Pretext tasks for images include rotation prediction where the model learns to predict the rotation angle of an image, jigsaw puzzle where the model arranges shuffled patches to learn spatial relationships, colorization where the model predicts colors from grayscale images to learn color semantics, and inpainting where the model fills masked regions to learn object structure and context. For text, pretext tasks include masked language modeling where the model predicts masked words as in BERT, next sentence prediction where the model determines if sentences are consecutive to learn discourse relationships, and autoregressive generation where the model predicts the next token as in GPT to learn language modeling capabilities.

### 15.3.2 Benefits

Self-supervised learning provides several key benefits including the ability to leverage unlabeled data that is often abundant and cheap to obtain, where the model can learn from vast amounts of data without requiring expensive human annotation. The learned representations are often general-purpose and transfer well to downstream tasks, where the model learns to capture the underlying structure of the data rather than task-specific patterns. Self-supervised learning often outperforms supervised pre-training on downstream tasks, where the learned representations are more robust and generalizable because they are not biased toward specific labeled examples.

## 15.4 Contrastive Learning ◆

Contrastive learning learns representations by contrasting positive and negative pairs, where the model learns to pull similar examples together and push dissimilar examples apart in the representation space, enabling the discovery of meaningful semantic structure without explicit supervision.

### 15.4.1 Core Idea

The core idea of contrastive learning is to maximize agreement between different views of the same data (positive pairs) while minimizing agreement with other data (negative pairs), where this approach forces the model to learn representations that are invariant to irrelevant transformations while being sensitive to semantic differences. This is achieved by training the model to distinguish between positive pairs that should have similar representations and negative pairs that should have different representations, where the contrastive loss encourages the model to learn a representation space where similar examples are close together and dissimilar examples are far apart. The key insight is that by learning to distinguish between positive and negative pairs, the model naturally learns to capture the underlying semantic structure of the data without requiring explicit labels.

### 15.4.2 SimCLR Framework

The SimCLR framework applies two random augmentations to each image, encodes both views using the same encoder network  $\mathbf{z}_i = f(\mathbf{x}_i)$  and  $\mathbf{z}_j = f(\mathbf{x}_j)$ , and minimizes the contrastive loss (NT-Xent) to learn invariances.

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{I}[k \neq i] \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}. \quad (15.1)$$

#### Variables.

- $\mathbf{z}_i, \mathbf{z}_j$ : embeddings of two augmented views of the same image.
- $\text{sim}(\cdot, \cdot)$ : cosine similarity between normalized embeddings.
- $\tau > 0$ : temperature scaling the softness of the distribution.
- $2N$ : number of views in a batch (two per original sample); negatives are the other  $2N - 2$  views.
- $\mathbb{I}[\cdot]$ : indicator that excludes the anchor itself from the denominator.

### 15.4.3 MoCo (Momentum Contrast)

MoCo (Momentum Contrast) uses a momentum encoder and a queue of negative samples for efficiency, where the momentum encoder is updated using exponential moving average of the main encoder parameters to maintain consistency in the representation space. The queue stores negative samples from previous batches to provide a large and diverse set of negative examples without requiring large batch sizes, where this approach enables efficient contrastive learning with limited computational resources while maintaining high-quality representations.

### 15.4.4 BYOL (Bootstrap Your Own Latent)

BYOL (Bootstrap Your Own Latent) surprisingly works without negative samples by using an online network that is updated by gradients, a target network that is updated using momentum, and a prediction head on the online network that learns to predict the target network's output. This approach avoids the need for negative samples by learning to predict the target representation from the online representation, where the target network provides a moving target that encourages the online network to learn consistent representations across different augmentations.

### 15.4.5 Applications

Contrastive learning has achieved state-of-the-art results in image classification where the learned representations transfer well to downstream classification tasks, object detection where the learned features enable accurate localization and classification of objects, and segmentation where the learned representations capture fine-grained spatial information needed for pixel-level predictions. The approach is particularly valuable for medical imaging with limited labels, where the learned representations can be fine-tuned with minimal labeled data while maintaining high performance on diagnostic tasks.

## 15.5 Real World Applications

Representation learning—automatically discovering useful features from raw data—is fundamental to modern deep learning success. Good representations make downstream tasks easier and enable transfer learning across domains.

### 15.5.1 Transfer Learning in Computer Vision

Transfer learning in computer vision reuses learned representations to save time and data, where medical imaging with limited data benefits from starting with representations learned from millions of general images like ImageNet and then fine-tuning on medical data, enabling networks that learned to recognize textures and shapes in everyday photos to adapt to recognize pathologies in X-rays with just a small medical dataset. Custom object detection for businesses allows retailers to detect their specific products on shelves and manufacturers to identify particular defects by using pre-trained vision models and fine-tuning with just hundreds of examples, where the learned representations of edges, textures, and objects transfer effectively, making custom vision systems practical for small businesses. Wildlife monitoring applications use camera traps to monitor endangered species, generating millions of images where transfer learning enables creating species classifiers with limited labeled examples, accelerating research without requiring biologists to manually label vast datasets.

### 15.5.2 Natural Language Processing

Learned language representations revolutionize text applications through multilingual models that learn representations capturing meaning across languages, where a model trained on English, Spanish, and Chinese text learns that "cat," "gato," and "mao" represent similar concepts, enabling zero-shot translation and allowing improvements in high-resource languages to benefit low-resource languages automatically. Domain adaptation applications use customer service chatbots that employ language models pre-trained on general text and then fine-tuned on company-specific conversations, where the general language understanding including grammar, reasoning, and world knowledge transfers while fine-tuning adds domain expertise, making sophisticated chatbots feasible without training from scratch. Sentiment analysis for brands allows companies to monitor social media sentiment about their products by using general text representations learned from billions of documents and then adapting to specific brand vocabulary, providing accurate sentiment analysis even for newly launched products.

### 15.5.3 Cross-Modal Representations

Cross-modal representations learn representations spanning multiple modalities, where image-text search systems like Google Images allow searching photos using text descriptions by requiring representations where images and text descriptions of the same concept are similar, where models learn joint representations by training on millions of image-caption pairs, enabling finding relevant images even for queries with no exact text matches. Video understanding applications use YouTube's recommendation and search systems that learn representations combining visual content, audio, speech transcripts, and metadata, where these multi-modal representations understand videos better than any single modality alone, improving search relevance and recommendations. Accessibility tools for visually impaired users generate descriptions of images on web pages using cross-modal representations trained on image-caption pairs, enabling generating relevant, helpful descriptions automatically and making the web more accessible.

### 15.5.4 Impact of Good Representations

Good representations matter because they provide data efficiency by enabling the solution of new tasks with less labeled data, where the learned representations capture the essential structure of the data and can be reused across different tasks. They enable generalization by providing better performance on diverse, real-world examples, where the learned representations are robust to variations in the data and can handle unseen examples effectively. Knowledge transfer allows expertise learned on one task to help others, where the learned representations can be fine-tuned for new tasks without starting from scratch. Semantic understanding captures meaningful structure in data, where the learned representations reflect the underlying semantic relationships in the data, enabling better understanding and manipulation of the data.

These applications demonstrate that representation learning is not just a theoretical concept—it's the foundation enabling practical deep learning with limited data and computational resources.

## Key Takeaways

### Key Takeaways 15

- **Representation quality** is judged by transfer and linear separability.
- **Self-supervision** shapes embeddings via predictive or contrastive signals.
- **Evaluation matters:** consistent probes and protocols enable fair comparison.

## Exercises

### Easy

*Exercise 15.1 (Encoder-Decoder Symmetry).* Why share architecture between  $q(z|x)$  and  $p(x|z)$ ?

**Hint:**

*Computation; conceptual symmetry.*

*Exercise 15.2 (KL in ELBO).* State the role of  $D_{KL}(q||p)$  in VAE training.

**Hint:**

*Regularisation; posterior matching.*

*Exercise 15.3 (Reparameterisation Trick).* Explain why reparameterisation enables gradient flow.

**Hint:**

*Sampling vs. deterministic path.*

*Exercise 15.4 (Prior Choice).* Justify Gaussian prior for VAE latents.

**Hint:**

*Tractability; simplicity; universality.*

### Medium

*Exercise 15.5 (ELBO Derivation).* Derive the ELBO from Jensen's inequality.

**Hint:**

$\log \mathbb{E}[X] \geq \mathbb{E}[\log X]$ .

*Exercise 15.6 (Beta-VAE).* Explain how  $\beta$ -VAE encourages disentanglement.

**Hint:**

*Weighted KL penalty; independence.*

**Hard**

*Exercise 15.7 (Posterior Collapse).* Analyse conditions causing posterior collapse and propose mitigation.

**Hint:**

*Strong decoder; KL annealing; free bits.*

*Exercise 15.8 (Importance-Weighted ELBO).* Derive the importance-weighted ELBO and show it tightens the bound.

**Hint:**

*Multiple samples; log-mean-exp.*

*Exercise 15.9 (Advanced Topic 1).* Explain a key concept from this chapter and its practical applications.

**Hint:**

*Consider the theoretical foundations and real-world implications.*

*Exercise 15.10 (Advanced Topic 2).* Analyse the relationship between different techniques covered in this chapter.

**Hint:**

*Look for connections and trade-offs between methods.*

*Exercise 15.11 (Advanced Topic 3).* Design an experiment to test a hypothesis related to this chapter's content.

**Hint:**

*Consider experimental design, metrics, and potential confounding factors.*

*Exercise 15.12 (Advanced Topic 4).* Compare different approaches to solving a problem from this chapter.

**Hint:**

*Consider computational complexity, accuracy, and practical considerations.*

*Exercise 15.13 (Advanced Topic 5).* Derive a mathematical relationship or prove a theorem from this chapter.

**Hint:**

*Start with the definitions and work through the logical steps.*

*Exercise 15.14 (Advanced Topic 6).* Implement a practical solution to a problem discussed in this chapter.

**Hint:**

*Consider the implementation details and potential challenges.*

*Exercise 15.15 (Advanced Topic 7).* Evaluate the limitations and potential improvements of techniques from this chapter.

**Hint:**

*Consider both theoretical limitations and practical constraints.*



# Chapter 16

## Structured Probabilistic Models for DL

This chapter covers graphical models and their integration with deep learning.

### ⌚ Learning Objectives

1. Directed and undirected graphical models and conditional independencies
2. Neural networks with graphical structures for hybrid models
3. Basic inference and when approximations are required
4. Learning objectives for structured prediction tasks

## Intuition

Graphical structure encodes assumptions about which variables interact, where neural components capture complex local relationships while the graph constrains global behavior, aiding data efficiency and interpretability. For example, in medical diagnosis, a graphical model can encode the causal relationships between symptoms, diseases, and treatments, where neural networks learn complex patterns in individual symptoms while the graph structure ensures that the diagnosis follows logical medical reasoning. A graphical model is like a city's traffic system, where the neural networks are the individual vehicles that can navigate complex local routes, while the graph structure provides the overall traffic rules and road network that ensures efficient and logical flow throughout the entire system.

### 16.1 Graphical Models ★

Graphical models provide a powerful framework for representing complex probability distributions using graph structures, where nodes represent random variables and edges encode probabilistic dependencies, enabling efficient representation and inference in high-dimensional spaces.

#### 16.1.1 Motivation

Graphical models represent complex probability distributions using graphs, where nodes represent random variables and edges encode probabilistic dependencies, enabling the efficient representation of high-dimensional joint distributions through factorization. This approach is particularly valuable in deep learning because it allows us to encode domain knowledge and structural assumptions directly into the model, where the graph structure provides inductive biases that guide the learning process and improve generalization. The graphical representation makes the model more interpretable by explicitly showing which variables interact and how, where this transparency is crucial for understanding model behavior and debugging complex

systems. Furthermore, graphical models enable efficient inference by exploiting the conditional independence structure encoded in the graph, where this computational efficiency is essential for practical applications with large-scale data.

### 16.1.2 Bayesian Networks

Bayesian networks use directed acyclic graphs (DAGs) to represent conditional dependencies, where the joint probability distribution factorizes as  $p(\mathbf{x}) = \prod_{i=1}^n p(x_i | \text{Pa}(x_i))$  with  $\text{Pa}(x_i)$  being the parents of  $x_i$ . In deep learning, Bayesian networks are particularly useful for modeling causal relationships and incorporating domain knowledge. The Naive Bayes classifier exemplifies this approach with

$$p(y, \mathbf{x}) = p(y) \prod_{i=1}^d p(x_i | y). \quad (16.1)$$

**Variables.**  $y$ : class label;  $\mathbf{x} = (x_1, \dots, x_d)$ : features;  $p(y)$ : class prior;  $p(x_i | y)$ : class-conditional likelihoods.

### 16.1.3 Markov Random Fields

Markov Random Fields use undirected graphs with potential functions. The joint distribution factorizes as

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c), \quad (16.2)$$

where  $\mathcal{C}$  are cliques,  $\psi_c$  are nonnegative potentials over variables in clique  $c$ , and  $Z$  is the partition function ensuring normalization.

## 16.2 Inference in Graphical Models ★

Inference in graphical models involves computing marginal probabilities and conditional distributions from the joint distribution, where the goal is to answer queries about the model's behavior given observed evidence, enabling prediction and decision-making in complex probabilistic systems.

### 16.2.1 Exact Inference

Exact inference methods compute exact marginal probabilities and conditional distributions, where variable elimination marginalizes variables sequentially by summing over their possible values, while belief propagation uses message passing on tree-structured graphs to efficiently compute marginals. The complexity of exact inference grows exponentially with the tree-width of the graph, making it often intractable for large, densely connected models, where this computational bottleneck motivates the need for approximate inference methods in practical applications. Despite their computational limitations, exact inference methods provide valuable theoretical foundations and are essential for understanding the structure of probabilistic models, where they serve as benchmarks for evaluating the quality of approximate methods.

### 16.2.2 Approximate Inference

Approximate inference methods provide computationally tractable alternatives to exact inference, where variational inference optimizes a tractable approximation to the true posterior distribution by minimizing the KL divergence between the approximation and the target distribution. Sampling methods use Monte Carlo approaches to approximate expectations and marginal probabilities by drawing samples from the distribution, where these methods are particularly useful for high-dimensional models where exact inference is intractable. Loopy belief propagation extends belief propagation to graphs with cycles by iteratively passing messages until convergence, where this approach provides approximate solutions for complex graphical models that would otherwise be computationally intractable.

## 16.3 Deep Learning and Structured Models ★

Deep learning and structured models combine the representational power of neural networks with the structural constraints of graphical models, where neural networks learn complex feature representations while graphical models enforce structural constraints on the output space, enabling the modeling of complex, structured data with both local and global dependencies.

### 16.3.1 Structured Output Prediction

Structured output prediction uses graphical models to model output structure, where the goal is to predict complex, structured outputs that satisfy constraints and dependencies. Conditional Random Fields (CRFs) provide a powerful framework with

$$p(\mathbf{y} | \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp \left( \sum_c \mathbf{w}^\top \phi_c(\mathbf{x}, \mathbf{y}_c) \right). \quad (16.3)$$

**Variables.**  $\mathbf{x}$ : input;  $\mathbf{y}$ : structured output;  $c$ : cliques/factors;  $\phi_c$ : feature functions on  $(\mathbf{x}, \mathbf{y}_c)$ ;  $\mathbf{w}$ : weights;  $Z(\mathbf{x})$ : partition function ensuring normalization.

### 16.3.2 Structured Prediction with Neural Networks

Structured prediction with neural networks combines the representational power of neural networks with the structural constraints of graphical models, where the neural network extracts rich feature representations from the input data while the graphical model enforces structural constraints on the output space. The approach typically involves feature extraction using CNNs or RNNs to learn complex patterns in the input data, followed by structured inference using a CRF layer that ensures the output satisfies the required structural constraints, where the entire system is trained end-to-end using backpropagation to optimize both the feature extraction and structured prediction components simultaneously. An example is CNN-CRF for semantic segmentation, where the CNN learns to extract visual features from images while the CRF layer ensures that the predicted segmentation labels form spatially coherent regions, resulting in more accurate and visually plausible segmentations.

### 16.3.3 Neural Module Networks

Neural Module Networks compose neural modules based on program structure for visual reasoning, where the model learns to decompose complex visual reasoning tasks into simpler sub-tasks that can be solved by specialized neural modules. This approach enables the model to handle compositional reasoning tasks by learning to combine different modules in a structured way, where each module is responsible for a specific type of reasoning operation such as object detection, spatial relationships, or attribute recognition. The modular design allows the model to generalize to new combinations of reasoning operations and provides interpretability by showing which modules are used for each step of the reasoning process.

### 16.3.4 Graph Neural Networks

Graph Neural Networks (GNNs) operate on data structured as graphs via iterative message passing. Let  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  with node features  $\mathbf{X} \in \mathbb{R}^{|\mathcal{V}| \times D}$  and adjacency matrix  $\mathbf{A}$ . At layer  $l$ , each node  $v$  aggregates neighbor information and updates its representation:

$$\mathbf{m}_{\mathcal{N}(v)}^{(l)} = \text{AGGREGATE}^{(l)}(\{\mathbf{h}_u^{(l-1)} : u \in \mathcal{N}(v)\}), \quad (16.4)$$

$$\mathbf{h}_v^{(l)} = \text{UPDATE}^{(l)}(\mathbf{h}_v^{(l-1)}, \mathbf{m}_{\mathcal{N}(v)}^{(l)}), \quad (16.5)$$

where AGGREGATE is permutation-invariant (e.g., sum/mean/max) and UPDATE is learnable (e.g., linear map + nonlinearity). A common layer is the GCN:

$$\mathbf{H}^{(l)} = \sigma(\tilde{\mathbf{D}}^{-1/2} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-1/2} \mathbf{H}^{(l-1)} \mathbf{W}^{(l)}), \quad (16.6)$$

with  $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ ,  $\tilde{\mathbf{D}}_{ii} = \sum_j \tilde{A}_{ij}$ ,  $\mathbf{W}^{(l)}$  learnable, and  $\sigma$  an activation. After  $L$  layers,  $\mathbf{h}_v^{(L)}$  aggregates information from the  $L$ -hop neighborhood.

## 16.4 Real World Applications

Structured probabilistic models capture dependencies and uncertainties in complex systems. These models enable reasoning under uncertainty and provide principled frameworks for decision-making in real-world applications.

### 16.4.1 Autonomous Vehicle Decision Making

Autonomous vehicle decision making requires reasoning about uncertainties in complex environments, where predicting pedestrian behavior involves capturing uncertainty about pedestrian intentions based on their position, posture, and gaze direction, enabling the vehicle to make conservative decisions by slowing down when a pedestrian might cross rather than assuming they won't. Sensor fusion with uncertainty combines cameras, radar, and lidar data using probabilistic graphical models that weigh each sensor according to reliability in current conditions, where cameras work poorly in fog while radar penetrates fog better, providing robust perception despite individual sensor limitations. Planning under uncertainty involves route planning that considers uncertain travel times due to traffic, weather, and road conditions, where structured probabilistic models help vehicles balance expected arrival time with reliability and provide realistic time estimates.

### 16.4.2 Medical Diagnosis and Treatment

Medical diagnosis and treatment require careful uncertainty quantification, where Bayesian diagnosis systems handle the uncertainty inherent in medical diagnosis by encoding relationships between symptoms, diseases, and test results using structured probabilistic models that compute probability distributions over possible diagnoses, helping doctors order appropriate tests and consider differential diagnoses systematically. Personalized treatment planning uses probabilistic models that integrate genetic markers, tumor characteristics, and treatment histories to estimate probability distributions over treatment outcomes, where doctors use these estimates to discuss risks and benefits with patients, enabling informed shared decisions about treatment options. Drug interaction modeling addresses the risks faced by patients taking multiple medications by using structured models that capture dependencies between drugs while considering individual patient factors like age, kidney function, and genetics to estimate risk probabilities, enabling safer prescribing especially for elderly patients on many medications.

### 16.4.3 Natural Language Understanding

Natural language understanding involves handling the inherent ambiguity and structure of language, where machine translation quality uses probabilistic models to capture ambiguity in words that have multiple possible translations depending on context, with structured models representing sentence structure helping to select appropriate translations and providing confidence estimates for different interpretations, enabling highlighting of uncertain translations for human review. Information extraction requires understanding relationships between entities to extract structured information about who did what to whom, when, and where, where probabilistic graphical models capture these dependencies and provide uncertainty estimates about extracted information, enabling news aggregators to reconcile potentially conflicting reports from multiple sources. Voice assistant intent recognition handles ambiguous queries like "Book a table for two" that could mean restaurant reservations or furniture arrangements, where structured models use conversation context and user history to estimate intent probabilities, asking clarifying questions when uncertainty is high rather than guessing incorrectly.

### 16.4.4 Social Media Network Analysis

Social media networks provide rich examples of how deep learning and graph neural networks are applied to understand complex social structures and behaviors. In social media platforms like Facebook, Twitter, and LinkedIn, GNNs are used to model user interactions, content propagation, and community detection, where the graph structure represents users as nodes and relationships like friendships, follows, or interactions as edges. These models can predict user behavior, identify influential

users, detect fake news propagation, and recommend relevant content by learning representations that capture both individual user characteristics and their position within the social network structure.

### 16.4.5 Value of Structured Models

Structured models provide several key advantages including interpretability where the model structure reflects domain knowledge and causal relationships, enabling users to understand how the model makes decisions and what factors influence the predictions. Uncertainty quantification provides principled probability estimates that are crucial for decision-making in high-stakes applications, where knowing not just what the model predicts but how confident it is in that prediction can be the difference between success and failure. Data efficiency is achieved through structure that reduces parameters and sample complexity, where the explicit modeling of dependencies allows the model to learn more effectively from limited data. Reasoning capabilities enable inference, prediction, and decision-making under uncertainty, where the structured approach allows the model to handle complex scenarios that would be difficult for unstructured models. These applications show how structured probabilistic models provide principled frameworks for dealing with uncertainty in safety-critical and high-stakes applications.

## Key Takeaways

### Key Takeaways 16

- **Structure** encodes independence, enabling efficient inference.
- **Hybrid models** leverage neural expressivity with graphical constraints.
- **Inference choices** depend on graph type and potential functions.

## Exercises

### Easy

*Exercise 16.1 (Generator Role).* Describe the generator's objective in GANs.

**Hint:**

*Fool the discriminator.*

*Exercise 16.2 (Discriminator Role).* Describe the discriminator's objective in GANs.

**Hint:**

*Distinguish real from fake.*

*Exercise 16.3 (Mode Collapse).* Define mode collapse and its symptoms.

**Hint:**

*Generator produces limited variety.*

*Exercise 16.4 (Training Instability).* Name two causes of GAN training instability.

**Hint:**

*Oscillation; gradient vanishing.*

## Medium

*Exercise 16.5 (Nash Equilibrium).* Explain why GAN training seeks a Nash equilibrium.

**Hint:**

*Minimax game; no unilateral improvement.*

*Exercise 16.6 (Wasserstein Distance).* State the advantage of Wasserstein distance over JS divergence.

**Hint:**

*Gradient behaviour with non-overlapping distributions.*

## Hard

*Exercise 16.7 (Optimal Discriminator).* Derive the optimal discriminator for fixed generator.

**Hint:**

*Maximise expected log-likelihood.*

*Exercise 16.8 (Spectral Normalisation).* Analyse how spectral normalisation stabilises GAN training.

**Hint:**

*Lipschitz constraint; gradient norms.*

*Exercise 16.9 (Advanced Topic 1).* Explain a key concept from this chapter and its practical applications.

**Hint:**

*Consider the theoretical foundations and real-world implications.*

*Exercise 16.10 (Advanced Topic 2).* Analyse the relationship between different techniques covered in this chapter.

**Hint:**

*Look for connections and trade-offs between methods.*

*Exercise 16.11 (Advanced Topic 3).* Design an experiment to test a hypothesis related to this chapter's content.

**Hint:**

*Consider experimental design, metrics, and potential confounding factors.*

*Exercise 16.12 (Advanced Topic 4).* Compare different approaches to solving a problem from this chapter.

**Hint:**

*Consider computational complexity, accuracy, and practical considerations.*

*Exercise 16.13 (Advanced Topic 5).* Derive a mathematical relationship or prove a theorem from this chapter.

**Hint:**

*Start with the definitions and work through the logical steps.*

*Exercise 16.14 (Advanced Topic 6).* Implement a practical solution to a problem discussed in this chapter.

**Hint:**

*Consider the implementation details and potential challenges.*

*Exercise 16.15 (Advanced Topic 7).* Evaluate the limitations and potential improvements of techniques from this chapter.

**Hint:**

*Consider both theoretical limitations and practical constraints.*



# Chapter 17

## Monte Carlo Methods

This chapter introduces sampling-based approaches for probabilistic inference and learning.

### Learning Objectives

1. Monte Carlo estimation and variance reduction techniques
2. MCMC algorithms (Metropolis – Hastings, Gibbs) and their diagnostics
3. Sampling to approximate expectations and gradients
4. Pitfalls: poor mixing and autocorrelation

## Intuition

When exact integrals are intractable, we approximate them with random samples. The art is to sample efficiently from complicated posteriors and to estimate uncertainty from finite chains.

For example, estimating the expected value of a complex function over a high-dimensional probability distribution becomes feasible by drawing random samples and averaging their function values. Like a pollster surveying a small random sample of voters to predict election outcomes, Monte Carlo methods use random sampling to approximate complex mathematical expectations that would be impossible to compute exactly.

Think of Monte Carlo methods as a sophisticated dart-throwing game where you're trying to estimate the area of an irregular shape by throwing darts randomly at a board. The more darts you throw, the better your estimate becomes, and the key is learning to throw the darts in the most informative regions rather than just anywhere on the board.

### 17.1 Sampling and Monte Carlo Estimators ★

These methods approximate complex expectations by drawing random samples and computing sample averages, providing a powerful framework for handling intractable integrals in high-dimensional spaces.

#### 17.1.1 Monte Carlo Estimation

Approximate expectations using samples:

$$\mathbb{E}_{p(x)}[f(x)] \approx \frac{1}{N} \sum_{i=1}^N f(x^{(i)}), \quad x^{(i)} \sim p(x) \tag{17.1}$$

This fundamental equation (17.1) states that we can approximate the expected value of any function  $f(x)$  under distribution  $p(x)$  by drawing  $N$  independent samples  $x^{(i)}$  from  $p(x)$  and computing their average. The law of large numbers guarantees that this estimate converges to the true expectation as  $N \rightarrow \infty$ .

In deep learning, Monte Carlo estimation enables us to approximate intractable expectations in Bayesian neural networks, where we need to compute expectations over the posterior distribution of network weights. This is particularly crucial for uncertainty quantification, where we estimate the expected prediction and its variance by sampling from the posterior distribution of model parameters.

### 17.1.2 Variance Reduction

Variance reduction refers to techniques that decrease the statistical variance of Monte Carlo estimators without increasing the computational cost, leading to more accurate estimates with fewer samples. The goal is to design estimators that converge faster to the true expectation by exploiting structure in the problem or using clever sampling strategies.

Rao-Blackwellization leverages the power of conditional expectations by analytically computing expectations over some variables while sampling others, effectively reducing the dimensionality of the sampling problem. Control variates work by subtracting correlated zero-mean terms that are highly correlated with our estimator, effectively canceling out some of the variance. Antithetic sampling uses negatively correlated samples to create a form of variance cancellation, where high and low values tend to be paired together, reducing the overall variance of the estimator. These techniques are particularly valuable in deep learning when estimating gradients or expectations in high-dimensional parameter spaces, where naive sampling can be prohibitively expensive.

## 17.2 Markov Chain Monte Carlo ★

MCMC methods construct dependent sequences of samples that eventually converge to the target distribution, enabling sampling from complex posterior distributions that cannot be sampled directly.

### 17.2.1 Markov Chains

A Markov chain is a sequence of random variables where the future state depends only on the current state, not the entire history. This is captured by the Markov property:  $p(x_t|x_{t-1}, \dots, x_1) = p(x_t|x_{t-1})$ , meaning the conditional distribution of  $x_t$  given all previous states depends only on the immediate previous state  $x_{t-1}$ .

The stationary distribution  $\pi(x)$  is a probability distribution such that if  $x_t \sim \pi$ , then  $x_{t+1} \sim \pi$ . This means that once the chain reaches the stationary distribution, it remains there, making it the target distribution we want to sample from.

In deep learning, Markov chains are fundamental to understanding how MCMC algorithms explore the parameter space of neural networks. When we want to sample from the posterior distribution of network weights, we construct a Markov chain whose stationary distribution is exactly the posterior we're interested in. This enables us to explore the complex, high-dimensional space of neural network parameters and understand the uncertainty in our model predictions.

### 17.2.2 Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm is a general framework for sampling from any target distribution  $p(x)$  using a proposal distribution  $q(x'|x_t)$ . The algorithm works by proposing new states and accepting or rejecting them based on a carefully designed acceptance probability.

The mathematical foundation lies in the acceptance probability:

$$A(x', x_t) = \min \left( 1, \frac{p(x')q(x_t|x')}{p(x_t)q(x'|x_t)} \right) \quad (17.2)$$

This formula ensures that the resulting Markov chain has the correct stationary distribution by balancing the probability of moving from the current state to the proposed state with the probability of the reverse move. The algorithm proceeds by: (1)

proposing a new state  $x' \sim q(x'|x_t)$  from the proposal distribution, (2) computing the acceptance probability  $A(x', x_t)$ , (3) accepting the proposal with probability  $A(x', x_t)$  or staying at the current state otherwise.

In deep learning, Metropolis-Hastings enables sampling from the posterior distribution of neural network parameters, which is crucial for Bayesian neural networks. This allows us to quantify uncertainty in predictions by sampling multiple sets of weights from the posterior and computing prediction distributions, providing a principled way to understand model confidence and make robust decisions under uncertainty.

### 17.2.3 Gibbs Sampling

Gibbs sampling is a special case of Metropolis-Hastings where we update one variable at a time by sampling from its conditional distribution given all other variables. The mathematical foundation is captured by the update rule:

$$x_i^{(t+1)} \sim p(x_i|x_{-i}^{(t)}) \quad (17.3)$$

This equation states that we sample the new value of variable  $x_i$  from its conditional distribution given all other variables  $x_{-i}$  at their current values. The algorithm cycles through all variables, updating each one in turn, which ensures that the resulting Markov chain has the correct stationary distribution.

In deep learning, Gibbs sampling is particularly useful for Bayesian neural networks where we can sample individual parameters or groups of parameters from their conditional distributions. This approach is especially powerful when the conditional distributions are tractable, such as in conjugate models or when using variational approximations. Gibbs sampling enables efficient exploration of the posterior distribution of network parameters, allowing us to understand the uncertainty in different parts of the model and make more robust predictions.

### 17.2.4 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) leverages gradient information to achieve more efficient exploration of the target distribution compared to random walk methods. The algorithm treats the parameters as positions in a physics simulation and introduces auxiliary momentum variables to guide the exploration.

The mathematical foundation involves simulating Hamiltonian dynamics, where the total energy  $H(x, p) = U(x) + K(p)$  consists of potential energy  $U(x) = -\log p(x)$  (the negative log-probability) and kinetic energy  $K(p) = \frac{1}{2}p^T M^{-1}p$  (where  $M$  is the mass matrix). The algorithm alternates between sampling momentum from a Gaussian distribution and simulating the Hamiltonian dynamics to propose new states, which are then accepted or rejected based on the Metropolis criterion.

In deep learning, HMC is particularly valuable for sampling from the posterior distribution of neural network parameters because it can efficiently navigate the complex, high-dimensional parameter space using gradient information. The momentum component allows the sampler to move more efficiently through the parameter space, avoiding the slow random walk behavior of traditional Metropolis-Hastings methods. This makes HMC especially useful for Bayesian neural networks where we need to explore the posterior distribution of weights to quantify uncertainty in predictions.

## 17.3 Importance Sampling ★

Importance sampling is a variance reduction technique that allows us to estimate expectations under a target distribution  $p(x)$  by sampling from a different, more convenient proposal distribution  $q(x)$ . The mathematical foundation is the importance sampling identity:

$$\mathbb{E}_p[f(x)] = \mathbb{E}_q \left[ \frac{p(x)}{q(x)} f(x) \right] \approx \frac{1}{N} \sum_{i=1}^N \frac{p(x^{(i)})}{q(x^{(i)})} f(x^{(i)}) \quad (17.4)$$

This equation shows that we can estimate the expectation under  $p(x)$  by sampling from  $q(x)$  and reweighting the samples by the importance weights  $w(x) = \frac{p(x)}{q(x)}$ . The key insight is that we can choose  $q(x)$  to be easier to sample from than  $p(x)$ , as long as we account for the difference in probability mass through the importance weights.

The method is most effective when the proposal distribution  $q(x)$  is easy to sample from and has heavier tails than the target distribution  $p(x)$ , ensuring that the importance weights don't become too large and cause high variance in the estimator. In deep learning, importance sampling is particularly valuable for estimating gradients in reinforcement learning, where we need to compute expectations over policy distributions that may be difficult to sample from directly. It also plays a crucial role in training generative models, where we can use importance sampling to estimate intractable likelihoods and gradients more efficiently.

## 17.4 Applications in Deep Learning ★

Monte Carlo methods have become indispensable tools in modern deep learning, enabling practitioners to handle uncertainty, explore complex parameter spaces, and train models that would otherwise be intractable. These applications span across multiple domains of machine learning, from Bayesian inference to reinforcement learning and generative modeling.

In Bayesian deep learning, Monte Carlo methods enable us to sample from the posterior distribution of network weights, providing a principled way to quantify uncertainty in predictions. Rather than learning a single set of parameters, we can sample multiple sets of weights from the posterior and compute prediction distributions, giving us confidence intervals and measures of model uncertainty that are crucial for safety-critical applications.

Reinforcement learning heavily relies on Monte Carlo methods for policy gradient estimation, where we need to compute expectations over trajectories and action distributions. Monte Carlo tree search, used in game-playing algorithms like AlphaGo, explores possible future states by sampling from the action space, enabling agents to make informed decisions in complex environments with large state spaces.

Generative models benefit from Monte Carlo methods in training energy-based models, where we need to sample from complex distributions to estimate gradients and likelihoods. These methods also enable sampling from learned distributions, allowing us to generate new data points that follow the patterns learned by our models, which is essential for applications like image synthesis, text generation, and data augmentation.

## 17.5 Real World Applications

Monte Carlo methods use random sampling to solve complex problems that would be intractable through direct computation. These techniques enable approximating difficult integrals, exploring high-dimensional spaces, and quantifying uncertainty.

### 17.5.1 Financial Risk Management

Financial institutions face the constant challenge of understanding and managing uncertainty in an ever-changing market landscape. Monte Carlo methods have revolutionized how financial professionals approach risk assessment, moving beyond simple point estimates to comprehensive probabilistic analysis that captures the full spectrum of possible outcomes.

Value at Risk (VaR) estimation represents one of the most critical applications, where banks must estimate potential losses to maintain adequate capital reserves. Traditional approaches relied on historical data and normal distribution assumptions, but Monte Carlo simulation generates thousands of possible market scenarios, computing portfolio values under each scenario. This approach provides distributions of potential losses rather than single-point estimates, helping banks understand risks across different market conditions and stress scenarios that historical data might not capture.

Option pricing demonstrates the power of Monte Carlo methods in handling complex financial instruments. Financial derivatives have values that depend on uncertain future asset prices, and while simple options have analytical solutions, exotic options with path-dependent payoffs require numerical methods. Monte Carlo methods simulate possible price paths, computing option values as averages over many scenarios, enabling pricing of complex derivatives that would be impossible to value analytically. Retirement planning showcases how Monte Carlo methods can transform personal financial decision-making. Financial advisors use Monte Carlo simulation to project retirement savings over decades, considering uncertainties in investment returns, inflation rates, and life expectancy. Rather than promising a single outcome, simulations show probability distributions—like “85% chance your savings last through age 95”—helping people make informed decisions about savings rates, investment allocations, and retirement timing based on their risk tolerance and financial goals.

### 17.5.2 Climate and Weather Modeling

The prediction of complex physical systems like weather and climate represents one of the most challenging applications of Monte Carlo methods, where the inherent chaos and uncertainty of atmospheric processes demand sophisticated probabilistic approaches rather than deterministic forecasts.

Ensemble weather forecasting has transformed meteorology by running multiple simulations with slightly different initial conditions, representing the inevitable measurement uncertainty in atmospheric observations. These Monte Carlo-style ensembles provide probability distributions for forecasts—like “70% chance of rain”—which are far more useful than deterministic predictions for decision-making. This probabilistic approach helps with everything from personal decisions like carrying an umbrella to critical infrastructure planning and disaster preparedness, where understanding the range of possible outcomes is essential.

Climate change projections face even greater challenges, as long-term climate models involve enormous uncertainty in cloud physics, ocean circulation patterns, and future human emissions. Monte Carlo sampling over parameter uncertainties generates probability distributions for future climate scenarios, providing policymakers with the probabilistic information needed to make informed decisions about emissions reductions and adaptation strategies. Rather than providing a single temperature projection, these methods give us probability distributions that help us understand the likelihood of different warming scenarios and their associated risks.

Hurricane path prediction exemplifies how Monte Carlo methods can save lives through better decision-making. The familiar “cone of uncertainty” in hurricane forecasts comes from Monte Carlo simulations exploring possible paths given current conditions and atmospheric uncertainties. This probabilistic approach helps emergency managers make evacuation decisions that balance safety against unnecessary disruption, understanding not just the most likely path but the full range of possible outcomes and their associated probabilities.

### 17.5.3 Drug Discovery and Design

The pharmaceutical industry faces the daunting challenge of exploring vast chemical and biological spaces to discover new drugs, where the complexity of molecular interactions and the high cost of experimental testing make computational approaches essential. Monte Carlo methods have become indispensable tools in this process, enabling researchers to explore possibilities that would be impossible to test experimentally.

Molecular dynamics simulation represents one of the most powerful applications, where understanding how proteins fold and bind to drug molecules requires simulating atomic movements over time. Monte Carlo methods sample possible molecular configurations, computing binding affinities and predicting which drug candidates are worth the expensive experimental testing that can cost millions of dollars per compound. This computational screening accelerates drug discovery while dramatically reducing costs, allowing researchers to focus experimental efforts on the most promising candidates.

Clinical trial design has been transformed by Monte Carlo simulation, where pharmaceutical companies use these methods to estimate statistical power under various scenarios and patient populations. Simulations help determine the sample sizes needed to detect treatment effects reliably, preventing under-powered trials that waste resources or miss effective treatments. This probabilistic approach to trial design ensures that clinical studies are properly sized to answer the questions they’re designed to address, improving the efficiency of the drug development process.

Dose optimization represents another critical application, where finding optimal drug dosages involves balancing efficacy and toxicity under the inherent variability of individual patients. Monte Carlo simulation explores dose-response relationships across diverse patient populations, identifying regimens that maximize treatment benefit while minimizing risks. This personalized approach to dosing helps ensure that patients receive the most effective treatment while avoiding dangerous side effects, ultimately improving patient outcomes and reducing healthcare costs.

### 17.5.4 Practical Advantages

Monte Carlo methods have become indispensable across diverse fields because they offer unique advantages that make them the method of choice for handling complex, uncertain problems. These advantages stem from their fundamental ability to approximate intractable problems through random sampling, providing solutions where analytical approaches fail completely.

The ability to handle complexity represents perhaps the most significant advantage, as Monte Carlo methods work when analytical solutions are impossible. Many real-world problems involve high-dimensional integrals, complex probability distributions, or nonlinear dynamics that cannot be solved exactly, but Monte Carlo methods can approximate these solutions through sampling. This makes them applicable to problems that would otherwise be completely intractable, opening up new possibilities for analysis and decision-making.

Quantifying uncertainty is another crucial advantage, as Monte Carlo methods provide probability distributions rather than just point estimates. In many applications, understanding the range of possible outcomes and their associated probabilities is more valuable than knowing a single "best" answer. This probabilistic approach enables decision-makers to assess risks, plan for contingencies, and make informed choices under uncertainty.

The natural scalability of Monte Carlo methods makes them particularly attractive for modern computing environments, as more samples improve accuracy predictably, and the independent nature of simulations enables efficient parallel computation. This scalability, combined with the ability to leverage modern computing resources, makes Monte Carlo methods practical for large-scale problems that would be impossible to solve using other approaches.

These applications demonstrate how Monte Carlo methods enable decision-making under uncertainty across finance, science, and healthcare—problems where exact answers are impossible but approximate probabilistic understanding is invaluable. The combination of theoretical rigor and practical applicability makes Monte Carlo methods one of the most powerful tools in modern computational science.

## Key Takeaways

### Key Takeaways 17

- **Monte Carlo** approximates expectations; variance control is essential.
- **MCMC** constructs dependent samples targeting complex posteriors.
- **Diagnostics** (ESS, R-hat) guide reliability of estimates.

## Exercises

### Easy

*Exercise 17.1 (Self-Attention Intuition).* Explain why self-attention captures long-range dependencies.

**Hint:**

*Direct pairwise interactions.*

*Exercise 17.2 (Positional Encoding).* Why do Transformers need positional encodings?

**Hint:**

*Permutation invariance of self-attention.*

*Exercise 17.3 (Multi-Head Attention).* State the benefit of multiple attention heads.

**Hint:**

*Different representation subspaces.*

*Exercise 17.4 (Masked Attention).* Explain the role of masking in causal attention.

**Hint:**

*Prevent future information leakage.*

## Medium

*Exercise 17.5 (Computational Complexity).* Derive the computational complexity of self-attention.

**Hint:**

$O(n^2d)$  for sequence length  $n$ , dimension  $d$ .

*Exercise 17.6 (LayerNorm vs. BatchNorm).* Compare LayerNorm and BatchNorm in Transformers.

**Hint:**

*Independence from batch; sequence-level statistics.*

## Hard

*Exercise 17.7 (Sparse Attention).* Design a sparse attention pattern and analyse complexity savings.

**Hint:**

*Local windows; strided patterns;  $O(n \log n)$  or  $O(n\sqrt{n})$ .*

*Exercise 17.8 (Attention Visualisation).* Propose methods to interpret attention weights and discuss limitations.

**Hint:**

*Attention rollout; gradient-based; correlation vs. causation.*

*Exercise 17.9 (Advanced Topic 1).* Explain a key concept from this chapter and its practical applications.

**Hint:**

*Consider the theoretical foundations and real-world implications.*

*Exercise 17.10 (Advanced Topic 2).* Analyse the relationship between different techniques covered in this chapter.

**Hint:**

*Look for connections and trade-offs between methods.*

*Exercise 17.11 (Advanced Topic 3).* Design an experiment to test a hypothesis related to this chapter's content.

**Hint:**

*Consider experimental design, metrics, and potential confounding factors.*

*Exercise 17.12 (Advanced Topic 4).* Compare different approaches to solving a problem from this chapter.

**Hint:**

*Consider computational complexity, accuracy, and practical considerations.*

*Exercise 17.13 (Advanced Topic 5).* Derive a mathematical relationship or prove a theorem from this chapter.

**Hint:**

*Start with the definitions and work through the logical steps.*

*Exercise 17.14 (Advanced Topic 6).* Implement a practical solution to a problem discussed in this chapter.

**Hint:**

*Consider the implementation details and potential challenges.*

*Exercise 17.15 (Advanced Topic 7).* Evaluate the limitations and potential improvements of techniques from this chapter.

**Hint:**

*Consider both theoretical limitations and practical constraints.*

# Chapter 18

## Confronting the Partition Function

This chapter addresses computational challenges in probabilistic models arising from intractable partition functions.

### Learning Objectives

1. Why partition functions are hard and where they arise
2. Strategies: importance sampling, AIS, and contrastive methods
3. Bias/variance trade-offs in partition function estimation
4. Practical estimators under compute constraints

## Intuition

Partition functions are normalizing constants that ensure probability distributions sum to one by dividing unnormalized probabilities by their sum over all possible states. They normalize probabilities by summing over exponentially many states, making exact computation intractable for high-dimensional models.

For example, in a binary image with 100 pixels, computing the partition function requires summing over  $2^{100}$  possible configurations, which is computationally impossible. Like trying to count every possible arrangement of a massive jigsaw puzzle where each piece can be in one of two orientations, partition functions represent the total "weight" of all possible configurations that must be computed to get proper probabilities.

Think of partition functions as the "denominator" in a fraction where the numerator is the probability of a specific configuration and the denominator is the sum of probabilities of all possible configurations. Just as you can't know what fraction of a pizza you're eating without knowing the total size of the pizza, you can't compute exact probabilities without knowing the partition function.

### 18.1 The Partition Function Problem ★

Many probabilistic models have the form:

$$p(\mathbf{x}) = \frac{1}{Z} \tilde{p}(\mathbf{x}) \quad (18.1)$$

This equation (18.1) shows that the normalized probability  $p(\mathbf{x})$  is obtained by dividing the unnormalized probability  $\tilde{p}(\mathbf{x})$  by the partition function  $Z$ . The partition function  $Z = \sum_{\mathbf{x}} \tilde{p}(\mathbf{x})$  (for discrete variables) or  $Z = \int \tilde{p}(\mathbf{x}) d\mathbf{x}$  (for continuous variables) is intractable because it requires summing or integrating over all possible configurations.

In deep learning, this problem arises in energy-based models like Restricted Boltzmann Machines and modern generative models, where we need to compute likelihoods for training but cannot evaluate the partition function exactly. This forces us to

use approximate methods like contrastive divergence, noise contrastive estimation, or score matching to train these models effectively.

### 18.1.1 Why It's Hard

Computing the partition function  $Z$  presents fundamental computational challenges that grow exponentially with the dimensionality of the problem. The core difficulty lies in the requirement to sum or integrate over all possible configurations, which becomes computationally prohibitive as the number of variables increases.

The exponential growth in dimensionality means that for a model with  $d$  binary variables, we must consider  $2^d$  possible configurations, making exact computation impossible for realistic model sizes. This exponential explosion affects not just the computational cost but also the memory requirements, as we need to store and process information about exponentially many states.

The solution to this intractability involves developing approximate methods that avoid computing the partition function directly. These approaches include contrastive divergence, which uses short Markov chains to approximate gradients; noise contrastive estimation, which transforms the problem into binary classification; and score matching, which works with gradients rather than probabilities. Each method trades off between computational efficiency and approximation accuracy, enabling practical training of complex probabilistic models.

### 18.1.2 Impact

The intractability of partition functions has profound implications for probabilistic modeling and machine learning, fundamentally limiting our ability to work with complex models in their most natural form. When we cannot compute the partition function, we lose the ability to evaluate exact likelihoods, which are essential for model comparison, parameter estimation, and uncertainty quantification.

This limitation prevents us from directly computing gradients needed for learning, forcing us to develop alternative training procedures that approximate the true gradients. The inability to evaluate exact likelihoods also makes it difficult to compare different models or assess their relative performance, as we cannot compute the standard likelihood-based metrics that would naturally arise from the probabilistic framework.

These challenges have driven the development of specialized techniques that work around the partition function problem, including approximate inference methods, contrastive learning approaches, and score-based training procedures. While these methods provide practical solutions, they often introduce bias or require careful tuning, highlighting the fundamental tension between theoretical elegance and computational feasibility in probabilistic modeling.

## 18.2 Contrastive Divergence ★

Contrastive divergence is a practical training algorithm that approximates the intractable gradients in energy-based models by using short Markov chains instead of requiring exact sampling from the model distribution.

### 18.2.1 Motivation

For Restricted Boltzmann Machines (RBMs), the joint probability distribution is given by:

$$p(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} \exp(-E(\mathbf{v}, \mathbf{h})) \quad (18.2)$$

This equation (18.2) shows that the probability of a configuration is proportional to the exponential of the negative energy, normalized by the partition function  $Z$ . The exact gradient for learning requires computing expectations under the model distribution, which involves the intractable partition function.

The gradient computation involves two terms:

$$\frac{\partial \log p(\mathbf{v})}{\partial \theta} = -\mathbb{E}_{p(\mathbf{h}|\mathbf{v})} \left[ \frac{\partial E}{\partial \theta} \right] + \mathbb{E}_{p(\mathbf{v}, \mathbf{h})} \left[ \frac{\partial E}{\partial \theta} \right] \quad (18.3)$$

This equation (18.3) shows that the gradient consists of a positive term (expectation under the data distribution) and a negative term (expectation under the model distribution). The second term requires sampling from the full model distribution, which is intractable due to the partition function, motivating the need for approximate methods like contrastive divergence.

### 18.2.2 CD-k Algorithm

The CD-k algorithm approximates the intractable second term in the gradient by using a short Markov Chain Monte Carlo (MCMC) chain. An MCMC chain is a sequence of samples generated by iteratively applying a transition operator that preserves the target distribution, allowing us to sample from complex distributions without computing the partition function.

The algorithm works by starting from the data and running a short chain of k Gibbs sampling steps, which provides a biased but computationally efficient approximation to the true gradient. The key insight is that we don't need to run the chain until convergence; even a few steps provide a useful approximation that enables effective learning. This approach trades off between computational efficiency and approximation accuracy, making it practical for training energy-based models on large datasets. Despite being biased, the CD-k algorithm works surprisingly well in practice because the bias tends to be in a direction that still provides useful gradient information for learning. The algorithm has been successfully applied to training Restricted Boltzmann Machines and other energy-based models, demonstrating that approximate methods can be highly effective even when they don't provide exact gradients.

## 18.3 Noise-Contrastive Estimation ★

Noise-Contrastive Estimation (NCE) transforms the intractable density estimation problem into a tractable binary classification problem by distinguishing between data samples and noise samples, avoiding the need to compute partition functions.

### 18.3.1 Key Idea

The key insight of Noise-Contrastive Estimation is to transform the intractable density estimation problem into a tractable binary classification problem. Instead of trying to compute the partition function directly, NCE trains a classifier to distinguish between samples from the data distribution and samples from a known noise distribution.

This approach cleverly avoids the partition function problem by focusing on the relative probabilities rather than absolute probabilities. By learning to distinguish data from noise, the model implicitly learns the structure of the data distribution without needing to normalize it. The noise distribution serves as a reference point that makes the classification problem well-defined and computationally tractable, enabling effective learning even when the true data distribution has an intractable partition function.

### 18.3.2 NCE Objective

The NCE objective function combines the log-likelihood of correctly identifying data samples with the log-likelihood of correctly identifying noise samples:

$$\mathcal{L} = \mathbb{E}_{p_{\text{data}}} [\log h(\mathbf{x})] + k \cdot \mathbb{E}_{p_{\text{noise}}} [\log(1 - h(\mathbf{x}))] \quad (18.4)$$

This equation (18.4) shows that the objective maximizes the probability of correctly classifying data samples as data while also maximizing the probability of correctly classifying noise samples as noise. The parameter  $k$  controls the relative importance of the noise samples in the objective.

The classifier function is defined as:

$$h(\mathbf{x}) = \frac{p_{\text{model}}(\mathbf{x})}{p_{\text{model}}(\mathbf{x}) + k \cdot p_{\text{noise}}(\mathbf{x})} \quad (18.5)$$

This equation (18.5) shows that  $h(\mathbf{x})$  represents the probability that a sample comes from the data distribution rather than the noise distribution. The key insight is that this ratio can be computed without knowing the partition function, as the normalization constants cancel out in the ratio, making the objective tractable to optimize.

### 18.3.3 Applications

Noise-Contrastive Estimation has found widespread applications in natural language processing and machine learning, particularly in scenarios where traditional maximum likelihood estimation is intractable due to partition function issues. The method has been particularly successful in training word embeddings, where the goal is to learn dense vector representations of words that capture their semantic relationships.

In language modeling, NCE enables training of neural language models without computing the full softmax over the entire vocabulary, which would be computationally prohibitive for large vocabularies. The method has also been applied to energy-based models, where it provides a practical alternative to contrastive divergence for training models with intractable partition functions. These applications demonstrate the versatility of NCE as a general-purpose method for handling partition function intractability across different domains and model architectures.

### 18.3.4 Notes and references

The development of Noise-Contrastive Estimation represents a significant milestone in addressing partition function intractability, with its introduction by Gutmann and Hyvärinen in 2010 marking a turning point in probabilistic modeling. The method's theoretical foundations were further developed through connections to maximum likelihood estimation and the analysis of its asymptotic properties, establishing NCE as a principled alternative to traditional approaches.

Key achievements include the successful application of NCE to word2vec embeddings, which demonstrated its practical effectiveness in large-scale natural language processing tasks. The method's extension to neural language models enabled training on massive vocabularies without computational bottlenecks, contributing to the development of modern language models. Recent work has explored connections between NCE and other contrastive learning methods, highlighting its broader impact on representation learning and its role in the development of modern generative models that avoid partition function computation entirely.

## 18.4 Score Matching ★

Score matching is a powerful technique that learns probability distributions by matching the gradients of the log-density (score function) rather than the densities themselves. The score function is defined as:

$$\psi(\mathbf{x}) = \nabla_{\mathbf{x}} \log p(\mathbf{x}) \quad (18.6)$$

This equation shows that the score function represents the gradient of the log-probability with respect to the input variables. The key insight is that the score function can be computed without knowing the partition function, as the normalization constant cancels out when taking the gradient.

The score matching objective minimizes the squared difference between the model's score function and the data's score function:

$$\mathcal{L} = \frac{1}{2} \mathbb{E}_{p_{\text{data}}} [\|\psi_{\theta}(\mathbf{x}) - \nabla_{\mathbf{x}} \log p_{\text{data}}(\mathbf{x})\|^2] \quad (18.7)$$

This approach avoids the partition function problem entirely because the normalization constant cancels out in the gradient computation, making the objective tractable to optimize. Score matching has been particularly successful in training energy-based models and has connections to modern generative modeling techniques like diffusion models.

## 18.5 Real World Applications

Confronting the partition function—computing normalizing constants in probabilistic models—is a fundamental challenge. Practical applications require approximations and specialized techniques to make inference tractable in complex models.

### 18.5.1 Recommender Systems at Scale

The challenge of providing personalized recommendations at scale represents one of the most compelling applications of partition function approximation techniques, where the need to score millions of items for billions of users makes exact probability computation completely infeasible.

YouTube's video recommendation system exemplifies this challenge, where the platform must score millions of videos for each user in real-time. Computing exact probabilities would require evaluating partition functions over all possible videos, which is computationally impossible. Instead, the system uses approximate methods like negative sampling and importance sampling to provide good recommendations efficiently, enabling real-time personalization for billions of users while maintaining acceptable computational costs.

E-commerce platforms face similar challenges when ranking products for users, where models must learn to score product-user compatibility without being able to compute exact probabilities. These systems employ contrastive learning methods that approximate partition functions by sampling negative examples, enabling practical deployment at scale while maintaining recommendation quality. The success of these approaches demonstrates how approximate methods can provide effective solutions even when exact computation is impossible.

Music streaming services create personalized playlists by modeling sequential song compatibility, where full probabilistic models would require intractable partition functions over all possible song sequences. Practical systems use locally normalized models and sampling-based approximations to generate engaging playlists efficiently, showing how partition function approximation techniques enable complex sequential modeling at scale.

### 18.5.2 Natural Language Processing

Natural language processing presents unique challenges for partition function computation, where the exponential growth in possible sequences makes exact probability computation intractable for realistic applications. The field has developed sophisticated approximation techniques that enable practical deployment of probabilistic models at scale.

Modern language model training exemplifies these challenges, where models must predict next words from vocabularies of 50,000+ tokens. Computing partition functions over all possible next words for every training example would be prohibitively expensive, making traditional maximum likelihood estimation impossible. Techniques like noise contrastive estimation and self-normalization have made training practical, enabling language models that power translation, autocomplete, and conversational AI systems that would otherwise be impossible to train.

Neural machine translation systems face similar challenges, where translation models generate target sentences word by word while considering vast numbers of possible continuations. Exact probability computation would require intractable partition functions over all possible translation sequences, making traditional approaches infeasible. Beam search with approximate scoring enables practical translation systems that produce high-quality translations in real-time, demonstrating how approximation techniques can maintain performance while avoiding computational bottlenecks.

Named entity recognition represents another critical application, where identifying people, places, and organizations in text involves structured prediction over exponentially many possible tag sequences. While conditional random fields require computing partition functions efficiently, the forward-backward algorithm provides exact computation for chain structures, enabling accurate entity extraction in applications ranging from news analysis to medical record processing. This demonstrates how careful model design can sometimes avoid partition function intractability entirely.

### 18.5.3 Computer Vision

Computer vision applications present unique challenges for partition function computation, where the need to model complex spatial and structural relationships often leads to intractable probability distributions. The field has developed specialized approximation techniques that enable practical deployment of probabilistic models for image understanding tasks.

Semantic segmentation exemplifies these challenges, where labeling every pixel in images requires modeling dependencies between neighboring pixels to ensure coherent segmentation results. Fully modeling these dependencies would involve intractable partition functions over all possible pixel labelings, making exact inference impossible. Practical systems use approximate inference methods like mean field approximation and pseudo-likelihood, or employ structured models with

tractable partition functions using chain or tree structures that maintain computational feasibility while preserving important spatial relationships.

Pose estimation represents another critical application, where estimating human body poses involves predicting joint locations subject to anatomical constraints such as arms connecting to shoulders and legs having limited range of motion. Models that encode these constraints naturally have complex partition functions, but approximate inference techniques enable real-time pose estimation for applications ranging from gaming to physical therapy. These systems demonstrate how careful approximation can maintain performance while avoiding computational bottlenecks.

Object detection systems face partition function challenges similar to recommendation systems, where detecting objects requires scoring countless possible bounding boxes. Models that learn to rank boxes face the same fundamental challenge of intractable partition functions, but techniques like contrastive learning and hard negative mining make training practical. These approaches enable accurate detection in applications from autonomous driving to retail analytics, showing how partition function approximation techniques can be adapted across different domains.

#### 18.5.4 Practical Solutions

The practical solutions to partition function intractability have evolved into a sophisticated toolkit of approximation techniques, each with its own strengths and trade-offs. These methods represent the collective wisdom of the machine learning community in addressing one of the most fundamental challenges in probabilistic modeling.

Approximation methods like Monte Carlo sampling and variational inference provide general-purpose approaches to handling intractable partition functions, offering different trade-offs between computational efficiency and approximation accuracy. Monte Carlo methods use random sampling to approximate expectations, while variational inference provides deterministic approximations that can be more computationally efficient but may introduce bias. The choice between these approaches depends on the specific requirements of the application and the available computational resources.

Negative sampling techniques have proven particularly effective in large-scale applications, where they approximate partition functions using sampled negatives rather than computing exact expectations. This approach has been successfully applied in recommendation systems, language modeling, and computer vision, demonstrating its versatility across different domains. The key insight is that we often don't need exact probabilities but rather relative rankings or scores that can be computed efficiently.

Structured models represent another important strategy, where careful design can sometimes avoid partition function intractability entirely. By constraining the model structure to use tractable components like chains or trees, we can maintain computational feasibility while preserving important dependencies. This approach has been particularly successful in sequence modeling and structured prediction tasks, where the structure of the problem naturally suggests tractable approximations.

These applications demonstrate that confronting the partition function is not just a theoretical concern but a practical challenge requiring clever approximations to deploy probabilistic models at scale. The success of these methods across diverse applications shows that partition function intractability, while fundamental, is not insurmountable when approached with the right combination of theoretical understanding and practical engineering.

## Key Takeaways

### Key Takeaways 18

- **Partition functions** create intractable normalisers in many models.
- **Estimators** (IS, AIS, contrastive) trade bias and variance differently.
- **Practicality** depends on proposal quality and compute budget.

## Exercises

### Easy

*Exercise 18.1 (MDP Definition).* Define the components of a Markov Decision Process.

**Hint:**

*States, actions, rewards, transition dynamics, discount factor.*

*Exercise 18.2 (Value Function).* Explain the difference between  $V(s)$  and  $Q(s, a)$ .

**Hint:**

*State value vs. action-value.*

*Exercise 18.3 (Policy Types).* Contrast deterministic and stochastic policies.

**Hint:**

*Mapping vs. distribution over actions.*

*Exercise 18.4 (Exploration vs. Exploitation).* Give two exploration strategies in RL.

**Hint:**

*$\epsilon$ -greedy; UCB; entropy regularisation.*

### Medium

*Exercise 18.5 (Bellman Equation).* Derive the Bellman equation for  $Q(s, a)$ .

**Hint:**

*Recursive relationship with successor states.*

*Exercise 18.6 (Policy Gradient).* Explain why policy gradient methods are useful for continuous action spaces.

**Hint:**

*Direct parameterisation; differentiability.*

### Hard

*Exercise 18.7 (Actor-Critic Derivation).* Derive the advantage actor-critic update rule.

**Hint:**

*Baseline subtraction; variance reduction.*

*Exercise 18.8 (Off-Policy Correction).* Analyse importance sampling for off-policy learning and its variance.

**Hint:**

*Likelihood ratio; distribution mismatch.*

*Exercise 18.9 (Advanced Topic 1).* Explain a key concept from this chapter and its practical applications.

**Hint:**

*Consider the theoretical foundations and real-world implications.*

*Exercise 18.10 (Advanced Topic 2).* Analyse the relationship between different techniques covered in this chapter.

**Hint:**

*Look for connections and trade-offs between methods.*

*Exercise 18.11 (Advanced Topic 3).* Design an experiment to test a hypothesis related to this chapter's content.

**Hint:**

*Consider experimental design, metrics, and potential confounding factors.*

*Exercise 18.12 (Advanced Topic 4).* Compare different approaches to solving a problem from this chapter.

**Hint:**

*Consider computational complexity, accuracy, and practical considerations.*

*Exercise 18.13 (Advanced Topic 5).* Derive a mathematical relationship or prove a theorem from this chapter.

**Hint:**

*Start with the definitions and work through the logical steps.*

*Exercise 18.14 (Advanced Topic 6).* Implement a practical solution to a problem discussed in this chapter.

**Hint:**

*Consider the implementation details and potential challenges.*

*Exercise 18.15 (Advanced Topic 7).* Evaluate the limitations and potential improvements of techniques from this chapter.

**Hint:**

*Consider both theoretical limitations and practical constraints.*

# Chapter 19

## Approximate Inference

This chapter explores methods for tractable inference in complex probabilistic models.

### Learning Objectives

1. Variational inference and sampling-based approaches
2. ELBO objectives and coordinate ascent updates for simple models
3. Amortized inference and its benefits and limitations
4. Approximation quality using diagnostics and bounds

## Intuition

Exact posteriors are rare. We instead optimize over a family of tractable distributions or draw dependent samples, trading bias and variance to approximate expectations we care about.

For example, when trying to understand a patient's disease risk from their genetic data, we need to integrate information across thousands of genes and environmental factors, which creates an intractable posterior distribution. Like a detective trying to piece together a complex case from scattered evidence, approximate inference methods help us make sense of overwhelming amounts of uncertain information by focusing on the most important patterns and relationships.

Think of approximate inference as using a simplified map instead of a detailed satellite image when navigating a city. While the simplified map might miss some details, it captures the essential information needed to reach your destination efficiently.

Similarly, approximate inference methods trade some accuracy for computational efficiency, enabling us to make decisions in complex probabilistic models that would otherwise be impossible to solve exactly.

## 19.1 Variational Inference ★

Variational inference transforms intractable posterior inference into an optimization problem by approximating the true posterior with a simpler, tractable distribution that can be efficiently optimized.

### 19.1.1 Evidence Lower Bound (ELBO)

For latent variable model with intractable posterior  $p(z|x)$ , we approximate with  $q(z)$ . The mathematical derivation shows how we can bound the log-evidence:

$$\log p(\mathbf{x}) = \mathbb{E}_{q(\mathbf{z})} [\log p(\mathbf{x})] \quad (19.1)$$

$$= \mathbb{E}_{q(\mathbf{z})} \left[ \log \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{z}|\mathbf{x})} \right] \quad (19.2)$$

$$= \mathbb{E}_{q(\mathbf{z})} \left[ \log \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \right] + D_{KL}(q(\mathbf{z}) \| p(\mathbf{z}|\mathbf{x})) \quad (19.3)$$

$$\geq \mathbb{E}_{q(\mathbf{z})} \left[ \log \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \right] = \mathcal{L}(q) \quad (19.4)$$

The key insight is that the KL divergence  $D_{KL}(q(\mathbf{z}) \| p(\mathbf{z}|\mathbf{x}))$  is always non-negative, so the ELBO  $\mathcal{L}(q)$  provides a lower bound on the log-evidence. The motivation behind this mathematical framework is that maximizing the ELBO simultaneously maximizes the log-evidence and minimizes the KL divergence between our approximation  $q(\mathbf{z})$  and the true posterior  $p(\mathbf{z}|\mathbf{x})$ . This transforms the intractable inference problem into a tractable optimization problem where we can use standard optimization techniques to find the best approximation.

### 19.1.2 Variational Family

The choice of variational family determines the expressiveness and computational tractability of our approximation. We must balance between capturing the complexity of the true posterior and maintaining computational efficiency.

**Mean field:** Fully factorized approximation assumes all variables are independent:

$$q(\mathbf{z}) = \prod_{i=1}^n q_i(z_i) \quad (19.5)$$

This equation shows that the joint distribution factors into a product of individual marginals, making computation tractable but potentially missing important dependencies between variables.

**Structured:** Allow some dependencies by grouping variables into cliques:

$$q(\mathbf{z}) = \prod_c q_c(\mathbf{z}_c) \quad (19.6)$$

This equation permits dependencies within each clique  $c$  while maintaining independence between cliques, providing a middle ground between mean field and full posterior approximation. The trade-off between expressiveness and tractability is fundamental to variational inference, as more expressive families can better approximate the true posterior but require more computational resources.

### 19.1.3 Coordinate Ascent VI

Coordinate ascent variational inference optimizes each factor of the variational distribution iteratively while keeping all other factors fixed. The update equation for each factor is:

$$q_j^*(z_j) \propto \exp (\mathbb{E}_{q_{-j}} [\log p(\mathbf{z}, \mathbf{x})]) \quad (19.7)$$

This equation shows that the optimal factor  $q_j^*(z_j)$  is proportional to the exponential of the expected log-joint probability, where the expectation is taken over all other factors  $q_{-j}$ . The key insight is that each factor can be optimized independently given the others, making the optimization problem tractable. This approach is guaranteed to converge to a local optimum of the ELBO, providing a principled way to find good approximations to the true posterior.

### 19.1.4 Stochastic Variational Inference

Stochastic variational inference addresses the scalability limitations of traditional variational inference by using stochastic gradients and mini-batch processing to handle large datasets efficiently. The method enables variational inference to scale to

massive datasets by processing only small subsets of data at each iteration, making it practical for modern machine learning applications.

The approach combines mini-batch data processing with Monte Carlo estimation of expectations, allowing the algorithm to work with large datasets without requiring the full dataset to be loaded into memory. The reparameterization trick provides a crucial variance reduction technique that enables stable optimization by expressing the stochastic gradients in a form that has lower variance than naive Monte Carlo estimation. This combination of techniques makes stochastic variational inference the method of choice for large-scale probabilistic modeling, enabling the deployment of sophisticated probabilistic models in production systems.

## 19.2 Mean Field Approximation ★

Mean field approximation is the simplest form of variational inference that assumes all latent variables are independent, providing a computationally efficient but potentially limited approximation to complex posterior distributions.

### 19.2.1 Fully Factorized Approximation

The fully factorized approximation assumes that all latent variables are independent, leading to the factorization:

$$q(\mathbf{z}) = \prod_{i=1}^n q_i(z_i) \quad (19.8)$$

This equation shows that the joint variational distribution factors into a product of individual marginals, where each  $q_i(z_i)$  represents the approximate posterior for variable  $z_i$ . This factorization makes the optimization problem tractable by allowing each factor to be optimized independently, but it may miss important dependencies between variables that exist in the true posterior.

### 19.2.2 Update Equations

The mean field update equations provide the optimal form for each factor when all other factors are held fixed. For each variable  $z_j$ , the optimal factor is:

$$\log q_j^*(z_j) = \mathbb{E}_{i \neq j}[\log p(\mathbf{z}, \mathbf{x})] + \text{const} \quad (19.9)$$

This equation shows that the log of the optimal factor  $q_j^*(z_j)$  is proportional to the expected log-joint probability, where the expectation is taken over all other variables  $i \neq j$ . The constant term ensures proper normalization. The algorithm iterates these updates until convergence, with each update guaranteed to increase the ELBO, leading to a local optimum of the variational objective.

### 19.2.3 Properties

Mean field approximation exhibits several important properties that make it both useful and limited in practice. The method tends to underestimate variance, leading to overconfident predictions that may not capture the full uncertainty in the posterior distribution. This bias arises from the independence assumption, which prevents the approximation from capturing correlations between variables that could lead to higher uncertainty estimates.

Despite these limitations, mean field approximation remains computationally efficient and often provides surprisingly good approximations in practice. The efficiency comes from the simple factorization that allows each factor to be optimized independently, making the algorithm scalable to high-dimensional problems. The practical success of mean field methods in many applications demonstrates that the independence assumption, while theoretically limiting, often captures enough of the important structure to be useful for decision-making and prediction tasks.

## 19.3 Loopy Belief Propagation ★

Loopy belief propagation extends the exact message passing algorithm from trees to general graphs with cycles, providing an approximate inference method that often works well despite lacking convergence guarantees.

### 19.3.1 Message Passing

Message passing algorithms compute marginal probabilities by iteratively passing messages between nodes in a graphical model. The message from node  $i$  to node  $j$  is computed as:

$$m_{i \rightarrow j}(x_j) = \sum_{x_i} \psi(x_i, x_j) \psi(x_i) \prod_{k \in N(i) \setminus j} m_{k \rightarrow i}(x_i) \quad (19.10)$$

This equation shows that the message  $m_{i \rightarrow j}(x_j)$  combines the local potential  $\psi(x_i, x_j)$  between nodes  $i$  and  $j$ , the node potential  $\psi(x_i)$  at node  $i$ , and all incoming messages from other neighbors  $k \in N(i) \setminus j$ . The algorithm iteratively updates these messages until convergence, providing an efficient way to compute approximate marginals in complex graphical models.

### 19.3.2 Beliefs

The beliefs represent the approximate marginal probabilities for each variable, computed from the incoming messages:

$$b_i(x_i) \propto \psi(x_i) \prod_{j \in N(i)} m_{j \rightarrow i}(x_i) \quad (19.11)$$

This equation shows that the belief  $b_i(x_i)$  at node  $i$  is proportional to the product of the local node potential  $\psi(x_i)$  and all incoming messages  $m_{j \rightarrow i}(x_i)$  from neighboring nodes  $j \in N(i)$ . The beliefs provide the final approximation to the marginal probabilities after the message passing algorithm has converged, serving as the output of the loopy belief propagation algorithm.

### 19.3.3 Exact on Trees

For tree-structured graphs, belief propagation converges to exact marginals in a finite number of iterations. The key insight is that trees have no cycles, so each message is computed exactly once during the forward and backward passes of the algorithm. The mathematical foundation relies on the fact that the joint distribution factors according to the tree structure, and the message passing equations correspond exactly to the marginalization operations needed to compute the true marginals. This makes belief propagation on trees both exact and efficient, with computational complexity linear in the number of nodes.

### 19.3.4 Loopy Graphs

When applied to graphs with cycles, loopy belief propagation faces several challenges that distinguish it from the exact algorithm on trees. The presence of cycles means that messages can circulate indefinitely, potentially preventing convergence to a fixed point. Despite this theoretical limitation, the algorithm often provides surprisingly good approximations in practice, making it a valuable tool for approximate inference in complex graphical models.

The success of loopy belief propagation in applications like error-correcting codes and computer vision demonstrates its practical utility, even when theoretical convergence guarantees are absent. The algorithm's ability to capture local dependencies and propagate information through the graph structure makes it particularly effective for problems where the true posterior has complex dependencies that would be difficult to capture with simpler approximation methods. This practical success has made loopy belief propagation one of the most widely used approximate inference methods in machine learning and computer vision applications.

## 19.4 Expectation Propagation ★

Expectation propagation approximates complex probability distributions by replacing each factor with a simpler distribution from an exponential family, enabling tractable inference while preserving important statistical properties. The mathematical foundation is:

$$p(\mathbf{x}) = \frac{1}{Z} \prod_i f_i(\mathbf{x}) \approx \frac{1}{Z} \prod_i \tilde{f}_i(\mathbf{x}) \quad (19.12)$$

This equation shows that the true distribution  $p(\mathbf{x})$  is approximated by replacing each complex factor  $f_i(\mathbf{x})$  with a simpler factor  $\tilde{f}_i(\mathbf{x})$  from a tractable family. The algorithm iteratively refines these approximations to match the moments of the true distribution, providing a principled way to approximate complex posteriors. Expectation propagation is particularly effective for multi-modal posteriors where mean field approximation would fail, as it can capture multiple modes by using more expressive approximating families.

## 19.5 Real World Applications

Approximate inference makes complex probabilistic reasoning practical. When exact inference is intractable, approximate methods enable deploying sophisticated probabilistic models in real-world systems requiring fast, scalable inference.

### 19.5.1 Autonomous Systems

Autonomous systems represent one of the most demanding applications of approximate inference, where real-time decision making under uncertainty is essential for safety and performance. These systems must process vast amounts of sensor data, maintain probabilistic beliefs about their environment, and make decisions that could have serious consequences if incorrect. Robot navigation in uncertain environments exemplifies these challenges, where robots operating in homes or warehouses face sensor noise and unpredictable obstacles that make exact inference impossible. Approximate inference methods like particle filters and variational methods enable real-time localization and mapping despite these uncertainties, allowing robots to continuously update their beliefs about position and surroundings. The robot makes navigation decisions based on approximate posterior distributions computed in milliseconds, demonstrating how approximate inference enables practical autonomous systems that would be impossible with exact methods.

Drone flight control presents similar challenges, where autonomous drones must track their position, velocity, and orientation while compensating for wind and sensor errors. Extended Kalman filters, which are a form of approximate inference, provide real-time state estimation that enables stable flight. This makes applications from package delivery to aerial photography practical, showing how approximate inference techniques can be adapted to different types of autonomous systems with varying requirements and constraints.

### 19.5.2 Personalized Medicine

Personalized medicine represents a transformative application of approximate inference, where the goal is to tailor treatments to individual patients based on their unique genetic, clinical, and lifestyle factors. This approach requires integrating vast amounts of heterogeneous data to make predictions about disease risk, treatment response, and optimal interventions.

Genomic data analysis exemplifies these challenges, where understanding disease risk from genetic variants requires integrating evidence across thousands of genes with complex interactions. Approximate inference in Bayesian models combines genetic data with clinical information, computing posterior probabilities for disease risk and treatment response that would be impossible to calculate exactly. This enables precision medicine decisions about preventive care and drug selection, demonstrating how approximate inference can handle the complexity of modern genomic medicine while providing actionable insights for patient care.

Real-time patient monitoring in intensive care units presents another critical application, where monitoring systems must track dozens of vital signs and detect deterioration early to prevent adverse events. Approximate inference in hierarchical models captures the normal variation versus concerning trends in patient data, enabling systems to trigger alerts while avoiding false

alarms that cause alarm fatigue among medical staff. This balance between sensitivity and specificity is crucial for patient safety and demonstrates how approximate inference can be tailored to specific clinical requirements.

### 19.5.3 Content Recommendation

Content recommendation systems represent one of the most visible applications of approximate inference, where the challenge is to provide personalized content to billions of users in real-time while handling the massive scale and complexity of modern digital platforms. These systems must balance personalization with exploration, handle new users with minimal history, and provide recommendations that are both relevant and diverse.

Real-time feed ranking exemplifies these challenges, where social media platforms must rank posts for billions of users continuously, requiring inference methods that can scale to massive user bases while capturing uncertainty in preference estimates. Approximate inference in probabilistic models estimates user preferences from sparse interactions, computing rankings in milliseconds using variational methods that enable scaling while maintaining the probabilistic reasoning benefits that make recommendations robust and interpretable.

The explore-exploit tradeoff represents another critical challenge, where recommendation systems must balance showing proven content that users are likely to enjoy versus trying new items that might lead to discovery. Approximate Bayesian inference maintains uncertainty estimates about item quality, enabling the implementation of principled exploration strategies like Thompson sampling that prevent recommendation systems from getting stuck showing only popular content. This balance is essential for maintaining user engagement and discovering new content that users might enjoy.

### 19.5.4 Natural Language Systems

Natural language systems represent one of the most complex applications of approximate inference, where the goal is to understand and generate human language at scale while handling the inherent ambiguity and complexity of linguistic communication. These systems must process vast amounts of text, maintain probabilistic beliefs about meaning and intent, and provide responses that are both accurate and contextually appropriate.

Document understanding exemplifies these challenges, where extracting structured information from documents like contracts, medical records, and scientific papers involves uncertain entity recognition and relation extraction that requires sophisticated probabilistic reasoning. Approximate inference in structured models provides confidence estimates that enable systems to flag uncertain extractions for human verification while automating clear cases, demonstrating how approximate inference can be integrated into human-AI collaborative workflows.

Conversational AI systems face similar challenges, where chatbots must maintain beliefs about conversation state and user intent through approximate inference that can handle the ambiguity inherent in natural language. This enables systems to handle uncertainty gracefully by asking clarifying questions when uncertain about user goals rather than guessing wrongly, showing how approximate inference can be used to create more robust and user-friendly conversational interfaces.

### 19.5.5 Why Approximation Is Essential

Approximate inference is not merely a computational convenience but a fundamental necessity for making probabilistic modeling practical in real-world applications. The benefits of approximate inference extend far beyond simple computational efficiency, encompassing scalability, speed, flexibility, and the preservation of uncertainty quantification that makes probabilistic reasoning valuable.

Scalability represents perhaps the most critical benefit, as approximate inference enables the deployment of complex probabilistic models on real-world data sizes that would be impossible to handle with exact methods. This scalability is essential for modern applications that must process massive datasets while maintaining the sophisticated modeling capabilities that make probabilistic approaches valuable. The ability to handle large-scale problems while preserving model complexity is what makes approximate inference indispensable for practical machine learning systems.

Speed is another crucial advantage, as approximate inference provides results fast enough for interactive applications where real-time decision making is essential. This speed enables the deployment of probabilistic models in applications ranging from autonomous systems to recommendation engines, where delays can have serious consequences for user experience or system

performance. The combination of speed and accuracy makes approximate inference the method of choice for many practical applications where exact inference would be too slow or impossible.

These applications demonstrate that approximate inference is not a compromise but rather what makes probabilistic modeling practical at scale. The success of approximate inference methods across diverse domains shows that they provide the right balance between theoretical rigor and computational feasibility, enabling the deployment of sophisticated probabilistic models in real-world systems that would otherwise be impossible to build.

## Key Takeaways

### Key Takeaways 19

- **VI vs. MCMC**: bias-variance trade-offs define suitability for different applications.
- **ELBO optimisation** turns intractable inference into tractable learning problems.
- **Amortisation** speeds inference but can underfit the posterior distribution.
- **Approximation quality** depends on the choice of variational family and optimization method.
- **Practical deployment** requires balancing accuracy, speed, and scalability constraints.

## Exercises

### Easy

*Exercise 19.1 (Why Approximate?). Explain why exact inference is intractable in many models.*

**Hint:**

*Partition function; high-dimensional integration.*

*Exercise 19.2 (ELBO Connection). Relate ELBO to KL divergence between  $q$  and  $p$ .*

**Hint:**

$\log p(x) = \text{ELBO} + D_{KL}(q||p)$ .

*Exercise 19.3 (Mean-Field Assumption). State the mean-field independence assumption.*

**Hint:**

*Factored variational distribution.*

*Exercise 19.4 (MCMC vs. VI). Compare MCMC and variational inference trade-offs.*

**Hint:**

*Asymptotic exactness vs. computational speed.*

## Medium

*Exercise 19.5 (Coordinate Ascent VI).* Derive the coordinate ascent update for a simple model.

**Hint:**

*Fix all but one factor; optimise w.r.t. remaining factor.*

*Exercise 19.6 (Importance Sampling).* Explain how importance sampling estimates expectations.

**Hint:**

*Reweight samples from proposal distribution.*

## Hard

*Exercise 19.7 (Amortised Inference).* Analyse the trade-offs of amortised inference in VAEs.

**Hint:**

*Amortisation gap; scalability.*

*Exercise 19.8 (Reparameterisation Gradients).* Derive the reparameterisation gradient for a Gaussian variational distribution.

**Hint:**

*$z = \mu + \sigma\epsilon$  where  $\epsilon \sim \mathcal{N}(0, 1)$ .*

*Exercise 19.9 (Advanced Topic 1).* Explain a key concept from this chapter and its practical applications.

**Hint:**

*Consider the theoretical foundations and real-world implications.*

*Exercise 19.10 (Advanced Topic 2).* Analyse the relationship between different techniques covered in this chapter.

**Hint:**

*Look for connections and trade-offs between methods.*

*Exercise 19.11 (Advanced Topic 3).* Design an experiment to test a hypothesis related to this chapter's content.

**Hint:**

*Consider experimental design, metrics, and potential confounding factors.*

*Exercise 19.12 (Advanced Topic 4).* Compare different approaches to solving a problem from this chapter.

**Hint:**

*Consider computational complexity, accuracy, and practical considerations.*

*Exercise 19.13 (Advanced Topic 5).* Derive a mathematical relationship or prove a theorem from this chapter.

**Hint:**

*Start with the definitions and work through the logical steps.*

*Exercise 19.14 (Advanced Topic 6).* Implement a practical solution to a problem discussed in this chapter.

**Hint:**

*Consider the implementation details and potential challenges.*

*Exercise 19.15 (Advanced Topic 7).* Evaluate the limitations and potential improvements of techniques from this chapter.

**Hint:**

*Consider both theoretical limitations and practical constraints.*



# Chapter 20

# Deep Generative Models

This chapter examines modern approaches to generating new data samples using deep learning.

## ⌚ Learning Objectives

1. VAEs, GANs, and flow/diffusion models conceptually and practically
2. Training objectives and sampling procedures for each model class
3. Generative model evaluation with proper metrics and qualitative checks
4. Trade-offs in likelihood, sample quality, and mode coverage

## Intuition

Generative models are machine learning systems that learn to create new data samples that resemble a given dataset. For example, after training on thousands of cat photos, a generative model can produce entirely new, realistic cat images that never existed before.

Think of generative models as digital artists who study masterpieces to understand artistic style, then create original works in that same style. Just as a painter learns brushstrokes and color palettes from studying great works, these models learn the underlying patterns and structures in data to generate novel samples.

Generative models learn data distributions in complementary ways: explicit likelihoods, implicit adversarial training, or score-based diffusion. Each chooses a tractable learning signal that captures structure while managing complexity.

## 20.1 Variational Autoencoders (VAEs) ★

(See also Chapter 14 for detailed VAE coverage.)

### 20.1.1 Recap

VAE learns latent representation  $\mathbf{z}$  and decoder  $p_{\theta}(\mathbf{x}|\mathbf{z})$ :

$$\max_{\theta, \phi} \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} [\log p_{\theta}(\mathbf{x}|\mathbf{z})] - D_{KL}(q_{\phi}(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})) \quad (20.1)$$

The VAE objective consists of two terms: the reconstruction loss  $\mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} [\log p_{\theta}(\mathbf{x}|\mathbf{z})]$  encourages the decoder to accurately reconstruct input data from latent codes, while the KL divergence term  $D_{KL}(q_{\phi}(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}))$  regularizes the encoder to produce

latent distributions close to the prior  $p(\mathbf{z})$ . This balance ensures both faithful reconstruction and meaningful latent representations.

### 20.1.2 Conditional VAEs

Generate conditioned on class or attributes:

$$\max \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, y)} [\log p(\mathbf{x}|\mathbf{z}, y)] - D_{KL}(q(\mathbf{z}|\mathbf{x}, y) \| p(\mathbf{z})) \quad (20.2)$$

Conditional VAEs extend the standard VAE framework by incorporating conditioning information  $y$  (such as class labels or attributes) into both the encoder and decoder. The encoder  $q(\mathbf{z}|\mathbf{x}, y)$  learns to map input data and conditioning information to latent codes, while the decoder  $p(\mathbf{x}|\mathbf{z}, y)$  generates samples conditioned on both the latent code and the conditioning variable. This enables controlled generation of specific types of content.

### 20.1.3 Disentangled Representations

$\beta$ -VAE: Increase KL weight for disentanglement

$$\mathcal{L} = \mathbb{E}_q [\log p(\mathbf{x}|\mathbf{z})] - \beta D_{KL}(q(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})) \quad (20.3)$$

The  $\beta$ -VAE introduces a hyperparameter  $\beta$  that controls the strength of the KL regularization term. When  $\beta > 1$ , the model is encouraged to learn more independent latent factors, where each dimension of  $\mathbf{z}$  corresponds to a distinct, interpretable attribute of the data. This disentanglement is achieved by penalizing the mutual information between latent dimensions, forcing the model to encode different aspects of variation in separate latent variables.

## 20.2 Generative Adversarial Networks (GANs) ★

GANs train two competing neural networks—a generator that creates fake data and a discriminator that tries to detect the fakes—in an adversarial game that eventually produces highly realistic synthetic samples.

### 20.2.1 Core Idea

The fundamental concept behind GANs is elegantly simple yet powerful. Imagine an art forger trying to create paintings so convincing that even expert art critics cannot distinguish them from genuine masterpieces. The forger (generator) continuously improves their technique based on feedback from the critics (discriminator), while the critics become more sophisticated at detecting fakes. This adversarial relationship drives both parties to become increasingly skilled—the forger learns to create more realistic art, while the critics develop sharper detection abilities.

In the GAN framework, the generator  $G$  takes random noise as input and transforms it into synthetic data samples that should be indistinguishable from real data. The discriminator  $D$  acts as a binary classifier, receiving both real and generated samples and learning to distinguish between them. As training progresses, the generator becomes increasingly adept at fooling the discriminator, while the discriminator becomes better at detecting subtle differences between real and fake samples.

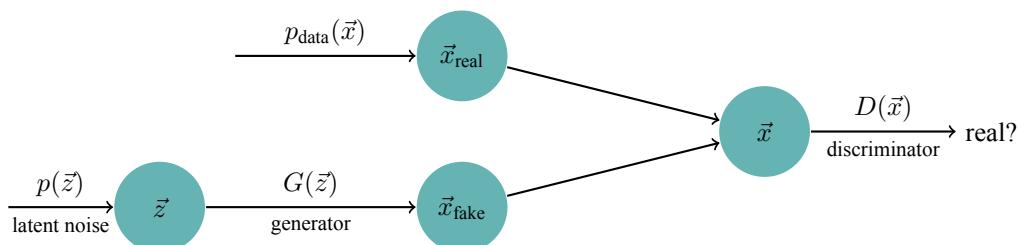


Figure 20.1: GAN architecture showing the adversarial relationship between generator and discriminator.

## 20.2.2 Objective

Minimax game:

$$\min_G \max_D \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))] \quad (20.4)$$

The GAN objective represents a minimax game where the discriminator aims to maximize its ability to distinguish real from fake samples, while the generator seeks to minimize the discriminator's accuracy. The first term  $\mathbb{E}_{\mathbf{x} \sim p_{\text{data}}} [\log D(\mathbf{x})]$  encourages the discriminator to assign high probability to real samples, while the second term  $\mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))]$  penalizes the discriminator for correctly identifying generated samples, simultaneously training the generator to produce more convincing fakes.

## 20.2.3 Training Procedure

The training process alternates between updating the discriminator and generator in a carefully orchestrated dance. During each training iteration, the discriminator is first updated to improve its ability to distinguish real from fake samples. This involves maximizing the objective  $\max_D \mathbb{E}_{\mathbf{x}} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}} [\log(1 - D(G(\mathbf{z})))]$ , which simultaneously rewards correct identification of real samples and penalizes misclassification of generated samples.

Following the discriminator update, the generator is trained to produce samples that are more likely to fool the updated discriminator. The generator's objective  $\min_G \mathbb{E}_{\mathbf{z}} [\log(1 - D(G(\mathbf{z})))]$  encourages it to generate samples that the discriminator will classify as real. This alternating optimization process continues until the generator produces samples that are indistinguishable from real data, at which point the discriminator can no longer provide meaningful gradients for further improvement.

## 20.2.4 Training Challenges

GAN training presents several formidable challenges that can derail the delicate balance between generator and discriminator. Mode collapse occurs when the generator discovers a small subset of samples that consistently fool the discriminator, causing it to generate only these limited variations rather than exploring the full diversity of the data distribution. This phenomenon is particularly problematic when the discriminator becomes too strong relative to the generator, creating a feedback loop where the generator finds it easier to repeatedly generate the same convincing samples rather than learning the full data manifold.

Training instability manifests as oscillatory behavior where the generator and discriminator continuously outmaneuver each other without reaching a stable equilibrium. The discriminator may become too powerful, providing vanishing gradients to the generator, or the generator may become too skilled, causing the discriminator to lose its ability to provide meaningful learning signals. This instability often leads to non-convergence, where the training process fails to reach a meaningful solution despite extensive optimization efforts.

## 20.2.5 GAN Variants

The evolution of GAN architectures has addressed many of the fundamental challenges in adversarial training through innovative design choices and theoretical insights. DCGAN introduced architectural guidelines that stabilized training by using strided convolutions, batch normalization, and ReLU activations, establishing a foundation for successful deep convolutional GANs. WGAN revolutionized the field by replacing the Jensen-Shannon divergence with the Wasserstein distance, providing more stable gradients and better convergence properties that significantly reduced training instability.

StyleGAN represents a breakthrough in high-quality image generation by introducing a style-based generator architecture that separates high-level attributes from stochastic variation, enabling fine-grained control over generated images. Conditional GANs extend the basic framework by incorporating auxiliary information such as class labels, allowing for controlled generation of specific types of content. CycleGAN addresses the challenge of unpaired image-to-image translation by introducing cycle consistency constraints, enabling style transfer between domains without requiring paired training data.

## 20.3 Normalizing Flows ★

Normalizing flows transform simple probability distributions through a series of invertible mappings to model complex data distributions while maintaining exact likelihood computation.

### 20.3.1 Key Idea

Transform simple distribution (e.g., Gaussian) through invertible mappings:

$$\mathbf{x} = f_{\theta}(\mathbf{z}), \quad \mathbf{z} \sim p_z(\mathbf{z}) \quad (20.5)$$

The fundamental insight of normalizing flows is that complex probability distributions can be constructed by applying a series of invertible transformations to a simple base distribution. Starting from a tractable distribution like a standard Gaussian  $p_z(\mathbf{z})$ , the flow model learns a bijective function  $f_{\theta}$  that maps samples from this simple distribution to the complex data distribution. The invertibility requirement ensures that we can compute exact likelihoods and perform exact sampling, making flows particularly attractive for applications requiring precise probability estimates.

### 20.3.2 Change of Variables

Density transforms as:

$$p_x(\mathbf{x}) = p_z(f^{-1}(\mathbf{x})) \left| \det \frac{\partial f^{-1}}{\partial \mathbf{x}} \right| \quad (20.6)$$

or equivalently:

$$\log p_x(\mathbf{x}) = \log p_z(\mathbf{z}) - \log \left| \det \frac{\partial f}{\partial \mathbf{z}} \right| \quad (20.7)$$

The change of variables formula is the mathematical foundation that enables normalizing flows to compute exact likelihoods. When transforming a random variable through an invertible function, the probability density changes according to the Jacobian determinant of the transformation. The first equation shows how the density at a point  $\mathbf{x}$  in the data space relates to the density at the corresponding point  $f^{-1}(\mathbf{x})$  in the latent space, scaled by the absolute value of the Jacobian determinant. The log-space formulation is computationally more stable and directly relates the log-likelihood of the data to the log-likelihood of the latent variable minus the log-determinant of the transformation.

### 20.3.3 Requirements

The success of normalizing flows depends critically on the design of the transformation function  $f$ , which must satisfy two fundamental requirements. First, the function must be invertible, meaning that for every output  $\mathbf{x}$ , there exists a unique input  $\mathbf{z}$  such that  $\mathbf{x} = f(\mathbf{z})$ . This bijectivity is essential for both likelihood computation and sampling, as it ensures a one-to-one correspondence between the latent and data spaces. Second, the function must have a tractable Jacobian determinant, as computing  $\log |\det \frac{\partial f}{\partial \mathbf{z}}|$  is required for likelihood evaluation. These constraints significantly limit the class of functions that can be used in practice, leading to the development of specialized architectures that satisfy both requirements while maintaining sufficient expressiveness to model complex distributions.

### 20.3.4 Flow Architectures

The design of invertible transformations has led to several innovative architectural paradigms that balance expressiveness with computational tractability. Coupling layers represent a particularly elegant solution by partitioning the input dimensions into two halves, where one half is transformed based on the other half, ensuring invertibility while allowing complex dependencies. This approach enables the modeling of intricate conditional relationships while maintaining the ability to compute exact likelihoods through the structured Jacobian.

Autoregressive flows extend this idea by modeling each dimension as a function of all previous dimensions, creating a natural ordering that facilitates both sampling and likelihood computation. This autoregressive structure allows for highly expressive

transformations while maintaining computational efficiency through the triangular nature of the resulting Jacobian matrix. Continuous normalizing flows represent a more recent innovation that leverages neural ordinary differential equations to model continuous-time transformations, providing a more flexible framework for learning complex flow dynamics while maintaining the theoretical guarantees of normalizing flows.

### 20.3.5 Advantages

Normalizing flows offer several compelling advantages that distinguish them from other generative modeling approaches. The ability to compute exact likelihoods provides a principled way to evaluate model performance and compare different architectures, addressing a fundamental limitation of GANs where likelihood estimation is intractable. This exact likelihood computation also enables applications requiring precise probability estimates, such as anomaly detection and uncertainty quantification.

The exact sampling capability ensures that generated samples are drawn from the true learned distribution, eliminating the approximation errors inherent in other methods. Unlike GANs, which require careful balance between generator and discriminator, normalizing flows provide stable training dynamics through standard maximum likelihood optimization. This stability makes flows particularly attractive for applications where reliable convergence is crucial, while their invertibility enables bidirectional mapping between latent and data spaces, opening up possibilities for data compression and representation learning.

## 20.4 Diffusion Models ★

Diffusion models learn to reverse a gradual noise corruption process, starting from pure noise and iteratively denoising to generate high-quality samples through a learned denoising network.

### 20.4.1 Forward Process

Gradually add noise over  $T$  steps:

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}) \quad (20.8)$$

Eventually  $\mathbf{x}_T \approx \mathcal{N}(\mathbf{0}, \mathbf{I})$ .

The forward process systematically corrupts clean data by adding Gaussian noise at each timestep, following a predefined noise schedule  $\{\beta_t\}_{t=1}^T$ . The parameterization  $\sqrt{1 - \beta_t} \mathbf{x}_{t-1}$  ensures that the signal-to-noise ratio decreases gradually, while  $\beta_t \mathbf{I}$  controls the amount of noise added at each step. This process is designed to be analytically tractable, allowing efficient sampling and training. The key insight is that after sufficient timesteps, the data becomes approximately Gaussian noise, providing a well-defined starting point for the reverse process.

### 20.4.2 Reverse Process

Learn to denoise (reverse diffusion):

$$p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \quad (20.9)$$

The reverse process represents the core learning challenge in diffusion models, where a neural network must learn to reverse the forward corruption process. The model  $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)$  predicts the parameters of a Gaussian distribution that should generate the previous timestep  $\mathbf{x}_{t-1}$  given the current noisy state  $\mathbf{x}_t$  and timestep  $t$ . The mean  $\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)$  and covariance  $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)$  are learned functions that capture the complex dependencies needed to reverse the noise corruption, effectively learning to "denoise" the data step by step.

### 20.4.3 Training

Predict noise  $\epsilon_\theta(\mathbf{x}_t, t)$  at each step:

$$\mathcal{L} = \mathbb{E}_{t, \mathbf{x}_0, \epsilon} [\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2] \quad (20.10)$$

The training objective simplifies the complex reverse process by focusing on noise prediction rather than directly modeling the reverse distribution. The loss function  $\mathcal{L}$  trains the network  $\epsilon_\theta(\mathbf{x}_t, t)$  to predict the noise  $\epsilon$  that was added to the clean data  $\mathbf{x}_0$  to produce the noisy observation  $\mathbf{x}_t$ . This approach leverages the fact that the forward process is analytically tractable, allowing efficient computation of the training targets. The expectation is taken over random timesteps  $t$ , clean data samples  $\mathbf{x}_0$ , and noise realizations  $\epsilon$ , ensuring the model learns to denoise across all noise levels and data types.

### 20.4.4 Sampling

Start from noise and iteratively denoise:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z} \quad (20.11)$$

The sampling process begins with pure Gaussian noise  $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  and iteratively applies the learned denoising function to generate clean samples. The sampling equation combines the predicted noise  $\epsilon_\theta(\mathbf{x}_t, t)$  with the current noisy state  $\mathbf{x}_t$  to estimate the previous timestep  $\mathbf{x}_{t-1}$ . The coefficients  $\alpha_t$  and  $\bar{\alpha}_t$  are derived from the noise schedule and ensure consistency with the forward process, while  $\sigma_t \mathbf{z}$  adds stochasticity to prevent the sampling from becoming deterministic. This iterative denoising process gradually transforms noise into realistic data samples.

### 20.4.5 The Algorithm

The diffusion model sampling algorithm can be expressed as follows:

---

#### Algorithm 8 Diffusion Model Sampling Algorithm

---

```

1: Input: Trained denoising network  $\epsilon_\theta(\mathbf{x}_t, t)$ , noise schedule  $\{\alpha_t, \bar{\alpha}_t\}_{t=1}^T$ 
2: Output: Generated sample  $\mathbf{x}_0$ 
3:
4: Sample initial noise:  $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
5: for  $t = T, T - 1, \dots, 1$  do
6:   Predict noise:  $\hat{\epsilon} = \epsilon_\theta(\mathbf{x}_t, t)$ 
7:   Compute denoised estimate:  $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \hat{\epsilon} \right)$ 
8:   if  $t > 1$  then
9:     Add stochasticity:  $\mathbf{x}_{t-1} = \mathbf{x}_{t-1} + \sigma_t \mathbf{z}$  where  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
10:  end if
11: end for
12: return  $\mathbf{x}_0$ 
```

---

### 20.4.6 Advantages

Diffusion models have emerged as the state-of-the-art approach for high-quality generative modeling, powering breakthrough applications like DALL-E 2, Stable Diffusion, and Midjourney. Their success stems from several key advantages that address fundamental limitations of previous approaches. Unlike GANs, diffusion models provide stable training dynamics without the delicate balance required between generator and discriminator, making them more reliable for large-scale applications.

The strong theoretical foundations of diffusion models provide principled guarantees about convergence and sample quality, addressing the theoretical gaps that plagued earlier generative approaches. The iterative denoising process naturally supports conditioning on various modalities, enabling text-to-image generation, image editing, and other controlled generation tasks. This flexibility, combined with their proven ability to generate photorealistic images and coherent text, has established diffusion models as the dominant paradigm for modern generative AI systems.

## 20.5 Applications and Future Directions ★

Generative models are transforming creative industries, scientific discovery, and everyday applications by enabling the creation of realistic synthetic content across multiple modalities.

### 20.5.1 Current Applications

The current landscape of generative model applications spans across multiple domains, each demonstrating the transformative potential of these technologies. Image generation has reached unprecedented levels of quality and controllability, with systems like DALL-E and Stable Diffusion enabling users to create photorealistic images from text descriptions. These models have revolutionized creative workflows, allowing artists and designers to rapidly prototype concepts, generate variations, and explore creative possibilities that would be impossible through traditional means.

Text generation has been equally transformative, with large language models like the GPT family demonstrating remarkable capabilities in natural language understanding and generation. These models have found applications in code generation, creative writing, and conversational AI, fundamentally changing how we interact with computers and access information. The ability to generate coherent, contextually appropriate text has opened new possibilities for content creation, education, and human-computer interaction.

Audio and speech generation have advanced significantly, with text-to-speech systems achieving near-human quality and music generation models creating original compositions in various styles. Voice conversion technologies enable personalized speech synthesis, while music generation models assist composers and content creators in exploring new musical ideas. Video generation represents the next frontier, with models capable of predicting future frames, synthesizing new video content, and creating animations that were previously impossible to generate automatically.

Scientific applications of generative models are particularly promising, with models being used to design novel molecules for drug discovery, predict protein structures, and discover new materials with desired properties. These applications demonstrate how generative models can accelerate scientific discovery by exploring vast spaces of possibilities that would be impractical to investigate through traditional experimental approaches.

### 20.5.2 Future Directions

The future of generative models promises even more sophisticated capabilities and broader applications. Controllability represents a key frontier, where researchers are developing methods for fine-grained control over generation processes, enabling users to specify not just what to generate, but how to generate it with precise control over style, composition, and semantic attributes. This enhanced controllability will make generative models more useful for professional applications where specific requirements must be met.

Efficiency improvements are crucial for making generative models more accessible and practical. Current models often require significant computational resources and time for generation, limiting their widespread adoption. Future developments will focus on faster sampling algorithms, more efficient architectures, and smaller models that can run on consumer hardware while maintaining high quality. Multi-modal integration represents another exciting direction, where unified models will seamlessly work across text, images, audio, and video, enabling more natural and intuitive human-computer interaction.

The incorporation of logical reasoning capabilities will enable generative models to produce more coherent and contextually appropriate outputs, while improved safety mechanisms will prevent the generation of harmful or inappropriate content. Better evaluation metrics will provide more reliable ways to assess generation quality, enabling more systematic progress in the field and better comparison between different approaches.

### 20.5.3 Societal Impact

The widespread adoption of generative models brings both tremendous opportunities and significant challenges that society must carefully navigate. Copyright and intellectual property issues have become increasingly complex as models are trained on vast datasets containing copyrighted material, raising questions about ownership, attribution, and fair use. The ability to

generate highly realistic content has also created new challenges around misinformation and deepfakes, requiring the development of robust detection methods and media literacy education.

The potential for job displacement in creative fields is a legitimate concern, as generative models become increasingly capable of producing professional-quality content. However, these tools also democratize access to creative capabilities, enabling individuals without traditional artistic training to express their ideas visually and musically. The environmental cost of training large generative models is substantial, with some models requiring enormous computational resources that contribute to carbon emissions, highlighting the need for more efficient training methods and renewable energy sources.

Ensuring equitable access to these powerful technologies is crucial for preventing the exacerbation of existing inequalities. The benefits of generative models should be available to all members of society, not just those with access to expensive hardware or specialized knowledge. Responsible development of these technologies requires careful consideration of these societal implications, balancing the tremendous potential for positive impact with the need to address legitimate concerns about misuse and unintended consequences.

## 20.6 Real World Applications

Deep generative models create new data samples, enabling applications from content creation to scientific discovery. Recent advances in GANs, VAEs, and diffusion models have made generation remarkably realistic and controllable.

### 20.6.1 Creative Content Generation

The realm of creative content generation has been fundamentally transformed by generative models, ushering in a new era of AI-powered creativity and design. AI art and design tools like Midjourney, DALL-E, and Stable Diffusion have democratized visual content creation, enabling anyone to produce professional-quality images from simple text descriptions. Designers leverage these tools for rapid prototyping, generating dozens of concept variations in minutes rather than hours of manual work, while artists use them as creative partners, combining AI-generated elements with traditional techniques to explore new artistic possibilities.

Music composition has been revolutionized by generative models that create original compositions in various styles, from background scores for videos to experimental pieces. These services generate royalty-free music customized to specific moods, tempos, and instrumentation, providing musicians with inspiration and content creators with affordable custom soundtracks. The ability to generate music on demand has opened new possibilities for personalized audio experiences and creative exploration. Architectural and product design have also benefited from generative models that can explore vast design spaces, proposing innovative variations on building layouts and product designs. Architects use these tools to generate floor plan alternatives that consider multiple constraints like lighting and space efficiency, while product designers can rapidly iterate through form variations, dramatically accelerating the creative process from initial concept to final prototype.

### 20.6.2 Scientific Discovery

The application of generative models to scientific discovery represents one of the most promising frontiers in AI research, with the potential to accelerate breakthroughs across multiple disciplines. Drug molecule design has been revolutionized by generative models that can propose novel drug candidates with specific desired properties, including optimal binding to target proteins, favorable safety profiles, and ease of synthesis. This approach explores chemical space far more efficiently than traditional trial-and-error synthesis methods, potentially accelerating drug discovery timelines and reducing costs for pharmaceutical companies developing treatments for cancer, infectious diseases, and other conditions.

Materials science has benefited tremendously from generative models that can design new materials with precisely specified properties, from stronger alloys for aerospace applications to more efficient batteries and solar cells for clean energy technologies. These models learn complex relationships between molecular structure and material properties, enabling researchers to propose novel materials for experimental validation that would be impossible to discover through traditional methods. This capability could dramatically accelerate the development of technologies needed for clean energy and sustainability.

Protein structure prediction and design represent another area where generative models are making transformative contributions. These models help predict how proteins fold into their functional three-dimensional structures and design proteins with novel functions for industrial processes, vaccine development, and therapeutic applications. The success of AlphaFold in protein structure prediction demonstrates how generative models can advance our understanding of biological systems and accelerate the development of new treatments and technologies.

### 20.6.3 Data Augmentation and Synthesis

The ability of generative models to create synthetic training data has become a crucial tool for addressing data scarcity and privacy concerns across multiple domains. Synthetic medical images represent a particularly important application, where generative models create realistic training data that doesn't correspond to real patients, enabling the development of better diagnostic models while protecting patient privacy. This approach is especially valuable for rare diseases where real data is limited, helping to address data imbalances that can bias machine learning models and improving their performance on underrepresented conditions.

Simulation for autonomous vehicles has emerged as another critical application, where generative models create realistic synthetic driving scenarios including rare but dangerous events like pedestrians jaywalking or vehicles running red lights. Self-driving cars can train on these synthetic scenarios to prepare for dangerous situations without risking real-world testing, addressing the "long tail" of rare but critical edge cases that are essential for safe autonomous operation. This approach enables more comprehensive training than would be possible with real-world data alone.

Video game content generation has been transformed by generative models that can create textures, terrain, character models, and even entire game levels. This capability reduces development costs and time while dramatically increasing content variety, enabling procedural generation that creates unique experiences for each player rather than requiring manual crafting of every asset. The result is more dynamic and engaging gaming experiences that can adapt to player preferences and behaviors.

### 20.6.4 Personalization and Adaptation

The personalization capabilities of generative models are creating new possibilities for customized content that adapts to individual preferences and needs. Avatar creation has become increasingly sophisticated, with apps generating personalized avatars from photos that maintain recognizable features while providing stylized or cartoon representations. These avatars appear in messaging apps, games, and virtual meetings, offering fun and privacy-conscious ways for users to represent themselves in digital spaces.

Text-to-speech personalization represents a particularly meaningful application, where generative models can create natural-sounding speech in a user's own voice from text input. This technology is invaluable for people who have lost their voice due to illness, allowing them to preserve their vocal identity and communicate naturally. It also enables personalized audiobook narration and accessible content delivery in preferred voices, making information more accessible to diverse audiences.

Style transfer and image editing applications have made sophisticated image manipulation accessible to everyone, from professional photographers to casual social media users. These tools can apply artistic styles to photos, change seasons in landscape photography, or realistically age or de-age faces, enabling creative expression and visual storytelling that was previously limited to those with advanced technical skills.

### 20.6.5 Transformative Impact

The transformative impact of generative models extends far beyond impressive technical demonstrations, fundamentally changing how we approach creative work, scientific discovery, and everyday applications. The democratization of creative tools represents perhaps the most significant shift, making sophisticated artistic and design capabilities accessible to everyone, not just those with years of training and expertise. This accessibility is breaking down barriers to creative expression and enabling new forms of artistic collaboration between humans and AI.

The acceleration of creative and scientific processes through rapid iteration and exploration of vast design spaces is another key transformation. Researchers and creators can now explore possibilities that would be impossible to investigate through traditional methods, dramatically reducing the time from concept to realization. This acceleration is particularly valuable in

scientific discovery, where generative models can explore complex domains like chemistry and biology to find solutions that might take years to discover through conventional approaches.

The synthesis capabilities of generative models, particularly their ability to create training data and simulations that would otherwise be unavailable, are opening new possibilities for machine learning and AI development. These capabilities are essential for addressing data scarcity, privacy concerns, and the need for diverse training examples across multiple domains. The practical impact of these applications demonstrates that generative models are not just impressive demonstrations but essential tools that are already transforming how we work, create, and discover.

## Key Takeaways

### Key Takeaways 20

- **Model families** differ in training signal and guarantees: VAEs use likelihood-based training with explicit density modeling, GANs employ adversarial training for implicit modeling, flows provide exact likelihoods through invertible transformations, and diffusion models learn to reverse noise corruption processes.
- **Evaluation** must consider likelihood, fidelity, diversity, and downstream utility: Different metrics capture different aspects of generation quality, and the choice of evaluation method should align with the intended application and user requirements.
- **Trade-offs** are inevitable: choose for the target application: Each approach offers distinct advantages—VAEs provide stable training and exact likelihoods, GANs excel at sample quality, flows enable exact sampling, and diffusion models achieve state-of-the-art results with strong theoretical foundations.
- **Training dynamics** vary significantly across approaches: VAEs and flows use standard maximum likelihood optimization, GANs require careful balance between generator and discriminator, while diffusion models benefit from stable training with principled noise schedules.
- **Applications** drive architectural choices: The specific requirements of creative content generation, scientific discovery, data augmentation, and personalization should guide the selection of appropriate generative modeling approaches and evaluation metrics.

## Exercises

### Easy

*Exercise 20.1 (VAE vs. GAN).* Compare the training objectives of VAEs and GANs.

**Hint:**

*Likelihood-based vs. adversarial.*

*Exercise 20.2 (Normalising Flow Invertibility).* Why must normalising flows be invertible?

**Hint:**

*Exact likelihood computation via change of variables.*

*Exercise 20.3 (Diffusion Process).* Describe the forward diffusion process in diffusion models.

**Hint:**

*Gradual addition of Gaussian noise.*

*Exercise 20.4 (Sampling Speed).* Compare sampling speed across VAEs, GANs, and diffusion models.

**Hint:**

*Single pass vs. iterative refinement.*

## Medium

*Exercise 20.5 (Flow Jacobian).* Derive the change-of-variables formula for a normalising flow.

**Hint:**

$$\log p(x) = \log p(z) + \log |\det \frac{\partial f}{\partial z}|.$$

*Exercise 20.6 (Denoising Score Matching).* Explain how diffusion models learn the score function.

**Hint:**

*Predict noise; connection to  $\nabla_x \log p(x)$ .*

## Hard

*Exercise 20.7 (Coupling Layer Design).* Design an invertible coupling layer and prove its properties.

**Hint:**

*Affine transformations; partition dimensions.*

*Exercise 20.8 (Guidance Trade-off).* Analyse the trade-off between sample quality and diversity in classifier-free guidance.

**Hint:**

*Guidance scale; conditional vs. unconditional scores.*

*Exercise 20.9 (Advanced Topic 1).* Explain a key concept from this chapter and its practical applications.

**Hint:**

*Consider the theoretical foundations and real-world implications.*

*Exercise 20.10 (Advanced Topic 2).* Analyse the relationship between different techniques covered in this chapter.

**Hint:**

*Look for connections and trade-offs between methods.*

*Exercise 20.11 (Advanced Topic 3).* Design an experiment to test a hypothesis related to this chapter's content.

**Hint:**

*Consider experimental design, metrics, and potential confounding factors.*

*Exercise 20.12 (Advanced Topic 4).* Compare different approaches to solving a problem from this chapter.

**Hint:**

*Consider computational complexity, accuracy, and practical considerations.*

*Exercise 20.13 (Advanced Topic 5).* Derive a mathematical relationship or prove a theorem from this chapter.

**Hint:**

*Start with the definitions and work through the logical steps.*

*Exercise 20.14 (Advanced Topic 6).* Implement a practical solution to a problem discussed in this chapter.

**Hint:**

*Consider the implementation details and potential challenges.*

*Exercise 20.15 (Advanced Topic 7).* Evaluate the limitations and potential improvements of techniques from this chapter.

**Hint:**

*Consider both theoretical limitations and practical constraints.*

## List of Abbreviations

**AdaGrad** Adaptive Gradient Algorithm

**Adam** Adaptive Moment Estimation

**AdamW** Adam with Decoupled Weight Decay

**BPTT** Backpropagation Through Time

**BFGS** Broyden-Fletcher-Goldfarb-Shanno

**CNN** Convolutional Neural Network

**CRF** Conditional Random Field

**DCGAN** Deep Convolutional Generative Adversarial Network

**ELU** Exponential Linear Unit

**GAN** Generative Adversarial Network

**GELU** Gaussian Error Linear Unit

**GRU** Gated Recurrent Unit

**KL** Kullback-Leibler (Divergence)

**L-BFGS** Limited-memory Broyden-Fletcher-Goldfarb-Shanno

**LSTM** Long Short-Term Memory

**MCMC** Markov Chain Monte Carlo

**MLP** Multilayer Perceptron

**MSE** Mean Squared Error

**NAG** Nesterov Accelerated Gradient

**NLP** Natural Language Processing

**PCA** Principal Component Analysis

**PReLU** Parametric Rectified Linear Unit

**ReLU** Rectified Linear Unit

**RNN** Recurrent Neural Network

**RMSProp** Root Mean Square Propagation

**SGD** Stochastic Gradient Descent

**SVM** Support Vector Machine

**VAE** Variational Autoencoder

**WGAN** Wasserstein Generative Adversarial Network

**XGBoost** eXtreme Gradient Boosting



---

# Glossary

**accuracy** Proportion of correct predictions over all predictions in a classification task.. 170

**attention mechanism** A technique that allows models to focus on relevant parts of the input when making predictions.. 159, 162

**backpropagation** An algorithm for training neural networks that computes gradients by propagating errors backward through the network.. 163

**BLEU** Bilingual Evaluation Understudy; an n-gram overlap metric for machine translation quality.. 173

**computer vision** A field of artificial intelligence that enables computers to interpret and understand visual information.. 1

**exploding gradient problem** A problem in deep networks where gradients become exponentially large as they propagate backward, causing unstable training.. 154

**gradient descent** An optimization algorithm that iteratively adjusts parameters in the direction of steepest descent of the loss function.. 126

**long short-term memory** A type of recurrent neural network architecture designed to overcome the vanishing gradient problem.. 155

**mini-batch** A small subset of the training data used in each iteration of gradient descent, typically containing 32-256 examples.. 125, 126

**natural language processing** A field of artificial intelligence that focuses on the interaction between computers and human language.. 1

**neural network** A computing system inspired by biological neural networks, consisting of interconnected nodes (neurons).. 1

**precision** Fraction of predicted positives that are true positives.. 170

**recall** Fraction of actual positives that are correctly identified.. 170

**recurrent neural network** A type of neural network designed for sequential data, where connections form directed cycles.. 151, 152, 158

**ROUGE** Recall-Oriented Understudy for Gisting Evaluation; a set of metrics for automatic summarization evaluation.. 173

**stochastic gradient descent** A variant of gradient descent that updates parameters using the gradient from a single example at a time.. 126

**vanishing gradient problem** A problem in deep networks where gradients become exponentially small as they propagate backward, making training difficult.. 154

---

# Index

- A/B test, 185
- ablation study, 176
- accuracy, 170
- activation function
  - GELU, 100
  - Leaky ReLU, 100
  - ReLU, 100
  - sigmoid, 100
  - Swish, 100
  - tanh, 100
- activation functions, 97
- AdaGrad, 128
- Adam, 128
- adaptive optimization, 128
- adversarial training, 120
- AlexNet, 142
- applications
  - anomaly detection, 227, 236
  - audio processing, 227
  - autonomous systems, 277
  - autonomous vehicles, 121, 147, 188, 251
  - chatbots, 121
  - compression, 236
  - computer vision, 268
  - content generation, 290
  - content moderation, 147
  - cross-modal learning, 244
  - data augmentation, 290
  - denoising, 236
  - drug discovery, 260
  - facial recognition, 227
  - financial forecasting, 165
  - financial risk, 260
  - fraud detection, 107
  - healthcare systems, 188
  - language models, 268
  - machine translation, 165
  - medical diagnosis, 107, 251
  - medical imaging, 121, 147
  - natural language processing, 244
  - natural language understanding, 251
- personalized medicine, 277
- recommendation systems, 107, 188, 277
- recommender systems, 268
- scientific discovery, 290
- text generation, 165
- transfer learning, 244
- voice assistants, 165
- weather forecasting, 260
- approximate inference
  - applications, 277
- artificial intelligence
  - deep learning, 2
- augmentation
  - audio, 113
  - text, 113
  - vision, 113
- autoencoders
  - applications, 236
- automatic differentiation, 104
- average pooling, 140
- backpropagation
  - algorithm, 104
- backpropagation through time, 154
- baseline, 174
- batch, 185
- batch normalization, 118, 176
- batch size, 179
- Bayesian optimization, 179
- beam search, 162
- bidirectional RNN, 162
- BLEU, 173
- calibration, 171
- chain rule
  - backpropagation, 104
- class imbalance, 169
- class weights, 181
- classical ML, 177
- computational graph, 104
- confusion matrix, 170
- convolution, 137

convolutional networks  
     applications, 147  
 cross-correlation, 137  
 cross-validation, 180  
 data augmentation, 113  
 data leakage, 181  
 deep learning, 177  
     introduction, 2  
 dilation, 139  
 drift, 185  
 dropout, 116  
 early stopping, 115, 133, 179  
 EfficientNet, 143  
 elastic net, 112  
 equivariance, 137, 140  
 feedforward network  
     introduction, 97  
 feedforward networks  
     applications, 107  
 focal loss, 181  
 forward propagation, 97  
 gated recurrent unit, 158  
 generative models  
     applications, 290  
 GoogLeNet, 143  
 gradient  
     computation, 104  
 gradient clipping, 120, 132, 133, 176  
 gradient descent  
     batch, 126  
     mini-batch, 106, 126  
 grid search, 179  
 hyperparameter tuning, 177  
 ImageNet, 194  
 Inception, 143  
 internal covariate shift, 118  
 invariance, 140  
 L-BFGS, 131  
 label smoothing, 120  
 learning rate, 126, 179  
 learning rate schedule, 129  
 learning-rate-schedule, 130  
 LeNet-5, 141  
 linear factor models  
     applications, 227  
 long short-term memory, 155  
 loss function  
     cross-entropy, 101  
     MSE, 101  
 loss functions, 97  
 machine learning  
     deep learning, 2  
 manual search, 179  
 mAP, 173  
 max pooling, 140  
 memory  
     neural, 152  
 metrics, 169  
 min-max scaling, 181  
 mini-batch, 125, 127  
     training, 106  
 mixup, 120  
 MobileNet, 143  
 model compression, 185  
 momentum, 127  
 Monte Carlo methods  
     applications, 260  
 multilayer perceptron, 97  
 multilayer perceptrons (MLPs), 97  
 natural gradient, 131  
 nDCG, 173  
 Nesterov, 127  
 network architecture, 97  
 network design  
     depth, 106  
     width, 106  
 neural network  
     feedforward, 97  
 neural networks  
     deep learning, 2  
 Newton's method, 131  
 non-linearity  
     activation functions, 100  
 normalization, 181  
 normalization layers, 118  
 online, 185  
 optimisation  
     key takeaways, 134  
 optimization  
     challenges, 131  
     gradient descent, 125

momentum, 127  
Nesterov, 128  
output layer  
classification, 101  
regression, 101  
oversampling, 181  
padding, 139  
parameter sharing, 139  
partition function  
applications, 268  
performance metrics, 185  
plateau, 126, 133  
pooling, 140  
practical methodology  
applications, 188  
precision–recall curve, 171  
random search, 179  
k, 173  
receptive field, 139  
recurrent networks  
applications, 165  
recurrent neural network, 151  
regression metrics, 172  
regularization  
applications, 121  
dropout, 116  
L1, 112  
L2, 112  
representation learning  
applications, 244  
ResNet, 141, 143  
RMSProp, 128  
ROC curve, 171  
ROUGE, 173  
saddle point, 126, 132  
second-order methods, 130  
sequence labeling, 153  
sequence modeling, 152  
sequence transduction, 153  
sequence types, 153  
sequence-to-sequence, 159  
stochastic gradient descent, 126  
shared parameters, 153  
sigmoid  
binary classification, 101  
silent failures, 174  
SMOTE, 181  
softmax  
multiclass classification, 101  
standardization, 181  
stochastic depth, 120  
stride, 139  
strided convolution, 141  
structured probabilistic models  
applications, 251  
teacher forcing, 162  
temporal dependency, 152  
two-stage training, 179  
undersampling, 181  
universal approximation theorem, 97  
unrolling in time, 153  
validation set, 115  
VGG, 142  
weight initialization  
He, 106  
Xavier, 106

---

# Bibliography

- [Ama98] Shun-ichi Amari. “Natural gradient works efficiently in learning”. In: *Neural Computation* 10.2 (1998), pp. 251–276.
- [BCB14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate”. In: *arXiv preprint arXiv:1409.0473* (2014).
- [Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [Cho+14] Kyunghyun Cho et al. “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. In: *arXiv preprint arXiv:1406.1078* (2014).
- [Dev+18] Jacob Devlin et al. “BERT: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).
- [DHS11] John Duchi, Elad Hazan, and Yoram Singer. “Adaptive subgradient methods for online learning and stochastic optimization”. In: *Journal of Machine Learning Research*. Vol. 12. 2011, pp. 2121–2159.
- [GBC16a] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL: <http://www.deeplearningbook.org>.
- [GBC16b] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning Book: Optimization for Training Deep Models*. <https://www.deeplearningbook.org/contents/optimization.html>. Accessed 2025-10-14. 2016.
- [GBC16c] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning Book: Sequence Modeling – Recurrent and Recursive Nets*. <https://www.deeplearningbook.org/contents/rnn.html>. Accessed 2025-10-14. 2016.
- [Goo+14] Ian Goodfellow et al. “Generative adversarial nets”. In: *Advances in neural information processing systems*. 2014, pp. 2672–2680.
- [He+16] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [HS97] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.
- [Hua+16] Gao Huang et al. “Deep Networks with Stochastic Depth”. In: *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 646–661.
- [IS15] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *arXiv preprint arXiv:1502.03167* (2015).
- [KB14] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).
- [KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.
- [LeC+89] Yann LeCun et al. “Backpropagation applied to handwritten zip code recognition”. In: *Neural computation* 1.4 (1989), pp. 541–551.

- [LN89] Dong C Liu and Jorge Nocedal. “On the limited memory BFGS method for large scale optimization”. In: *Mathematical Programming* 45.1-3 (1989), pp. 503–528.
- [Nes83] Yurii Nesterov. “A method for solving the convex programming problem with convergence rate  $O(1/k^2)$ ”. In: *Soviet Mathematics Doklady* 27 (1983), pp. 372–376.
- [Pol64] Boris T. Polyak. “Some methods of speeding up the convergence of iteration methods”. In: *USSR Computational Mathematics and Mathematical Physics* 4.5 (1964), pp. 1–17.
- [Pri23] Simon J. D. Prince. *Understanding Deep Learning*. MIT Press, 2023. URL: <https://udlbook.github.io/udlbook/>.
- [Rad+19] Alec Radford et al. “Language models are unsupervised multitask learners”. In: *OpenAI blog* 1.8 (2019), p. 9.
- [RM51] Herbert Robbins and Sutton Monro. “A stochastic approximation method”. In: *The Annals of Mathematical Statistics* 22.3 (1951), pp. 400–407.
- [RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [RHW86] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. “Learning representations by back-propagating errors”. In: *Nature* 323.6088 (1986), pp. 533–536.
- [Sil+16] David Silver et al. “Mastering the game of Go with deep neural networks and tree search”. In: *Nature* 529.7587 (2016), pp. 484–489.
- [Sri+14] Nitish Srivastava et al. “Dropout: a simple way to prevent neural networks from overfitting”. In: vol. 15. 1. 2014, pp. 1929–1958.
- [TH12] Tijmen Tieleman and Geoffrey Hinton. *Lecture 6.5—RMSProp: Divide the gradient by a running average of its recent magnitude*. COURSERA: Neural Networks for Machine Learning, 2012.
- [Vas+17] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.
- [Wik25a] Wikipedia contributors. *Attention (machine learning)*. [https://en.wikipedia.org/wiki/Attention\\_\(machine\\_learning\)](https://en.wikipedia.org/wiki/Attention_(machine_learning)). Accessed 2025-10-14. 2025.
- [Wik25b] Wikipedia contributors. *Recurrent neural network*. [https://en.wikipedia.org/wiki/Recurrent\\_neural\\_network](https://en.wikipedia.org/wiki/Recurrent_neural_network). Accessed 2025-10-14. 2025.
- [Yun+19] Sangdoo Yun et al. “CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 6023–6032.
- [Zha+24a] Aston Zhang et al. *Dive into Deep Learning: Attention Mechanisms and Transformers*. [https://d2l.ai/chapter\\_attention-mechanisms-and-transformers/index.html](https://d2l.ai/chapter_attention-mechanisms-and-transformers/index.html). Accessed 2025-10-14. 2024.
- [Zha+24b] Aston Zhang et al. *Dive into Deep Learning: Optimization*. [https://d2l.ai/chapter\\_optimization/index.html](https://d2l.ai/chapter_optimization/index.html). Accessed 2025-10-14. 2024.
- [Zha+24c] Aston Zhang et al. *Dive into Deep Learning: Recurrent Neural Networks*. [https://d2l.ai/chapter\\_recurrent-neural-networks/index.html](https://d2l.ai/chapter_recurrent-neural-networks/index.html). Accessed 2025-10-14. 2024.
- [Zha+18] Hongyi Zhang et al. “mixup: Beyond Empirical Risk Minimization”. In: *International Conference on Learning Representations (ICLR)*. 2018.