

Explanatory Notes for 6.390

Shaanticlair Ruiz (Current TA)

Fall 2024

Contents

0 Prerequisites - Explanatory Notes	17
0.1 Multi-variable Calculus (MIT's 18.02)	17
0.2 Vectors and Matrices (MIT's 18.02)	18
0.3 Linear Algebra (18.06)	19
0.4 Programming (6.100A, 6.1010)	20
0.5 Algorithms (6.1210)	20
0.6 Probability	21
0.7 Notation: Sets	22
0.8 Notation: Numbers and functions	24
0.9 Notation: Vectors Spaces	25
0.10 Optional	26
1 Introduction - Explanatory Notes	27
1.0.1 What is machine learning?	28
1.0.2 Why do Machine Learning: The Benefits	28
1.0.3 The role of humans	28
1.0.4 What's the plan?	30
1.0.5 The Problem	30
1.0.6 Solution Setup: What is a model?	31
1.0.7 The Solution	32
1.1 Problem Class	33
1.1.1 Supervised vs. Unsupervised	33
1.1.2 How do we store our data?	33
1.1.3 Supervised Learning	34
1.1.4 Unsupervised Learning	35
1.1.5 Other Types of Learning	35
1.1.6 Types of Learning not covered in this class (Optional)	36

1.2	Assumptions	37
1.2.1	An assumption about data	37
1.2.2	Is our data representative?	37
1.2.3	How do we compare data?	37
1.2.4	Identically Distributed Data	38
1.2.5	Independence (Review)	38
1.2.6	Independent and Identically Distributed	39
1.2.7	Estimation and Generalization	39
1.2.8	Other Assumptions	40
1.3	Evaluation Criteria	41
1.3.1	What is a loss function?	41
1.3.2	Examples of Loss Functions	41
1.3.3	How to use loss	42
1.4	Model Type	43
1.4.1	No Model	43
1.4.2	Models using Parameters	43
1.4.3	Prediction Rule	44
1.4.4	Hypothesis Notation	44
1.4.5	Fitting	45
1.4.6	Overfitting	46
1.5	Model Class	47
1.5.1	Hypothesis Class	47
1.5.2	Expressiveness	47
1.5.3	Choosing Model Classes	47
1.5.4	Our Linear Model	48
1.5.5	Linear Model: Vector Form	48
1.5.6	Linear Model: Cleaning Up	49
1.5.7	Other Models	49
1.6	Algorithm	51
1.7	Overview of the Course	52
1.8	Terms	53
2	Regression	54
2.1	Problem Formulation	54
2.1.1	Hypothesis (Review)	54
2.1.2	The Problem of Regression	58
2.1.3	Converting our data	59
2.1.4	Our dataset	61
2.1.5	Training our model	62
2.1.6	Learning to Generalize	63
2.2	Regression as an optimization problem	65
2.2.1	Objective Function	65
2.2.2	The Regularizer	66
2.2.3	More on the Objective Function	67

2.2.4	Minimization Notation	69
2.2.5	Optimal Value Notation	70
2.3	Linear Regression	72
2.3.1	The Linear Model, 1-D	72
2.3.2	The Linear Model, 2-D	72
2.3.3	The Linear Model, d-D	73
2.3.4	The Linear Model using Vectors	73
2.3.5	Regression Loss	75
2.3.6	Our Goal: Ordinary Least Squares	76
2.3.7	Visualizing our Model	77
2.3.8	Another Interpretation	79
2.4	The stupidest possible linear regression algorithm	80
2.5	Analytical solution: ordinary least squares	81
2.5.1	Trying to Simplify	81
2.5.2	Combining θ and θ_0	81
2.5.3	Combining data points	83
2.5.4	Many data points in a matrix	83
2.5.5	Objective Function in matrix form	86
2.5.6	Alternate Notation	87
2.5.7	Optimization in 1-D - Using Calculus	89
2.5.8	Optimizing for multiple variables	90
2.5.9	Gradient Notation	91
2.5.10	Matrix Calculus	92
2.6	Regularization	94
2.6.1	Coincidences, and fake patterns	94
2.6.2	Ridge Regression	96
2.6.3	λ , our regularization constant	97
2.6.4	Why not regularize θ_0 ?	100
2.6.5	Ridge Regression Solution	102
2.6.6	Invertibility	103
2.6.7	Uniqueness of θ^*	104
2.6.8	Error Amplification	107
2.6.9	Error Amplification Example (Optional)	108
2.6.10	Regularizer justification: Prior Knowledge (Optional)	109
2.7	Evaluating Learning Algorithms	111
2.7.1	What λ should we choose?	111
2.7.2	Tradeoffs: Estimation Error	111
2.7.3	Tradeoffs: Structural Error	113
2.7.4	Tradeoffs of λ	114
2.7.5	Evaluating Hypotheses	115
2.7.6	λ 's purpose: learning algorithms	116
2.7.7	Comparing Hypotheses and Learning Algorithms	116
2.7.8	Evaluating our Learning Algorithm	117

2.7.9	Validation: Evaluating with lots of data	117
2.7.10	Our Problem: When data is less available	118
2.7.11	Cross-Validation	118
2.7.12	Hyperparameter Tuning	119
2.7.13	How to tune our algorithm	120
2.7.14	Hyperparameter Tuning: Two kinds of optimization	120
2.7.15	Pseudocode Example	121
2.8	Terms	122
3	Gradient Descent	124
3.0.1	Why do we need gradient descent?	124
3.0.2	How do we improve?	125
3.0.3	The name: "gradient descent"	126
3.0.4	Input Space vs. Parameter Space	127
3.1	Gradient Descent in One Dimension	129
3.1.1	Derivatives (Review)	129
3.1.2	Optimize with Derivatives: 1-D	129
3.1.3	Convergence	130
3.1.4	Convergence: A little more formally (Optional)	131
3.1.5	Step size	131
3.1.6	Step size η	133
3.1.7	Our procedure	134
3.1.8	Termination Conditions	135
3.1.9	Convergence Theorem	136
3.1.10	Concavity	136
3.1.11	Local minima	137
3.2	Multiple Dimensions	139
3.2.1	Multivariable Local Approximation (Review)	139
3.2.2	2-D: One dimension at a time	141
3.2.3	Gradient Descent in n-D	144
3.2.4	The Gradient	144
3.2.5	The Plane Approximation	145
3.2.6	The Optimal Direction: The Gradient	146
3.2.7	Termination Condition	147
3.2.8	Another explanation of gradient (OPTIONAL)	148
3.3	Application to Regression	150
3.3.1	Ordinary Least Squares	150
3.3.2	Ridge Regression	150
3.3.3	Computational Gradient	151
3.3.4	Problems with Gradient Descent	153
3.4	Stochastic Gradient Descent	155
3.4.1	Another problem with gradient descent	155
3.4.2	A better way: stochastic GD	155
3.4.3	Ensuring Convergence	156

3.5 Terms	157
4 Classification	158
4.0.1 Regression (Review)	158
4.1 Classification	158
4.1.1 Motivation: Putting things into classes	158
4.1.2 What is classification?	159
4.1.3 Important Facts about Classes	160
4.1.4 Binary Classification	161
4.1.5 Classification Performance	161
4.2 Linear Classifiers	163
4.2.1 1-D Linear Classifiers	163
4.2.2 1-D classifiers in 2-D	163
4.2.3 A second 1-D separator, and our problem	165
4.2.4 The 2-D Separator: What vector do we use?	166
4.2.5 2D Separator - Matching components	168
4.2.6 The Dot Product (Review)	169
4.2.7 Using the dot product	170
4.2.8 Introducing our offset	171
4.2.9 How does the offset affect our classifier?	172
4.2.10 Distance from the Origin to the Plane	175
4.2.11 Extending to higher dimensions	176
4.2.12 IMPORTANT: A difference between regression and classification . .	178
4.2.13 3d plot of 2d separator	182
4.2.14 Separable vs Non-separable data	182
4.3 Linear Logistic Classifiers	184
4.3.1 The problem	184
4.3.2 The real problem: $\text{sign}(u)$ is flat	184
4.3.3 The sigmoid function	185
4.3.4 Sigmoid as a probability	187
4.3.5 Logistic Regression	187
4.3.6 Prediction Threshold	188
4.3.7 Linear Logistic Classifier	189
4.3.8 Modifying our sigmoid	190
4.3.9 Viewing our sigmoid in 3D	192
4.3.10 LLCs and LCs have the same boundary	192
4.3.11 Learning LLCs: Loss Functions	193
4.3.12 Building our new loss function	193
4.3.13 Loss Function for Multiple Data Points	195
4.3.14 Simplifying our expression - Piecewise	196
4.3.15 Getting rid of the product	196
4.3.16 Negative Log Likelihood	197
4.3.17 LLCs and overfitting	198
4.4 Gradient Descent for Logistic Regression	201

4.4.1	Summary	201
4.4.2	The problem: Gradient Descent	201
4.4.3	Getting the gradient: Chain Rule	202
4.4.4	Getting our individual derivatives	202
4.4.5	Simplifying our chain rule	203
4.5	Handling Multiple Classes	205
4.5.1	Approaches to multi-class classification	205
4.5.2	Extending our Approach: One-Hot Encoding	205
4.5.3	Probabilities in multi-class	207
4.5.4	Turning sigmoid multi-class	208
4.5.5	Our Linear Classifiers	209
4.5.6	Softmax	211
4.5.7	NLLM	212
4.5.8	A side comment: Sigmoid vs. Softmax	213
4.6	Prediction Accuracy and Validation	216
4.7	Terms	217
5	Feature Representation	218
5.1	Gaining intuition about feature transformations	224
5.1.1	Transforming our separator	225
5.1.2	Transforming our data	226
5.1.3	Positive vs. Negative	227
5.2	Systematic feature construction	229
5.2.1	Polynomial Basis	229
5.2.2	Radial Basis	236
5.3	Hand-constructing features for real domains	239
5.3.1	Discrete Features	239
5.3.2	Text	248
5.3.3	Numeric values	249
5.4	Terms	253
6	Neural Networks 1 - Neurons, Layers, and Networks	254
6.0.1	Machine Learning Applications	254
6.0.2	Neural Network Perspectives: The brain	255
6.0.3	Neural Network Perspectives: Classification and Regression	255
6.0.4	Building up a basic neural network	256
6.0.5	Neural Network Perspectives: Predictions with Big Data	258
6.1	Basic Element	258
6.1.1	What's in a neuron: The Linear Component	259
6.1.2	Weights and Biases	259
6.1.3	Linear Diagram	261
6.1.4	Adding nonlinearity	263
6.1.5	Nonlinear Diagram	263
6.1.6	Putting it together	264

6.1.7	Neuron Diagram	265
6.1.8	Our Loss Function	266
6.1.9	Example: Linear Regression	267
6.1.10	Example: Linear Logistic Classifiers	268
6.2	Networks	270
6.2.1	Abstraction	270
6.2.2	Some limitations: acyclic networks	270
6.2.3	How to build networks	271
6.2.4	Layers	272
6.2.5	The Basic Structure of a Neural Network	274
6.2.6	Single Layer: Visualizing our Components	275
6.2.7	Single Layer: Visualizing our Inputs	276
6.2.8	Dimensions of a layer	278
6.2.9	The known objects of our layer	279
6.2.10	The other variables of our layer: weights and offsets	280
6.2.11	Pre-activation	282
6.2.12	Summary of a layer	283
6.2.13	The weakness of a single layer	285
6.2.14	Adding a second layer	287
6.2.15	Many Layers	289
6.2.16	Our Complete Neural Network	290
6.3	Choices of activation function	293
6.3.1	Trying out linear activation	293
6.3.2	Linear Layers: An example	294
6.3.3	The problem with linear networks	295
6.3.4	Example of Activation Functions	296
6.4	Loss functions and activation functions	301
6.4.1	Other Considerations	301
6	Neural Networks 2 - Back-Propagation and Training	304
6.5	Error back-propagation	304
6.5.1	Review: Gradient Descent	304
6.5.2	Review: Gradient Descent with LLCs	305
6.5.3	Review: LLC as Neuron	306
6.5.4	LLC Forward-Pass	307
6.5.5	LLC Back-propagation	308
6.5.6	Summary of neural network gradient descent: a high-level view	310
6.5.7	A two-neuron network: starting backprop	311
6.5.8	Continuing backprop: One more problem	313
6.5.9	Finishing two-neuron backprop	314
6.5.10	Many layers: Doing back-propagation	316
6.5.11	What do these derivatives equal?	318
6.5.12	Activation Derivatives	320
6.5.13	Loss derivatives	322

6.5.14	Many neurons per layer	323
6.5.15	The chain rule: Matrix form	324
6.5.16	How the Chain Rule changes in Matrix form	327
6.5.17	Relevant Derivatives	328
6.6	Training	331
6.6.1	Comments	331
6.6.2	Pseudocode	331
6.7	Optimizing neural network parameters	333
6.7.1	Mini-batch	333
6.7.2	Adaptive Step Size - Challenges	337
6.7.3	Vanishing/Exploding Gradient	338
6.8	Regularization	340
6.8.1	Methods related to ridge regression	340
6.8.2	Dropout	343
6.8.3	Batch Normalization	345
6.9	Terms	351
7	Convolutional Neural Networks	354
7.0.1	Fully Connected Networks	354
7.0.2	The drawbacks of fully-connected networks	355
7.0.3	Intro to Image Processing	357
7.0.4	Spatial Locality	358
7.0.5	Translation Invariance	359
7.1	Filters	361
7.1.1	Motivating the Filter	361
7.1.2	Windowing	362
7.1.3	1-D case	363
7.1.4	1D Example	365
7.1.5	Convolution	367
7.1.6	Convolution Output Size	369
7.1.7	Padding	369
7.1.8	2D Filter	371
7.1.9	2-D convolution	373
7.1.10	Dot Product Generalization	374
7.1.11	Filter Banks	376
7.1.12	Tensor Filters	379
7.1.13	Tensor Filters: All channels	380
7.1.14	Convolution is Linear	381
7.1.15	RGB colors (Optional)	382
7.1.16	Adding Convolution to our Neural Networks	384
7.1.17	Training our Convolutional Layer	386
7.1.18	Benefits of Convolution	386
7.1.19	Our NN dimensions	389
7.1.20	Stride	389

7.1.21	Output shape	391
7.2	Max-pooling	393
7.2.1	Aggregating information	393
7.2.2	Deriving max-pool	394
7.2.3	Max-pool stride	395
7.2.4	Clarifications on max-pooling	397
7.2.5	Max-pool: A functional layer	398
7.2.6	Max-pool: Some problems with "translation invariance".	399
7.3	Typical architecture	400
7.4	Backpropagation in a simple CNN	403
7.4.1	Our Simplest Example	403
7.4.2	Chain rule to get full derivative	405
7.4.3	Easy, Familiar Derivatives	405
7.4.4	ReLU Derivative	405
7.4.5	Filter Derivative	408
7.4.6	Maxpool derivative	410
7.4.7	Maxpool derivative: somewhat similar to sign function (Optional) .	411
7.5	Terms	412
8	Transformers	413
8.0.1	CNNs	413
8.0.2	The problem with locality	414
8.0.3	RNNs	415
8.0.4	Transformers	417
8.1	Vector embeddings and tokens	418
8.1.1	One-hot encoding isn't enough	418
8.1.2	Word Embeddings: Similarity between words	418
8.1.3	Vector Similarity: Dot Products	419
8.1.4	Word2vec	421
8.1.5	Probability	421
8.1.6	"Adding" words together	424
8.1.7	Tokenization	425
8.2	Attention	426
8.2.1	The Attention Mechanism: queries, keys	427
8.2.2	The Attention Mechanism: attention weights	428
8.2.3	Scaling factor for softmax	431
8.2.4	The Attention Mechanism: values, attention	432
8.2.5	Why we need context	437
8.2.6	Why we need <i>attentive</i> context	437
8.2.7	Self-attention	439
8.2.8	Self-attention in matrix form	439
8.2.9	Positional Encoding	442
8.2.10	Masking	442
8.2.11	Attention Heads	444

8.3	Transformers	446
8.3.1	How to create embeddings	446
8.3.2	Attention Heads	447
8.3.3	Residual Connections	450
8.3.4	Layer Normalization	451
8.3.5	Feed Forward	453
8.3.6	Transformer Block	455
8.3.7	Translation Task: training	457
8.3.8	Encoder + Decoder Structure	459
8.3.9	Predicting a token	462
8.3.10	Training Process	464
8.3.11	Variations	464
8.4	Terms	465
9	Non-parametric Methods	467
9.0.1	Parametric Methods	467
9.0.2	Non-parametric methods	468
9.0.3	Why learn about non-parametric methods?	469
9.1	Nearest Neighbor	470
9.1.1	Nearest Neighbors: An example	471
9.1.2	Simplified Voronoi Diagram (Optional)	472
9.1.3	k-Nearest Neighbors	474
9.1.4	Locally weighted regression	476
9.1.5	Tradeoffs	476
9.1.6	Distance metrics (Optional)	477
9.2	Tree Models	480
9.2.1	An example in 2D space	482
9.2.2	Partitioning: Formalizing our Tree	484
9.2.3	Regression	486
9.2.4	Regression Loss	487
9.2.5	Greedy algorithms	490
9.2.6	How to be greedy	490
9.2.7	Tree regression pseudocode	493
9.2.8	Pruning	494
9.2.9	Classification	496
9.2.10	Classification Loss: Misclassification Error	497
9.2.11	"Purity" of child nodes: Empirical Probability	498
9.2.12	Classification Loss 2: The Gini Index	500
9.2.13	Information 1: Uncertainty (Optional)	501
9.2.14	Information 2: Entropy (Optional)	503
9.2.15	Classification Loss 3: Entropy	505
9.2.16	Which Loss function to use?	507
9.2.17	Bagging: General Concept	509
9.2.18	Bagging: Bootstrapping (Optional)	510

9.2.19 Bagging: Completed	513
9.2.20 Random Forests	515
9.2.21 Other types of tree models	516
9.2.22 Benefits of Trees	516
9.3 Terms	517
10 Markov Decision Processes 0 - State Machines	519
10.1 State Machines	519
10.1.1 How to Model Time	519
10.1.2 States	520
10.1.3 How states are stored	520
10.1.4 State examples	521
10.1.5 Input	523
10.1.6 Transition	523
10.1.7 Transition Examples	524
10.1.8 Output	526
10.1.9 Output Function	526
10.1.10 Output Examples	527
10.1.11 A Completed State Machine	529
10.1.12 Using a State Machine	530
10.1.13 Example Run-Through of a State Machine	531
10.1.14 State Machine Diagram	532
10.1.15 Finite State Machines	534
10.1.16 State Transition Diagrams	534
10.1.17 Simplified state transition diagrams: One-input graphs	536
10.1.18 Linear Time-Invariant Systems (LTI)	538
10 Markov Decision Processes 1 - Value Functions, Policies	540
10.0.1 A new perspective: the "outside world"	541
10.0.2 Making "decisions"	541
10.0.3 Transitioning between States	542
10.0.4 Introducing Rewards	545
10.0.5 Markov Decisions Processes	547
10.1 Definition and Value Functions	549
10.1.1 States and Actions in our MDP	549
10.1.2 Transition Model	550
10.1.3 Comments on our Transition Function	551
10.1.4 State-Transition Diagram: Review	552
10.1.5 State-Transition Diagram: Probabilistic	553
10.1.6 Transition Matrix	555
10.1.7 Reward Function	558
10.1.8 MDP Formalized	559
10.1.9 Policies	560
10.1.10 Value Functions	563

10.1.11 Finite Horizon	565
10.1.12 Finite Horizon Value Function	566
10.1.13 Finite Horizon, $H = 1$	567
10.1.14 Finite Horizon, $H = 2$	569
10.1.15 Finite Horizon, $H = 3$ and beyond	572
10.1.16 Finite Horizon MDP Solution	576
10.1.17 Finite-Horizon, using our Blanket Example (Optional)	578
10.1.18 Infinite Horizon	580
10.1.19 Discounting	581
10.1.20 Discount factor: Termination	582
10.1.21 Lifespan of our MDP	583
10.1.22 Infinite Horizon Value Function	586
10.1.23 Solving the Infinite-Horizon Value Function	587
10.1.24 Infinite-Horizon, using our Blanket Example (Optional)	589
10 Markov Decision Processes 2 - Optimal Policies, Q-Values	591
10.2 Finding policies for MDPs	591
10.2.1 Optimal Policies – Finite Horizon, $H = 0, 1$	591
10.2.2 Finite Horizon: $h = 2$	593
10.2.3 Q-Values	595
10.2.4 $H = 2$ completed	596
10.2.5 $H = 2$ Extended Solution (Optional)	597
10.2.6 $H = 3$ and beyond	598
10.2.7 Finite-Horizon Q-Value MDP solution	600
10.2.8 Dynamic Programming	600
10.2.9 Dynamic Programming Performance (Optional)	602
10.2.10 Optimal Policies – Infinite Horizon	604
10.2.11 Finding an Optimal Policy: Value Iteration	605
10.2.12 Convergence of Value Iteration	608
10.2.13 Value Iteration: Termination Condition	609
10.2.14 Convergence Theorems	612
10.3 Terms	614
11 Reinforcement Learning	616
11.0.1 MDP Review	616
11.0.2 What if we don't know as much?	617
11.0.3 Learning about our MDP	618
11.0.4 Reinforcement Learning	619
11.0.5 Supervised vs. Unsupervised vs. RL	621
11.1 Reinforcement Learning Algorithms Overview	622
11.1.1 Evaluating RL algorithms	622
11.1.2 Different types of RL models	623
11.1.3 Types of Reinforcement Learning	624
11.2 Model-free methods	626

11.2.1 Q-learning: Computing Q from new data	627
11.2.2 Q-learning: Making an update rule	628
11.2.3 Selecting our action: ϵ -greedy	630
11.2.4 Q-learning	632
11.2.5 Initialization	634
11.2.6 Action and state space	634
11.2.7 An alternate view of Q-learning (Optional)	635
11.2.8 Problems with Q-learning: Slow Convergence	637
11.2.9 Deep Q-learning	641
11.2.10 Catastrophic Forgetting	643
11.2.11 Experience Replay	644
11.2.12 Fitted Q-learning	646
11.2.13 Policy Search	649
11.3 Model-based RL	651
11.3.1 Computing \hat{T}	651
11.3.2 The Laplace Correction	652
11.3.3 Computing \hat{R}	655
11.3.4 Solving our MDP	655
11.4 Bandit Problems	657
11.4.1 Slot machines	657
11.4.2 Formalizing the Bandit Problem	657
11.4.3 k-armed bandit problem	658
11.4.4 Exploration vs. Exploitation	659
11.4.5 Contextual Bandit Problems	660
11.5 Terms	661
12 Clustering	663
12.0.1 Why do clustering?	663
12.1 Clustering Formalisms	665
12.1.1 Unsupervised Learning	665
12.1.2 What is clustering?	665
12.2 The k-means formulations	667
12.2.1 Defining a cluster: The mean	667
12.2.2 k-means	667
12.2.3 k-means loss	668
12.2.4 One-cluster loss	669
12.2.5 Building up to k clusters	669
12.2.6 k-mean loss: final form	670
12.2.7 Making further use of the indicator function (Optional)	671
12.2.8 Initializing the k-means algorithm	672
12.2.9 First step: moving our cluster means	673
12.2.10 Second step: Reassign data points	674
12.2.11 The cycle continues	674
12.2.12 The k-means algorithm	675

12.2.13 Pseudocode	676
12.2.14 Using gradient descent: minimizing distance to μ	676
12.2.15 Getting labels	677
12.3 How to evaluate clustering algorithms	678
12.3.1 Initialization	678
12.3.2 Choice of k	679
12.3.3 Subjectivity of k	680
12.3.4 Hierarchical Clustering	681
12.3.5 k-means in feature space	682
12.3.6 Solutions: Validation	682
12.3.7 Solutions: Consistency	682
12.3.8 Solutions: Ground Truth	683
12.3.9 Applications: Visualization and Interpretability	683
12.3.10 Applications: Downstream Tasks	684
12.3.11 A benefit of clustering	685
12.3.12 Weaknesses of k-means	686
12.4 Terms	687
13 Autoencoders	688
13.0.1 Unsupervised Learning	688
13.0.2 Autoencoders: Compression	689
13.0.3 Training	690
13.1 Autoencoder Structure	691
13.1.1 Visualization	691
13.1.2 Anatomy of an Autoencoder	693
13.1.3 One layer encoder/one layer decoder	694
13.1.4 Autoencoders in general	695
13.2 Autoencoder Learning	697
13.3 Evaluating an Autoencoder	699
13.3.1 Dimensionality of a	699
13.3.2 Data Analysis	700
13.3.3 Downstream Tasks	703
13.4 Linear Encoders and Decoders	705
13.4.1 Principle Component Analysis (Optional)	706
13.4.2 Low-variance: less important (Optional)	706
13.4.3 Different axes (Optional)	707
13.4.4 General example (Optional)	709
13.4.5 Non-linear encoders (Optional)	709
13.5 Advanced Encoders and Decoders (Optional)	711
13.5.1 Generative Networks (Optional)	711
13.5.2 Adversarial Optimization (Optional)	715
13.5.3 Generative Adversarial Networks (Optional)	717
13.5.4 De-noising (Optional)	719
13.5.5 Attention (Optional)	720

13.5.6 Transformer Networks (Optional)	721
13.6 Terms	724
A Matrix Derivatives	725
A.1 Introduction and Review	725
A.1.1 Partial Derivatives	726
A.1.2 Thinking about derivatives	728
A.1.3 Derivatives: Approximation	729
A.2 Derivative: Scalar/Vector (Gradient)	731
A.2.1 Finding the scalar/vector derivative	731
A.2.2 Review: Planar Approximation	732
A.2.3 Our completed scalar/vector derivative	735
A.3 Derivative: Vector/Scalar	738
A.3.1 Working with the vector derivative	738
A.4 Derivative: Vector/Vector	741
A.4.1 The vector/vector derivative	741
A.5 General derivative (Vector/Vector)	743
A.5.1 More about the vector/vector derivative	746
A.6 Derivative: matrix/scalar	748
A.7 Derivative: scalar/matrix	750
A.8 Tensors	752
A.8.1 Other Derivatives	752
A.8.2 Dimensions (Optional)	754
A.8.3 Dealing with Tensors	756
A.9 Chapter 7 Derivatives	759
A.9.1 The loss derivative	759
A.9.2 The weight derivative	759
A.9.3 Linking Layers $\ell - 1$ and ℓ	764
A.9.4 Activation Function	766
A.9.5 Element-wise multiplication	768
A.10 Terms	770
B Optimizing Neural Networks	771
B.1 Strategies towards adaptive step-size	771
B.1.1 Momentum	771
B.1.2 Adadelta	779
B.1.3 Adagrad	783
B.1.4 Adam	784
B.2 Batch Normalization Details	788
B.2.1 Applying batch normalization to backprop	788
C Recurrent Neural Networks	794
C.0.1 Review: Neural Networks So Far	794
C.0.2 Time in a Neural Network	795

C.1	State Machines	796
C.2	Recurrent Neural Networks	796
C.2.1	Building up RNNs	796
C.2.2	Offset	796
C.2.3	Activation Function	798
C.2.4	Shape	799
C.2.5	Complete RNN	801
C.2.6	RNN as a "network"	802
C.2.7	RNN fully unpacked	803
C.2.8	RNN Example 1 (Optional)	806
C.2.9	RNN Example 2 (Optional)	809
C.3	Sequence-to-sequence RNN	812
C.3.1	The sequence-to-sequence perspective	812
C.3.2	Sequence length	813
C.3.3	Training data	815
C.3.4	Training and Evaluation	816
C.3.5	Activation Functions	817
C.4	RNN as a language model	819
C.4.1	Tokens	819
C.4.2	Predicting tokens	820
C.4.3	Start token and end token	821
C.4.4	Why we might use RNNs for language	824
C.4.5	Why RNNs don't work (well) for language	824
C.5	Terms	826
D	Word2vec – Skipgram Approach	827
D.1	Vector embeddings and tokens	827
D.1.1	One-hot encoding isn't enough	827
D.1.2	Word Embeddings: Similarity between words	828
D.1.3	Vector Similarity: Dot Products	829
D.1.4	Semantic Similarity and Word Frequency	830
D.1.5	Clarifying our probability	831
D.1.6	Computing predicted probabilities	834
D.1.7	Skip-gram approach: Training our word2vec model	836
D.1.8	Issues with skip-gram	841

CHAPTER 0

Prerequisites - Explanatory Notes

This course assumes knowledge of several topics. Here, we'll outline them: hopefully, this will make it easier to get up to speed if you have a gap in your background, and to know what you're looking for.

This is designed to be somewhat comprehensive, so if you've taken a class, you can likely skip the corresponding section.

If a class has its own prerequisite, then we assume the understanding of that class as well. For example, we assume you know multi-variable calculus, so single-variable is assumed as well.

0.1 Multi-variable Calculus (MIT's 18.02)

You will need most of the differential aspects of multi-variable calculus:

- The concept of partial derivatives and how to find them
- The **multivariable** chain rule
- An intuition for the gradient as the **direction** of greatest increase in a function, and the **magnitude** of that increase.

It is helpful to able to visualize a surface created by a function. You won't need to memorize the shapes of specific surfaces; just the 3D intuition in general.

You should also be able to imagine "zooming in" to that surface, and seeing it "locally" as a **plane**, just like how we zoom in to a one-variable function, and see the **tangent line**.

Sometimes, in this class, we will also do derivatives **numerically**, where we approximate the derivative with finite steps, of the form

$$\frac{dy}{dx} \approx \frac{\Delta y}{\Delta x} \quad (1)$$

You should also be comfortable with some basic ideas of "infinity": what happens in the "limit as we approach infinity", for example.

You will not need double/triple/line integrals, curl, divergence, or greens/stokes theorem.

0.2 Vectors and Matrices (MIT's 18.02)

You will need an understanding of vectors:

- You need to know what a vector is, with two interpretations:
 - an **ordered list** of numbers
 - an object in some "space" with **magnitude** and **direction**
- You should know that the **length** or **dimension** of a vector is just how many numbers(**or elements**) that vector contains.
- You should know that a **scalar** is just a number, or in some perspectives, a 1-element vector.
- You need to be able to **add** or **subtract** pairs of vectors. You should also be able to **scale** them (in other words, multiply by a **scalar**).
- You should know how to take the **derivative** of a vector \vec{v} , with respect to a scalar x , in the form

$$\frac{d\vec{v}}{dx} \quad (2)$$

You will also need to understand the **dot product**, and its intuition as the **similarity** between vectors.

You will need to understand matrices:

- You should know what a matrix is, with three perspectives:
 - a 2D grid of numbers (in a "rectangle")
 - an **ordered list** of equal-length vectors.
 - a **transformation** of vectors.

- You should understand the **dimensions** of a matrix, and the common notation (# of rows \times # of columns)
- You need to be able to multiply two matrices, or a matrix times a vector.
 - You need to understand when you are able to multiply matrices, based on dimensions.
 - The dimensions of the new matrix after multiplication.
 - How to do the calculation of multiplying matrices by hand.

You should understand the **determinant**: both how to calculate it, and the intuition behind it.

Finally, you should know what a matrix **inverse** is, and that a **zero-determinant** matrix has **no inverse**.

0.3 Linear Algebra (18.06)

Currently, linear algebra is a **new** prerequisite.

You need all of the concepts mentioned in the "vectors and matrices" segment.

You should also understand **independence** between vectors, the **rank** of a matrix, and how to take the **transpose** of a matrix.

Linearity is a really nice (and important!) property , where a function doesn't get in the way of some simple operations: **addition** or **scalar multiplication**. The order doesn't matter.

Definition 1

For **linear** function/operator \mathbb{L} ,

Addition (of any kind) has the same effect, before or after the function.

$$\mathbb{L}(x + y) = \mathbb{L}(x) + \mathbb{L}(y)$$

Multiplication by a **scalar** also has the same effect.

$$\mathbb{L}(3z) = 3\mathbb{L}(z)$$

Example: The **derivative** is **linear**:

$$\frac{d}{dx}[f + g] = \frac{d}{dx}[f] + \frac{d}{dx}[g] \quad (3)$$

$$\frac{d}{dx}[10h] = 10 \frac{d}{dx}[h] \quad (4)$$

Often, we talk about **operators**. They're like functions, where they have an input and an output. But sometimes, we use this word when we have **another** function as an input.

Definition 2

An **operator** often takes in a **function** as an input, and gives another **function** as an output.

Example: The **derivative** is also an **operator**. If you input $f(x) = x^2$, the output is $\frac{d}{dx}[f(x)] = 2x$: another **function**.

An operator doesn't have a really "unique" definition in math: it's used for convenience.

Thus, we can call the derivative a **linear operator**.

The **visual** intuition of a matrix as a spatial transformation is useful in this class.

It would be helpful to understand **nullspace**, **column space**, and **vector spaces** in general.

A good reference for intuition is the linear algebra series by YouTube channel, 3blue1brown! Each video averages 11 minutes, and has been helpful for many past students.

0.4 Programming (6.100A, 6.1010)

For 6.100A:

You should be familiar with object-oriented programming in **Python**: you will be implementing various classes and simple algorithms in this class.

You should have a basic understanding of **time complexity** and big-O notation. Ease with reading basic pseudocode would be helpful as well.

For 6.1010:

Either 6.1010 or 6.1210 are counted as a prereq: they are not equivalent, and neither is individually required to understand the course, but either will make your work in this class much easier.

6.1010 offers more coding experience, which makes the process of implementation much smoother.

Misc:

Prior understanding of numpy is not mandatory, but would be very helpful. Pytorch would also be helpful, but is not used until much later in the course.

0.5 Algorithms (6.1210)

Either 6.1010 or 6.1210 are counted as a prereq: they are not equivalent, and neither is individually required to understand the course, but either will make your work in this class much easier.

6.1210 is about **algorithms**, and thus makes it easier to understand our discussions of different algorithms throughout this class.

Concepts of dynamic programming, complexity, and reading/writing pseudocode are all helpful for this course.

0.6 Probability

You don't need to have taken a full probability course, but there are some core concepts you must understand:

- **Probability** (or chance) is the relative **frequency** of a particular outcome, if you were to run many trials - specifically, it is the proportion of those trials that gave this particular outcome.

For example, if $p = .4$, then you should expect to have that event occur 40 times for every 100, on average.

- Probabilities p are between 0 and 1:
 - If $p = 0$, the event **will not** occur
 - If $p = 1$, the event **will definitely** occur.
 - If $p = .5$, the event has a 50% chance of happening if you try once.
 - Any other probability between 0 and 1 will give some corresponding percentile chance of occurring ($100 * p\%$)
- You can represent probability of event A as

$$P(A) \quad (5)$$

Treating P as a function that returns the **probability**.

- If you want two events to both occur, you can write that with an **and** statement:

$$P(A \text{ and } B) = P(A \cap B) \quad (6)$$

- If you want at least one of two events to occur, you can write that with an **or** statement:

$$P(A \text{ or } B) = P(A \cup B) \quad (7)$$

- The **conditional probability** is the probability of an event **given** that another event has already occurred:

$$P(B|A) \quad (8)$$

This is read as "The probability of B **given** A".

- We can use this to get the probability of two events at the same time.

$$P(A \text{ and } B) = P(A|B)P(B) \quad (9)$$

- Two events are **independent** if knowing the outcome of one event does not affect the odds of another. You can write this as

$$P(A|B) = P(A) \quad (10)$$

You can equivalently use the next bullet point's definition.

- The chance of two **independent** events both happening is their odds multiplied.

$$P(A \text{ and } B) = P(A)P(B) \quad (11)$$

- The **sum** of the probabilities of all outcomes **must add** to 1: otherwise, there's a chance of getting none of the listed outcomes.
- The chance of a particular event not occurring is called the **complement**, and has a chance of $1 - p$.

0.7 Notation: Sets

There are some common definitions and notations you should be familiar with (though they may be introduced if necessary).

If you understand an equation, move on to the next one: each explanation is mostly basic.

Definition 3

An **element** is a single object in a collection of objects.

This definition is often linked to **sets**.

Definition 4

A **set** is a collection of distinct elements with no given order: if you shuffle the elements in a different order, you have the same set.

- We can **define** a set by listing out its elements. For example, this says, "The set A

contains the numbers 1, 2, and 3"

$$A = \{1, 2, 3\} \quad (12)$$

- Shows that an element is **in** a set. The following says "x is an element of the set A".

$$x \in A \quad (13)$$

- Shows that an element is **not in** a set. The following says "x is **not** an element of the set A".

$$x \notin A \quad (14)$$

- Natural numbers**, or the counting numbers.

$$N = \{1, 2, 3, 4, 5, \dots\} \quad (15)$$

- Real numbers**, or all of the numbers on the number line (including the full space between integers).

$$\mathbb{R} \quad (16)$$

- We can also define a set by starting with another set, and then listing a **restriction**:

The following says, "Include each natural number n".

$$\{n \in N\} \quad (17)$$

This seems redundant, but now, we can choose to only include natural numbers **larger than 10**:

$$\{n \in N \mid n > 10\} = \{11, 12, 13, 14, \dots\} \quad (18)$$

- A set A is a **subset** of B if all elements in A are contained in B. B is then said to be a **superset** of A.

For example, the set of natural numbers is a **subset** of the set of real numbers: every natural number is also a real number.

Here, this says " \mathbb{N} is a subset of \mathbb{R} ".

$$\mathbb{N} \subseteq \mathbb{R} \quad (19)$$

Notice that this symbol looks similar to \leq : that's intentional! This is because the two sets could be the same set, while one is a subset of the other.

- The previous symbol allowed for the two sets to be the same. But if we know they aren't, we can use a **proper subset**.

Here, this says " \mathbb{N} is a **proper** subset of \mathbb{R} ".

$$\mathbb{N} \subset \mathbb{R}$$

(20)

Since we know that some real numbers are not natural numbers, they can't be the same.

0.8 Notation: Numbers and functions

- **Sum notation:** adding up elements in a sequence. For example:

$$\sum_{n=1}^6 n^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2 \quad (21)$$

- **Product notation:** multiplying elements in a sequence. For example:

$$\prod_{n=1}^5 n = 1 \times 2 \times 3 \times 4 \times 5 \quad (22)$$

- **Rounding up:** round up real numbers. The following says, "round 2.5 up to the nearest whole number"

$$\lceil 2.5 \rceil = 3 \quad (23)$$

- **Rounding down:** round down real numbers. The following says, "round 2.5 down to the nearest whole number"

$$\lfloor 2.5 \rfloor = 2 \quad (24)$$

- **Function** notation: shows the name of the function, the set of inputs, and the set of outputs.

For example, this below says, "the function f takes real numbers as inputs, and outputs natural numbers."

$$f : \mathbb{R} \longrightarrow \mathbb{N} \quad (25)$$

- If you want to get the **maximum** or **minimum** output of a function, you use the function with the corresponding name: max or min.

For example:

$$\min_{x \in \mathbb{R}} x^2 = 0 \quad (26)$$

$$\max_{x \in \mathbb{R}} \sin x = 1 \quad (27)$$

Below the max or min declaration you can denote the domain over which to find the maximum or minimum, respectively.

- Sometimes, you don't want the minimum or maximum output: you want to know the **input** that gives you the minimum or maximum output. If the domain can be inferred from context, it may be omitted.

So, you pick an **argument** (input variable) and get the **argmax** or **argmin**

The following says, "x = 1 **gives you** the minimum output for $f(x) = (x - 1)^2$ ".

$$\arg \min_x (x - 1)^2 = 1 \quad (28)$$

The following says, "f(x) = 0 **is** the minimum output for $f(x) = (x - 1)^2$ ".

$$\min_x (x - 1)^2 = 0 \quad (29)$$

Make sure to keep track of the **difference** between min and argmin, or max and argmax!

0.9 Notation: Vectors Spaces

Here, we'll build up some notation for representing sets of vectors, by representing them as ordered sequences.

- Often, we care about **ordered sequence** of numbers. Maybe you want to return the entire sequence.

We start with ordered pairs of numbers: you can represent every pair of elements from two sets with \times .

For example, here we have "every pair of two natural numbers":

$$\mathbb{N} \times \mathbb{N} = \{(1, 1), (1, 2), (2, 1), (2, 2)\dots\} \quad (30)$$

Notice that this can be used to fill in an grid of numbers: with real numbers, you can fill in the whole space with no gaps. Note that since sets do not contain duplicate elements, (1, 1) is not included twice.

- If you want **more than two** elements, you can simply use more **crosses** \times .

For example, here is every trio of natural numbers:

$$\mathbb{N} \times \mathbb{N} \times \mathbb{N} = \{(1, 1, 1), (1, 1, 2), (1, 2, 1)\dots\} \quad (31)$$

- Here, we introduce a shorthand, because writing every cross (example: $\mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}$) can get tiring.

We can compress multiplication with **exponents**, so we'll do the same here:

$$\mathbb{R}^5 = \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \quad (32)$$

- Because one of our perspectives on **vectors** is as "an ordered list of numbers", we can represent all of our desired vectors using this notation.

In general, the set of all length-n vectors can be represented as

$$\mathbb{R}^n \quad (33)$$

0.10 Optional

Here, we list concepts that could be **helpful** for understanding this course more easily, but are entirely **not required**, as we'll be teaching what we need.

- Tensors
- Convolution
- Examples of Optimization (Least-Squares, etc.)
- Markov Chains
- Probability and Statistics (Expectation, Variance, Distributions...)
- Mathematical maturity (from upper-level math courses, etc.)
- Matrix Calculus

CHAPTER 1

Introduction - Explanatory Notes

These are the explanatory course notes produced by **Shauntclair Ruiz**, a TA as of Spring 2023. They are intended to be **supplementary** to the official lectures notes.

The official course lecture notes are designed to be **minimal**, and present what the instructors think that you absolutely **need to know** to understand and interact with the current state of machine learning.

These notes, by contrast, are designed to provide more thorough **explanations**. We **explain** certain logical leaps, **break down** concepts into smaller parts, and try to make the notes more **accessible** to students who find the primary notes too dense.

These notes cover the **same** topics as the primary notes, just with a different **presentation**. Most of the explanations in this document are a reaction to **difficulties** that students have had in previous semesters.

If the concepts in these explanatory note chapters are **familiar** to you, or if you find them to be too **drawn-out**, you can **skim** sections that you're not concerned about.

We, again, stress that neither set of notes is more "advanced", as they cover the **same material**. They simply reflect **different** learning styles and backgrounds.

It may be helpful to refer to these explanatory notes as you digest the official lecture notes: the main section numbers (1.1, 1.2, 1.3...) should **match** with the official notes.

If you have any concerns or points of **confusion**, feel free provide **feedback** on this ongoing project.

1.0.1 What is machine learning?

Why study machine learning? To answer that, let's learn what machine learning really *is*. Fortunately, it's all in the name.

Machine learning is a broad field. We use **machines**, or computers, and give them data to **learn** from.

Why are we teaching machines? Same as why we want **people** to learn: so they can use that learning to make good **decisions**.

So, in short, we can say:

Concept 5

The main focus of **machine learning** is making **decisions** or **predictions** based on **data**.

1.0.2 Why do Machine Learning: The Benefits

Why use machine learning? What is it good for?

The techniques used in machine learning have many applications. It has become the best way to handle many different problems:

- Facial detection
- Speech recognition
- Language processing
- Many problems that involve data or signal processing

Based on speed, time to develop, "robustness", etc.

Different ML techniques have become the best way to handle many problems in many fields. As a result, it has become very popular!

1.0.3 The role of humans

If machines can solve all of these problems, where do humans play a role?

Well, these machines aren't (yet) able to **set themselves up** to solve these problems: humans have to set up the system so the machines can succeed. We call this **framing** the problem.

We'll use an example to help explain.

- A human has to **recognize** that there is a problem to solve.
 - **Example:** You want to have self-driving cars. Your problem: those cars need to be able to **watch** the road.

- They have to decide what kind of **solutions** you want to try, and use that as the basis for training.
 - **Example:** You decide to create a **model** that can replicate vision for our car.
 - This **model** represents the kinds of **solutions** you expect to work: a particular model will allow for a certain approach to a situation.
- They have to **gather** data to train with.
 - **Example:** You might gather **videos** from dashcam footage, or create a virtual simulator for your car to drive in.
- They have to choose the **algorithms** we'll use for learning: what **instructions** do we give our computer?
 - **Example:** To "train" your model, you could need to adjust it to perform **better**. How do you adjust it, using the videos?
- They have to look at the final result and **validate** whether it's a good enough solution to use.
 - **Example:** You **test** out your model in a car: does it notice obstacles?
- They have to consider the possible **ethics** or other consequences of this solution.
 - **Example:** What's the most "responsible" way of driving? When should a car prioritize its own safety, or the safety of pedestrians? How much control should the user have?

This is over-simplified, but it gives us a high-level view of what we'll need going forward.

These are all important steps, and they require the human in question to make smart and responsible choices. That's why you need to learn machine learning: in order to use it **effectively**, you have to **understand** it!

We want to understand machine learning, so we'll break it down into different parts. We'll do this by **asking** ourselves a couple **questions**, and thinking about machines in the broadest sense we can: as the **solution** to a **problem**.

This breakdown is different from the one above!

1.0.4 What's the plan?

We know that, in machine learning, we want to make **decisions** or **predictions** using **data**.

Let's frame this more generally: we want to **solve** a **problem** presented to us, using our **machine** and some **data**.

This brings up some questions:

- What exactly is our **problem**?
- And **solution** do we want to use?

The answer depends on the situation, but we can break down these questions into simpler, easier ones.

1.0.5 The Problem

Simply put, our goal is to create a **machine** that **takes in** data and **spits out** some kind of results.

In that way, our machine is just a **function**.

The **problem**, then, is to reach that goal: to get our desired output from our input.

That means we're focused on what's **outside** of our machine - here, we don't know or care how the machine works, we just know what we've got (input), and what we want (output).

- **Assumptions:** What do we **assume** about our **problem**? What do we expect about our **data**, or our possible **solutions**? How do we use this knowledge?

– This step is important because these assumptions can allow us to **simplify** the problem, and often, our approach **depends** on them.

– **Example:** We might be looking at our patients (several adorable puppies), and **assume** that they are all the same **age**: we can simplify by not including age as a variable.

We often use these assumptions to come up with solutions: if they aren't true, your approach may fail!

- **Problem Class:** What are the **needs** of our particular problem? What **kind** of inputs and outputs are expected?

– In this situation, "class" means, "set of things with something in **common**". So, our "problem class" tells us, "what **kind** of problem do we have?"

"Which **group** of problems does ours **fit into**?"

– This is important for choosing our solution: our solution follows from the problem.

In order to answer a question, you need to know what you're being asked!

- * We might also use **existing** solutions to similar **problems** as inspiration for our own work.
- **Example:** Our inputs are weight, blood pressure, and breed. Our output is a number: how long do they have to live? This will be a real number, in years.
- **Evaluation Criteria:** What is our goal? We know the **kind** of output we want (structure, type, etc.), but how do we measure the **quality** of an answer?
 - This evaluation criteria is crucial, both for telling our machine how to **improve**, and to **show** other humans how well it **performs**.

Example: We could use the absolute difference between the lifetime predicted, and the lifetime the puppies actually experience.

These aspects together make up our problem, that we now need a **solution** for.

1.0.6 Solution Setup: What is a model?

Remember: our goal is to create a **machine** that **takes in** data and **spits out** some kind of results.

The **solution** is what's **inside** the machine - how do we do it? What approach do we use?

First, let's dig a little into what a **solution** is: we've mentioned before that our solution will often rely on a **model**, but what exactly *is* a model?

For our purposes, a model is a way to **simplify reality**: we strip away everything that doesn't matter, and just leave a system that can work *well enough*, in the ways that matter.

In machine learning, we sometimes care less about how **realistic** the model is, than its ability to get **good results**. That means our model is not always structured to match reality.

Definition 6

A **model** is a way of mathematically **representing** a **system**.

This system is **simplified** to only include the **details** we care about and give us the level of **accuracy** we want.

We do this sometimes because we don't *know* the true model, and sometimes because simulating the true model is too expensive and time-consuming.

We boil down a **system** into the values we **care about**, and how those values **affect** each other (in terms of math equations).

Example: A planetary model that simulates **gravity** between Mars and the sun may not account for the density of the planet, or everything that happens on the surface... but that might be good enough to predict the **length of a year** on Mars.

However, in this example, we knew all of the values of the model (the weight of the planet and sun, the distance from the sun...). We have no need for machine learning: the model is already **complete**.

Again, we emphasize that a model doesn't have to be structured to match reality - but if we know the true model, this can help.

In the problems we face, we **don't know** those values, or even always what **model** will work best. That's where the techniques we will learn come in.

1.0.7 The Solution

So now, we have a vague idea of what our solution might look like. So, let's break it into parts, like we did for the problem.

- **Model Type:** Will we make a model? What kinds of **data** will we **include** in our model?
 - Sometimes, a model isn't necessary: do we really need it? If we do, how do we **use** that model?
- **Model Class:** What **kind** of model will we use? What sort of **variables** will we use, and what **structure** will our math use?
 - Just like with problem classes, a model class is a set of models: a collection of models with similar structure.
 - We will spend much of this class exploring **different** model classes: each has benefits in different circumstances.
- **Algorithm:** Once we have a model, how do we "teach" it what we want it to know? We'll need a **procedure** for this - an algorithm.
 - Which algorithms we choose will affect how well our machine can learn: how quickly will it learn, and how good is the end result?

Now, we take a deeper dive into each aspect listed about, starting with our **base assumptions**.

1.1 Problem Class

"Problem class" is, from the name, the type of problem you are presented with: what inputs and outputs are expected?

But there is a second aspect of the problem we haven't discussed - what **data** is does the machine have available when it is **training**?

1.1.1 Supervised vs. Unsupervised

We know that, to train our computer, we have to give it **input** data, but how does the machine know whether it's doing well? We could, for example, give it an "**answer key**": the correct outputs we expect from it.

If we do, we call that **supervised learning**.

Or, do we find a different way to measure its success? We break it down into a few common cases:

In a way, we're "supervising" it by giving it the right answer: we're guiding it and making sure it does what we want it to!

- **Supervised Learning** is when we train our machine using a set of **inputs** and the correct matching **outputs**.
 - **Example:** You show your machine a bunch of **pictures** (inputs), and then **label** what is in each picture: like a dog (output).
- **Unsupervised Learning** is when we **don't** give our machine the answers, and it has to guess without having a "correct" answer.
 - This is often used in cases where we don't know a "correct" answer in advance. For example, we might want to find some kind of **pattern** in our data, and we have no way of predicting that!
 - **Example:** You look at a bunch of animals (input) and try to invent species for groups of animals.
- There are other cases that we will save for the end of this section.

We'll list some common problem classes for each of these types. You don't have to memorize these types, but they will come back later in the course.

You don't need to know the species "cow" and "pig" to figure out that they're different from each other!

But first, one more detail.

1.1.2 How do we store our data?

A data point usually represents a single "thing". It's helpful to be able to use **multiple** facts about this one thing.

To store these facts, a data point requires **multiple** values: pieces of information you want to draw **conclusions** from.

We often standardize the information our machine receives by storing this information in a **vector**.

Being consistent makes it easier to develop the techniques we need!

Our models are usually made up of **equations**, so we want to be able to **compute** with these values. So, each variable will be represented with a real number, or multiple if necessary.

Thus, one data point is a **vector of real numbers**. Specifically, a **column vector**.

Notation 7

x is our **vector of inputs**.

It is a column vector. Its matrix shape is $(d \times 1)$.

Example: Suppose we have a data point x that's a vector of 3 numbers: its shape is (3×1) . We write this as:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad (1.1)$$

Since this is so common, we introduce some notation.

The real numbers are represented with \mathbb{R} . Since we're combining **multiple** real numbers, we use **exponent** notation to represent this.

Notation 8

The **set of length-d vectors** is written as \mathbb{R}^d .

Example: \mathbb{R}^2 represents all of the length-2 vectors: all of the vectors/points on the 2D plane.

So, we might say a data point $x \in \mathbb{R}^d$ if all we know is that x is a length-d vector.

1.1.3 Supervised Learning

- **Regression** takes in a vector of numbers, and predicts some **real number** as an output. Our goal is to correctly guess the desired output.

– **Example:** You want to predict how much a worker makes based on their job, where they live, and how many years they've worked.

Notice that "where they live" isn't usually represented as a number: we often have to convert certain data types.

- **Classification** also takes in a vector of numbers, but outputs a **label**: we have a set of **classes**, and we want to **label** each data point as a member of one class.

– This means our output is **discrete**: each class is separate output value, and we have k classes.

As before, a "class" is just another word for a group of related things.

- **Example:** You have several documents and want to **label** which **language** each is written in.

1.1.4 Unsupervised Learning

- **Density Estimation** takes in data, and tries to approximate the **distribution** of that data: what is the chance of getting a new data point x ? _____

- **Example:** You want to get the **distribution** of human **heights** in a particular city.

We define "distributions" in section 1.2.

- **Clustering** is when you want to sort data points into groups of similar points, without knowing the groups in advance.

- **Example:** You want to sort patients with a disease into groups, where each group might need different treatments.

- **Dimensionality Reduction** is a bit different: the goal is to take a vector, and reduce the length of the vector, while still keeping the information that's important.

- You may not need every dimension to store the information you need, so you can save on space and time by storing it in a smaller vector.

- **Example:** You find out height has no effect on income, so you ignore height. Or maybe you find that having both education and literacy is redundant. _____

- Notice that what information is relevant depends on what you're using it for.

Since we often automate this process, real examples might not be so simple!

1.1.5 Other Types of Learning

Now, we turn to some types of learning that are, arguably, neither supervised nor unsupervised.

- **Reinforcement Learning** is used when you have an "environment" you can interact with. Different choices will change what that environment looks like, and may reward or punish you. _____

- The goal is to pick the actions that give you the best rewards.

We represent the current environment with something called a **state**.

- **Example:** You have a robot on Mars, and you want to move your robot (action) to reach certain goals (rewards)

- This isn't **supervised** because you **don't know** the correct action. But it isn't fully unsupervised because you **do know** when you get a reward.

- **Sequence Learning** is used to take one sequence, and turn it into another. In these sequences, each output depends on all of the previous inputs.

- This means you need to store information about previous inputs using a **state**.

- **Example:** Predicting the next word in a sentence, based on the words so far. You **predict** one word for each new one you receive, so you return a **sequence**.
- We're partly "supervised" by being given the output sequence, but we don't know what our states need to look like.

1.1.6 Types of Learning not covered in this class (Optional)

These will not be covered, but are worth mentioning.

- **Semi-supervised Learning** gives us some supervised training data that has been labelled, but also some that has not.
- **Active Learning** gives our computer the ability to **choose** which data points it receives: this is used when data is **expensive**, and we want to learn efficiently.
- **Transfer Learning** is used when we apply learning from one task to another, related task. That way, the new task can be learned faster.

1.2 Assumptions

Let's look at our underlying assumptions: the rest of this class relies on these assumptions.

1.2.1 An assumption about data

Let's return back to our original goal: we want to use **data** to teach our machine to give us **results** we want. Just like how a person might learn from their **experience** and use it to make **judgments**.

However, there's an **assumption** built in to this statement, one we need to look at more closely: we are assuming that **past** data allows us to predict **future** data.

This may seem obvious, but it isn't always: past data may not be **representative** of the future, for example.

- **Example:** We can't use the weather over the month of July to predict the weather in the month of December.

This often called the problem of **induction**: using the past to predict the future.

1.2.2 Is our data representative?

First, let's solve the problem presented above:

- **Example:** We got our weather from a **different** month than we're trying to predict.

So, it seems our problem is that our **data** and what we're trying to **predict** are from **two different sources**.

We want them to come from the **same source**, then. In this case, we could say we want them to be from the **same** month. Great. But how do we say this in general?

1.2.3 How do we compare data?

We got down to the real problem: we want our new data to be from a similar source to the old data. One month couldn't **represent** another, because they **behave** differently.

- **Example:** For different months, we get different rainy days, different temperature ranges, so on: they can't be compared.

In general, we need a way to describe what we mean by "different": what describes one of these months?

- **Example:** To us, all that matters is the weather: how **likely** are we have a rainy day, for example? In fact, we'd like to know how **likely** every outcome is.

We represent this with something called a **distribution**. A distribution gives us exactly what we just described: **how likely** different events are to occur. _____

This is how our system "behaves", in a way.

Definition 9

A **distribution** is a **function** that gives us the **probability** of different **outcomes**.

Example: The **distribution** of outcomes on a coin is 50% chance of heads, 50% chance of tails.

Notice that distributions are **probabilistic**: outcomes have a certain **chance** of occurring. Otherwise, these problems would be simple.

Why is it called a distribution? Well, we're taking the **odds**, and spreading them out (or **distributing** them) over multiple different outcomes!

1.2.4 Identically Distributed Data

We can think of this distribution as a **simplified** view of the **source** of our data. Each "outcome" is a data point; one we can use to **learn**.

We want our **past** data we **learn** from, and our **future** data with **test** with, to have the **same** distribution.

We also want different points in the **same** dataset (past or future) to be from the same **distribution**: if they aren't, then why are we lumping them together?

We want them to be the "same", or **identical**: they have **exactly** the same chances for each outcome.

We want to focus on one problem at a time - one distribution.

In other words: we want our sets of data to be **identically distributed**.

Definition 10

If two **data points** (or datasets) are **identically distributed**, then they have the **same** underlying **distributions**.

In other words, they have the **same probabilities** for each possible **outcome**.

Example: Two fair coins will behave the same as each other: they both have 50-50 odds. Thus, they're **identically distributed**.

1.2.5 Independence (Review)

There's a second assumption that is just as important: when we draw two different data points, we are also **assuming** that the results of one do not **affect** the other.

If one point **depended** on another, then there's no **new** information: you could have used the last point to guess this one.

This means you're **not learning**, which is a problem: you need many experiences to come to a good **conclusion**, that will apply well in the future.

Because we don't want the result of one data point to **depend** on another, we call this assumption **independence**.

Definition 11

Two **data points** are **independent** if **knowledge** of the outcome for one data point does not affect the **probabilities** for the other.

Example: If you flip two coins, knowing that one coin comes up heads does not tell you anything about the other coin: the two coin tosses are **independent**.

This definition is a bit informal: the proper definition is to say that, for two events A and B, $P(A)P(B) = P(A \text{ and } B)$

1.2.6 Independent and Identically Distributed

We combine both of these assumptions into our final result: we want our data points and data sets to be both **independent** and **identically distributed**.

Definition 12

IID, or **Independent and Identically Distributed**, means that if you draw two data points, they

- Come from the **same distribution**: they have the same **probabilities** for each outcome,
- They **aren't related** in any other way: they are **independent**, meaning the **outcome** of one **does not** affect the other.

Example: Based on the two examples above, flipping two coins (or rolling a die twice) is IID.

We shorten this to one acronym, which tells you how important it is: it is the base assumption in many different statistics, inference, and machine learning settings.

We will assume this to be true, and use that assumption throughout the class. We expect our data to be IID in most cases.

1.2.7 Estimation and Generalization

In this section, the main theme has been applying knowledge about **training** data to **new**, unfamiliar situations, like our **testing** data.

We have a word for this that we haven't used so far: **generalization**.

Definition 13

Generalization is the **problem** of applying **current** knowledge to **new** situations we've never seen before.

We want to be able to take the **specific** case of our training data, and apply it to the more **general** case of any of the possible **new** data.

A second problem is the **nature** of our training data: because we **randomly** select it, we don't have a perfect idea of what the true distribution looks like.

The randomness means that our sample will look a bit different each time we generate it.

This creates some **noise**: something that interferes with what we're trying to focus on.

The problem of using our sample to **estimate** the true distribution, despite imperfect, "noisy" data, is **estimation**.

Just like how background **noise** can make it harder to listen to a phone call!

Definition 14

Estimation is the **problem** of taking **imperfect** data and using it to **estimate** the "true" information we're looking for.

1.2.8 Other Assumptions

There are some other assumptions we will make, that will not go into as much detail on:

- We know the set of possible answers: the type of answer we should give back, whether number, label, making a choice...
 - If we don't know what kind of answer we're supposed to give, how can we build a model to give back that answer?
- Our problem is solvable: the "true" model can be represented and answered using our computer.

Imagine if you were supposed to write an essay, but could only answer with real numbers between 0 and 1 - this is what we want to avoid.

Here are some more which are less universal.

- The data might be generated by a Markov chain.
- The data might be **adversarial**: designed to specifically exploit weaknesses in the machine.

If you don't know what this is, don't worry! We come back to it later.

Some of these assumptions are required in order to move forward at all. Others narrow down the options we have to work with, so we can find a good solution in a reasonable amount of time.

1.3 Evaluation Criteria

1.3.1 What is a loss function?

In order to solve our task, we want to be able to measure how our machine is performing. We do this by creating a measure of success or failure, called a **loss function**.

Definition 15

A **loss function** measures how **poorly** your machine is **performing** on a **task**.

The output is a **real number**. If your machine is performing **well**, then you will have a **low** output. And vice versa: if it is doing **badly**, it will have a **high** output.

Example: If you counted the number of questions you got **wrong** on a test, that could be a **loss function**.

A loss function usually has the **correct** and **predicted** guesses as inputs: it has to compare them to know how well it's doing.

Notation 16

Often, we will use g to represent our **guess** as to the correct answer: this is the output of our model; our **prediction**.

The **true answer** is often represented by either a or y .

Our **loss** is the function \mathcal{L} , so altogether, our computed loss is $\mathcal{L}(g, a)$.

1.3.2 Examples of Loss Functions

Different loss functions are useful for different situations.

- **0-1 Loss** is a simple kind of loss: if our answer is correct, the value is 0. If our answer is incorrect, the value is 1.

$$\mathcal{L}(g, a) = \begin{cases} 0 & \text{if } g = a \\ 1 & \text{otherwise} \end{cases}$$

This matches our earlier example of "number of questions wrong on a test".

- This kind of loss is often used for **discrete** situations, where there are k options and one is correct - like on a multiple-choice test.
- **Linear loss** is the **absolute difference** between your answer and the correct one.

$$\mathcal{L}(g, a) = |g - a| \quad (1.2)$$

- **Square loss** is the **square difference**.

$$\mathcal{L}(g, a) = (g - a)^2 \quad (1.3)$$

- Because the slope increases as you get further away from 0, it punishes large errors more aggressively than small errors.
- **Asymmetric Loss** punishes some outcomes more than others. It may be worse to miss a heart attack, than to expect one and be wrong.

$$\mathcal{L}(g, a) = \begin{cases} 1 & \text{if } g = 1 \text{ and } a = 0 \\ 10 & \text{if } g = 0 \text{ and } a = 1 \\ 0 & \text{otherwise} \end{cases}$$

1.3.3 How to use loss

We want to reduce loss as much as we can: in other words, **minimize** it. But there are lots of ways to do that.

In this class we will minimize the **expected loss**: the average loss we would *expect* based on the probability of each outcome.

We do this because, over the **long-term**, the **expected loss** should reflect what we actually get.

We could the "worst-case" loss, or the average loss, etc...

Concept 17

In most machine learning problems, we want to **minimize** our **expected loss**.

This is called the **law of large numbers**: if you have a large number of trials, the average value should be close to the expected value.

But we also need to be careful when choosing our loss function: if we get to choose how we're grading ourselves, then we need to pick an accurate way to measure progress!

1.4 Model Type

Now, we start on the solution. Do we choose to use a model? If we do, there are some other details we have to consider.

1.4.1 No Model

A model allows us to **simplify** what we learn from our data. So, if we don't use a model, we have to use our data **directly**.

One way to do this is to simply average some known data points that "seem" similar to the newest query. This is called the **nearest neighbor** approach.

Example: You measure a chemical's physical properties, and **label** it based on which one you've seen before is the most **similar**.

When we say "similar", there are multiple ways to interpret this, but often we use distance in \mathbb{R}^d space.

1.4.2 Models using Parameters

These days, we're much more likely to **use** a model to make our prediction.

But, as we mentioned before, a model can be adjusted, and for almost any problem, we'll need to adjust it to fit our needs. This is the process of **training**.

How do we adjust a model? Our models will be a **function**, that has several values it uses to do calculations on our inputs. For example, here's a simple model:

$$f(x) = A \sin(Bx) + C \quad (1.4)$$

In this case, we have one input variable, x . And we have three values that don't change based on the input: A , B , and C . These values are called **parameters**.

Definition 18

Parameters are the **non-input variables** in a model that can be **adjusted** to adjust the model.

~~~~~  
**Parameters** tell you about your **model**, while the **input** variables describe one piece of **data**.

**Example:** When using the linear equation  $f(x) = mx + b$ ,  $m$  and  $b$  are your parameters.

You can think of a parameter as a dial on a machine that you can "tune" to different values, like a radio.

Adjusting these parameters will change how the model behaves - different outputs for each input - but it keeps the same overall **structure**.

By structure, we mean the formula: the way variables and parameters **interact**. The above model will (almost) always be **different** from \_\_\_\_\_

"Almost" because, if  $A = B = 0$  for both cases, they're both the constant function  $f(x) = C$ .

$$f(x) = Ax^2 + Bx + C \quad (1.5)$$

They both have three parameters, and one input, but they are different models.

### 1.4.3 Prediction Rule

Our goal is to use one of these equations to **directly** calculate our prediction. For that reason, we call this equation our **prediction rule**, but more often, we will call it our **hypothesis**.

#### Definition 19

A **hypothesis** is the **function** that defines our model, using a fixed number of **parameters**.

The **output** of our hypothesis is typically the **prediction** our model is designed to create.

### 1.4.4 Hypothesis Notation

For simplicity's sake, we often lump all of our **parameters** into a single **vector**,  $\theta$ . Just like for  $x$ , we'll use a **column vector**.

#### Notation 20

$\theta$  is our **vector of parameters**.

It is a column vector. Its matrix shape is  $(d \times 1)$ .

**Example:** Here is a vector  $\theta$  with 4 parameters.

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{bmatrix} \quad (1.6)$$

Often, a vector is represented by bolding a variable ( $x$ ), or putting an arrow over it ( $\vec{x}$ ). Since we work with vectors so often in this class, we will omit this notation.

Similarly, we lump all of our inputs into a single **vector**,  $x$ .

But, if we have **multiple** data points, we need to label them **separately**.

**Notation 21**

$x^{(i)}$  is the  $i^{\text{th}}$  **data point**, represented as a vector.

Sometimes, you may instead see the notation  $x_i$ .

$x$  is the **input** to our hypothesis  $h$ , but since  $\theta$  (our parameters) can be **adjusted**, we can think of it as a **second** "type" of input.

To represent this, we use  $f(a; b)$  notation:  $a$  is our input to a single **function**, but  $b$  allows us to describe a whole **family** of functions (by adjusting parameters).

**Notation 22**

Our **hypothesis** is shown in the form  $h(x; \theta)$ .

$x$  is our main input, while  $\theta$  is used to **define** our function (using parameters).

**Example:** Consider every linear function  $y = mx_1 + b$ .

The input is a single value  $x_1$ , while the parameters are  $m$  and  $b$ : we can put those into  $\theta$ .

$$x = \begin{bmatrix} x_1 \end{bmatrix} \quad \theta = \begin{bmatrix} m \\ b \end{bmatrix} \quad h(x; \theta) = mx_1 + b$$

### 1.4.5 Fitting

The process of **adjusting** our model (i.e. its **parameters**) to match our data is called **fitting**.

As we mentioned before, our goal is typically to **minimize** expected loss. But this expected loss is based on knowing the **true** distribution of our data. We call this loss our **test error**.

Since we usually don't know the true distribution, we have to settle for our best guess - the **training data** that we've gathered.

We call it this because we're "testing" our machine in the real world.

Instead, we could minimize the **training** error: we average it out, to see our performance. Let's write that out.

The loss for our  $i^{\text{th}}$  data point is  $\mathcal{L}(g^{(i)}, a^{(i)})$ . So, we average out  $n$  of those points:

$$\frac{1}{n} \sum_{i=1}^n \mathcal{L}(g^{(i)}, a^{(i)}) \tag{1.7}$$

Let's write this in terms of  $x$  and  $y$ .  $a$  is just another name for  $y$ .

Our guess is given by the hypothesis, so  $g^{(i)} = h(x^{(i)}; \theta)$ .

In this equation, we'll leave off  $\theta$ , to allow for non-parametric hypotheses.

**Key Equation 23**

The **expected loss** for a hypothesis is:

$$\frac{1}{n} \sum_{i=1}^n \mathcal{L}(h(x^{(i)}), y^{(i)})$$

This is the equation we would **minimize** with  $\theta$ .

### 1.4.6 Overfitting

But, we have to be careful - we mentioned before that the randomness of our sampling can introduce **noise**.

If we too heavily emphasize the current values, we may not **generalize** well to new data. This problem is called **overfitting**, and we will talk about it a lot in this course.

**Definition 24**

**Overfitting** happens when we fit **too strongly** to a particular dataset.

Because we focus too much on that **dataset**, our machine **learns** incorrect facts about the overall distribution.

This makes our model worse at **generalizing** to new situations.

**Example:** You want to know what cats are like. By coincidence, you see three black cats in a row. You assume all cats are probably black: you've **overfit** to your data.

In this course, we will discuss many ways to tackle **overfitting**.

## 1.5 Model Class

In this section, we'll assume we're using a model.

### 1.5.1 Hypothesis Class

As we mentioned before, changing our **parameters** will change the specific model we have, but it will have the same overall **structure**.

Models with the same equation are put in the same **model class**. Since our models are defined by their **hypothesis**, we will often talk about the **hypothesis class**.

#### Definition 25

A **hypothesis class** is a collection of **hypotheses** with the **same type of equation**: the only difference between them is the **value** of their **parameters**.

Another description:

- The **hypothesis class** represents all of the **possibilities** for a model class: we can get every option based on our **parameters**.

**Example:** Every hypothesis of the form  $mx + b$  is in the same **hypothesis class**.

Another way to say "same type of equation" is "same functional form".

### 1.5.2 Expressiveness

Note that some hypothesis classes are capable of things that others are **not**. For example, our linear function  $mx + b$  could never produce a **parabola**  $x^2$ .

That means if our problem **requires** a more complicated model, then we can't ever get a good result!

This can be summarized by **expressiveness** or "richness" of a hypothesis class.

#### Definition 26

If one **hypothesis class** is more **expressive** than another, it can represent a **larger** collection of possible hypotheses.

Sometimes, if a problem can't be solved in one model class, it might be solvable in a more **expressive** one.

**Example:** Quadratic equations ( $Ax^2 + Bx + C$ ) are more expressive than **linear** equations ( $mx + b$ ). Every linear equation can be **created** using quadratics, but not the other way around.

### 1.5.3 Choosing Model Classes

So, the question is - which model class should you use for a given problem?

Your first instinct might be to use the most **expressive** one you can. However, this can become very **expensive** to compute, because there are many more options you have to explore.

Often, it is already **known** what kinds of models work well for what kinds of problems - we'll explore some of those options in this class.

It's also more likely to overfit! We'll discuss why another time.

As an ML researcher gains more **experience**, they can use that experience to make **educated** guesses: they may look at multiple possible models, and pick one based on theory or practice.

Choosing the **class** of model we want is called the **model selection** problem. Choosing the **parameters** for our model, on the other hand, is **model fitting**.

Research on this is ongoing: we continue to develop new model classes to try to better handle new and old problems!

## 1.5.4 Our Linear Model

We will start this class off using one of the simplest models we know: one that only uses **addition** and **scalar multiplication**.

We have our input variables,  $x_1, x_2, x_3 \dots$  that we can combine using these two operations. We can add them together, add a constant, or multiply by a constant.

We can write this in general as:

$$h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \dots + \theta_d x_d \quad (1.8)$$

Where  $\theta_i$  are our parameters.

## 1.5.5 Linear Model: Vector Form

$\theta$  and  $x$ , both being vectors, are being multiplied in a way that looks similar to the **dot product**: multiplying together elements, and then adding.

$$h(x) = \theta_0 + \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \quad (1.9)$$

So, we can rewrite it more compactly this way:

$$h(x) = \theta_0 + \theta \cdot x \quad (1.10)$$

Note that this looks very similar to the  $y = mx + b$  formula, our original linear function!

Note that, in order for the dot product to work,  $x$  and  $\theta$  must have the same **shape**.

**Concept 27**

When using a linear model,  $x$  and  $\theta$  must have the **same shape**. They both have length  $d$ .

Meaning, they are both  $(d \times 1)$  column vectors.

### 1.5.6 Linear Model: Cleaning Up

Unfortunately, we had to leave  $\theta_0$  out to make it work: if we want to talk about **all** parameters, we'll instead use the symbol  $\Theta$ .

**Notation 28**

We represent the **parameters** of our **linear** equation as  $\Theta = (\theta, \theta_0)$

We'll swap out the dot product for matrix multiplication: using matrices will make things easier (later in the class!)

$$h(x) = \theta_0 + \begin{bmatrix} \theta_1 & \theta_2 & \dots & \theta_d \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \quad (1.11)$$

Finally, we condense our vectors into symbols.

In order to make the matrix multiplication work, we have to take the **transpose**  $\theta^T$ .

**Key Equation 29**

The **linear model** has a hypothesis of the form

$$h(x) = \theta^T x + \theta_0$$

This is the form you will use through a significant portion of this course - it's good to get used to it!

### 1.5.7 Other Models

We will explore several different kinds of models in this course.

In general, we will assume that we have a fixed, finite number of parameters. Models that don't have this restriction are called **non-parametric** models. We will use them sparingly in this class.

Instead, we will focus on our **linear** model, in stages:

- We'll upgrade the linear model with a non-linear function, so we can solve **non-linear** problems.
- We will combine many of these "non-linear units" to create **neural networks**.

Arguably, neural networks are the most **powerful** tool in the ML arsenal, and key to machine learning's modern explosion in usage.

Many of the modern models used in complex and high-performing system are **variations** on neural networks, so we will give them all the attention they need.

We'll explore some neural network variants on later on:

- Convolutional Neural Networks
- Recurrent Neural Networks

## 1.6 Algorithm

Finally, once we have our model class and a tool for evaluating our model, we can finally begin the process of **fitting** our model.

This is where the problem of developing an **algorithm** comes in - we need to decide on what set of instructions will best help us find a good model.

Our problem will typically boil down to a kind of optimization: minimizing a loss function, or more often, a modified loss function called an **objective function**.

Different problems will require different algorithms and techniques: some are general-purpose optimizers, others are specially tailors for the needs of machine learning.

One of our most powerful tools will be **gradient descent**; so much so that it has its own devoted chapter.

But, we will leave that to the next chapters.

## 1.7 Overview of the Course

Here is a short summary of each chapter.

- **Introduction:** an introduction to the basic concepts of the course, and what to expect going forward.
- **Regression:** using our linear model to learn to make numeric predictions about future data.
- **Gradient Descent:** learning to use the gradient, our "multivariable derivative", to optimize functions, like loss.
- **Classification:** using our model to sort data into different classes, and introducing some non-linear functions into that model.
- **Feature Representation:** transforming the data we receive, both to make them usable by a computer, and expanding our hypotheses to non-linear functions.
- **Neural Networks:** showing how you can combine multiple non-linear functions, to create a much more powerful function for new, exciting problems.
- **Convolutional Neural Networks:** building on neural networks with convolution, making it easier to handle images, signals, and other problems.
- **Sequential Models:** introducing "states", a way to store information over time, and how to do decision-making using that information.
- **Recurrent Neural Networks:** We combine neural networks with states to build up a sequence of outputs over time, allowing us to do some language processing.
- **Reinforcement Learning:** making decisions in a changing environment, where some states and choices reward you more than other.
- **Non-parametric methods:** introducing some different tools, which are often cheaper to develop and sometimes just as effective as more complex methods.
- **Clustering:** trying to find hidden patterns and structures in data, and making that data easier to visualize for human usage.

## 1.8 Terms

- Machine Learning
- Problem Class
- Model
- Model Class
- Distribution
- Identically Distributed
- Independence
- IID
- Induction
- Generalization
- Estimation
- Supervised Learning
- Unsupervised Learning
- Regression
- Classification
- Loss Function
- Expected Loss
- Parameter
- Non-Parametric Model
- Hypothesis
- Fitting
- Overfitting
- Hypothesis Class
- Expressiveness
- Linear Model

# CHAPTER 2

---

## Regression

---

Machine learning, as demonstrated in the first chapter, is an incredibly broad subject.

- The best way to start, then, is with an example.

We'll start with a simple model: **regression**.

- As we go through, we'll develop some broader **concepts**: we'll use these throughout the rest of the class.

### 2.1 Problem Formulation

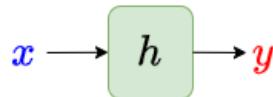
#### 2.1.1 Hypothesis (Review)

In last chapter, we broke up our model into two parts: **problem** and **solution**.

- Our problem: using our input  $x$  to **predict** output  $y$ .

$$x \rightarrow \boxed{?} \rightarrow y$$

- Our solution: we use a function, called a **model**.
  - This is also called our **hypothesis**  $h$ .



Our hypothesis reads an input  $x$ , and predicts the output  $y$ .

### Definition 30

A **hypothesis** is a **function** we use to predict  $y$ , based on  $x$ .

$$y = h(x)$$

This is also called a **model** in machine learning.

There are many models we could use: this makes it hard to search for a good one.

- One solution is to restrict ourselves to only a certain **class** of model.

Each of these is a different kind of model we could try:

$$h_A(x) = \theta_1 x + \theta_0$$

$$h_B(x) = \theta_2 x^2 + \theta_1 x + \theta_0$$

$$h_C(x) = \theta_2 \sin(\theta_1 x) + \theta_0$$

Each of these formats represents a **hypothesis class**, or a "model class".

### Definition 31

A **hypothesis class**  $\mathcal{H}$  is a **set** of possible hypotheses.

- Typically, we include all of the hypotheses with the **same equation format**.

**Example:** Let's consider the hypothesis class, represented by  $h_A$ : \_\_\_\_\_

$$\mathcal{H}_A = \left\{ h(x) : h(x) = \theta_1 x + \theta_0 \right\} \quad (2.1)$$

Here are a few example functions from  $\mathcal{H}_A$ :

Not familiar with set notation?  $\mathcal{H}_A$  is:

"the set of every function that looks like  $\theta_1 x + \theta_0$ ".

$$h_1(x) = 1x + 5 \quad h_2(x) = 3x - 9 \quad h_3(x) = -10x + 10 \quad h_4(x) = e^2 x - \pi$$

This makes things easier: rather than searching all possible hypotheses, we're searching  $\mathcal{H}$ .

### Concept 32

Our goal is to **search**  $\mathcal{H}$ , to find a good **hypothesis**  $h$ .

Within a hypothesis class, every hypothesis has the **same structure**.

- What makes a hypothesis different? The **constants**  $\theta_i$ .
- We call these **parameters**.

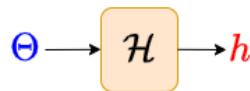
### Definition 33

A **parameter**  $\theta_i$  is a **number** that is plugged into the hypothesis.

- Each **hypothesis**  $h \in \mathcal{H}$  has a **unique list of parameters**  $\Theta$ .

If we know our model class  $\mathcal{H}$ ,  $\Theta$  **fully defines** our hypothesis  $h$ .

Typically,  $\theta_i$  is a real number, for our purposes.



By plugging in our  $\Theta$  values into the formula for  $h$ , we get a particular hypothesis.

For example:

$$\begin{bmatrix} 2 \\ -7 \end{bmatrix} \longrightarrow h(x) = \theta_1 x + \theta_0 \longrightarrow h_1(x) = 2x - 7$$

This  $\Theta$  is what we'll modify, and use to find a **better model**.

### Concept 34

Often, we already know which hypothesis class  $\mathcal{H}$  we're using.

- So, it's very easy to convert back-and-forth between  $\Theta$  and  $h$ .

So, we often consider the **parameters** and **hypothesis** interchangeable, or equivalent.

- "Finding a good hypothesis" and "finding good parameters" are, more or less, the same problem.

Notice that we have **two separate steps** of plugging in, when we use a model  $h$ :

- When choosing our model  $h$ , we "plug in" **parameters**  $\Theta$  to create our equation.

$$h(x) = \theta_1 x + \theta_0 \quad \xrightarrow{\Theta = \begin{bmatrix} 2 \\ -7 \end{bmatrix}} \quad h_1(x) = 2x - 7 \quad (2.2)$$

- When we want to use our model to predict  $y$ , we "plug in" our **input value**  $x$ .

$$h_1(x) = 2x - 7 \quad \xrightarrow{x=5} \quad h_1(5) = 2(10) - 7 = 13 \quad (2.3)$$

We'll introduce some new notation to keep the difference clear.

### Notation 35

We can write the same hypothesis  $h$  **two different ways**:

$$h(\textcolor{red}{x}) \qquad \qquad h(\textcolor{red}{x}; \Theta)$$

- The first notation is denser and **simpler** to read.
- The second notation includes  $\Theta$ , acknowledging that we had to "plug in" **parameters** to create  $h$ .

We **distinguish** between "input variables" and "parameters" by separating them with a **semicolon** ;.

## 2.1.2 The Problem of Regression

Our hypothesis is a function that solves a **problem**. What kind of problem are we dealing with?

We distinguish different types of problems based on two things:

- **Inputs:** what kind of data do we have to work with?
- **Outputs:** what are we trying to predict?

The notation for functions reflects this idea: what matters most is, "what comes in, and what goes out".

### Notation 36

A **function** is notated based on what sorts of **inputs** it can take, and the **outputs** it can return.

A function  $f$  is written like this:

$$f : \text{set of inputs} \rightarrow \text{set of outputs}$$

Functions, of course, weren't specifically designed for machine learning. But the same idea applies to other STEM disciplines.

- **Example:** Suppose that the input is "**income**" (real number  $r \in \mathbb{R}$ ) and the output is "**number of hats owned**" (natural number  $n \in \mathbb{N}$ ). The function for this would be

$$f : \mathbb{R} \rightarrow \mathbb{N}$$

Sometimes, we'll call our set of inputs, the **input space**.

### Definition 37

The **input space** is the set of all **possible inputs** to our hypothesis  $h$ .

Technically, a **space** is a set "with **added structure**", which is about as broad as it sounds.

With that out of the way, let's talk about **regression**:

- In regression, we receive data as a **real-valued vector**, and converting it into a **real number**.

Writing this a little more formally:

Remember the notation for real-numbered vectors we introduced in the last chapter!

**Definition 38**

**Regression** is a **machine learning problem** where we use a **vector of real numbers** to predict a **real-valued number**.

In other words, we want a **hypothesis**  $h$  of the form:

$$h : \mathbb{R}^d \rightarrow \mathbb{R}$$

**Example:** If you have **3 values** in your input vector (height, weight, age) and **1 real output** (life expectancy), you would need a hypothesis

$$h : \mathbb{R}^3 \rightarrow \mathbb{R}$$

A more visual example:



In this example, you have one input  $x \in \mathbb{R}$  (x-axis), and you want to **predict** the output  $y \in \mathbb{R}$  (y-axis) based on that. These points are the dataset you want to **learn** to match.

### 2.1.3 Converting our data

Often, our data **wont fit** this format: maybe we have a car brand, or a color as a variable.

- This requires **converting** this data into real numbers. We do this using something called a **feature** transformation.

**Definition 39**

A **feature** is one distinct piece of **information** in our input.

A **feature transformation** takes those pieces of information, and **transforms** them - often, a more **useful** data type.

- In other cases, we use it on data that's already in the right format, to find **new patterns** in data (we'll return to this idea in a later chapter).

**Example:** You have three car brands. Instead of representing them normally, you instead turn them into vectors:

$$\text{Brand A} \rightarrow \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \text{Brand B} \rightarrow \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \text{Brand C} \rightarrow \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad (2.4)$$

We do our feature transformation with a function: we often denote this function as  $\varphi(x)$ .

This particular feature transformation is called **one-hot encoding!** We'll return to it later.

**Notation 40**

The **function** we use to do a **feature transformation** is typically written as  $\varphi$ .

- If our input data is  $x$ , the **transformed** data is written as  $\varphi(x)$ .

**Example:** For our above car brand example:

$$\varphi(\text{Brand A}) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad (2.5)$$

There are many different feature transformations for different needs. We will come back to this in a later chapter.

- For now, we will simply assume that all of our inputs  $x$  are **already** in  $\mathbb{R}^d$  (vectors of real numbers).

## 2.1.4 Our dataset

Now, we want to find a hypothesis that solves our *particular* regression problem well.

- But in order to predict results, we need **data**.

### Concept 41

**Regression** is a **supervised problem**:

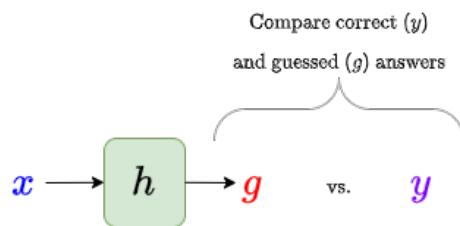
This means that, when we're trying to predict the correct answer  $y$ , we **already have** that answer.

- That way, we can check how good/bad our model's **prediction** was.

**Example:** A student practicing "supervised learning" has a practice exam, with all of the solutions.

- They try to do the exam **without** looking at the solutions.
- After they finish, they **check** the solutions, to see what they did wrong, and how they can do better.

In this analogy, each data point includes a "**question**" (input  $x$ ) and the "**answer**" (output  $y$ ) to that question.



Our model doesn't actually produce  $y$ : it produces a guess  $g$ , which we hope is similar to  $y$ .

We want to **pair up** inputs with their correct outputs, so we'll write our first data point as

$$(x, y) \quad (2.6)$$

But we have many data points we need to sort, like this.

- We'll distinguish each data point using  $x^{(i)}$  notation, from last chapter:

**Notation 42**

$x^{(i)}$  is the  $i^{\text{th}}$  **data point**, represented as a vector.

- Sometimes, you may instead see the notation  $x_i$ .

We can rewrite this as:

$$(x^{(1)}, y^{(1)})$$

Repeating this for all of our data, we have a set of  $n$  data points,  $\mathcal{D}_n$ :

$$\mathcal{D}_n = \left\{ (x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)}) \right\}$$

### 2.1.5 Training our model

We'll use this **training data** to train our model: hopefully, if our model performs well with this limited data, it'll perform well with new data.

In order to train our model, we need to know how **good** or bad it is, on the data we can see.

But as we'll see below, that's not always the case.

- We'll measure the "badness" of our model as **loss**, from last chapter.

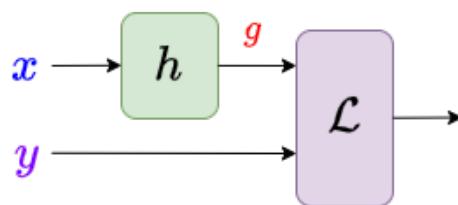
**Definition 43**

(Review from Introduction Chapter)

A **loss function** measures how **poorly** your machine is **performing** on a **task**.

- The output is a **real number**.

If your machine is performing **well**, then you will have a **low** output. And vice versa: if it is doing **badly**, it will have a **high** output.



Our loss function gives us our tool for "comparing"  $y$  to  $g$ .

We'll take the **average** loss, across all of our **training data**. This is our **training error**.

**Key Equation 44**

**Training Error**  $\mathcal{E}_n$  is written as:

$$\mathcal{E}_n(h) = \frac{1}{n} \sum_{i=1}^n \text{Loss}^{(i)} = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(h(x^{(i)}), y^{(i)})$$

This is the **expected** (average) loss over all of our training data.

Note that this is a bit weird: we're using  $h$  as an input.

- We want  $\mathcal{E}_n$  to **evaluate** and compare different hypotheses  $h$ : it has to receive that hypothesis in order to evaluate it.

The hypothesis, in turn, takes the training data as input.

**Clarification 45**

$h$  is a **function** that takes a **variable**,  $x$ , as its input. This is the usual format.

$\mathcal{E}_n$  is also a function, but it takes in  $h$ , a **different function**, as an input!

- That means one function is using **another function** as an input. This can sometimes cause confusion.

We do this because  $\mathcal{E}_n$  observes our hypothesis, and outputs, "how bad" in this particular hypothesis at **predicting** this data".

"Training", in the simplest case, boils down to reducing our error as much as possible.

- But wait: our goal **isn't** to improve performance on our **training data**.
- Our goal is to perform well on **new data**!

Below, we'll add another component, that makes this more complicated.

### 2.1.6 Learning to Generalize

Above, our idea was, "get better at practice data, get better at future data". But this has a problem:

- Even though our training data and future data should be from the **same distribution** (IID), **randomness** means they won't be exactly the same.

So if we perfectly matched our training data, we'd be inaccurate to the real distribution.

In the last chapter, we introduced a solution: a second dataset, that we **test** our data on.

- We want our machine to handle **new situations** it hasn't seen before: testing data allows us to try out a "new situation".

**Definition 46**

**Generalization** is the ability to take something **specific**, and apply it to something more **broad**.

- In machine learning, we want our model to look at some **limited data**, and be able to perform well on a much larger body of **future data** it hasn't seen before.

So, let's find out how well our model generalizes. We'll define **test error**: our performance on the **testing data**. This time, we have  $m$  new data points.

$$\mathcal{E}(h) = \frac{1}{m} \sum_i \mathcal{L}(h(x^{(i)}), y^{(i)}) \quad (2.7)$$

We'll start counting from  $n+1$  because we've already used the first  $n$  points when training.

**Key Equation 47**

**Testing Error**  $\mathcal{E}$  is written as:

$$\mathcal{E}(h) = \frac{1}{m} \sum_{i=n+1}^{n+m} \mathcal{L}(h(x^{(i)}), y^{(i)})$$

We could start over from  $i = 1$ , but that would be a bit confusing. If you said "the 10th data point", someone might ask, "from the training, or testing data? This avoids this problem.

Because we want to generalize, we want to minimize **test error**.

- But, because we want our model to do well on "data it **hasn't seen** before", we can't use it during our training process.

For now, the next best thing after "minimize test error" is "**minimize training error**", while using techniques to improve how we **generalize**.

How can we "generalize" if we can't see all of that extra data? We'll get into that below.

## 2.2 Regression as an optimization problem

We want to make our **loss** (error) as low as we can: we want to **minimize** it. This is a form of **optimization** - getting the best results from our system.

Here, we'll introduce some of the terms and notation of optimization.

Most of computer science boils down to some kind of optimization.

### 2.2.1 Objective Function

Now, we confront a major challenge: we have two different priorities.

- We want to perform well on **training data**: our model will learn some insights about the true distribution of data.
- But we also want our model to **generalize** well: we want it to do well on data it has never seen before.

Currently, our approach focuses on one priority: we measure our success using **training error**.

Because our training data gives us a limited view of the true distribution.

$$\frac{1}{n} \sum_{i=1}^n \mathcal{L}(h(x^{(i)}), y^{(i)}) \quad (2.8)$$

- This function represents "what we consider **important**": it mathematically describes what we want to improve.
- We call this an **objective function**.

#### Definition 48

An **objective function**  $J$  is the function that tells us what we want to improve, or **optimize**:

- Usually, this means that our goal is **minimizing** it.

We minimize our objective function by adjusting our model, via **parameters**  $\Theta$ . So, we take that as our input:  $J(\Theta)$ .

Since we are focusing more on  $\Theta$  than before, we'll replace  $h(x^{(i)})$  with  $h(x^{(i)}; \Theta)$ :

$$J(\Theta) = \text{Training Error} = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(h(x^{(i)}; \Theta), y^{(i)})$$

## 2.2.2 The Regularizer

If our objective function describes "what we care about", and we care about **generalization**, shouldn't we **include** it in our objective function?

- Let's do that: we'll call it a **regularizer** term.

$$J(\Theta) = \text{Training Error} + \text{Regularizer} \quad (2.9)$$

We continue using our training error from before:

$$J(\Theta) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(h(x^{(i)}; \Theta), y^{(i)}) + \text{Regularizer} \quad (2.10)$$

Later, we'll construct our regularizer so that minimizing it will create a **more general** model  $\Theta$ .

- Because we want to make  $\Theta$  more **general**, we'll use it in our regularizer: we'll call it  $R(\Theta)$ .

$$J(\Theta) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(h(x^{(i)}; \Theta), y^{(i)}) + R(\Theta) \quad (2.11)$$

Our strategy is to **minimize**  $J(\Theta)$ : by reducing training loss and  $R(\Theta)$ , we hope to make a better model.

We're missing one thing: how much do we want to prioritize  $R(\Theta)$ , compared to **training error**?

- We'll **scale** our regularizer by a **constant**  $\lambda \geq 0$ .
- The larger  $\lambda$  is, the more we care about **generalizing** to new data.

### Key Equation 49

In general, we write the **objective function** as:

$$J(\Theta) = \left( \frac{1}{n} \sum_{i=1}^n \mathcal{L}(h(x^{(i)}; \Theta), y^{(i)}) \right) + \lambda R(\Theta)$$

- The left term is the **loss**: how well we perform on training data.
- The right term is the **regularization**: we hope that **minimizing** it will make our model more "general" (good for new situations).

This is the function we want to **minimize**. What does our **regularizer** look like?

- We'll come back to this later. we'll go one step at a time, and first learn how to minimize **training error**.

### 2.2.3 More on the Objective Function

Notice that our objective function **depends** on our training data  $\mathcal{D}$  as well: the same model will work better for some problems, than others.

- **Example:** A model trained in political science learns different things compared to one trained in geology.
- We'd expect it to perform pretty badly on a geology exam.

" $\mathcal{D}$  affects  $J$ , but isn't the main input" is **similar** to our previous idea: " $\Theta$  affects hypothesis  $h$ , but we usually **don't** think of it as the **input**".

- We made this distinction with some **notation**:  $h(x; \Theta)$ .

#### Notation 50

Just like how we can use ";" when writing  $h(x; \Theta)$ , we'll use the **same notation** for  $J$ :

$$J(\Theta; \mathcal{D})$$

$$J(\Theta; \mathcal{D}) = \left( \frac{1}{n} \sum_{i=1}^n \mathcal{L}(h(x^{(i)}; \Theta), y^{(i)}) \right) + \lambda R(\Theta)$$

$\Theta$  is our "main" input variable, but data  $\mathcal{D}$  is important for computing  $J$ .

One more comment:

**Clarification 51**

Students often get confused by the fact that our **objective function**  $J$  is a function of  $\Theta$ , while **training error**  $\mathcal{E}_n$  is a function of  $h$ .

$$\overbrace{J(\Theta)}^{\text{Uses } \Theta} = \overbrace{\mathcal{E}_n(h)}^{\text{Uses } h} + \lambda R(\Theta)$$

The difference is that **training error**  $\mathcal{E}_n(h)$  is **more general** than the **objective function**  $J(\Theta)$ , and  **$h$**  is **more general** than  **$\Theta$** .

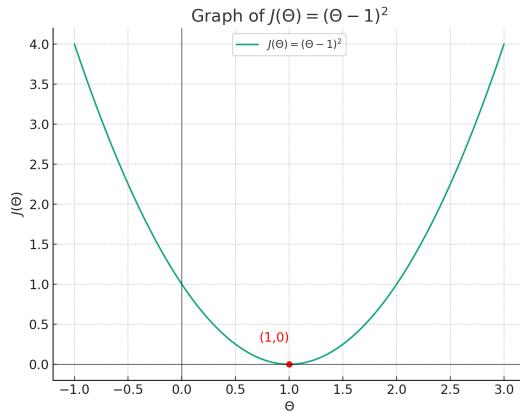
- By "more general", we mean that there are more situations where we can use  $h$  than  $\Theta$ .
  - $\Theta$  is a list of parameters: we only use it if we have a **parametric model**.
  - $h$  is used if we have **any kind of model**.

So, let's compare  $\mathcal{E}_n(h)$  and  $J(\Theta)$ :

- **Training error** can be used for any model  $h$ , so we don't want to use  $\Theta$ : our model could be **non-parametric**.
- Our **objective function assumes** we have parameters  $\Theta$ : we know our model class  $H$ , we just want to optimize  $\Theta$ .

## 2.2.4 Minimization Notation

Our goal is to **minimize**  $J$  by adjusting  $\Theta$ . To demonstrate, we'll use the following example:



Take  $J(\Theta) = (\Theta - 1)^2$ . The minimum output is 0, which happens at  $\Theta = 1$ . So, we have a minimum at  $(1, 0)$ .

Our goal is to find this **minima**. There are two questions we're interested in:

- What is the **minimum value of  $J$**  we can find by **adjusting  $\Theta$** ?
- Which **model  $\Theta$**  gives us the minimal  $J$ ? In other words: which model performs best?

We define a distinct function for answering each of these questions.

First:

- What is the **minimum value of  $J$**  we can find by **adjusting  $\Theta$** ?

### Notation 52

The **min function** gives you the **minimum output** of a function we get by adjusting one chosen **variable**.

$$\min_{\Theta} J(\Theta)$$

The **function we want to minimize** is written to the right, while the **variable we adjust** is written below.

**Example:**

$$\min_{\Theta} (\Theta - 1)^2 = 0 \quad (2.12)$$

0 is the minimum value of  $J$  we can find by adjusting  $\Theta$ .

Next:

- What is the **minimum value of  $J$**  we can find by **adjusting  $\Theta$** ?

#### Notation 53

The **argmin function** tells you the value of the **input variable** that gives the **minimum output**.

$$\arg \min_{\Theta} J(\Theta)$$

The **function we want to minimize** is written to the right, while the **variable we adjust** is written below.

**Example:**

$$\arg \min_{\Theta} (\Theta - 1)^2 = 1 \quad (2.13)$$

1 is the value of  $\Theta$  which gives the minimum  $J$ .

#### Clarification 54

Why is it called "**argmin**"?

"**Argument**" is used as another word for "**input variable**".

And our argmin function returns the **argument** with the **minimum** output. Hence, **arg min**.

## 2.2.5 Optimal Value Notation

Our goal is to find the best model, represented by some  $\Theta$ . We'll call this "optimal" model,  $\Theta^*$ .

#### Notation 55

We add a **star** \* to indicate the **optimal** variable choice.

If that variable is  $z^*$ , you would say it as "z-star".

**Example:**

$$\Theta^* = 1 \text{ for the above example.} \quad (2.14)$$

So, if we want optimal  $\Theta$ , we're looking for:

**Key Equation 56**

Our **optimal parameter** vector is written as

$$\Theta^* = \arg \min_{\Theta} J(\Theta)$$

## 2.3 Linear Regression

Now that we understand the problem of **regression**, and the concept of **optimizing** over it, we'll introduce our **hypothesis class**.

We want a function that can use information to **predict** outputs.

### 2.3.1 The Linear Model, 1-D

We'll start off small: we have **one variable**, and something we want to predict. And we'll pick the simplest pattern we can:

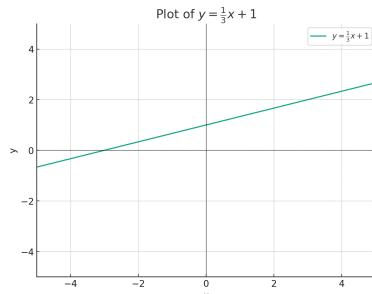
$$y = mx + b \quad (2.15)$$

A **linear** equation.

- $m$  tells us **how much** our input  $x$  **affects** our output  $y$ .
- $b$  accounts for everything **unrelated** to  $x$ : what is  $y$  when  $x = 0$ ?

$b$  and  $m$  are our **parameters**: that means they're part of  $\Theta$ . We'll rename them  $b = \theta_0$  and  $m = \theta_1$ .

$$h(x) = \theta_1 x + \theta_0 \quad (2.16)$$



Here's  $\frac{1}{3}x + 1$ . We call this 1D because there's only one input dimension,  $x$ . But we plot it in 2D to see the output, too!

### 2.3.2 The Linear Model, 2-D

We want to have **multiple** input variables:  $x$  will be a **vector**, not a number.

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (2.17)$$

So, for our above example, we'll replace  $x$  with  $x_1$ .

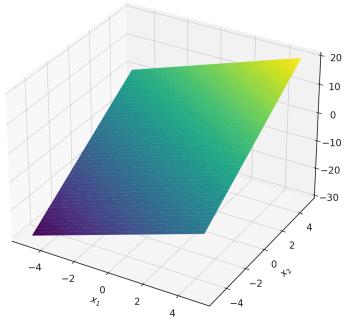
$$h(x) = \theta_1 x_1 + \theta_0 \quad (2.18)$$

The simplest way to include  $x_2$  by just **adding** it. We have a scaling factor  $\theta_1$  for  $x_1$ , so we'll give  $x_2$  its own **parameter**,  $\theta_2$ :

If  $\theta_1$  is the "slope" for  $x_1$ ,  $\theta_2$  is the "slope" for  $x_2$ .

$$h(x) = \theta_2 x_2 + \theta_1 x_1 + \theta_0 \quad (2.19)$$

3D plot of  $y = 2x_1 + 3x_2 - 5$



Here's a 2d regression: it's a plane. The height represents the output.

### 2.3.3 The Linear Model, d-D

You can **expand** this to  $d$  dimensions by **adding more terms**:

This is the "dimension" of our input space: the **number** of input variables we have.

$$h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \dots + \theta_d x_d \quad (2.20)$$

We need  $n + 1$  dimensions to plot an  $n$ -dim regression, so... we can't plot  $n > 2$ .

### 2.3.4 The Linear Model using Vectors

We **multiply** components of  $x$  and  $\theta$  together, then **add** them together. This looks like a **dot product**:

$$h(x) = \theta_0 + \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \quad (2.21)$$

If we write this symbolically, we get:

$$h(x) = \theta_0 + \theta \cdot x \quad (2.22)$$

$\theta$  includes all of our parameters, **except for**  $\theta_0$ .

- $\theta$  is used for our **dot product**,  $\Theta$  includes **all** parameters.

### Notation 57

We represent the **parameters** of our **linear** equation as  $\Theta = (\theta, \theta_0)$ .

This formula looks similar to  $y = mx + b$  again! Only this time, we have **vectors** instead.

We'll swap out the dot product for **matrix multiplication**.

### Key Equation 58

A dot product  $a \cdot b$  can be written as **matrix multiplication** instead:

$$a \cdot b = a^T b$$

In this class, we'll usually find it more useful to work with matrix multiplication.

### Definition 59

The **linear regression** hypothesis is  $h(x) = \theta \cdot x + \theta_0$ , or

$$h(x) = \theta^T x + \theta_0$$

Remember that, when written out, this looks like:

Make sure you know what  $\theta^T$  is: it's the **transpose** of  $\theta$ .

$$h(x) = [\theta_1 \ \theta_2 \ \theta_3 \ \dots \ \theta_d] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_d \end{bmatrix} + \theta_0 \quad (2.23)$$

This is the **hypothesis class** of **linear hypotheses** we will reuse throughout the class.

### 2.3.5 Regression Loss

We need to decide on our **loss function** for regression: how **badly** is our model is performing?

- Our goal is for our **guess**  $g$  to be close to the **real** output  $y$ .
- The more **different** they are, the worse.

We could use "absolute difference"  $|g - y|$ , but **squared difference** tends to be much more useful:

Why? We discuss in the Concept box below.

$$\mathcal{L}(g, y) = (g - y)^2 \quad (2.24)$$

We call this **square loss**. It punishes high and low guesses equally, and the punishments become more **severe** as the **difference** increases.

Our slope  $\frac{d}{dx}x^2 = 2x$  gets larger as we move away from  $x = 0$ .

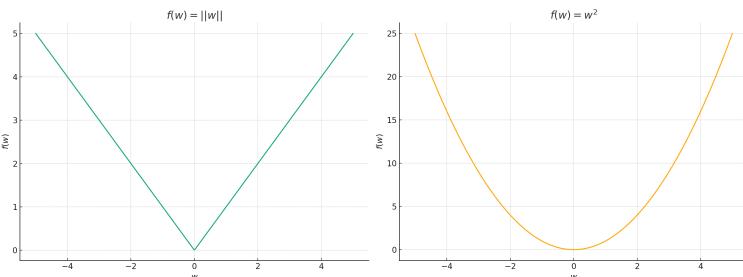
#### Concept 60

We use **square distance** for a few good reasons:

- High and low guesses are treated **equally**.
- It works well with **matrix multiplication**:

$$\|w\|^2 = w^T w$$

- $\|w\|$  is **not smooth**: it doesn't always have a derivative!  $\|w\|^2$  is smooth.
  - $\|w\|$  doesn't have a derivative at  $w = 0$ .
- The **slope** becomes **small** when you get closer to the correct answer: you'll know when you're getting close.



What's the derivative of  $\|w\|$  at 0? We don't have one!

A "stronger" version of smoothness requires that **every derivative** ( $f'$ ,  $f''$ ,  $f'''$ ...) is **continuous**.

We just care if the derivative exists everywhere, though.

### 2.3.6 Our Goal: Ordinary Least Squares

Now, we have the concepts we need.

- We want to **minimize** loss  $\mathcal{L}$  on our data set, using the **linear** model  $\Theta(h)$ .
- We'll use that linear model to **predict** the outputs of our data points.

This goal can be turned into an **objective function**:  $J(\Theta) = J(\theta, \theta_0)$

$$J(\theta, \theta_0) = \text{Training Loss} \quad (2.25)$$

Let's go step-by-step:

- Training loss is our **expected loss**, averaged over each data point.  $g^{(i)}$  is our prediction for  $y^{(i)}$ .

Remember that  $y^{(i)}$  is the "correct" answer for our  $i^{\text{th}}$  data point.

$$J(\theta, \theta_0) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(g^{(i)}, y^{(i)}) \quad (2.26)$$

- We'll use **squared loss** to evaluate each data point:

$$J(\theta, \theta_0) = \frac{1}{n} \sum_{i=1}^n (g^{(i)} - y^{(i)})^2 \quad (2.27)$$

- We use our **hypothesis**  $h(x^{(i)}; \Theta)$  to make our guess  $g^{(i)}$ .

$$J(\theta, \theta_0) = \frac{1}{n} \sum_{i=1}^n (h(x^{(i)}; \Theta) - y^{(i)})^2 \quad (2.28)$$

- Our hypothesis is a **linear model**  $\theta^T x^{(i)} + \theta_0$ .

#### Key Equation 61

The **ordinary least squares objective function** for **linear regression** is written as

$$J(\theta, \theta_0) = \frac{1}{n} \sum_{i=1}^n ((\theta^T x^{(i)} + \theta_0) - y^{(i)})^2$$

If we break this into parts:

$$J(\theta, \theta_0) = \underbrace{\frac{1}{n} \sum_{i=1}^n}_{\text{Averaging}} \left( \underbrace{(\theta^T x^{(i)} + \theta_0)}_{\text{guess}} - \underbrace{y^{(i)}}_{\text{answer}} \right)^2 \quad (2.29)$$

Now, this is an **optimization** problem. We need to find the model  $(\theta, \theta_0)$ , that gives us the best (minimal)  $J$ .

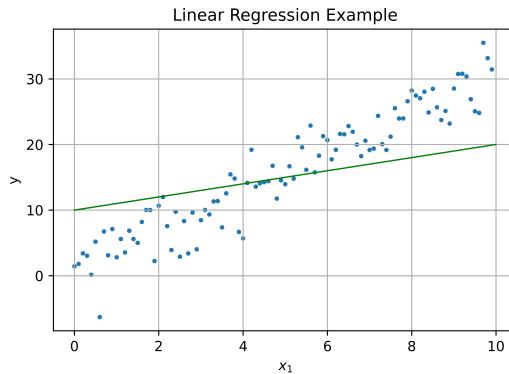
$$\theta^*, \theta_0^* = \arg \min_{\theta, \theta_0} J(\theta, \theta_0)$$

We now have two parameters in our argmin function, but aside from listing both of them, the notation is the same. We just substituted  $\Theta = (\theta, \theta_0)$

### 2.3.7 Visualizing our Model

We'll start with the **one-variable** case. With one input, one output, we use a 2D plot to graph our data.

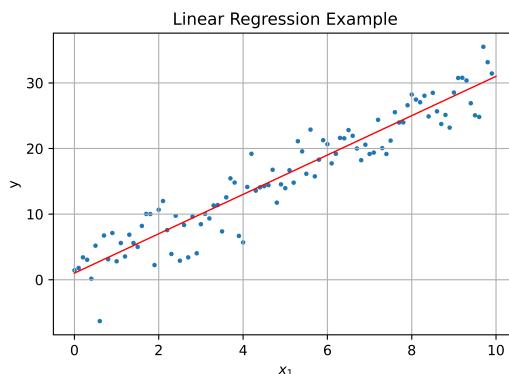
- Each piece of data is a simple  $(x, y)$  pair: a **point**.
- Meanwhile, our hypothesis is a **line**: for each  $x$ , it predicts a different  $y$ .



This linear model doesn't fit our data very well:  $(\theta_0 = 10, \theta_1 = 1)$

We're trying to get our line as **close as possible** to the points, hoping to find a linear pattern.

- We call this "**fitting**" our line to the data.



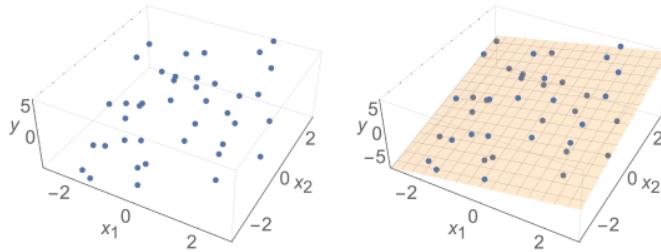
This line is much better fitted to the data:  $(\theta_0 = 1, \theta_1 = 3)$

How do we know our line doesn't fit our data? Because it isn't "close" to the shape of the data.

We want our model to really represent the data it comes from: they should look similar.

What does this look like if we have **two variables**? You need a 3D space, with 2 dimensions for the input.

- Each piece of data is a **point**  $(x_1, x_2, y)$ .
- Our hypothesis is a **plane**: for each pair  $(x_1, x_2)$ , it predicts a different  $y$ .



This plane is **fitted** the same way our line was. Notice that  $y$  is our **height**: this is the **output** of our regression.

Earlier, we mentioned that we can't really **visualize** higher dimensions. \_\_\_\_\_

Looking at a 4D "plane" would be a headache.

- So, instead, we don't even try to. We'll think of them in terms of our math. When we need intuition, we'll rely on the 2D plane.
- Because they're a higher-dimensional version of a **plane**, we call it a **hyperplane**.

### Definition 62

A **hyperplane** is a **higher-dimensional version** of a **plane** - a **flat** surface that continues on forever.

We use it to represent our **linear** hypothesis for the purpose of **regression**.

- We have  $d$  dimensions ( $d$  variables) in our input.
- To represent our output, we need one additional,  $(d + 1)^{\text{th}}$  dimension.

Visually, the "**height**" of our plane represents the **output** of  $h(x)$ .

- Our line was a **1-D** object in a **2-D** plane.
- Our plane was a **2-D** object in a **3-D** space.

So, our **hyperplane** is a  $d$  dimensional object in a  $d + 1$  dimensional space.

- Our goal is the same: we want our **hyperplane** to be as **close** to all of our data points as it possibly can.

### 2.3.8 Another Interpretation

So far, we've generally interpreted our model similarly to  $mx + b$ .

$$h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \dots + \theta_d x_d \quad (2.30)$$

- Using our  $mx + b$  analogy, we can see  $\theta_k$  as the "slope" of  $x_k$ .

$$\frac{\partial h}{\partial x_k} = \theta_k \quad (2.31)$$

- In other words,  $\theta_k$  tells us how much  $x_k$  affects the **output**.

#### Concept 63

The **larger**  $\|\theta_k\|$  is, the **more important**  $x_k$  is to our output.

- If we **increase**  $\|\theta_k\|$ , then  $x_k$  will have a **bigger effect** on  $h(x)$ .

$$\underbrace{\frac{\partial h}{\partial x_k}}_{\text{Effect of } x_k \text{ on } h} = \theta_k$$

This is a simple **pattern**: "as we change  $x_k$ , we change  $h(x)$ ".

We can also **compare** each  $\theta_k$  term to each other.

- If  $\theta_2$  is larger than  $\theta_1$ , then increasing  $x_2$  would affect the output **more** than increasing  $x_1$ .

$$h(x) = 2x_1 + \underbrace{10000x_2}_{x_2 \text{ has greater effect}} \quad (2.32)$$

- We could say that  $x_2$  has a stronger effect on the output than  $x_1$ : it **weighs more heavily** in the calculation.

Because of this, we sometimes call  $\theta_k$  the **weight** for  $x_k$ .

#### Definition 64

A **weight** is a **parameter** that tells us how **strongly** a variable influences our **output**.

It is usually a **scalar** that we **multiply** by our variable.

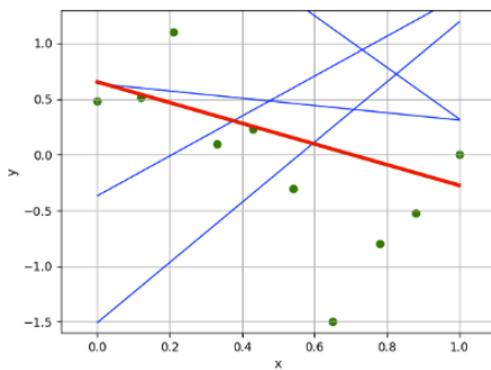
- We say that  $\theta_0$  is an "offset", and the other  $\theta_i$  terms are "weights".

## 2.4 The stupidest possible linear regression algorithm

So, now we want to try to **optimize**  $J$  based on  $\theta$  and  $\theta_0$ . How do we do that? Let's start as **simple** as we possibly can.

We can't try **every** possible  $\Theta$ , because there are an **infinite** number of them. Rather than thinking too hard about a possible pattern, or an **algorithm**, let's just **randomly** try options.

We'll try **random** values for  $\theta$  and  $\theta_0$ , and **pick** whichever option gives us the **best result**. Seems simple, if inefficient.



Each line (hypothesis) was randomly generated. We focus on the red one: this is the best model, out of all the ones that we tried.

Why introduce such a silly algorithm? For a few reasons:

- It gives us an **example** of an optimization algorithm that's very **simple**.
- **Randomly** generated results create a good **baseline** - more intelligent algorithms can be compared to this one, to see how well we're doing.

Sometimes, you might come up with a clever technique, only to find out it isn't better than a random model! It happens more than you'd think.

## 2.5 Analytical solution: ordinary least squares

We can do **better** than randomly **generate** parameters, though. In fact, in this rare case, we can actually **solve** for optimal parameters!

### 2.5.1 Trying to Simplify

Our solution will involve a lot of algebra. Because of that, it's worth it to **simplify** our formula as much as possible beforehand.

$$J(\theta, \theta_0) = \frac{1}{n} \sum_{i=1}^n ((\theta^T x^{(i)} + \theta_0) - y^{(i)})^2 \quad (2.33)$$

By algebra, we just mean "shuffling around math symbols in an equation".

Most parts of this equation can't really be **simplified**:  $y$  and  $x$  are just variables, and we can't do anything with the **sum** without knowing our data points.

- But, there's one notable detail: we **separated**  $\theta_0$  from our other  $\theta_k$  terms.
- If we can **include**  $\theta_0$  in the dot product, our math will be easier.

### 2.5.2 Combining $\theta$ and $\theta_0$

Let's go back to our **original** equation for  $(\theta^T x + \theta_0)$ , before we switched to **vectors**.

$$h(x) = \theta_0 + \theta \cdot x = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \dots + \theta_d x_d \quad (2.34)$$

We drop the <sup>(i)</sup> notation whenever it isn't necessary, to de-clutter the equations.

Let's just say we've picked one random data point.

- We simplified our notation with a **dot product**: each  $\theta_k$  term is **multiplied** by an  $x_k$  term.
- This is, of course, **excluding**  $\theta_0$ .
  - We would end up with a simpler result if we could include  $\theta_0$  in the  $\theta$  vector. But,  $\theta_0$  would need to be **multiplied by an  $x_0$  term**.

$$h(x) = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{bmatrix} \cdot \underbrace{\begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_d \end{bmatrix}}_{\text{We need } x_0} \quad (2.35)$$

We have a trick: let's factor out  $x_0 = 1$ .

You can always factor out 1 without changing the value!

$$h(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \dots + \theta_d x_d \quad (2.36)$$

So, this means we just have to **append** a 1 to our vector  $x$ . At the **same time**, we'll append  $\theta_0$  to  $\theta$ !

$$x = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{bmatrix}, \quad h(x) = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{bmatrix} \cdot \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \quad (2.37)$$

We'll write that symbolically, and then apply a transpose.

$$h(x) = \theta \cdot x = \theta^T x \quad (2.38)$$

### Concept 65

Sometimes, to simplify our algebra, we can **append**  $\theta_0$  to  $\theta$ .

To make this possible, we **choose**  $x_0 = 1$ .

- This requires **appending** a value of 1 to  $x$ .

Once we do this, we can **write**

$$h(x) = \theta^T x$$

We **have** to append this 1 to every single  $x^{(i)}$  in order for this to **work**. But, now we can treat our **parameters** as **one vector**.

### 2.5.3 Combining data points

There's another place we can clean things up:

- Currently, when using our **objective** function, we have to **sum** over **every** single data point.

Note that, for convenience, we've included  $\theta_0$  in  $\theta$ .

$$J(\theta) = \frac{1}{n} \sum_{i=1}^n (\theta^T x^{(i)} - y^{(i)})^2 \quad (2.39)$$

- This is a bit of a hassle - we have to consider every  $(x^{(i)}, y^{(i)})$  term separately.
- Is there a better way?

We've solved this kind of problem before: using vectors above, we were able to work with **many parameters**  $\theta_k$  and **many variables**  $x_k$  at the same time.

$$h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \dots + \theta_d x_d \xrightarrow{\text{vector form}} h(x) = \theta^T x$$

- This made it easier to do lots of math quickly.

#### Concept 66

One of the biggest benefits of **matrices** is being able to do **lots of math at the same time**.

- In particular, **matrix multiplication** allows you to do **addition** and **multiplication** on as many elements as you want.

Can we do the **same** here - combining **many data points** into one object?

### 2.5.4 Many data points in a matrix

We want to combine all of our data points into a single matrix.

- We're already using **rows** to represent **multiple dimensions** of a data point.

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \quad k^{\text{th}} \text{ dimension in row } k \quad (2.40)$$

A reminder: a "dimension"(feature) is one aspect of our input data. For an animal, it might be height, weight, age, etc.

Each dimension stores one piece of information.

- We'll need different notation to separate **data points**: we'll use **columns**.

$$X_{1D} = \underbrace{\begin{bmatrix} x^{(1)} & x^{(2)} & x^{(3)} & \dots & x^{(n)} \end{bmatrix}}_{i^{\text{th}} \text{ data point in column } i} \quad (2.41)$$

**Definition 67**

We use **rows** to indicate the different **dimensions**  $x_k$  of a single data point, and **columns** to indicate each **data point**  $x^{(i)}$ .

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \quad X_{1D} = \begin{bmatrix} x^{(1)} & x^{(2)} & \dots & x^{(n)} \end{bmatrix}$$

- Note that the capitalized X matrix is used for **all of our data points**  $x^{(i)}$ .

These formats are both useful, but limited:

- $x$  can handle **many dimensions**, but represents **one data point**.
- $X_{1D}$  represents **many data points**, but with only **one dimension** for each.

Our solution? Combine them into a single object:

**Key Equation 68**

$X$  is our **input matrix** in the shape  $(d \times n)$  contains information **n data points** with **d dimensions each**.

$$X = \left\{ \begin{bmatrix} x_1^{(1)} & \dots & x_1^{(n)} \\ \vdots & \ddots & \vdots \\ x_d^{(1)} & \dots & x_d^{(n)} \end{bmatrix} \right\} \text{d dimensions}$$

**Example:** Consider 3 data points in 2 dimensions:  $[1, 2]^T, [9, 5]^T, [10, 11]^T$ .

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 9 \\ 5 \end{bmatrix}, \begin{bmatrix} 10 \\ 11 \end{bmatrix} \rightarrow X = \begin{bmatrix} 1 & 9 & 10 \\ 2 & 5 & 11 \end{bmatrix} \quad (2.42)$$

If we want to replace  $\theta^T x + \theta_0$  with  $\theta^T x$ , we can include 1's at the top: \_\_\_\_\_

Or the bottom, if we want.

$$X = \underbrace{\begin{bmatrix} 1 & \cdots & 1 \\ x_1^{(1)} & \cdots & x_1^{(n)} \\ \vdots & \ddots & \vdots \\ x_d^{(1)} & \cdots & x_d^{(n)} \end{bmatrix}}_{\text{n data points}}}_{\text{d + 1 dimensions}} \quad (2.43)$$

We can do the same for Y: combine all of the data points into one matrix.

### Key Equation 69

Y is our **output matrix** in the shape  $(1 \times n)$  that contains all data points.

$$Y = [y^{(1)} \ \dots \ y^{(n)}]$$

Why is this a row vector, not a matrix?

This is a **regression** problem: each output is a scalar, not a vector!

## 2.5.5 Objective Function in matrix form

Now that we can use all of our **data points** at the same time, we can condense our objective function.

$$J(\Theta) = \frac{1}{n} \sum_{i=1}^n (\theta^T x^{(i)} - y^{(i)})^2 \quad (2.44)$$

We can simultaneously compute  $(\theta^T x^{(i)} - y^{(i)})$  for all of our data points at the same time:

$$\theta^T X - Y = \begin{bmatrix} \theta^T x^{(1)} - y^{(1)} & \theta^T x^{(2)} - y^{(2)} & \dots & \theta^T x^{(n)} - y^{(n)} \end{bmatrix}$$

This isn't exactly what we want, though: we want  $(\theta^T x^{(i)} - y^{(i)})^2$ : multiplied by itself. We can do this with a **dot product**:

$$(\theta^T x^{(i)} - y^{(i)})^2 \rightarrow (\theta^T X - Y) \cdot (\theta^T X - Y) \quad (2.45)$$

How do we write this dot product with matrix multiplication?

### Clarification 70

When  $a$  and  $b$  were **column vectors**, we could take their dot product as  $a^T b$ .

$$a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \implies a \cdot b = a^T b$$

But we have to do the opposite for **row vectors**  $p$  and  $q$ : we get  $p q^T$ .

$$p = \begin{bmatrix} p_1 & p_2 & \cdots & p_m \end{bmatrix} \implies p \cdot q = p q^T$$

$$q = \begin{bmatrix} q_1 & q_2 & \cdots & q_m \end{bmatrix}$$

We could show that this matrix multiplication gives the results we want, with calculation.

But this is a little tedious, and doesn't teach us much, so we recommend trying it yourself if you're unconvinced.

Because  $(\theta^T X - Y)$  is a **row vector**, we'll have to write it in the  $p q^T$  format:

$$(\theta^T x^{(i)} - y^{(i)})^2 \rightarrow (\theta^T X - Y)(\theta^T X - Y)^T \quad (2.46)$$

This formula represents our original objective function:

Here, we're comparing  $a^T b$  to  $ab^T$ .

$$\frac{1}{n} (\theta^T X - Y) (\theta^T X - Y)^T = \frac{1}{n} \sum_{i=1}^n (\theta^T x^{(i)} - y^{(i)})^2 \quad (2.47)$$

**Key Equation 71**

Using  $X$ ,  $Y$ , and  $\theta$  can write our **objective function** for **multiple** variables and **multiple** data points as

$$J(\theta) = \frac{1}{n} (\theta^T X - Y) (\theta^T X - Y)^T$$

Because  $n$  is a constant, we sometimes ignore it when we're minimizing  $J(\Theta)$ .

It is important to **remember** the **shape** of our objects, as well.

**Concept 72**

Our matrices have the shapes:

- $X$ :  $(d \times n)$  - matrix
- $Y$ :  $(1 \times n)$  - row vector
- $\theta$ :  $(d \times 1)$  - column vector
- $\theta_0$ :  $(1 \times 1)$  - scalar
- $J$ :  $(1 \times 1)$  - scalar

If we combine  $\theta_0$  into  $\theta$ , replace every use of  $d$  with  $d + 1$ .

These shapes are worth **memorizing**.

Notice that these shapes make sense for our above equation! Try working through the matrix multiplication to verify this.

## 2.5.6 Alterate Notation

One side problem: some ML texts use the **transpose** of  $X$  and  $Y$ .

**Notation 73**

Some subjects use **different notation** for **matrices**. The main difference is that  $X$  and  $Y$  use their **transpose**, which we'll notate as

$$\tilde{X} = X^T \quad \tilde{Y} = Y^T$$

Thus, our equation above becomes

$$J = \frac{1}{n} (\tilde{\mathbf{x}}\theta - \tilde{Y})^T (\tilde{\mathbf{x}}\theta - \tilde{Y})$$

## 2.5.7 Optimization in 1-D - Using Calculus

Now that we've sorted our data, we can start **optimizing**  $\theta$ . Our goal is to modify  $\theta$ , to find the minimal  $J$ .

$$J(\theta) = \frac{1}{n} \sum_{i=1}^n (\theta^T x^{(i)} - y^{(i)})^2 \quad (2.48)$$

We're gonna focus on one data point/dimension at a time, so we don't need our matrix notation.

We'll start with a simplified case, and build our way up:

- First, we limit our attention to one data point ( $n = 1$ ):

$$J(\theta) = (\theta^T x - y)^2 \quad (2.49)$$

- And we'll assume  $\theta$  and  $x$  are one-dimensional ( $d = 1$ ).

$$J(\theta) = (\theta x - y)^2 \quad (2.50)$$

If we use  $\theta$  as a **variable**, this is an ordinary single-variable function! How would we find the **minimum**?

- Using **calculus!** Anywhere there's a local **minimum**, we typically know the **derivative is 0**.

Assuming a "smooth" surface...

### Concept 74

If our function  $J(\theta)$  has **one variable**, we find possible **local minima**  $\theta$  wherever the **derivative**  $\partial J / \partial \theta$  is zero.

$$\frac{\partial J}{\partial \theta} = 0$$

- To make sure it's a minimum (not a maximum), we also need to make sure that the **second derivative** is positive ( $J''(\theta) > 0$ ).
- We already know this is true for **squared loss**, so we won't bother with this step.

- Note that we're taking  $\frac{\partial}{\partial \theta}$ , **not**  $\frac{\partial}{\partial x}$ .
- This is because our goal is to modify  $\theta$  (our model), not  $x$  (our data).

Let's do this for our simple example:

Why do we need  $\partial J / \partial \theta = 0$ ?

If  $\partial J / \partial \theta > 0$ , then decreasing  $\theta$  slightly would reduce  $J$ : there's a lower point nearby, so this isn't a minimum!

If  $\partial J / \partial \theta < 0$ , increasing  $\theta$  has the same effect.

Last Updated: 09/03/24 03:53:41

We want to train our model to match our data, not the other way around!

$$J'(\theta) = 2x(\theta x - y) = 0 \quad (2.51)$$

We just find where the slope  $J'(\theta)$  is 0, and solve for  $\theta$ !

Because this is the optimal  $\theta$ , we call it  $\theta^*$ .

$$\theta^* = \frac{y}{x} \quad (2.52)$$

## 2.5.8 Optimizing for multiple variables

This time, we'll do **one data point**, having **d dimensions**.

- This gets a bit tricky, because we have to do our math with **vectors**.

Because we only have one data point, we'll omit the <sup>(i)</sup> notation.

$$J(\theta) = (\theta^T x - y)^2 \quad (2.53)$$

We'll **optimize** this. In the **one-dimensional** case, we wanted to set the **derivative** of  $J$  to **zero**, using a single  $\theta$  variable.

- Now, we have **multiple variables**  $\theta_k$  that we could optimize.
- How do we know when we've optimized all of them?

Well, if we consider each dimension separately,  $\theta_k$  would be optimized if

$$\frac{\partial J}{\partial \theta_k} = 0 \quad (2.54)$$

So, maybe it would be reasonable to just set **every** derivative to **zero**?

- It turns out, the answer is **yes**!

If every individual derivative  $\partial J / \partial \theta_k$  is zero, we have a potential minimum.

We can use the reasoning that we used for 1D:

If one of our derivatives  $\partial J / \partial \theta_k > 0$ , then we could decrease  $\theta_k$  to find a nearby point which is "lower" than our current point: we don't have a minimum.

Thus, every derivative must be 0.

**Concept 75**

If our function  $J(\theta)$  has **d+1 parameters**, we find possible **local minimum**  $\theta$  anywhere that obeys the system of equations

$$\frac{\partial J}{\partial \theta_0} = 0 \quad \frac{\partial J}{\partial \theta_1} = 0 \quad \frac{\partial J}{\partial \theta_2} = 0 \quad \dots \quad \frac{\partial J}{\partial \theta_d} = 0$$

Or in general,

$$\frac{\partial J}{\partial \theta_k} = 0 \quad \text{for all } k \text{ in } \{0, 1, 2, \dots, d\}$$

- Why  $d+1$  parameters instead of  $d$ ? Because we're including  $\theta_0$  as one additional parameter.

The **solution** to this system of equations will be our **desired list of parameters**,  $\theta^*$ .

Again, we ignore the second requirement of making sure this isn't a **maximum or saddle point**.

$$\theta^* = \begin{bmatrix} \theta_1^* \\ \theta_2^* \\ \vdots \\ \theta_d^* \end{bmatrix} \quad (2.55)$$

## 2.5.9 Gradient Notation

Above, we wrote our derivative **element-wise**: each derivative got its own equation.

- We can make things easier by storing our derivatives in a **column vector**, just like how  $\theta$  stores  $\theta_k$  terms.

$$\frac{\partial J}{\partial \theta} = \begin{bmatrix} \partial J / \partial \theta_0 \\ \partial J / \partial \theta_1 \\ \vdots \\ \partial J / \partial \theta_d \end{bmatrix} \quad (2.56)$$

How do we know that this is a column vector, and not a row vector?

Check out the matrix derivative notes for a complete explanation.

We'll think of this as a bigger, **multivariable** derivative, called the **gradient**.

We'll give more conceptual intuition for the gradient in the next chapter. Look forward to it!

**Key Equation 76**

The **gradient** of  $J$  with respect to  $\theta$  is

$$\nabla_{\theta} J = \frac{\partial J}{\partial \theta} = \begin{bmatrix} \partial J / \partial \theta_0 \\ \partial J / \partial \theta_1 \\ \vdots \\ \partial J / \partial \theta_d \end{bmatrix}$$

It has the same dimensions ( $d \times 1$ ) as  $\theta$ .

Now, we can rewrite our previous rule, "every derivative is 0":

$$\nabla_{\theta} J = \begin{bmatrix} \partial J / \partial \theta_0 \\ \vdots \\ \partial J / \partial \theta_d \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} = \vec{0} \quad (2.57)$$

**Concept 77**

If  $\theta$  is a **vector**, we can find possible **local minima**  $\theta$  of  $J(\theta)$  anywhere that the **gradient**  $\nabla_{\theta} J$  is zero.

$$\nabla_{\theta} J = \vec{0}$$

This is the general equation we **solve** to find  $\theta^*$ .

Why is  $\partial J / \partial \theta$  a  $(d \times 1)$  matrix instead of  $(1 \times d)$ ?

We don't have a special reason: it's a convention we've picked, that works well with the rest of our math.

In fact, some other texts might do differently. To learn more about our rules, check the Matrix Derivatives notes.

Once again: we should check if it's a minimum. But we continue to ignore this caveat.

### 2.5.10 Matrix Calculus

Now, we return to the general case: **n data points**, each having **d dimensions**.

$$J(\theta) = \frac{1}{n} (\theta^T X - Y) (\theta^T X - Y)^T$$

Using the "alternate" notation":

$$J = \frac{1}{n} (\tilde{X}\theta - \tilde{Y})^T (\tilde{X}\theta - \tilde{Y}) \quad (2.58)$$

We will **not** show how to take this matrix derivative. But our result is:

Want to know how?  
Check out A.4 in the official appendix.

$$\nabla_{\theta} J = \frac{2}{n} \tilde{X}^T (\tilde{X}\theta - \tilde{Y}) = 0 \quad (2.59)$$

**Clarification 78**

Note that matrix derivatives often look **similar** to traditional, single-variable derivatives. However, they are **not the same**.

- Often, this can result in **shape errors**: we end up with the wrong matrix shape.

From here, we just solve for  $\theta$ , using matrix multiplication rules.

How do we take a matrix derivative in general? Check out an explanation in the Matrix Derivatives chapter.

**Key Equation 79**

The **solution** for **OLS optimization** is

$$\theta = \underbrace{(\tilde{X}^T \tilde{X})^{-1}}_{d \times d} \underbrace{\tilde{X}^T}_{d \times n} \underbrace{\tilde{Y}}_{n \times 1}$$

Or, in our **original** notation,

$$\theta = \underbrace{(XX^T)^{-1}}_{d \times d} \underbrace{X}_{d \times n} \underbrace{Y^T}_{n \times 1}$$

- If  $\theta_0$  is included in  $\theta$ , then dimension  $d$  is replaced with  $d + 1$ .

We've finished with "ordinary least squares".

- We call it "ordinary" because this is the **simple** version of the problem.
- We'll make it more complex by introducing **regularization**.

Note that this requires that  $XX^T$  is invertible: we need to compute that inverse  $(XX^T)^{-1}$  in the equation, after all.

This technique doesn't work if that's not the case. But, we'll introduce a different technique to solve this problem: regularization!

## 2.6 Regularization

The above solution gives us the **best** model for matching our **training data**.

- But earlier, we mentioned that we want our model to be able to **generalize** to new data.

We need some math to represent this goal of "generality".

Because our training data doesn't perfectly reflect all of our future data.

- This equation won't measure our performance on training data, but instead encourages us to do well on **future data**.

We call this type of function a **regularizer**.

### Definition 80

A **regularizer** is a term to our **objective function** that helps measure how **general** our hypothesis is.

- By **optimizing** this term, we hope to create a model that works better with **new data** we didn't train with.

This function takes in our **vector of parameters**  $\Theta$  as an input:  $R(\Theta)$ .

But how do we make a model "more general?"

- First, we need to **understand** the problem: **what's wrong** with only using our training data?

### 2.6.1 Coincidences, and fake patterns

The goal of our model  $\theta$  is to look for **patterns**. This is a concept we discussed earlier (section 2.3.8):

**Concept 81**

(Review)

The **larger**  $\|\theta_k\|$  is, the **more important**  $x_k$  is to our output.

- If we **increase**  $\|\theta_k\|$ , then  $x_k$  will have a **bigger effect** on  $h(x)$ .

$$\underbrace{\frac{\partial h}{\partial x_k}}_{\text{Effect of } x_k \text{ on } h} = \theta_k$$

This is a simple **pattern**: "as we change  $x_k$ , we change  $h(x)$  by this much".

$\theta$  is trying to identify **which variables** have an effect on our output, and by how much.

- If  $\|\theta_k\|$  is large, our models thinks that  $x_k$  is important to our output.

Our goal is to modify each  $\theta_k$  term until it matches the real distribution.

$\theta_k < 0$  doesn't mean that  $x_k$  doesn't matter: it just means that it affects  $h(x)$  by decreasing it, instead of increasing it.

But seeing a pattern ( $x_k$  affecting  $h(x)$ ) doesn't mean that it's real: **random chance** can cause us to see patterns that don't really exist.

- **Example:** You take 20 quizzes during a semester. Every time you did **well**, you happened to be wearing a **red shirt**.
- You might come to believe that red shirts help you study **better**, even though it was just luck.

Isn't this kind of coincidence a bit unlikely? Maybe.

- But what happens if we have 10 possible coincidences? It's less likely that we avoid all of them.
- In other words: the more **opportunities** there are for something rare to happen, the **more likely** it is to happen.

Suppose the chance of one coincidence is  $p$ .

The chance of **one** coincidence not occurring is  $(1 - p)$ . The chance of **ten** coincidences not occurring is  $(1 - p)^{10}$ .

As we get more chances, we're more likely to see a coincidence.

**Concept 82**

The more **patterns** we're looking for, the more likely that **at least one** of the them shows up by **coincidence**.

- Our data can, by chance, match a pattern that **isn't even real**.

For some entertaining examples, search the phrase "spurious correlation".

This really becomes a problem for our model: often, we **don't know** what data matters, so we include **everything** we possibly can.

- But the more data we include, the more likely that something looks **important** on **accident!**

This is an interesting situation, where **learning more** about our training data can cause our model to perform **worse**: we're **overfitting**.

### Definition 83

(Review from Introduction chapter)

Learning more about our training data isn't necessarily the same as learning more about the true distribution that data came from!

**Overfitting** occurs whenever we **learn** ("fit") our training data too exactly, and it causes problems when we see **new data**.

- Often, we **memorize** very specific patterns, that don't actually hold up in general: they only appeared in our randomly sampled training data.

**Example:** You flip a coin 10 times, and it comes up heads 8 times.

- You've decided that the coin has an 80% chance of coming up heads.
- But it's a fair coin: the 'training data' (10 coin flips) doesn't match the 'true distribution' (50% chance of heads).

Now, we understand our problem. Let's come up with a solution.

## 2.6.2 Ridge Regression

We're worried about our model **finding false patterns** based on weak evidence.

- If our model finds a **pattern**, then it'll increase  $\|\theta_k\|$ : it considers  $x_k$  to be important for predicting  $h(x)$ , because the data coincidentally makes it **look** important.

If our model is "too eager" to find patterns, we can make it **more skeptical** of patterns it might see.

- In other words, we'll **punish** our model for increasing  $\|\theta_k\|$  too easily.

It still need to look for *some* (hopefully real) patterns, so that it can make good predictions  $h(x)$ .

We'll actually use  $\theta_k^2$ , for the same reasons we use squared loss:

Smoother, works well for positive and negative  $\theta_k$ , etc.

**Concept 84**

We want to **discourage** our model from prematurely deciding that  $x_k$  has a **effect** on  $h(x)$ .

- Thus, we'll **discourage** our model from increasing  $\|\theta_k\|$ :  $\theta_k$  represents this "**effect**".

$$R(\theta_k) = \theta_k^2$$

- Our algorithm will **minimize**  $R(\Theta)$ , so we'll make this  $\theta_k^2$  **small**.

We can repeat this process for all of our  $\theta_k$  terms, and combine them into a vector  $\theta$ :

$$R(\Theta) = \sum_k \theta_k^2 = \|\theta\|^2 = \theta \cdot \theta$$

This is our **ridge regression** model.

Why "ridge regression"?  
We'll get into that later.

**Key Equation 85**

Our **regularizer for regression** will be given by **square magnitude** of  $\theta$ :

$$R(\Theta) = \|\theta\|^2 = \theta^\top \theta$$

- In this model, we are biasing  $\|\theta\|$  towards 0: the farther from 0 we are, the more we punish the model.

This approach is called **Ridge Regression**.

### 2.6.3 $\lambda$ , our regularization constant

We can now create an "**objective function**" that includes training loss, **and** regression.

$$J(\theta) = \frac{1}{n} \sum_{i=1}^n \left( \underbrace{(\theta^\top x^{(i)} + \theta_0)}_{\text{guess}} - \underbrace{y^{(i)}}_{\text{answer}} \right)^2 + \underbrace{\|\theta\|^2}_{\text{Regularizer}} \quad (2.60)$$

Notice that these terms "compete" with each other:

**Concept 86**

Our **training loss** and **regularizer** compete with one another, affecting our model in opposing ways:

- Our regularizer wants us keep  $\theta$  small, being more **suspicious** of predicting patterns.
- But in order to make good predictions for our **training data**, we need to find the **real** patterns: some  $\theta$  values need to be bigger, to predict our data.

We need a **balance** between training loss and regularization. It would be useful to have a way to **control** that balance.

- That way, we can decide, "how much do we care about matching training data, versus avoiding coincidental patterns?"

This is why we don't end up with model  $\theta = \vec{0}$ : while this model would have low **regularization**  $R(\Theta)$ , it'll have high **training loss**.

We have a tool for this: we'll represent "how much we care about **regularization**" with a **constant  $\lambda$** .

**Definition 87**

**Lambda**, or  $\lambda$ , is the constant ( $\lambda \geq 0$ ) we **scale** our **regularizer** by.

$$\text{Total Regularization} = \lambda R(\Theta) = \lambda \|\theta\|^2$$

It represents **how strongly** we want to regularize: the larger it is, the more strongly we try to **generalize** our model.

Why is  $\lambda \geq 0$ ?

**Clarification 88**

We keep  $\lambda \geq 0$ , because  $\lambda < 0$  would encourage our model to make  $\|\theta\|$  **bigger**, no matter what.

- There's a limit to how small  $|\theta|$  can be, but there's **no limit** on how big it can be: it would just keep getting bigger forever.

Finally, we have our **completed** objective function:

**Key Equation 89**

The **objective function** for **ridge regression** is given as

$$J(\Theta) = \frac{1}{n} \sum_{i=1}^n \left( \underbrace{(\theta^T x^{(i)} + \theta_0)}_{\text{guess}} - \underbrace{y^{(i)}}_{\text{answer}} \right)^2 + \underbrace{\lambda \|\theta\|^2}_{\text{Regularizer}}$$

Once again, we note the contrast between "training loss" and "regularization".

- We've only made one change:  $\lambda$  can increase or decrease our focus on regularization.

This  $\lambda$  is crucial to our model: the larger it is, the more we punish our model for increasing  $\|\theta\|$ .

Readers might catch that our regularizer doesn't include  $\theta_0$ . There's a good reason for that! We'll discuss it below.

**Concept 90**

The more **regularization** (large  $\lambda$ ) we have, the more we're **focused** on keeping  $|\theta|$  small.

- If we make  $\lambda$  **too big**, then  $\|\theta\|$  becomes **very small**: our model doesn't learn enough information.
- If we make  $\lambda$  **too small**, then  $\|\theta\|$  becomes **very big**: our model learns **all** the patterns of our data, even the ones that come from random noise.

## 2.6.4 Why not regularize $\theta_0$ ?

Note that when we regularize with  $\lambda \|\theta\|^2$ , we're **not including**  $\theta_0$  in our vector:

$$R(\Theta) = \lambda \theta^T \theta \quad \theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{bmatrix} \quad (2.61)$$

This suggest that we're **not regularizing**  $\theta_0$ . Why is that?

### Concept 91

We **do not regularize**  $\theta_0$ .

- We **allow**  $\theta_0$  to take whatever value fits best.

The first reason:  $\theta_0$  works **differently** from our other  $\theta_k$  terms.

- A term  $\theta_k$  tell us how **important** one variable  $x_k$  is to our output  $h(x)$ .
- $\theta_0$  works almost the opposite way: it **ignores** our input  $x$ , and gives a baseline output.
  - In other words: if we **remove** the effect of all of our variables ( $x = \vec{0}$ ), what do we expect to see?

Naturally, regularization has very different effects:

- If you regularize  $\|\theta_k\|$ , your model will emphasize the **effect** of  $x_k$  by a smaller amount.
- If you regularize  $\|\theta_0\|$ , you've just **shifted** every output, by the same amount.

If our input data is centered on  $x = \vec{0}$  (equally above and below on each axis),  $\theta_0$  is the **average** output we expect to see.

### Concept 92

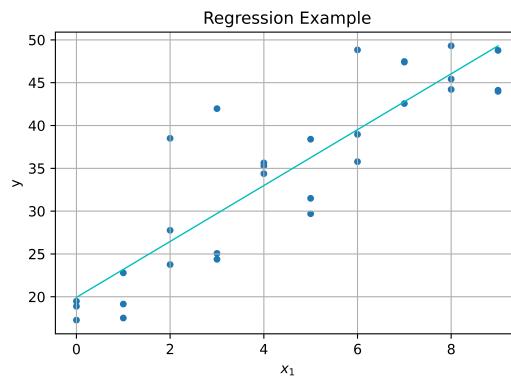
Unlike  $\theta_k$ ,  $\theta_0$  doesn't find **patterns** in the data: it just tells us roughly how **large** our output is.

- In fact,  $\|\theta_0\|$  affects all of our data in the **exact same way**.
- Decreasing  $\|\theta_0\|$  would shift all outputs equally: this doesn't do anything to make it more **accurate** for future data.

Because of this difference, it's not necessary to **regularize**  $\theta_0$ .

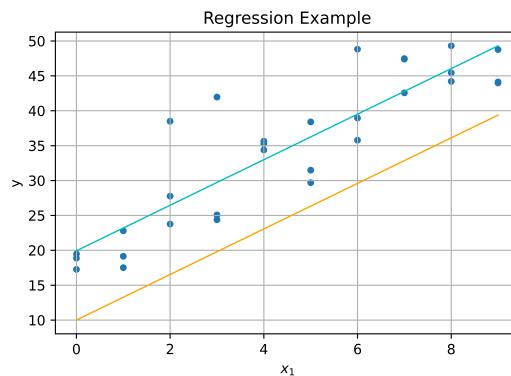
Another reason we don't regularize  $\theta_0$  is that it's often **necessary** for good predictions. This is best shown with a **visual** example.

- Let's take an example with one input,  $x_1$ . So, we have a **linear** function:  $h(x) = \theta_1 x_1 + \theta_0$ .



Our regression example.

Let's suppose we **regularize**, or shrink, our offset  $\theta_0$ , while keeping everything else the same:



Reducing our offset pulls our line further away from all of our data! This is a serious problem.

This shows that we **need** our offset!

- We use it to **shift** our hyperplane around the space: otherwise, otherwise, it's difficult to fit data **far** from the origin.

### Concept 93

$\theta_0$  gives us a baseline for the **size** of our output values.

- If we reduce  $\theta_0$ , all of our outputs will be reduced equally: they'll be **less accurate**.

Imagine that we have three data points at (1, 1001), (2, 1002), and (3, 1003).

It's clear that our data is offset by roughly 1000: we need  $\theta_0$  to address this.

## 2.6.5 Ridge Regression Solution

Now, we have our regression loss function,

$$J(\Theta) = \frac{1}{n} \sum_{i=1}^n \left( \underbrace{(\theta^T x^{(i)} + \theta_0)}_{\text{guess}} - \underbrace{y^{(i)}}_{\text{answer}} \right)^2 + \underbrace{\lambda \|\theta\|^2}_{\text{Regularizer}} \quad (2.62)$$

Which we can express in matrix form:

$$J(\theta) = \frac{1}{n} (\theta^T X - Y) (\theta^T X - Y)^T + \lambda \theta^T \theta \quad (2.63)$$

We can now **optimize** this for  $\theta$ . We'll do some matrix calculus, omitting the steps here:

$$\nabla_{\theta} J = \frac{2}{n} \tilde{X}^T (\tilde{X}\theta - \tilde{Y}) + 2\lambda\theta = 0 \quad (2.64)$$

Finally, we **solve** (using some linear algebra – multiplying by inverses, distributive property, add/subtracting, etc.):

We're gonna cheat a bit, and ignore  $\theta_0$ .

One way to do this is to subtract a constant from every value in  $Y$ , so that the data is centered on  $y = 0$ : you don't need an offset  $\theta_0$ .

But in some problems, we might just include  $\theta_0$  in  $\theta$ , to make our problem simpler, even though we're accidentally regularizing  $\theta_0$ .

### Key Equation 94

The **solution** to the **ridge regression** problem allows us to find the **optimal** model  $\theta^*$ .

$$\theta^* = (\tilde{X}^T \tilde{X} + n\lambda I)^{-1} \tilde{X}^T \tilde{Y}$$

or, in our original notation:

$$\theta^* = (X^T X + n\lambda I)^{-1} X^T Y$$

Where  $I$  is the  $(d \times d)$  identity matrix.

Review: an identity matrix is a square matrix with 1's on its diagonal, and 0's everywhere else.

In general,  $AI = IA = A$ .

## 2.6.6 Invertibility

This solution is great! We just have one problem:

- It requires that our **inverse** of  $(XX^T + n\lambda I)^{-1}$  exists.

Thankfully, this is true so long as  $\lambda > 0$ !

### Concept 95

If  $\lambda > 0$ , we can be sure that  $(XX^T + n\lambda I)$  has an **inverse**.

- And thus, we have a **unique solution** to  $\theta$  for our **ridge regression problem**.

You **do not need to know** the linear algebra that justifies this statement. You just need to know that it's true.

This, by the way, presents one more justification for regularization:

### Concept 96

Another benefit of **regularization** is that we can **always** find an **analytical solution** for  $\theta$ .

- $(\tilde{X}^T \tilde{X} + n\lambda I)$  is invertible, thus, we can compute

$$\theta^* = (XX^T + n\lambda I)^{-1} XY^T$$

Sometimes, we call non-invertible matrices **singular**.

The short version of the justification is:

$XX^T$  is positive semi-definite:  $v^T XX^T v \geq 0$ .

If you add positive elements on the diagonal ( $A = XX^T + n\lambda I$ ), then you can only increase  $v^T Av$ : now it's  $v^T Av > 0$ .

Thus,  $(XX^T + n\lambda I)$  is positive definite: this means it has an inverse.

If that doesn't make any sense, don't worry about it.

### Definition 97

All of the following statements about square matrix  $A$  are equivalent:

- $\det(A) = 0$ .
- $A$  has **no inverse**.
- $A$  is **singular**.
- $A$  is **not full rank**.

A few more definitions of singular:

- $A$  has at least one eigenvalue of 0.
- $A$  has linearly dependent rows and columns.
- $Ax = 0$  has a solution other than  $x = 0$ .

Sometimes, you say this as, "Ax = 0 has a non-trivial solution".

- $A$  is positive definite.

### 2.6.7 Uniqueness of $\theta^*$

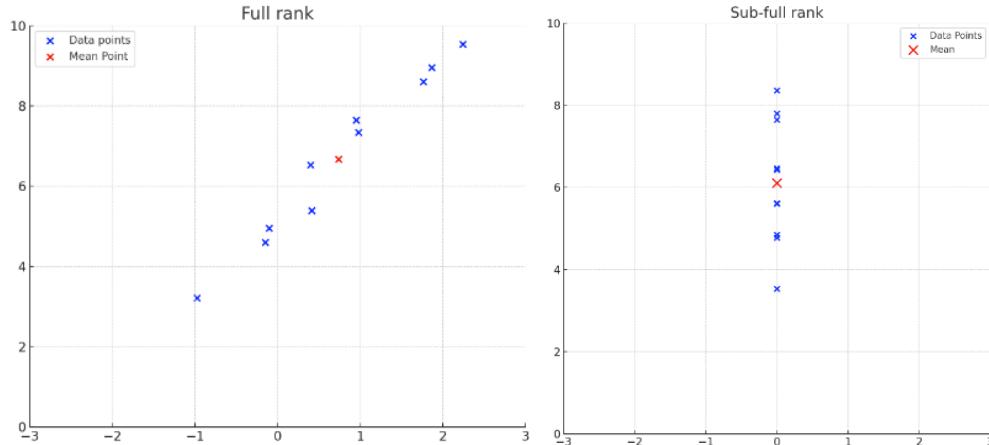
This seems suspicious: why on earth would **regularizing**  $\theta$  allow us to find a solution?

- In order to understand this, we need to understand **what happens** when  $XX^T$  is **singular**.

If  $XX^T$  isn't full rank, that means that  $X$  **isn't full rank**, either. \_\_\_\_\_

What does this mean? Let's consider an example in 1d: we'll compare a "full rank" version, to one that isn't full-rank.

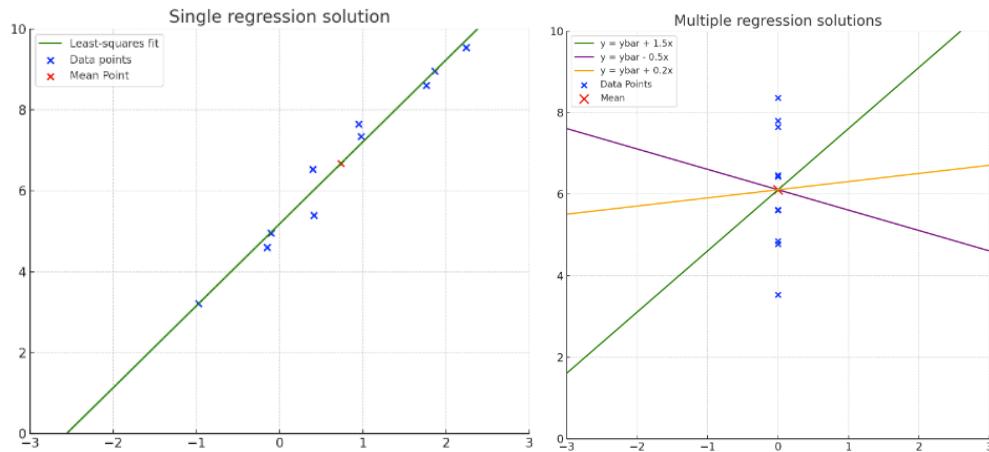
Showing that this is true is more a linear algebra problem than an ML problem, so we'll defer it here.



Typically (left plot), we expect our data to be **full rank**: our input space is 1-D, our input data occupies 1-D.

But sometimes (right plot), we might have data which is **less than full rank**: in this case, the input data is 0-D: only occupying  $x = 0$ .

Sure, this data looks weird, but why is this problematic? Because it creates **multiple optimal solutions**:



Our "full rank" (left plot) data has one optimal solution.  
Our "sub-full rank" (right plot) data has many!

This is the real reason why, if  $\mathbf{X}\mathbf{X}^T$  isn't invertible, we can't find an analytical solution: there are actually **many** possible solutions!

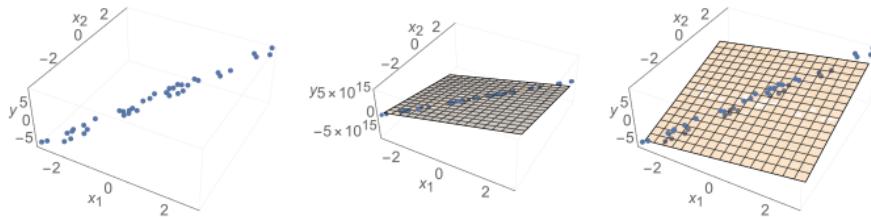
### Concept 98

If  $\mathbf{X}\mathbf{X}^T$  isn't invertible, we **cannot** use our formula to find an analytical (formula-based) **solution** for  $\theta$ .

- That's because there are **many optimal solutions**.
- In fact, there are **infinitely many of them!**

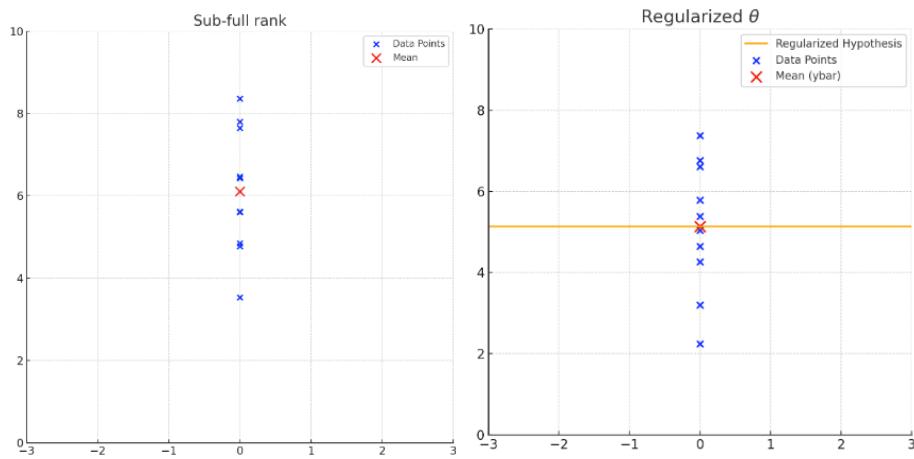
We can see this in higher dimensions too: below, we have a 2-d input space (3-d plot), but our inputs only occupy a **line**.

We call this kind of problem **collinearity**: our input data sit on a lower-dimensional "linear" surface.



There are many possible planes that go through that line: each of these is an **equally good** solution for regression.

We don't have this problem if we use regularization: among all of our **equivalent** options, we just pick the one with the **minimal**  $\|\theta\|$ .



Now, we have one solution: we can get this analytically!

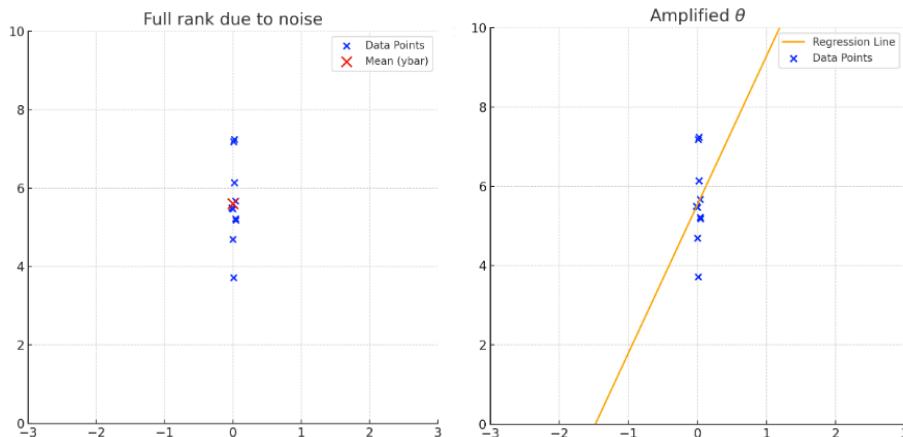
## 2.6.8 Error Amplification

Surely, this is a really rare situation: our data typically won't line up **perfectly**.

- Maybe, but that's part of the problem: let's see what happens if our data is **not** perfectly aligned. We say that  $XX^T$  is **ill-conditioned** or "nearly singular".

### Definition 99

We call a matrix **ill-conditioned** or **nearly singular** if it is very close to a singular/non-invertible matrix.



You can compute whether a matrix is ill-conditioned using something called a "condition number", but we'll omit that point.

Our data has a **large slope** now! This is a problem:  $x$  has **no effect** on  $y$ , we just added a little noise to the  $x$ -axis.

Our model notices, "**very small** change in  $x$ , **moderate** change in  $y$ ", and assumes the slope  $\theta$  should be **large**.

- This is wrong: our change in  $y$  isn't actually explained by  $x$ , it's explained by some "**randomness**" in the output (which is common in real data).

Another way to see this: technically, if you wanted to draw a line through the data, you'd draw a vertical line: "infinite" slope.

### Concept 100

One problem with data that *almost* falls on a line, is **error amplification**.

- If  $x_i$  varies by a small amount, while  $y$  varies by a larger amount, our model may assume  $\theta_i$  is **very large**.
- That means that, if you had a much larger  $x_i$  value, the model will predict  $y$  is **way larger**.

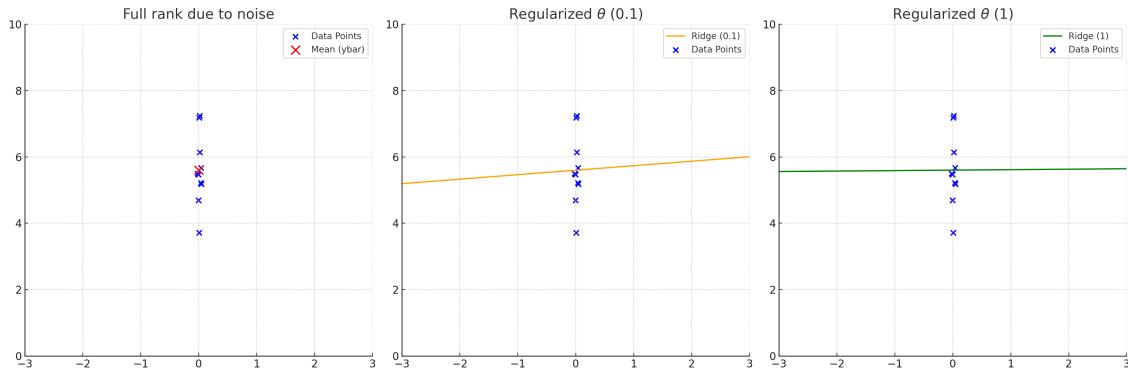
Thus, our error gets **amplified** as we move further away.

If our problem is that  $\theta$  is too large, then we can solve this problem by regularizing  $\theta$ .

Last Updated: 09/03/24 03:53:41

Sometimes, you might hear people mention "numerical instability" of the inverse when a matrix has a small determinant: these are, mathematically, the same problem.

If  $\det(A) = 0.01$ , then  $\det(A^{-1}) = 100$ . The



With  $\lambda = 0.1$ , our slope is already much less sharp. with  $\lambda = 1$ , we end up with (mostly) the same solution as we did in the singular case.

### Concept 101

Ridge Regression helps **improve** our model by

- Making our model more **general** and resistant to **overfitting**
- Making sure **solutions** are **unique**
- Keeping our matrix  $XX^T$  **invertible**, so we can find a **solution**.

## 2.6.9 Error Amplification Example (Optional)

Here's a very simple computational example:

- **Example:** Suppose that two inputs are **almost exactly the same**:  $x_1^{(1)} = 0$ , and  $x_1^{(2)} = 0.01$ .
- But, the outputs are **somewhat** different. The outputs are  $y^{(1)} = 25$ ,  $y^{(2)} = 26$ .

If we assume **all** of the change in the output is a result of the the **input**, then it looks like a **tiny** change in  $x_1$  has a **huge** effect on the output  $y$ .

$$\frac{dy}{dx} = \frac{\Delta y}{\Delta x} = \frac{1}{.01} = 100 \quad (2.65)$$

This suggests that  $x_1$  has a **100x** effect on our output! Even though, it could just be that  $x$  has small variation, and  $y$  has larger variation.

$$100x_1 + 25 = h(x) \quad (2.66)$$

Imagine that  $x_1 = 10$ : suddenly, the prediction is 1025. That's why we call it **error amplification**: a small error near  $x = 0$ , becomes huge as we get further away.

### 2.6.10 Regularizer justification: Prior Knowledge (Optional)

One more way to justify our regularizer applies to a lot of broader statistics: the concept of a **prior belief**.

- Suppose we have prior expectations about what our model should look like: based on theory, or prior experience.
- We might consider a model **more different** from that past one,  $\Theta_{\text{prior}}$ , to be **suspicious**, and less likely to be good.

So, we can **punish our model** for being too different from that expectation.

- This makes our model more **conservative**: it avoids creating a very "extreme" model, without strong justification from the training data.

Our data has to "convince" us that it's worth trying a different model.

#### Concept 102

If we have a **prior** hypothesis  $\Theta_{\text{prior}}$  to work with, we might improve our **new** model by encouraging it to be **closer** to the old one.

$$R(\Theta) = \|\Theta - \Theta_{\text{prior}}\|^2$$

We measure how **similar** they are using **square distance**.

**Example:** You have a **pretty good** model for **predicting** company profits, but it isn't perfect. You decide to train a **better** one, but you expect it to be **similar** to your old one.

In our case, we **don't have** a prior hypothesis  $\Theta_{\text{prior}}$ . We have no clue of what a **good solution** looks like.

We'll take a neutral stance:

- When we know nothing, we're **equally likely** to expect  $\theta_k$  to be positive, or negative.
- In other words, we don't know if  $x_i$  is likely to increase or decrease  $y$ .

So, our guess should average out to 0: the most likely effect is **none**.

- Another way to justify this: if we pick a bunch of variables randomly, we might expect a lot of them to be **irrelevant**.
- **Example:** Without knowing anything, we probably don't expect your birth date to affect your academic performance, positive or negatively.

Thus, we treat  $\Theta_{\text{prior}} = \vec{0}$ .

**Concept 103**

We can interpret **ridge regression** as expecting each  $\theta_k$  terms to be close to 0.

- In other words,  $\Theta_{\text{prior}} = \vec{0}$ .

$$R(\Theta) = \|\theta - \vec{0}\|^2 = \|\theta\|^2$$

This is the same formula we arrived at earlier.

## 2.7 Evaluating Learning Algorithms

Now, we have successfully developed an **algorithm** for **learning** from our data. But, did our algorithm make a **good** hypothesis? How do we do **better**?

### 2.7.1 What $\lambda$ should we choose?

There's something we ignored earlier: how do we pick the **best** value of  $\lambda$ ? We didn't go into detail, but that value of  $\lambda$  will affect our algorithm's **performance**.

- We mentioned that different  $\lambda$  values have different **tradeoffs**, so we need to figure out which  $\lambda$  value is best for our problem.
- This  $\lambda$  adjusts exactly how we learn: how do we balance learning from **data** against the need to **generalize**?

So, we need to **optimize** our  $\lambda$  value. Let's figure out how to go about that.

### 2.7.2 Tradeoffs: Estimation Error

High and low  $\lambda$  values have benefits and drawbacks. These tradeoffs can be loosely divided into **two categories**.

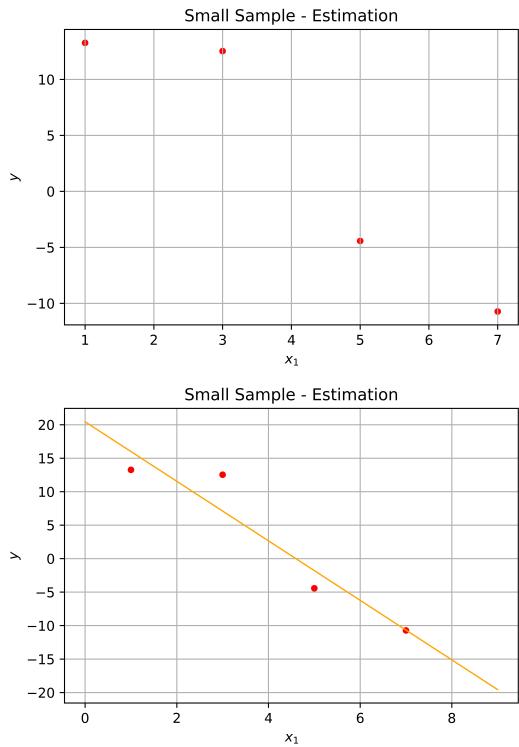
When we generalize, we're trying to avoid **estimation error**: we incorrectly guess the overall distribution we're trying to fit. We **estimate** poorly if we **generalize** poorly.

#### Definition 104

**Estimation error** is the error that results from poorly **estimating** the **solution** we're trying to find.

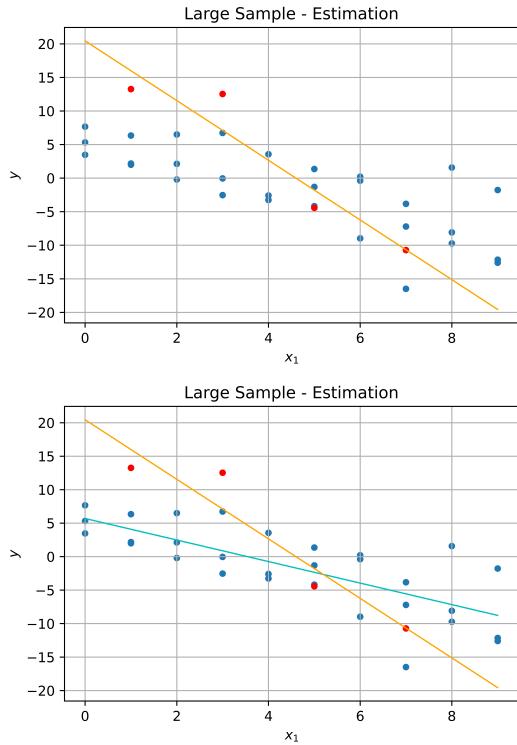
This can be caused by **overfitting**, getting a bad (**unrepresentative**) sample, or not having enough **data** to come to conclusion.

**Example:** Let's try a regression problem, but we'll use only 4 points to make our plot.



This is the regression solution we get based on our small dataset.

We might be suspicious. One way to reduce **estimation error** is to increase our number of data points (though this isn't always an option, or sufficient!)



Our regression from before doesn't look so good on this model... We make an updated regression, and get a more accurate result.

#### Clarification 105

$\lambda$  doesn't lower **estimation error** in the **same way** that increasing **sample size** does, but the problem is **similar**.

### 2.7.3 Tradeoffs: Structural Error

However, not all problems are caused by estimation error: sometimes, it **isn't even possible** to get a good result - you chose the wrong **model class**.

This means the **structure** of your model is the problem, not your method of **estimation**. Thus, we call this **structural error**.

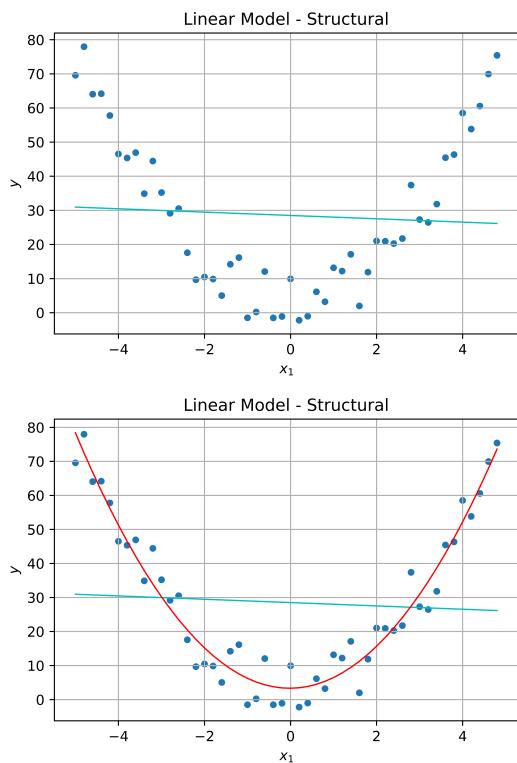
#### Definition 106

**Structural error** is the error that results from having the wrong **structure** for the **task** you are trying to accomplish.

This can result from the **wrong class** of model, but sometimes, your model class doesn't have the **expressiveness** it needs for a complex problem.

It can also happen if your algorithm **limits** the available models in some way, like how  $\lambda$  does.

**Example:** If the **true shape** of a distribution is a parabola  $x^2$ , there is **no** linear function  $mx + b$  that can match that: this creates **structural error**.



Our **linear** model isn't able to represent a quadratic function... so, we switch to a more expressive model: a **quadratic** equation.

#### Clarification 107

Note that  $\lambda$  does not restrict our model class **as severely** as **switching polynomial order**, like above.

But,  $\lambda$  **limits** the use of larger  $\theta$ , which does make it **unable** to solve some problems. So, the **structural error** problem is similar.

Remember that **expressiveness** is about how many possible models you have: if you have more models, you can solve more problems.

#### 2.7.4 Tradeoffs of $\lambda$

Based on these two categories, we can discuss the tradeoffs of  $\lambda$  more easily.

As we mentioned, regularization **reduces** estimation error:

If we overfit to our current data, we are poorly **estimating** the distribution, because the training data may not perfectly **represent** it.

**Concept 108**

A large  $\lambda$  means **more regularization**: we more strongly push for a more **general** model, over a more **specific** one.

This results in...

- **Reduced** estimation error
- **Increased** structural error

However, **regularization** also **limits** the possible models we can use - those it views as less "general", it **penalizes**.

- That means the scope of possible models is **smaller** - some models are no longer **acceptable**. What if the only valid solution was in that space we **restricted**? Well, then we can't **find** it.
- That means there are certain **structural** limits on our model: that means that regularization **increases** structural error!

**Concept 109**

A small  $\lambda$  means **less regularization**: we care less about a more **general** model, allowing more **specific** data to come into play.

This results in...

- **Increased** estimation error
- **Reduced** structural error

## 2.7.5 Evaluating Hypotheses

So, we know that we have these **two** types of **error**. But it's **difficult to measure** them separately.

So instead, we just want to measure the **overall performance** of our hypothesis.

We do this using our **testing error**: this tells us how good our hypothesis is **after** training.

$$\mathcal{E}(h) = \frac{1}{m} \sum_{i=n+1}^{n+m} (h(x^{(i)}) - y^{(i)})^2 \quad (2.67)$$

Note that, before, we were using **regularization**. This is so we can **make** a more **general** model.

But here, we've **removed** it, because training is **done**: we're **not** going to make our hypothesis **better**. We just care about how **good** it came out.

We're already measuring the **generalizability** by using **new data**!

**Clarification 110**

When we **evaluate a hypothesis** using **testing error**, we are **done training**: our hypothesis will not change.

Because of this, we **do not** include the **regularizer** when **evaluating** our hypothesis.

## 2.7.6 $\lambda$ 's purpose: learning algorithms

Notice that we **removed** regularization when we were **evaluating** our hypothesis: regularization was used to **create** our hypothesis, but it is not **part** of that hypothesis.

That's because  $\lambda$  is part of our **algorithm**: it determines how we find our hypothesis. So, let's talk about that.

Our hypothesis only includes the parameters  $\Theta$ : not  $\lambda$ !

**Definition 111**

A **learning algorithm** is our procedure for **learning** from data. It uses that data to create a **hypothesis**. We can diagram this as:

$$\mathcal{D}_n \longrightarrow \boxed{\text{learning alg } (\mathcal{H})} \longrightarrow h$$

In a way, it's a function that takes in **data**  $\mathcal{D}_n$ , and outputs a **hypothesis**  $h$ .

We're choosing **one hypothesis**  $h$  from the hypothesis class  $\mathcal{H}$ : this is why  $\mathcal{H}$  appears in the notation above.

We can write this as  
 $h \in \mathcal{H}$

## 2.7.7 Comparing Hypotheses and Learning Algorithms

We can take our learning algorithm

$$\mathcal{D}_n \longrightarrow \boxed{\text{learning alg } (\mathcal{H})} \longrightarrow h$$

And compare it to our hypothesis  $h$ :

$$x \rightarrow \boxed{h} \rightarrow y$$

In a way, our learning algorithm is a function, that outputs another function!

This is similar to  $\mathcal{E}_n$ , which instead takes a function as **output**!

- Our **hypothesis** can be adjusted with our **parameter**  $\Theta$ : if we change  $\Theta$ , we change our **performance**.
- Our **learning algorithm** depends on  $\lambda$ : so,  $\lambda$  is like a **parameter**. But, it's different from  $\Theta$ :  $\Theta$  **is** our model,  $\lambda$  controls how we **choose** our model.

- So, it's a parameter ( $\lambda$ ) that affects other parameters ( $\Theta$ ). Because of that, we call it a **hyperparameter**.

It affects our hypothesis by pressuring it to have lower magnitude!

#### Definition 112

**Parameters** are **variables** that adjust the behavior of **our model**: our hypothesis.

A **hyperparameter** is a **variable** that can adjust **how we make models**: our learning algorithm.

The **only** hyperparameter we have for now is  $\lambda$ , but the **development** of hyperparameters is an ongoing area of **research**.

#### Concept 113

**Lambda**, or  $\lambda$ , is a **hyperparameter**: it controls our **learning algorithm**.

## 2.7.8 Evaluating our Learning Algorithm

So, while we can evaluate each **hypothesis**, it's also important to measure how our **learning algorithm** is performing.

How do we measure it? Well, the job of our **learning algorithm** is to **pick good hypotheses**.

#### Concept 114

We can **evaluate** the performance of a **learning algorithm** using **testing loss**: a good learning algorithm will create **hypotheses** with low testing loss.

You could think of this as measuring the **skill** of a **teacher** (the learning algorithm) by the **success** of their **student** (the hypothesis) on a **test** (testing loss).

## 2.7.9 Validation: Evaluating with lots of data

When we were creating hypotheses, **randomness** caused some problems: you might not get **training data** that matched the **testing data** very well.

The **same** can happen here, when **evaluating your algorithm**: maybe your model happened to create a bad (or unusually good!) hypothesis because of **luck**.

The easy solution to **randomness** is to add **more data**: we get more **consistency** that way.

So, we **repeatedly** get new training data and test data. For each, we train a **different hypothesis**. We can **average** their performance out, and use that to **estimate** the quality of our algorithm.

**Definition 115**

**Validation** is a way to **evaluate a learning algorithm** using **large amounts of data**.

We do this by **running** our algorithm **many times** with new data, and **averaging** the testing error of all the hypotheses.

- This process is often requires having **lots of data** to train with, but is a **provably** good approach.

### 2.7.10 Our Problem: When data is less available

As mentioned, this takes up **lots of data**. What if our data is limited?

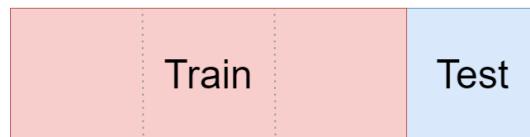
In this case, we'll assume that have some **finite** data,  $\mathcal{D}_n$ . We **can't get more**.

Data is often **expensive**. It might even be impossible to get more!

Previously, we solved validation by using **more data**, and generating **multiple hypotheses**.

- One set of data gives us one **hypothesis**.
- But, what if, rather than using **completely** new data for each hypothesis, we used **slightly different** data each time?

First, need to break  $\mathcal{D}_n$  into a chunk for training, and a chunk for testing.



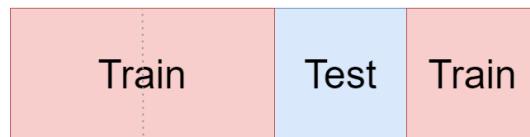
How do we get more hypotheses from this dataset?

### 2.7.11 Cross-Validation

We mentioned that we want **different** hypotheses. Our hypotheses depend on our **training data**. So we want to **change** our training data.

We can't **add** data to it, because then we **lose** testing data. We shouldn't **remove** training data, because then we're just making a hypothesis that's **less well-informed**.

Instead, we'll **swap** some of the training data for testing data.



This will create a new hypothesis, and the data is partially different! In fact, we can do this for each of our chunks:



We now have **four different hypotheses** for the price of one!

#### Definition 116

**Cross-validation** is a way to **evaluate** a learning algorithm using **limited data**.

- We do this by **breaking** our data into **chunks** to create **multiple hypotheses** from one dataset.
- For each **chunk**, we train one dataset on all the data **not in that chunk**. We get our **test error** using the chunk **we left out**.

For  $k$  chunks, we end up with  $k$  hypotheses. By **averaging** out their performance, we can **approximate** the quality of our algorithm.

This approach is much **less expensive**, and very common in machine learning!

But, some of the theoretical **benefits** of validation are not **proven** to be true for cross-validation.

#### Clarification 117

Note that the goal of validation and cross-validation is **not** to evaluate **one hypothesis**.

Instead, it is instead meant to evaluate a **learning algorithm**. This is why we have to create **many** hypotheses: we want to see that our algorithm is **generally** good!

## 2.7.12 Hyperparameter Tuning

Now, we know how to **evaluate** a learning algorithm, just like how we **evaluate** a hypothesis.

Once we knew how to evaluate a hypothesis, we started optimizing our **parameters** for the **best** hypothesis. So, we could do the same for our **learning algorithm**.

How do we **optimize** a learning algorithm?

Each  $\lambda$  value creates a slightly **different** learning algorithm: we can **optimize** this **hyperparameter** to create the **best** learning algorithm.

### 2.7.13 How to tune our algorithm

When we were **optimizing** our hypothesis, we started by **randomly** trying hypotheses. Then, we used an **analytical** approach. \_\_\_\_\_

We don't always have **simple** equations to work with: with all of our data, it's hard to come up with **manageable** equations. So, we **won't** try doing it **analytically**.

By "analytical", we mean directly creating an equation, and solving it.

So, we could **randomly** try  $\lambda$  values and pick the **best** one. This is pretty **close** to what we usually end up doing. For each value we pick, we'll use **cross-validation** to evaluate.

For now, we'll systematically go through  $\lambda$  values:  $\lambda = .1, .2, .3 \dots$

## Concept 118

**Hyperparameter tuning** is how we **optimize** our **learning algorithm** to create the **best** hypotheses.

The simplest way to do this is to try **multiple** different values of  $\lambda$ . For each value, we use **cross-validation** to evaluate that learning algorithm.

Finally, we pick whichever  $\lambda$  gives you the **best** algorithm, and thus the **best** hypotheses.

### 2.7.14 Hyperparameter Tuning: Two kinds of optimization

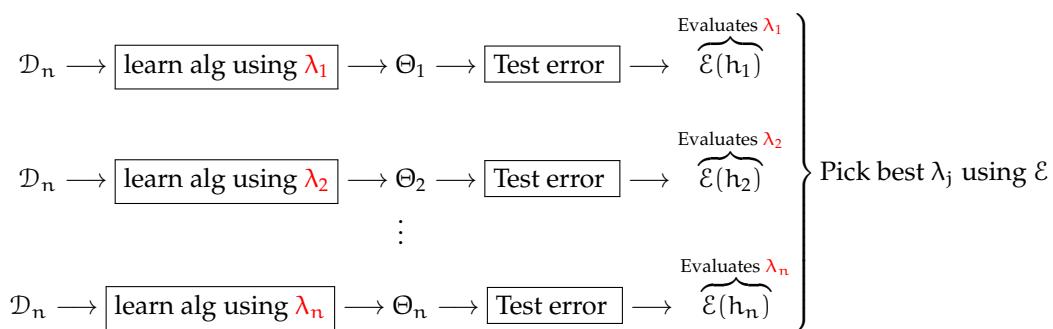
There's something often confusing about hyperparameter tuning to students:

- When we're **optimizing**  $\lambda$ , we try many values  $\lambda_j$ . Each  $\lambda_j$  creates a **learning algorithm** we have to evaluate.
  - But a single learning algorithm is already an optimization problem: the learning algorithm is supposed to find  $\Theta$ .
    - So, we have to **optimize**  $\Theta$ , while we're in the middle optimizing  $\lambda$ .

In case the word "optimization" starts to look like gibberish in this section, remember: it just means, "find the best option".

That means, **every time** we try a different  $\lambda$  value, we have to do one optimization problem. Optimizing  $\Theta$  many times, lets you optimize  $\lambda$  once.

Remember that, in this situation,  $\Theta$  and  $h$  are almost(but not quite) the same thing.



That means we have **two layers** of optimization!

#### Clarification 119

We **optimize**  $\lambda$  by trying many values.

- But, for each  $\lambda$  value, we have to **optimize**  $\Theta$ .

So, we have to optimize  $\Theta$  **repeatedly** in order to optimize  $\lambda$  **once**! This gives us  $\lambda^*$ .

Once we've found our best hyperparameter  $\lambda^*$ , we can use it to get our best parameters:  $\theta^*$ .

### 2.7.15 Pseudocode Example

This technique is **not** limited to regression. Thus, we'll be a bit more **general**: we won't assume an **analytical** solution. Instead, we **optimize** by just trying different  $\Theta$  values.

We can represent this in pseudocode:

```
LAMBDA-OPTIMIZATION(D, lambda_values, theta_values)
1  for λ in lambda_values      #Try lambda values
2    for Θ in theta_values      #Try theta values
3      Calculate J(Θ)          #Compare values
4      Choose best theta value Θ*  #Best for each lambda
5  Choose best lambda value λ*
6
7  return λ*
```

If this pseudocode isn't helpful to you, don't worry! Some students like it, some don't.

To reiterate: this  $\lambda^*$  will then we used to get our final result,  $\theta^*$ .

## 2.8 Terms

- Hypothesis
- Theta ( $\Theta$ )
- Input Space
- Regression
- Feature
- Feature Transformation
- Training Error
- Test Error
- Objective Function
- Min function
- Argmin function
- Star Notation ( $\theta^*$ )
- Linear Regression
- Hypothesis Class
- Square Loss
- Ordinary Least Squares (OLS) Problem
- OLS Objective Function
- Hyperplane
- Weight
- Input Matrix
- Output Matrix
- Gradient
- OLS Solution
- Regularization
- Regularizer
- Regularizer for Regression
- Lambda ( $\lambda$ )

- Ridge Regression
- RR Objective Function
- RR Solution
- Invertibility
- Estimation Error
- Structural Error
- Expressiveness
- Learning Algorithm
- Hyperparameter
- Validation
- Cross-Validation
- Chunk (Cross-Validation)
- Hyperparameter Tuning

# CHAPTER 3

---

## Gradient Descent

---

### What is gradient descent?

#### 3.0.1 Why do we need gradient descent?

In the last chapter, we used an **analytical** approach to solve the OLS and RR problems.

By "analytical", we mean we got an **explicit** answer: an equation we can use to directly compute the correct answer.

The trouble is, we can't always do this:

- Sometimes the problem or the loss function can't be **rearranged** into a simple **equation**.
- Or, we have **too much** data, and directly computing the answer would take way **too long**.

#### Concept 120

Most **problems** we come across cannot be solved **analytically**.

Well, if we can't **directly** find the **best** answer, what's the next best thing? Finding a **better** solution than your current one.

So, our mission is to gradually try to find a better and better answer. This type of approach has a couple benefits:

- It's **quicker** to see if we're using a good model: if we're making very little progress, we can **quit** early and try something else.
- If we don't need **all** of our data to get the answer, we don't need to spend as much time. If our answer is **good** and not getting better, we can **stop**.
- It's easier to find a **better** answer than the **best** answer: our equations will be **simpler**. In some case, it might not have even been **possible** without this gradual approach!

### Concept 121

When we can't reasonably find a **best** answer, it's often easier to find a **better** answer and gradually **improve**.

**Gradient descent** follows this philosophy: we gradually **update** our solution to make it better and better.

### 3.0.2 How do we improve?

So, now, the question is: how do we **improve** our hypothesis? We'll be modifying our hypothesis  $\theta$  by some amount: \_\_\_\_\_

$$\theta_{\text{new}} = \theta_{\text{old}} + \Delta\theta \quad (3.1)$$

We'll do the same for  $\theta_0$ , but we'll do it separately. We'll come back to that.

### Notation 122

In equations, we'll often use  $\theta_{\text{old}}$  and  $\theta_{\text{new}}$  to represent **before** and **after** we take a step.

We will use this notation **elsewhere** in the class.

So, we are interesting in  $\Delta\theta$ : how do we plan to change  $\theta$ ? What does  $\Delta\theta$  look like?

Well, we want to modify

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{bmatrix} \quad (3.2)$$

So, we want to modify each of those terms.

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{bmatrix} + \begin{bmatrix} \Delta\theta_1 \\ \Delta\theta_2 \\ \vdots \\ \Delta\theta_d \end{bmatrix} \quad (3.3)$$

So, we have our total change!

$$\Delta\theta = \begin{bmatrix} \Delta\theta_1 \\ \Delta\theta_2 \\ \vdots \\ \Delta\theta_d \end{bmatrix} \quad (3.4)$$

Notice that the shape of this change matches the shape of  $\theta$ : ( $d \times 1$ ).

### Concept 123

We need a **separate** term  $\Delta\theta_i$  for each  $\theta_i$  we want to **improve**.

So, a vector of the **total** change,  $\Delta\theta$ , needs to have the **same shape** as  $\theta$ : ( $d \times 1$ ).

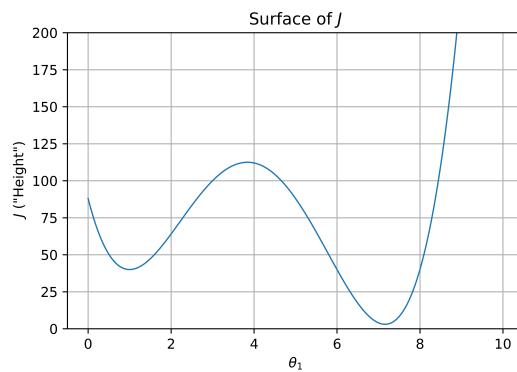
### 3.0.3 The name: "gradient descent"

Our goal is to gradually **decrease**  $J$ , step-by-step. We do this using the **gradient**, hence "gradient descent". Why the gradient? We'll discuss that later.

But why the word "**descent**"?

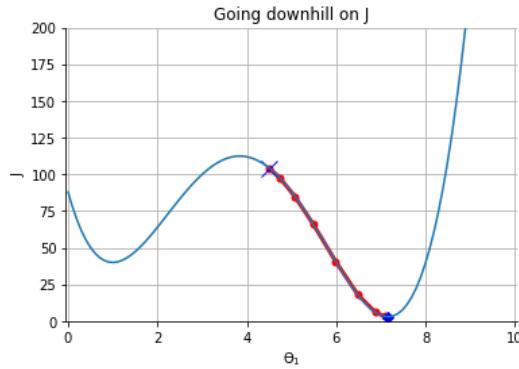
Our intuition is to imagine  $J$  as having a **height** at every input value. If you combine all of these different points, you get a **surface**, like the surface of a hill.

Reminder: Why are we "decreasing"  $J$ ? Because  $J$  represents the "badness" of our model. We want it to be, simply put, "less bad".



You can imagine this like some hills we want to "descend".

Then, decreasing  $J$  is moving **down** the the function, similar to rolling a ball down a hill. In other words, we **descend** the hill.



Like this! Starting from the blue "X", moving 'downhill'.

### 3.0.4 Input Space vs. Parameter Space

One more thing to note: we have two similar situations.

- $J$  is a **function** with  $\theta$  as an **input**:  $J(\theta)$ .
- $h$  is a **function** with  $x$  as an **input**:  $h(x)$ .

In both cases, we can imagine the **output** as the "**height**" of our function: the **hill** we mentioned before. This **physical** intuition is useful to **gradient descent**.

But, what about **input** to our function? That's the  $x$ -axis our hill is floating above:

- With  $h(x)$ , our  $x$ -axis was our **input space**, all possible  $x_1$  values: the "space" containing all of our possible inputs.
- With  $J(\theta)$ , our  $x$ -axis is the **parameter space**, all possible  $\theta$  values. We also called this our "**hypothesis space**".

#### Definition 124

The **parameter space** is our set of all **possible** parameter combinations.

This is the same as the **hypothesis space**, because our parameters **define** our hypothesis.

When we **optimize** our hypothesis, we are "**exploring**" the hypothesis space.

We're assuming 1-D right now for simplicity. If we were 2-D, we'd have an entire 2D grid under our hill!

- This can be seen as the "collection of **all possible models** in our model class".
- Why do we call it a "parameter **space**", not a "parameter set"?
  - It's the **structure**: the fact that some hypothesis are "closer" to each other:  $\theta = 1$  is closer to  $\theta = 2$  than  $\theta = 10$

We mentioned one useful feature: we have a concept of which hypotheses are "similar": those which are **closer** in parameter space. \_\_\_\_\_

This is the **space** we're exploring, as we try to move **downhill**.

We've already used this fact! When normalizing towards the previous hypothesis,  $\theta_{\text{old}} = 0$ , we minimized  $\|\theta - \theta_{\text{old}}\|$ .

#### Clarification 125

Pay attention to your **axes**!

Sometimes, we're doing a 2-D or 3-D plot of  $J$ , and our inputs are  $\theta_k$ . Other times, we're plotting hypothesis  $h$ , with our input axes  $x_i$ .

These two plots could have the same surface, but they **represent** completely different things.

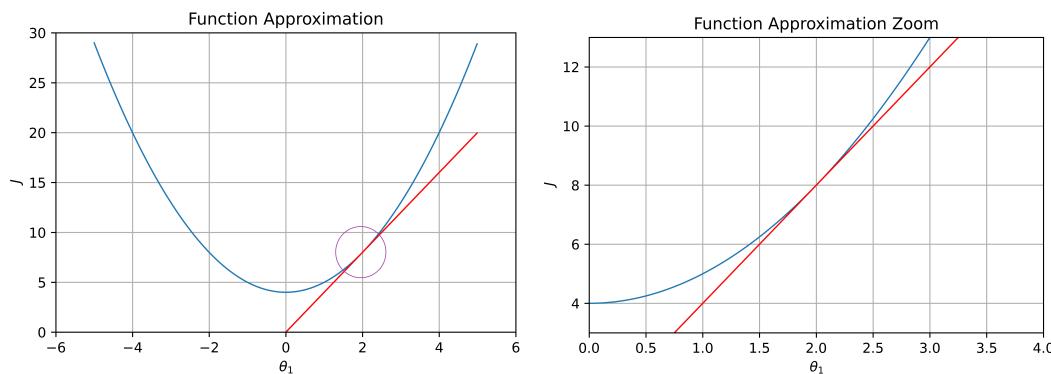
## 3.1 Gradient Descent in One Dimension

### 3.1.1 Derivatives (Review)

Here, we'll use some concepts from **calculus**.

We'll make improvements in small **steps**. And, we measure our improvement against the **loss function**,  $J$ : that's what we want to **optimize**.

In calculus, we found that, over **small** enough steps, you can **approximate** a smooth function as a straight line.



It looks more like a line as we zoom in: hence the **local** approximation.

#### Concept 126

A **smooth** (enough) function can be **approximated** with a **straight line** if you **zoom** in on it enough.

Looking at it this way is called a **local** view.

### 3.1.2 Optimize with Derivatives: 1-D

This gives us the **slope** of the function locally. Last chapter, we used  $\frac{dJ}{d\theta} = 0$  to get our **minimum**.

But, let's not get too greedy - we want to **improve** our hypothesis, **not** immediately try to find the **best** one.

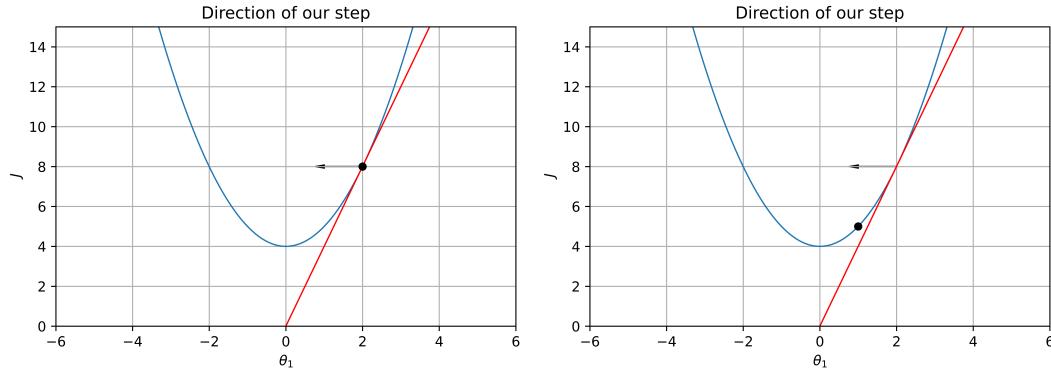
Well, what does our slope tell us? It tells us:

Because the best one might be expensive to find this way!

- How quickly  $J$  changes
- Whether it **increases** or decreases as we change  $\theta$

That second one tells us *how to change*  $\theta$ : we want to move in the direction that **decreases**  $J$ .

If the slope is **positive**, then we want to **decrease**  $\Delta\theta$ : the sign of  $\Delta\theta$  is the opposite of our desired change!



Our slope is **positive**. We want to **decrease** our function, so we move in the **negative** direction, and "fall down" the surface.

And so, for now, we have

$$\Delta\theta = -\frac{dJ}{d\theta} \quad (3.5)$$

### Concept 127

In **1-D**, you can use the **derivative** to **optimize** our function  $J$ .

The **derivative** tells us how to immediately adjust  $\theta_i$  to **improve** our  $J$  **locally**: we move in the **opposite direction**.

This gives us a procedure for optimizing  $J$ : get the derivative  $J'(\theta)$ , and repeatedly adjust  $\theta$  in the opposite direction until you're satisfied.

There's a certain way this feels like we're moving "**downhill**": we're moving "down" the slope, to try to find a local **minimum**.

We'll need to pick a condition for being satisfied, but we'll get to this later

### 3.1.3 Convergence

If you do this procedure with the above equation, though, you'll often run into **problems**. Why is that?

Well, because each of your steps is too **big** or too **small**: we won't be able to find a **stable** answer, i.e. **converge**!

What does it mean to **converge**?

It means we get a **single answer** after repeated steps: given enough time, we'll get **close as we want** to one number, and **stay there**.

**Definition 128**

If a sequence **converges**, then our result gets as **close as we want** to a **single number**, without going **further away**.

**Example:** The numbers  $1/n: \{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots\}$  converges to 0.

If our answer **doesn't converge**, then it **diverges**. We can see why this might be bad: if we never **approach** a single answer, how do we know what value to **pick**?

### 3.1.4 Convergence: A little more formally (Optional)

Let's be more specific. Our sequence  $S$  will converge to  $r$ .

$$S = \{s_1, s_2, s_3, s_4, \dots\} \quad (3.6)$$

"As close as we want": let's say we want the maximum distance to be  $\epsilon$ . That means, no matter what  $\epsilon > 0$ , we'll get closer at some point:  $|m - s_i| < \epsilon$

$$|m - s_i| < \epsilon \text{ for some } i \quad (3.7)$$

"And stay there": at some time  $k$ , we never move further away again:

**Definition 129**

If a sequence  $S$  **converges** to  $m$ , then for all  $\epsilon > 0$ , we can say

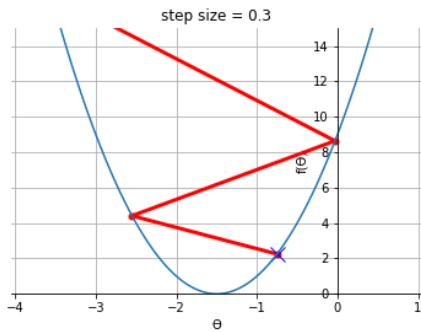
$$|m - s_i| < \epsilon \text{ for all } i > k \quad (3.8)$$

This is a "formal" definition of convergence.

### 3.1.5 Step size

If your steps are too **big**, your result might **diverge**: you make such big jumps, you move **away** from the minimum, and get worse.

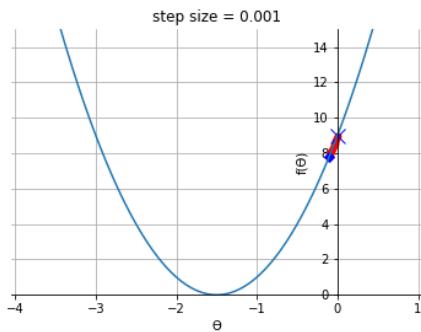
Remember, if it **diverges**, it never **approaches** a single value!



We start at the blue "x" mark. Notice that, even though we try to move toward the minimum, we go too far and accidentally get further and further!

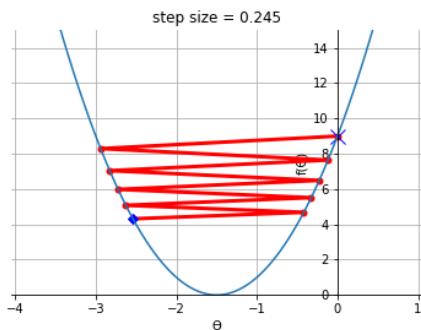
If they're too **small**, you might **converge** too slowly: it'll take way **too long** to make progress.

**Converging** means it successfully **approaches** an answer!



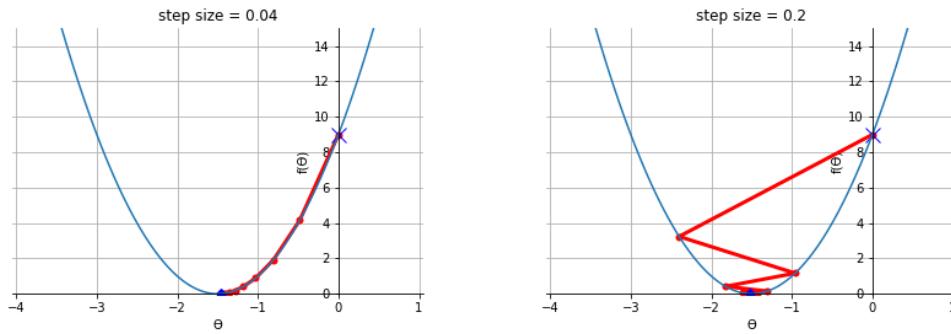
Our step size is too small: this is going to take too long!

In-between, it might converge, but **oscillate** a bunch: this can slow down getting an answer!



Most of our step is spent undoing the last step... we get better very slowly.

But, if we get the right step size, it'll converge nice and reasonably!



Both of these look pretty good! One of them oscillating a bit is fine.

One question you might ask is, "**how much** oscillation is too much? Am I converging **fast enough**?"

This is a good question, but the simple answer is that there is **no objective answer**: it depends on what you **need** and how much **time** you have. But you should strive to do **better** when you can!

#### Concept 130

Using the **wrong** step size can cause:

- Slow convergence
- Strong Oscillation
- Divergence

Which is why we **adjust** the step size using  $\eta$ .

### 3.1.6 Step size $\eta$

Right now, our step size is at the mercy of  $J'(\theta)$ . But, we don't have to be: we could **scale** our step size up or down.

We do this with our **scaling** factor (also called a **learning rate**),  $\eta$ .

So, we can rewrite our **change** in  $\theta$  as:

$$\Delta\theta = -\eta \frac{dJ}{d\theta} = -\eta J'(\theta) \quad (3.9)$$

**Definition 131**

Our step size parameter  $\eta$ , or **eta**, **scales** how large each of our optimization steps are.

If  $\eta$  is bigger, we might **learn** faster, but we also risk **diverging**.

Different values of  $\eta$  are good for **different situations**.

### 3.1.7 Our procedure

So, we have our parameter **update**,  $\Delta\theta$ . We'll start at  $t = 0$ .

Before, we represented the  $i^{\text{th}}$  **data point** with  $x^{(i)}$ . We'll reuse this **notation**.

**Notation 132**

Here, we're changing  $\theta$  over **time**: each step happens at  $t = \{1, 2, 3, \dots\}$  so we need **notation** for that.

We'll **reuse** the notation from  $x^{(i)}$ , for the  $i^{\text{th}}$  data point.

In this case, we'll do  $\theta^{(t)}$ : the value of  $\theta$  after  $t$  **steps** are taken.

Earlier, we **introduced**  $\theta_{\text{old}}$  and  $\theta_{\text{new}}$ : these are  $\theta^{(t-1)}$  and  $\theta^{(t)}$ .

**Example:** After **10 steps** of 1-D gradient descent, we have gone from  $\theta^{(0)}$  to  $\theta^{(10)}$ .

So, we move the **first** time using  $J'(\theta^{(0)})$ .

Once we've moved in parameter space **one** time, though, our **derivative** has changed: we're in a different part of the **surface**.

So, we'll take a **second** step with a **new** derivative,  $J'(\theta^{(1)})$ .

We want to do this **repeatedly**. We'll take our equation

$$\theta_{\text{new}} = \theta_{\text{old}} + \Delta\theta \quad (3.10)$$

And combine it with our **chosen** step size.

**Key Equation 133**

In **1-D, Gradient Descent** is implemented as follows:

At each time step  $t$ , we **improve** our hypothesis  $\theta$  using the following rule:

$$\theta_{\text{new}} = \theta_{\text{old}} - \eta J'(\theta_{\text{old}})$$

Using  $\theta^{(t)}$  notation:

$$\theta^{(t)} = \theta^{(t-1)} - \eta J'(\theta^{(t-1)})$$

We repeat until we reach whatever our chosen **termination condition** is.

We can also write it as:

$$\theta_{\text{new}} = \theta_{\text{old}} - \eta \cdot \left( \frac{dJ}{d\theta} \Big|_{\theta=\theta_{\text{old}}} \right)$$

We've got our gradient descent **update** rule in 1-D!

### 3.1.8 Termination Conditions

When do we **stop**? We can't let it run forever.

We have some options:

- Stop after a **fixed**  $T$  steps.
  - This has the advantage of being **simple**, but how do you know what the **correct** number of steps is?
- Stop when  $\theta$  **isn't changing** much:  $|\Delta\theta| < \epsilon$ , for example.
  - If our  $\theta$  isn't changing much, our algorithm isn't **improving** our hypothesis much. So, it makes sense to stop: we've stabilized.
- Stop when the **derivative is small**:  $|J'(\theta)| < \epsilon$ .
  - Mathematically **equivalent** to our last choice. But a different **perspective**: if the slope is small, our surface is relatively **flat**, and we're near a **minimum** (probably).
  - "The derivative is **small**" is weaker, but in the same spirit as "the derivative is **zero**",  $J'(\theta) = 0$ , from last chapter.

### 3.1.9 Convergence Theorem

It turns out, if our function is **nice** enough, and we pick the **right** value of  $\eta$ , we can guarantee convergence!

#### Theorem 134

##### Gradient Descent Convergence Theorem

We want to optimize function  $J$ . If  $J$  is

- Smooth enough
- Convex

And

- $\eta$  is small enough

Then gradient descent **will** converge to the **global minimum**!

- ~~~~~
- "Small enough" seems vague, but it basically means, "if the first **two statements** are true, then there **exists** a choice of  $\eta$  that converges."

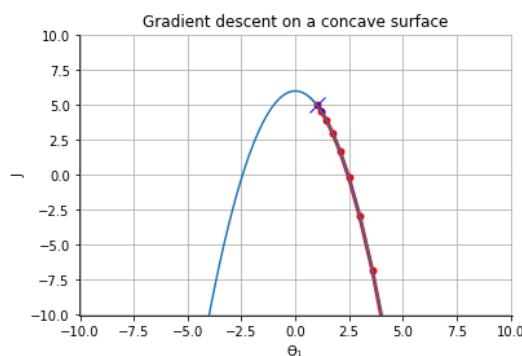
Or, if your  $\eta$  is too **big**, you can keep trying **smaller** ones, until it works.

This is amazing! We can **guarantee** a best solution in some cases!

### 3.1.10 Concavity

One requirement we haven't focused on " $J$  is **convex**". Why do we need  $J$  to be convex?

Well, if it's **concave**, there is no **global minimum**: it goes down forever!



Our gradient just leads us downhill forever.

**Concept 135**

If our function  $J$  is **concave**, then our result will not **converge**: it will continue to **decrease** more and more indefinitely.

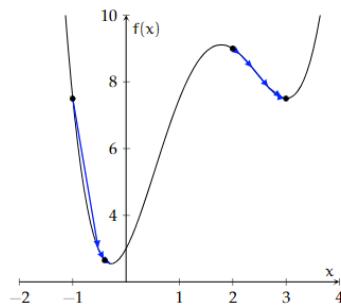
So, for future problems, let's assume it **doesn't** go down forever: if it was, then there is no best solution! We don't have a **valid** problem.

### 3.1.11 Local minima

Even if we don't have that problem, we have a **different** one:

Gradient descent **gradually** improves our solution until it reaches one it's **satisfied** with. But, what if there are **multiple** solutions we could reach?

Are they all equally good?



Depending on your starting position (**initialization**), you could find a different local minimum!

Maybe not! So, if our function isn't **always convex**, we can end up with **multiple** "valleys", or **local** minima.

**Definition 136**

A **global** minimum is the **lowest** point on our entire function: the one with the lowest **output**.

A **local** minimum is one that is the **lowest** point among those points that are **near** it.

- For **local minima**, if you add or subtract a **tiny** amount  $\epsilon$  to the input, the output will **increase**.

So, we **won't** necessarily end up with the **global** minimum, even with a *small*  $\eta$ .

This shows that **initialization matters**!

**Definition 137**

**Initialization** is our "starting point": when we first **start** our algorithm, what are our **parameters** set to?

If we have a **different** starting position, we can find a **different** local minimum.

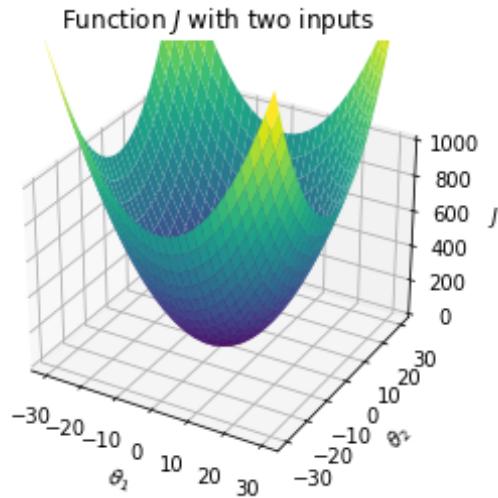
**Concept 138**

**Gradient descent** often finds **local** minima near the initialization, not necessarily **global** minima.

This means, if our function has **multiple local minima** (not fully convex), our **initialization** can affect our **solution**.

## 3.2 Multiple Dimensions

Now that we've handled the 1-D case, we'll move into 2-D: now, we have **two** parameters,  $\theta_1$  and  $\theta_2$ , as the input to  $J$ .



The "height" of your plot in 3D, is, again, your output! You want to move **downhill**.

### 3.2.1 Multivariable Local Approximation (Review)

Again, we rely on **calculus**. We want to move up to having more parameters: more **dimensions**.

Before, in 1-D, we found that, if you **zoomed** in enough on a function (using a "**local view**"), we could **approximate** it as a **straight line**, and move up or down that slope.

There are **two** ways we can view our **approximation** in 2-D: \_\_\_\_\_

- First, we could turn it back into 1-D: we remove one variables.

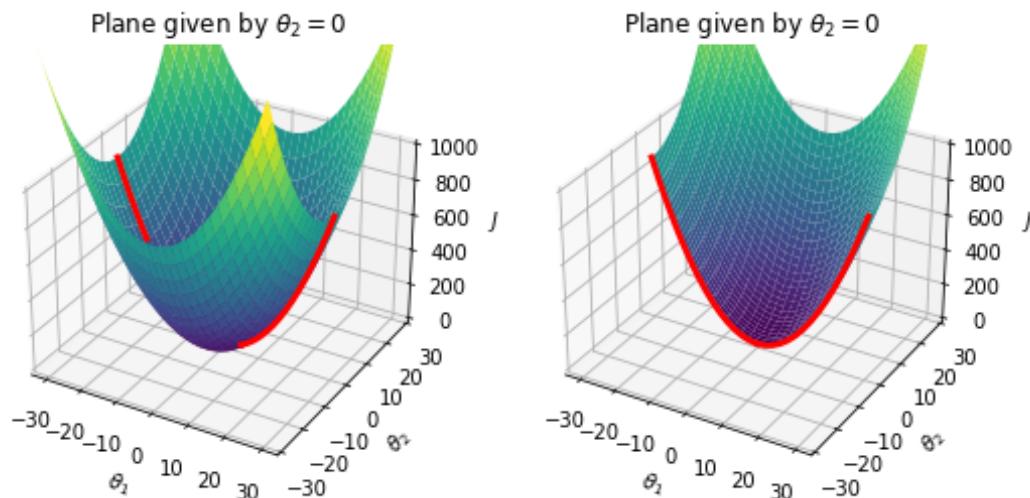
We do this by turning one variable constant: take  $\theta_2 = 0$ . Now, we have one free variable  $\theta_1$ . Same as 1-D.

Remember that, by 2-D, we mean two **parameters**/inputs to  $J$ . If we add in the **height** of our function, that means our plot will **look** like 3-D!

#### Concept 139

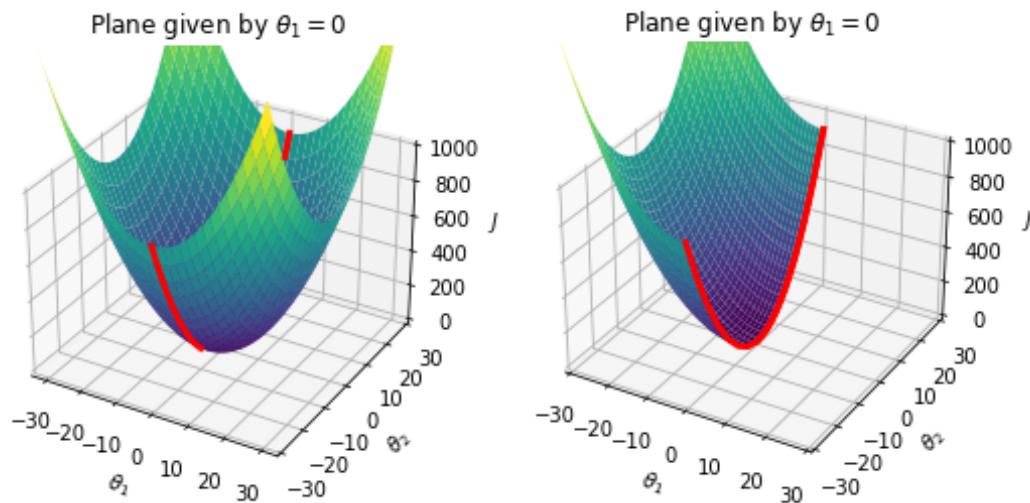
We can **reduce** the number of **variables** we have to work with, by holding some of them **constant**. That way, we have a **simpler** problem to work with.

This is **the same** as taking a single 2-D plane in a 3-D plot.



If we focus on a single plane of this surface, we end up with a **parabola**.

We can do the same the other way: we take  $\theta_1 = 0$ , and now we have a 1-D problem in  $\theta_2$ .



We can slice along the other axis as well!

Along each **axis**,  $\theta_1$  and  $\theta_2$ , you can **approximate** our function as **two** different straight lines. Which leads into our next point...

- Second way: if we take the two perpendicular **lines** we got from each dimension, we can combine them into a **plane**.

#### Concept 140

If we have **two input variables** (a 2-D problem), we can **approximate** our surface as a **plane** if we **zoom** in enough.

If you look closely enough at any smooth surface in 3D space, it will look roughly "flat".

**Example:** The earth tends to look flat up close, even though it's a sphere.

These **approximations** will allow us to **optimize**.

### 3.2.2 2-D: One dimension at a time

How do we **improve** our function  $J$ ? Now that we have **two** dimensions, we have to store our change  $\Delta\theta$  in a **vector**:

$$\Delta\theta = \begin{bmatrix} \Delta\theta_1 \\ \Delta\theta_2 \end{bmatrix} \quad (3.11)$$

This **complicates** things: we have two different things to consider **at once**.

Well, the **simplest** way would be to treat it as a **1-D** problem, and do exactly what we did **before**.

$$\Delta\theta_1 = \frac{\partial J}{\partial\theta_1} \quad (3.12)$$

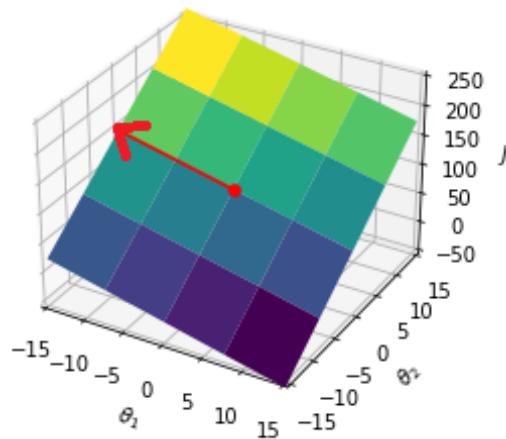
Note that we switched to **partial** derivatives, because we have **multiple** input variables  $\theta_i$ .

Writing this in our **new** notation, we get:

$$\Delta\theta = -\eta \begin{bmatrix} \partial J / \partial\theta_1 \\ 0 \end{bmatrix} \quad (3.13)$$

And then we would take a **step**, moving along the  $\theta_1$  **axis**.

Movement in  $\theta_1$  on  $J$



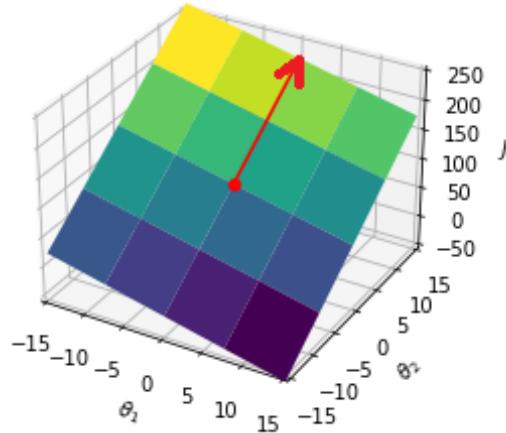
We can move along  $\theta_1$  just like on a line.

What if we treated this as a 1-D problem for the **other** variable,  $\theta_2$ ?

$$\Delta\theta = -\eta \begin{bmatrix} 0 \\ \partial J / \partial \theta_2 \end{bmatrix} \quad (3.14)$$

With this equation, we would be **moving** along the  $\theta_2$  axis.

Movement in  $\theta_2$  on  $J$



We can do the same with  $\theta_2$ .

Why not move in **both** directions **at once**? We can **combine** our two derivatives: we'll add up our two steps.

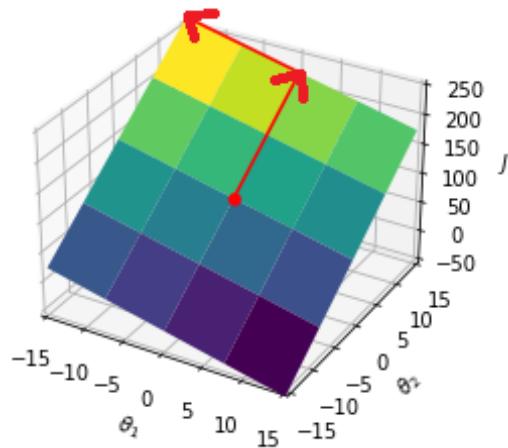
**Linearity** means that I can **add** them up without anything **weird** happening.

$$\Delta\theta = -\eta \begin{bmatrix} \partial J / \partial \theta_1 \\ 0 \end{bmatrix} - \eta \begin{bmatrix} 0 \\ \partial J / \partial \theta_2 \end{bmatrix} \quad (3.15)$$

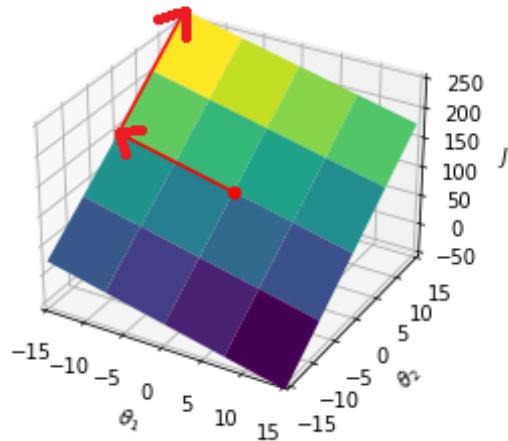
The relevant linearity rule:  $L(x + y) = L(x) + L(y)$ . In other words: taking two separate steps is the same as one big step.

These can be combined because we're treating our function as a **flat** plane: if I move in the  $\theta_1$  direction first, it doesn't change the  $\theta_2$  slope, and vice versa.

Combining two movements



Combining two movements



Our plane being flat means we can take both operations, back-to-back! Notice that the order doesn't matter.

$$\Delta\theta = -\eta \begin{bmatrix} \partial J / \partial \theta_1 \\ \partial J / \partial \theta_2 \end{bmatrix} \quad (3.16)$$

So, let's use that to optimize:

**Key Equation 141**

In **2-D**, you can optimize your function  $J$  using this rule:

$$\theta_{\text{new}} = \theta_{\text{old}} - \eta \begin{bmatrix} \frac{\partial J}{\partial \theta_1} \\ \frac{\partial J}{\partial \theta_2} \end{bmatrix}$$

Using  $\theta_{\text{old}}$

This is our **gradient descent** rule for 2-D.

This sort of approach makes some **sense**: if  $\frac{\partial J}{\partial \theta_1}$  is **bigger** than  $\frac{\partial J}{\partial \theta_2}$ , that means that you can get **more benefit** from moving in the  $\theta_1$  direction than  $\theta_2$ .

So, in that case, your step will move more in the  $\theta_1$  direction: it's a more **efficient** way to get a **better** hypothesis!

But for now, we **don't know** that this is necessarily the **optimal** way to change  $\theta$  - we'll explore that later.

**3.2.3 Gradient Descent in n-D**

This idea can be built up in **any number** of dimensions: each variable  $\theta_k$  creates a **different** line we can use to **approximate**.

And, we can combine them into a **flat hyperplane** : so, we can **add up** all of the different **derivatives**.

A hyperplane is just the equivalent of a plane in a higher dimensional space.

**Key Equation 142**

In **n-D**, you can optimize your function  $J$  using this rule:

$$\theta_{\text{new}} = \theta_{\text{old}} - \eta \begin{bmatrix} \frac{\partial J}{\partial \theta_1} \\ \frac{\partial J}{\partial \theta_2} \\ \vdots \\ \frac{\partial J}{\partial \theta_d} \end{bmatrix}$$

Using  $\theta_{\text{old}}$

In 3-D space, a plane fills one "slice" of 2-D space. In 4-D space, the hyperplane fills up one "slice" of 3-D space.

This is our **generalized gradient descent** rule.

**3.2.4 The Gradient**

We call this **gradient** descent because that right term we just invented **is** the gradient!

**Definition 143**

The gradient can be written as

$$\nabla_{\theta} J = \begin{bmatrix} \frac{\partial J}{\partial \theta_1} \\ \frac{\partial J}{\partial \theta_2} \\ \vdots \\ \frac{\partial J}{\partial \theta_d} \end{bmatrix} = \frac{dJ}{d\theta}$$

So, our rule can be rewritten (for the last time) as:

**Key Equation 144**

The **gradient descent** rule can be generally written as:

$$\theta_{\text{new}} = \theta_{\text{old}} - \eta \nabla_{\theta} J(\theta_{\text{old}})$$

$\theta_{\text{old}}$  is the input to  $\nabla_{\theta} J$ , not multiplication!

$$\theta_{\text{new}} = \theta_{\text{old}} - \eta J'(\theta_{\text{old}})$$

Now, is  $\nabla_{\theta} J$  the **optimal** way to improve(optimize)  $\theta$ ? Let's find out.

### 3.2.5 The Plane Approximation

So, what is the best direction? Which way will increase/decrease  $J$  **fastest**?

Is it the gradient? Let's explore a bit to figure that out. Let's look at our plane, and see what hints it might provide: \_\_\_\_\_

For explanation purposes, we'll assume 2-D, but the explanation extends to n-D.

**Concept 145**

Assume your function is, at least locally, a **flat plane**.

- A **flat plane** has only **one** direction of **maximum increase**: this is the direction you might call, "directly **uphill**" if you think of elevation.
- The **opposite** direction is the direction of **maximum decrease**, or "**downhill**".
- If you move at a **right angle** to the "best" direction (maximum increase/decrease), the function **will not change**. In elevation, you stay at the **same height**!

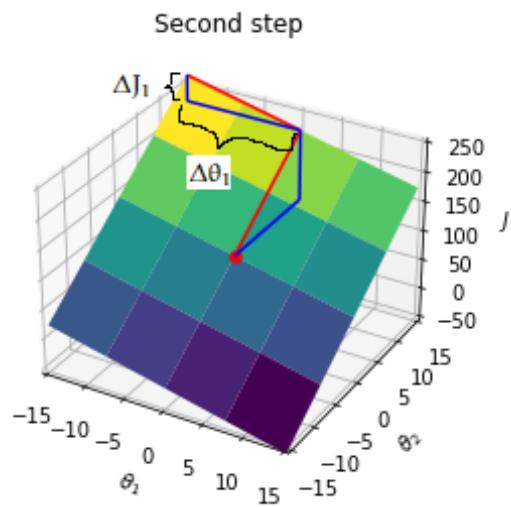
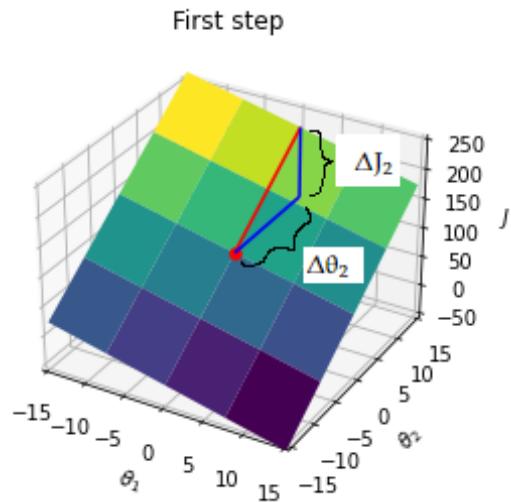
This is useful! We can **break down** any direction into the part that **affects** our function  $J$ , and the part that **doesn't**: \_\_\_\_\_

In the n-D case, we have **more** perpendicular directions. But, all of them have **no effect**!

### 3.2.6 The Optimal Direction: The Gradient

How do we get the optimal direction?

The **total** change in  $J$  is gotten by just **adding** the change in each direction (planes are simple, which makes this possible!):



You can add up the results of our two steps:  $\Delta J_2$  and  $\Delta J_1$ .

$$\Delta J \approx \Delta J_1 + \Delta J_2 \quad (3.17)$$

Let's convert that using derivatives:

$$\Delta J \approx \Delta \theta_1 \frac{\partial J}{\partial \theta_1} + \Delta \theta_2 \frac{\partial J}{\partial \theta_2} \quad (3.18)$$

Now we've got a useful equation: the total change. As a bonus we can see a clear **pattern**

$(\Delta\theta_i$  matches  $i^{\text{th}}$  derivative).

So, **condense** this pattern, like we did for our linear model: using a **dot product**.

$$\Delta J \approx \begin{bmatrix} \Delta\theta_1 \\ \Delta\theta_2 \end{bmatrix} \cdot \begin{bmatrix} \partial J / \partial \theta_1 \\ \partial J / \partial \theta_2 \end{bmatrix} = \Delta\theta \cdot \nabla_{\theta} J \quad (3.19)$$

The **gradient** shows up! Interesting. But what does that **mean**?

Well, we want to **maximize** (or minimize!) our  $\Delta J$ . How do we maximize a **dot product**?

- A dot product  $a \cdot b$  is maximized when the directions of  $a$  and  $b$  are **the same**!
- The direction of the gradient was given to us by our calculations.
- So, we want  $\Delta\theta$  to match the **gradient**! That way, they're in the same direction.

Maximizing this dot product means maximizing  $\Delta J$ , which is our goal.

So, we just demonstrated that the **gradient** gives us the **best** direction for  $\Delta\theta$ .

So, all we have to do is to **flip** the sign to **minimize**  $\Delta J$ .

And so, gradient descent is complete!

#### Concept 146

The **gradient**  $\nabla J$  is the **direction of greatest increase** for  $J$ .

That means the opposite direction  $-\nabla J$  is the **direction of greatest decrease** in  $J$ .

This is the single **most important concept** in this entire chapter!

### 3.2.7 Termination Condition

We can still use our termination conditions from before, but we need to be careful to make sure they extrapolate to n-D.

- Stop after a fixed  $T$  steps.
  - Nothing to change here.
- Stop when  $\|\theta\|$  isn't changing much:  $\|\Delta\theta\| < \epsilon$ , for example.
  - We just had to replace **absolute value** with **magnitude**.
- Stop when the derivative is small:  $|J'(\theta)| < \epsilon$ 
  - Nothing to change here.

We don't use this one often, though!

### 3.2.8 Another explanation of gradient (OPTIONAL)

Some students may not like the first explanation given for why gradient is the **direction of greatest increase**. So here, we use a slightly **different** approach, one that's more **geometric**.

Feel free to skip this section if you are not interested.

We look at a random 2-D vector,  $\Delta\theta$  - no assurances about how good or bad it is.

Currently, our vector **components** are based on  $\theta_1$  and  $\theta_2$ . But, it can be useful to **switch** perspectives.

Our vector can **also** be broken up into **parts** based on whether it **affects**  $J$ . This will let us take a **look** at the "best direction" we're trying to **find**.

- Uphill: the "best" direction  $\hat{u}_{best}$  (magnitude  $\Delta B$ )
- Same height: the direction with no effect,  $\hat{u}_{none}$  (magnitude  $\Delta N$ )

As we established before, these two directions are perpendicular on the plane.

$$\Delta\theta = \underbrace{\hat{u}_{best}}_{\text{Full effect on } J} + \underbrace{\hat{u}_{none}}_{\text{No effect on } J} = \Delta B * \hat{u}_{best} + \Delta N * \hat{u}_{none} \quad (3.20)$$

So, all of the change in  $J$  just comes from  $\hat{u}_{best}$ . We **don't care** about the other direction!

What about higher dimensions?

If we have  $k$  more dimensions, we just include  $k$  more unit vectors which add nothing to  $\Delta J$ .

#### Concept 147

In a local planar approximation, the **only** component of  $\Delta\theta$  that **affects**  $J$  is the **direction of greatest increase**,  $\hat{u}_{best}$ .

So, we can determine  $\Delta J$  using **only that component**.

Thus,  $\hat{u}_{best}$  gives us the "direction of greatest increase": if we rotate our vector, we replace some of  $\hat{u}_{best}$  with  $\hat{u}_{none}$ .

- In which case, we're losing some  $\Delta J$ . So, we don't want to change direction like that.

However, this is the same kind of behavior as the dot product:

$$\Delta J \approx \Delta\theta \cdot \nabla_\theta J \quad (3.21)$$

- When we do the dot product  $a \cdot b$ , we take the **projection** of  $a$  onto  $b$ : we only take the component of  $a$  in the same direction as  $b$ .
- In this case, we only include the **component** of  $\Delta\theta$  which matches  $\nabla_\theta J$ .

So, let's compare the two.

- The  $u_{best}$  component of  $\Delta\theta$  is the only part that matters for  $\Delta J$ .
- The  $\nabla_\theta J$  component of  $\Delta\theta$  is the only part that matters for the **dot product**, which equals  $\Delta J$ .

They're the same!

### 3.3 Application to Regression

One nice thing about **gradient descent** is that it is **easy** to switch the kind of problem you're applying it to: all you need is your **parameters**(s)  $\theta$ , and a function to optimize,  $J$ .

From there, you can just **compute** the gradient.

#### 3.3.1 Ordinary Least Squares

Our **loss** function is

$$J(\theta, \theta_0) = \frac{1}{n} \sum_{i=1}^n \left( (\theta^T x^{(i)} + \theta_0) - y^{(i)} \right)^2 \quad (3.22)$$

Or, in **matrix** terms,

Including the appended row of 1's from before.

$$J = \frac{1}{n} (\tilde{X}\theta - \tilde{Y})^T (\tilde{X}\theta - \tilde{Y})$$

Our gradient, according to **matrix derivative** rules, is

$$\nabla_{\theta} J(\theta) = \frac{2}{n} \tilde{X}^T (\tilde{X}\theta - \tilde{Y}) \quad (3.23)$$

Before, we set it equal to **zero**. But here, we can instead take **steps** towards the solution, using **gradient descent**.

We could use the **matrix** form, but sometimes it's easier to use a **sum**. Fortunately, derivatives are easy with a sum. If so, here's **another** way to write it:

$$\nabla_{\theta} J(\theta) = \frac{2}{n} \sum_{i=1}^n (\theta^T x^{(i)} - y^{(i)}) x^{(i)} \quad (3.24)$$

Either way, we use gradient descent **normally**:

Remember that  $\theta_{\text{old}}$  is an **input** to the gradient, not multiplied by it!

$$\theta_{\text{new}} = \theta_{\text{old}} - \eta \nabla_{\theta} J(\theta_{\text{old}})$$

Using  $\theta^{(t)}$  notation:

$$\theta^{(t)} = \theta^{(t-1)} - \eta \nabla_{\theta} J(\theta^{(t-1)})$$

#### 3.3.2 Ridge Regression

Ridge regression is similar.

$$J(\theta) = \frac{1}{n} \sum_{i=1}^n \left( \underbrace{(\theta^T x^{(i)} + \theta_0)}_{\text{guess}} - \underbrace{y^{(i)}}_{\text{answer}} \right)^2 + \underbrace{\lambda \|\theta\|^2}_{\text{Regularizer}}$$

However, we have to treat  $\theta_0$  as **separate** from our other data points, because of **regularization**: remember that it **doesn't** apply to  $\theta_0$ .

For  $\theta$ :

$$\nabla_{\theta} J_{\text{ridge}}(\theta, \theta_0) = \frac{2}{n} \sum_{i=1}^n \left( (\theta^T x^{(i)} + \theta_0) - y^{(i)} \right) x^{(i)} + 2\lambda\theta \quad (3.25)$$

For  $\theta_0$ :

$$\frac{\partial J_{\text{ridge}}(\theta, \theta_0)}{\partial \theta_0} = \frac{2}{n} \sum_{i=1}^n \left( (\theta^T x^{(i)} + \theta_0) - y^{(i)} \right) \quad (3.26)$$

Notice that we used a **gradient** for our vector  $\theta$ , but since  $\theta_0$  is a single variable, we just used a **simple derivative**!

#### Concept 148

The **gradient**  $\frac{dJ}{d\theta}$  must have the **same shape as  $\theta$** : this shape-matching is why we can easily **subtract** it during gradient descent.

$$\underbrace{\theta_{\text{new}}}_{(d \times 1)} = \underbrace{\theta_{\text{old}}}_{(d \times 1)} - \eta \underbrace{\nabla_{\theta} J(\theta_{\text{old}})}_{(d \times 1)}$$

The derivative  $\frac{dJ}{d\theta_0}$  is based on  $\theta_0$ , a constant. So, the shape must be  $(1 \times 1)$ .

$$\theta_0^{(t)} = \theta_0^{(t-1)} - \eta \frac{dJ}{d\theta_0} \Big|_{\theta_0=\theta_0^{(t-1)}}$$

### 3.3.3 Computational Gradient

Sometimes, we **can't** easily find the **equation** for our gradient: maybe our loss isn't a simple **equation**, or we have some **other** kind of problem. So, rather than getting the **exact** gradient, we **approximate** it.

But how do we **approximate** the gradient? Well, first, we could **reference** how we approximate a **simple derivative**.

The definition of the **derivative** can be gotten as

A derivative is just a 1-D gradient, after all!

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} \quad (3.27)$$

But, what if we can't take the **limit**? Or, we just don't **want** to?

We can **approximate** by taking  $h$  to be a small, **finite** number.

Instead of  $h$ , we'll call this  $\delta$ .

### Concept 149

When **approximating** the derivative, we can choose a **small** finite width to measure, called  $\delta$ , so that

$$\frac{df}{dx} \approx \frac{f(x+\delta) - f(x)}{\delta}, \quad \delta \ll 1$$

So, let's **extend** that to the **gradient**:

$$\nabla_{\theta} J = \begin{bmatrix} \partial J / \partial \theta_1 \\ \partial J / \partial \theta_2 \\ \vdots \\ \partial J / \partial \theta_d \end{bmatrix} \quad (3.28)$$

Luckily, the **gradient** is just a bunch of derivatives **stacked** in a **vector**!

So, we can just **compute** each of them **separately**, and then put them together.

Let's show how we'd **write** that in **vector** form, for just one of them. We want something like

$$J'(\theta) \approx \underbrace{\frac{J(\theta + \delta) - f(\theta)}{\delta}}_{\text{Not correct, but closer}} \quad (3.29)$$

This isn't quite right, because a **scalar**  $\delta$  would **add to every term**.

We **only** want to shift **one** variable at a time, so we can do a **simple** derivative.

Let's say we want  $dJ/d\theta_1$ . We would **only** want to add  $\delta$  to  $\theta_1$ : the other parameters are **unchanged**.

So, we **can't** add a **scalar**. Instead, we need a  $(d \times 1)$  vector: one term to **separately** add to each  $\theta_k$  term.

$$\Delta\theta = \begin{bmatrix} \Delta\theta_1 \\ \Delta\theta_2 \\ \vdots \\ \Delta\theta_d \end{bmatrix} \quad (3.30)$$

We want most terms **unchanged**, so we'll **add 0** to each of them, and we'll add  $\delta$  to the one term we want to **edit**.

$$\Delta\theta = \begin{bmatrix} \delta \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \quad (3.31)$$

We'll **create** one of these vectors for each **dimension**. We'll give them a special **name**:  $\delta_k$ , for the  $k^{\text{th}}$  dimension.

$$\delta_1 = \begin{bmatrix} \delta \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \quad \delta_2 = \begin{bmatrix} 0 \\ \delta \\ \vdots \\ 0 \\ 0 \end{bmatrix} \quad \delta_{d-1} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \delta \\ 0 \end{bmatrix} \quad \delta_d = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \delta \end{bmatrix} \quad (3.32)$$

Finally, we'll **divide** by  $\delta$ . We have what we need for our full equation:

### Key Equation 150

In order to **computationally find the gradient**, you need to find the **partial derivative** for each term  $\theta_k$ .

$$\frac{dJ}{d\theta_k} \approx \frac{J(\theta + \delta_k) - J(\theta)}{\delta}$$

Where

- $\delta$  is a small positive number
- $\delta_k$  is the  $(d \times 1)$  **column vector** with a  $\delta$  in the  $k^{\text{th}}$  row, and a 0 in every other row.

### 3.3.4 Problems with Gradient Descent

Gradient descent is very handy, but it's important to be aware of some of its **problems**.

We've discussed a couple: diverging, oscillating, and converging slowly. We also have to

worry about **local minima** that aren't as good as other answers.

But there's also a **requirement**: our loss function has to be **smooth** and **differentiable**. If it isn't, we can't take the **gradient** of it.

**Concept 151**

**Gradient descent** requires for your **functions** to be (at least mostly) **smooth and differentiable**.

Our **answer** is also only as good as our **loss function**: if our loss function is not good for what we actually want to **accomplish**, then we can easily create a **bad** model.

## 3.4 Stochastic Gradient Descent

### 3.4.1 Another problem with gradient descent

Some **benefits** of gradient descent come from the fact that GD **gradually** improves:

- You can pause early to check progress, or quit early if your model is good enough. We save on computation time.

But we can improve this feature: currently, we use a **sum** of all data points to get our gradient. Meaning, we have to compute all of our data before we take a single step.

### 3.4.2 A better way: stochastic GD

Instead, why wait until we have **added** up over all the data? We could just **compute** the gradient over **one** data point **at a time**. In fact, to be fair, we'll do it **randomly**.

But wait, this **seems** like it would be **less** effective - after all, how much does **one** data point tell you?

To compensate for using less data, our steps will have to be **smaller**!

Well, even if it isn't much, this isn't very **different** from adding them up all at **once**: in **theory**, taking lots of **little** steps should average out to the **same** information as if we do it all at once.

#### Definition 152

**Stochastic Gradient Descent (SGD)** is the process of applying **gradient descent** on **randomly** selected data points.

This should **average** out to being **similar** to regular (batch) gradient descent, but the **randomness** often lets it improve **faster** and **avoid** some common problems.

There are more possible benefits, too: **randomly** choosing data points adds some **noise**, and random movement might be able to pull us out of local minima we don't want.

Stochastic is just a very mathematically precise word for "random".

This sort of **noise** and **randomness** can make it hard for our model to **perfectly** fit the training data: this can reduce **overfitting**, too!

We mean "noise" in the signals sense: random **variation** in our data. Randomly choosing data points is more unpredictable than using all of our data.

The random selection makes the data "look different" each time, so it's hard to perfectly match it.

**Concept 153**

There are many **benefits** to **SGD** (Stochastic Gradient Descent) over regular BGD (Batch Gradient Descent).

- SGD can sometimes **learn** a good model **without** using all of our **data**, which can **save us time** when data sets are **too large**.
  - It can also let us address problems **early** if the model **isn't** improving.
- The noise produced by the random sampling in SGD can sometimes help it **avoid local minima**.
  - This is because the model might be moved in a **random direction** in **parameter space**, and randomly **pulled out** of that minimum, even if BGD would have gotten **stuck**.
- The noise also **reduces overfitting**, because it's **harder** for the model to **memorize** the exact details of the **distribution**.

### 3.4.3 Ensuring Convergence

How do we make sure that our SGD method converges? We need some kind of termination criteria. Thankfully, there's a useful theorem on the matter:

**Theorem 154**

SGD **converges** with *probability one* to the **optimal**  $\Theta$  if

- $f$  is convex

And our step size(learning rule)  $\eta(t)$  follows these rules:

$$\sum_{t=1}^{\infty} \eta(t) = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \eta(t)^2 < \infty$$

Why these rules? Let's see:

- The **first** rule is for the **same** reason as for regular BGD: if it isn't **convex**, we can get stuck in **local minima**, or if it's **concave**, decrease **forever**.
- The **second** rule means that your steps need to add up to an **infinite distance**: this allows you to reach **any** possible point in your **parameter space**.
- The **third** one is a bit **trickier**, but basically means the steps need to get **smaller**, so we can approach the **minimum** (otherwise we might **diverge**!)

One option is  $\eta(t) = 1/t$ . But often, we use rules that **decrease more slowly**, so that it doesn't take as **long**.

In this case, though, we're often **no longer** guaranteed convergence.

## 3.5 Terms

- Gradient Descent
- Parameter Space
- Local view (Calculus)
- Linear Function Approximation
- Planar Function Approximations
- Convergence
- Divergence
- Oscillation
- Step size
- Termination Condition
- Concavity/Convexity
- Global Minimum
- Local Minimum
- Initialization
- Gradient (Direction of Maximum Increase)
- Gradient Descent Rule
- Gradient Shape
- Gradient Approximation
- Stochastic Gradient Descent
- Batch Gradient Descent
- BGD Convergence Theorem
- SGD Convergence Theorem

Why? Because of our third condition,  $\sum_t \eta(t) < \infty$ . If  $\eta(t)$  decreases too slowly, this sum goes to infinity, and our GD algorithm might **diverge**.

# CHAPTER 4

---

## Classification

---

### 4.0.1 Regression (Review)

In chapter 2, we handled the problem of **Regression**: taking in lots of data (stored as a **vector of real numbers**), and returning another single **real number**.

Remember that we used a  $(d \times 1)$  column vector for our data points  $x^{(i)}$ .

$$h_{reg} : \mathbb{R}^d \rightarrow \mathbb{R} \quad (4.1)$$

This was good for when we wanted to predict some **numeric** output: stock prices, height, life expectancy, and so on.

But, this isn't the **only** type of problem we might have to deal with.

## 4.1 Classification

### 4.1.1 Motivation: Putting things into classes

We don't *always* want a **real number** output: sometimes, we just have a different kind of question.

Often, it's more useful to, rather than give numeric values, instead sort things into **categories**, or what we will call **classes**.

#### Definition 155

A **class** is **set** of things that have something relevant in **common**.

**Example:** A beagle and a golden retriever could both be put in a **class** called "dog". This is useful if you just want to know whether you have a dog or not!

### 4.1.2 What is classification?

This is the goal of **classification**: we want to take lots of **information**, and use them to **predict** what **class** a data point belongs in.

#### Definition 156

**Classification** is the **machine learning problem** of sorting items into different, **discrete** classes.

In this setting, we take **real-valued data**, stored in a  $(d \times 1)$  **vector**, and return one of our **classes**.

$$h : \mathbb{R}^d \rightarrow \{C_1, C_2, C_3, \dots C_n\}$$

Where  $\{C_1, C_2, C_3, \dots C_n\}$  are all **classes**. Sometimes, we call the value we return a **label** instead.

**Example:** Suppose you want to classify different **animals** as a bird, a mammal, or a fish. You are given (as input) 5 pieces of useful data to **classify** with.

As a refresher, the function notation here just says, "take in a  $d$ -dimensional vector, and output one of our  $n$  discrete classes."

$$h : \mathbb{R}^5 \rightarrow \{\text{Bird, Mammal, Fish}\} \quad (4.2)$$

**Classification** can be useful for lots of situations:

- **Deciding** which **action** to take in a difficult situation
- **Diagnosing** a patient, and determining the best **treatment**
- **Sort** information to be **processed** later
- And more!

Just like with regression, we can depict our **hypothesis** as the function

$$x \rightarrow [h] \rightarrow y \quad (4.3)$$

**Concept 157**

**Classification** is also **supervised**: meaning, you have **training** data  $\mathcal{D}_n$  with the **correct** answers given:

$$\mathcal{D}_n = \left\{ \left( x^{(1)}, y^{(1)} \right), \dots, \left( x^{(n)}, y^{(n)} \right) \right\}$$

In **unsupervised** problems, you're not told the "correct" answer and have to just guess one!

### 4.1.3 Important Facts about Classes

There's a few important things we should remember about classes moving forward.

- Classes are **discrete**: each class is a distinct "thing", **separate** from other classes.
  - This is unlike real numbers, which are **continuous**: you can **smoothly** transition between them.
- This isn't **always** true, but usually, classes are **finite**: there are only so many of them, which we write as  $n$ .
  - Meanwhile, there are **infinitely many** real numbers.
- These classes may not have a natural **order**: is there a correct way to order "[Bird, Mammal, Fish]"? Not really.
  - The real numbers are ordered, too.
- In some problems, you get to **decide** what classes you choose. Do you want to compare dogs vs. cats, mammals vs. fish, color of fur? What goes in different classes, or the same?
  - You can change units (lb vs kg), but you don't "decide" how the real numbers work.

**Concept 158**

**Classes** are

- **discrete**
- **finite** (usually)
- **not** necessarily **ordered**
- often **defined** based on your **needs**

#### 4.1.4 Binary Classification

So, how do we get **started**? Well, we want to create the **simplest** case, and maybe we can get the **general** idea.

**Two** is the **smallest** number of useful classes: often, this boils down to a **yes-or-no** question. Typically, we **represent** these two choices as  $+1$  and  $-1$ , respectively.

##### Definition 159

**Binary classification** is the **problem** of sorting elements into one of **two categories**.

Often, these categories are defined by a "**yes-or-no**" question.

$$h : \mathbb{R}^d \rightarrow \{-1, +1\}$$

**Example:** You could look at a person and say, "are they sick?" or, "is that a dog"? You can **classify** data in a binary way **based** on those questions.

#### 4.1.5 Classification Performance

And how do we measure how well this model is doing? The easiest way might be, "count the number of wrong guesses".

This is captured by **0-1 Loss**:

##### Definition 160

**0-1 Loss** is a way of measuring **classification** performance: if you get the **wrong** answer, you get a loss of **1**. If you're **right**, then **0** loss.

$$\mathcal{L}(g, a) = \begin{cases} 0 & \text{if } g = a \\ 1 & \text{otherwise} \end{cases}$$

This type of loss is as **simple** as we can get: similar to counting how many wrong answers you get on a **multiple-choice** test.

If we want to get our training error, we'll just average over the data points:

$$\mathcal{E}_n(h) = \frac{1}{n} \sum_{i=1}^n \begin{cases} 0 & \text{if } g_i = a_i \\ 1 & \text{otherwise} \end{cases}$$

Just like before, we care about **testing loss** more than **training loss**: we want our model to **generalize**.

This relies on our typical IID assumption from chapter 1.

$$\mathcal{E}(h) = \frac{1}{m} \sum_{i=n+1}^{n+m} \begin{cases} 0 & \text{if } g = a \\ 1 & \text{otherwise} \end{cases}$$

Next, we figure out what **model** we use to do our classification.

## 4.2 Linear Classifiers

If you wanted to break up your data into two parts (+1 and -1), how might you do it? Let's explore that question.

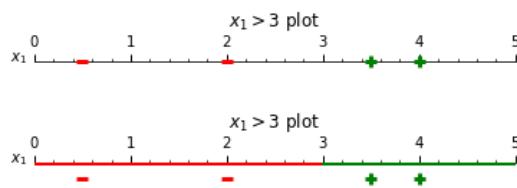
### 4.2.1 1-D Linear Classifiers

As usual, we'll start with the **simplest** case we can think of: 1-D. So, we only have one variable  $x_1$  to **classify** with.

The simplest version might be to just **split** our space in **half**: those above or below a certain **value**. This is our parameter,  $C$ .

$$x_1 > C \quad \text{or} \quad x_1 - C > 0 \quad (4.4)$$

**Example:** For the below data (where green gives positive and red gives negative), could classify positive as  $x_1 > 3$ .



We plot everything above  $x = 3$  as **positive**, and **negative** otherwise.

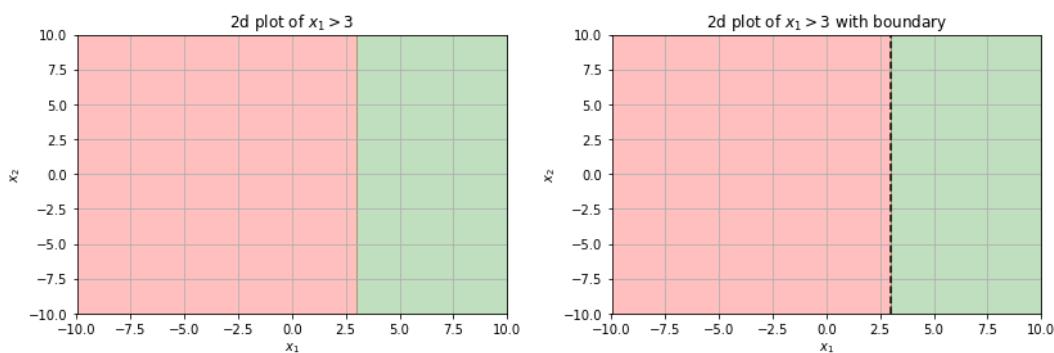
We could also call it  $\theta_0$ , in the spirit of our  $\theta$  notation for parameters.

$$x_1 + \theta_0 > 0 \quad (4.5)$$

### 4.2.2 1-D classifiers in 2-D

Let's add a variable and see how our classifier looks on a 2-D plot.

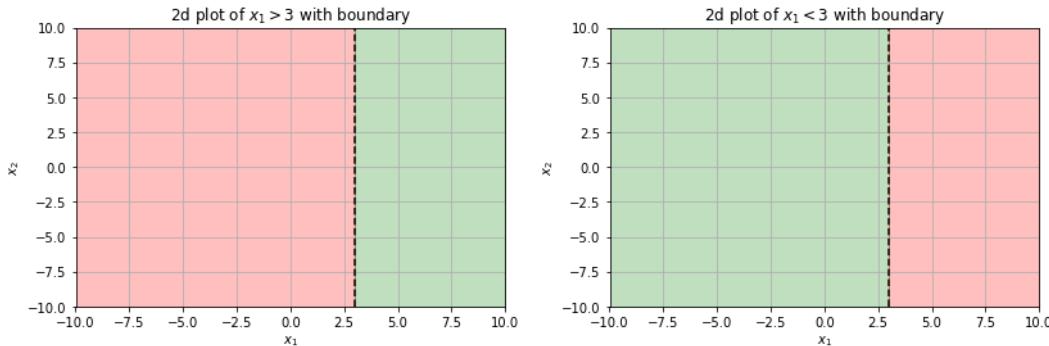
We'll omit the data points for now.



On the right, we've drawn the **dividing** line between our two regions.

Interesting - the **boundary** between positive and negative is defined by a **vertical line**.

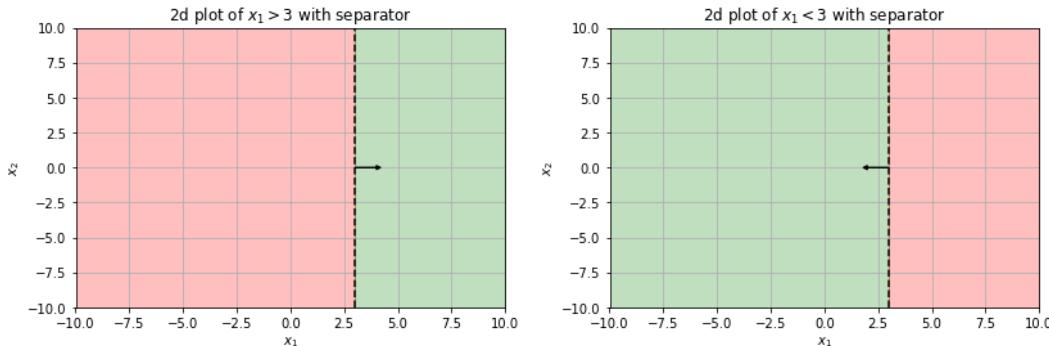
A vertical line is missing some information, though: compare  $x_1 > 3$  and  $x_1 < 3$ :



These two plots have the same line, but have their sides flipped.

So, we have a **line** that gives us the boundary, but we **also** need to include information about which way is the **positive** direction.

What tool best represents **direction**? We could use angles, but we haven't used that much so far. Instead, let's use a **vector** to **point** in the right direction.



Now, it's clear which plot is which, just using our **line** and **vector**!

The object that represents our classification is called a **separator**!

Since our variables are  $x_1$  and  $x_2$ , this is a separator in **input space**.

### Definition 161

A **separator** defines how we **separate** two different classes with our **hypothesis**.

It includes

- The **boundary**: the **surface** where we **switch** from one **class** to another.
- The **orientation**: a **description** of which **side** of the boundary is assigned to **which class**.

For example, let's take our specific separator from above.

### Concept 162

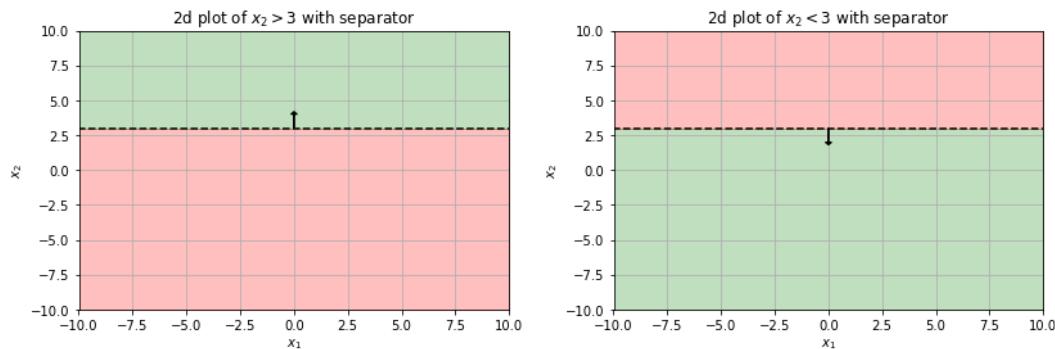
We can define our **1-D separator** using

- The **boundary** between the **positive** and **negative** regions: in 2-D input space, this looks like a vertical or horizontal **line**.
- A **vector** pointing towards whichever side is given a **+1 value**.

We call it "orientation" because you could imagine "flipping over" the space, so the positive and negative regions are swapped.

### 4.2.3 A second 1-D separator, and our problem

What if we use  $x_2$  to **separate** our data?

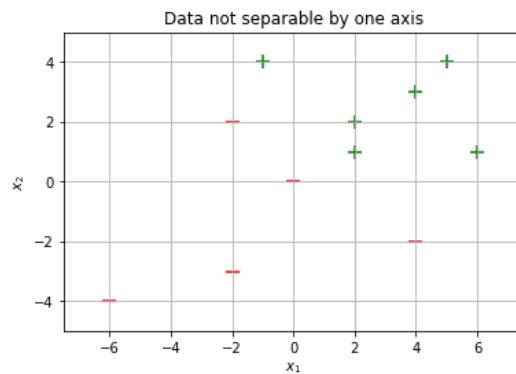


Instead of having a vertical separator, we have a **horizontal** one.

We get the same sort of plot along the **other axis**!

So, this is cool so far, but it's not a very **powerful** model: we can only handle a situation where the data is evenly divided by **one axis**.

And if that's the case, what's the point of our **other** variable?



There's no vertical or horizontal line we can use to split this space!

#### 4.2.4 The 2-D Separator: What vector do we use?

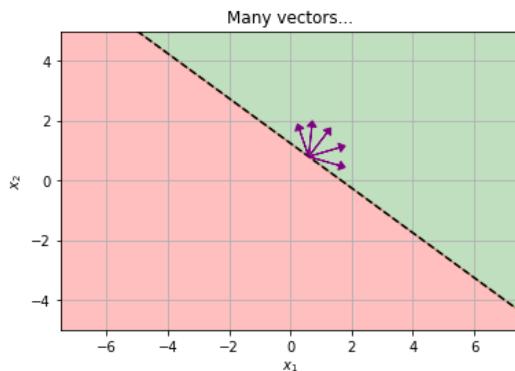
Just looking at our example, we might wonder, "well, if we can use **vertical** lines or **horizontal** lines, can't we just use a line in **another** orientation?"

It turns out, we **can**!



If we allow lines at an angle, we can classify all of our data correctly!

So, we've got our **boundary**. But we still need a vector to tell us which side is **positive**. But there are **many** possible vectors we could choose:



All of these vectors point towards the **correct** side of the plane. Is there a **best** one to use?

Above, we used the vector that was **vertical** or **horizontal**. This makes sense: if we're doing  $x_1 > 3$ , it seems reasonable to have the arrow **point** in the positive- $x_1$  direction.

But this vector also happened to be **perpendicular** to our **line**: this is the line's **normal vector**,  $\hat{n}$ . This vector has a couple nice properties:

- It is **unique**: in 2-D, there is only 1 **normal** direction. The opposite side is just  $-\hat{n}$ .
- It points directly **away** from the plane.
- If our plane is at the **origin**, any point with a **positive**  $\hat{n}$  component is on the **positive** side. This will be important later!

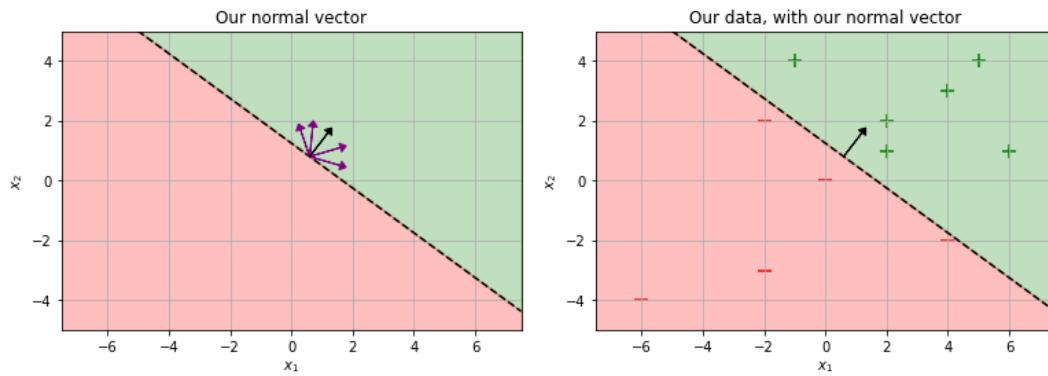
So, we have a **unique** vector that tells us which side is **positive**. Let's go with that!

### Concept 163

Every **line** in 2-D has a **unique normal vector** that can be used to **define** the **angle/direction** of the line.

The **direction** the vector is "facing" is also called the **orientation**.

Our normal vector for the above separator:



We can define our plane using the **normal vector**!

It's clear that this vector in some way is a **parameter**: if we change this vector, we get a different **orientation**, and a different **classifier**.

We have **represented** parameters in the past using  $\theta$ . We need **two** different  $\theta_k$ : one for the  $x_1$  component, another for the  $x_2$  component.

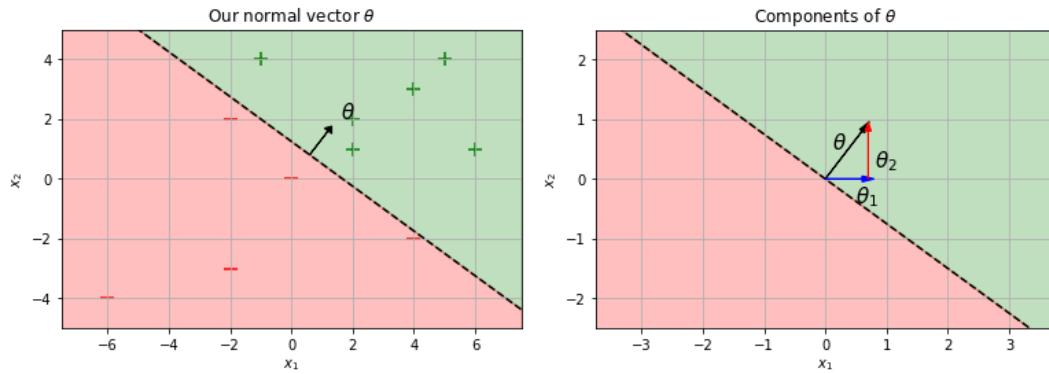
So, we'll use that.

### Notation 164

The vector  $\theta$  represents the **normal vector** to our line in 2D.

$$\hat{\theta} = \theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$$

We add this to our diagram:

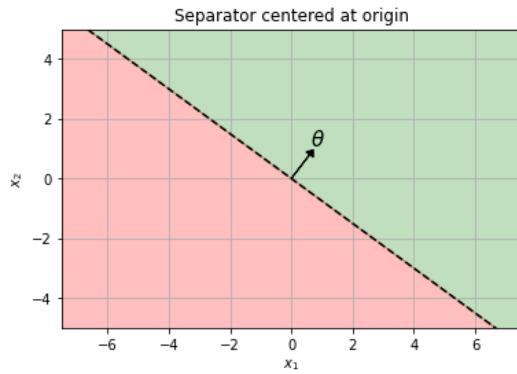


$\theta$  is our normal vector!

Nice work so far. The next question is: how do we describe this separator **mathematically**?

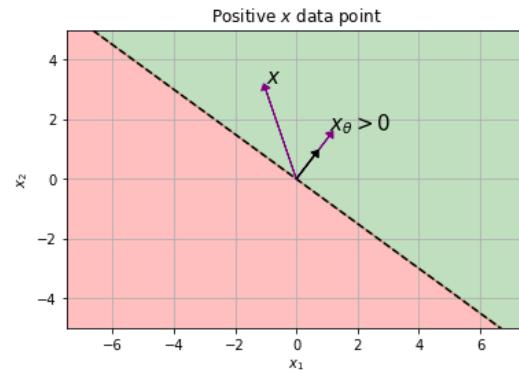
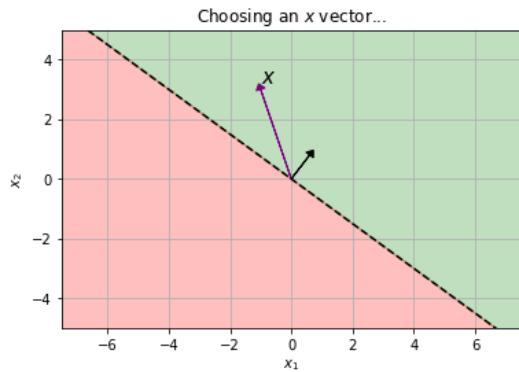
#### 4.2.5 2D Separator - Matching components

As always, we'll **simplify** the problem to make it more manageable: for now, we'll assume our **separator** is centered at the **origin**.

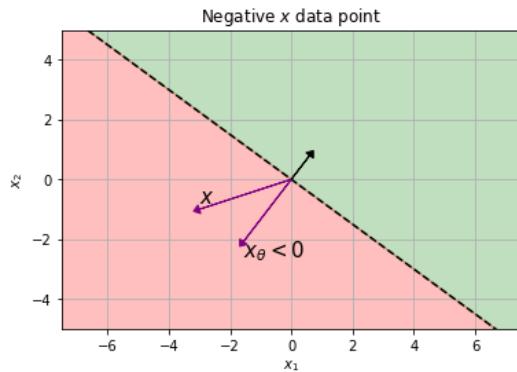


So, we have our vector,  $\hat{n}$ . As we mentioned above, anything on the **same** side as  $\hat{n}$  is **positive**, and anything on the **opposite** side is **negative**.

For a line on the origin, "On the same side of the line" can be interpreted as "has a positive  $\hat{n}$  component". We'll find that component next.



This vector has a **positive** component in the  $\theta$  direction.



This vector has a **negative** component in the  $\theta$  direction.

How do we represent "on the same side" mathematically? How do we **find** whether the component is **positive or negative**? We use the **dot product**.

#### 4.2.6 The Dot Product (Review)

How to calculate the dot product should be familiar to you, but we'll talk about some **intuition** that you may not be exposed to.

**Concept 165**

You can use the **dot product** between unit vectors to measure their "similarity": if two vectors are more **similar**, they have a **larger** dot product.

In the most clear cases, take unit vectors  $\hat{a}$  and  $\hat{b}$ :

- If they are in the **exact same** direction,  $\hat{a} \cdot \hat{b} = 1$
- If they are in the **exact opposite** direction,  $\hat{a} \cdot \hat{b} = -1$
- If they are **perpendicular** to each other,  $\hat{a} \cdot \hat{b} = 0$

Remember, **unit vectors** have a length of 1.

What about non-unit vectors?

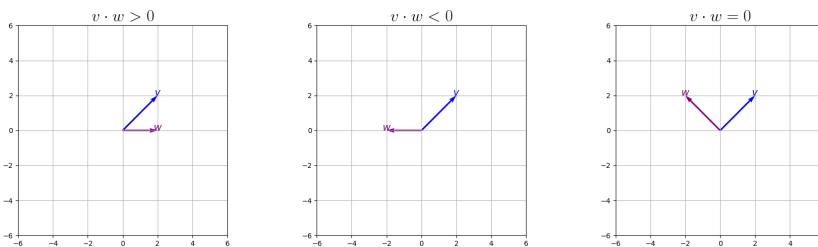
These unit vectors are then scaled up by the **magnitude** of each of our vectors. Because magnitudes are **always positive**, the dot product sign doesn't change.

**Concept 166**

You can use the **dot product** between non-unit vectors to measure their "similarity" **scaled by their magnitude**.

If two vectors are more **similar**, they have a **larger** dot product.

- If the vectors are **less** than  $90^\circ$  apart, they are more similar: they will share a **positive** component:  $\vec{a} \cdot \vec{b} > 0$
- If the vectors are **more** than  $90^\circ$  apart, they will share a **negative** component:  $\vec{a} \cdot \vec{b} < 0$
- If they are **perpendicular** ( $90^\circ$ ) to each other,  $\vec{a} \cdot \vec{b} = 0$

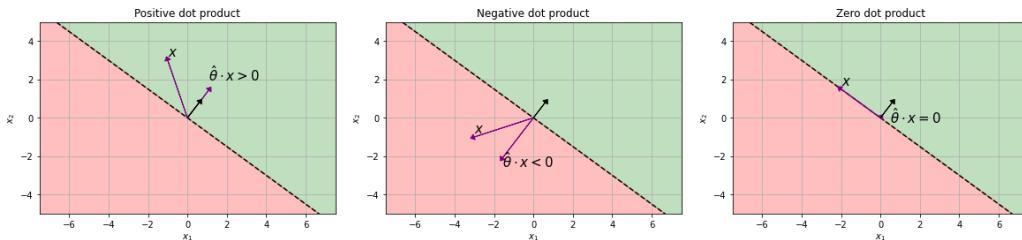


We can see here how  $v$  and  $w$  are "more similar" when  $v \cdot w > 0$ .

#### 4.2.7 Using the dot product

So, the **sign** of the dot product is a useful tool. If a point is on the line, it is **perpendicular** to  $\theta$ , our **normal vector**.

So, if a point has a **positive** dot product, it is on the **same side** as  $\theta$ , and if it's **negative**, it's on the **opposite side**.



Our dot product can tell us which region of space we're in.

So, we can classify things based on the **sign** of it. Written as an equation, we can define the sign function:

### Key Equation 167

For a **linear separator** centered on the **origin**, we can do **binary classification** using the hypothesis

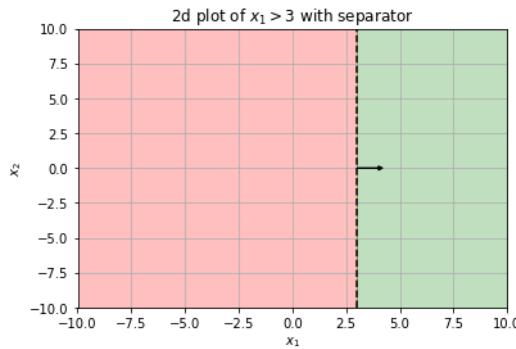
$$h(x; \theta) = \text{sign}(\theta \cdot x) = \begin{cases} +1 & \text{if } \theta \cdot x > 0 \\ -1 & \text{otherwise} \end{cases}$$

- Note that we assigned  $-1$  if  $\theta \cdot x = 0$ .
- This is an **arbitrary** convention: we could also have assigned  $+1$  for  $\theta \cdot x = 0$ .
- All that matters is to have a decision, and be consistent about it.

### 4.2.8 Introducing our offset

Now that we have handled the case where our linear separator is on the **origin**, we want to **shift** our separator **away** from it.

In our **1-D** case, we easily **shifted** away from the origin: any separator  $x_1 > C$  where  $C$  **isn't zero**, we shift by  $C$  units.



By making our inequality  $x_1 > 3$  **nonzero**, we moved away from the origin by 3 units!

We could make our inequality **nonzero**, then! That could move us **away** from the origin, just in a different **direction**.

Or, we could equivalently do this... Note:  $A \iff B$  means A and B are equivalent!

$$x_1 > 3 \iff x_1 - 3 > 0 \quad (4.6)$$

So, instead, we could just add a constant to our expression, which we will call  $\theta_0$ .

We'll also switch out  $\theta \cdot x = \theta^T x$ .

### Key Equation 168

A general **linear separator** can do **binary classification** using the hypothesis

$$h(x; \theta) = \text{sign}(\theta^T x + \theta_0) = \begin{cases} +1 & \text{if } \theta^T x + \theta_0 > 0 \\ -1 & \text{otherwise} \end{cases}$$

- Again, we assigned  $-1$  if  $\theta \cdot x = 0$ .
- Again, this choice is **arbitrary**. Consistency is what matters most.

Notice that this looks very similar to what we did in regression! We'll get into that in a bit.

First, a quick look at the components of our equation:

### Concept 169

For **binary classification**,  $\theta$  and  $\theta_0$  entirely **define** our **linear separator**.

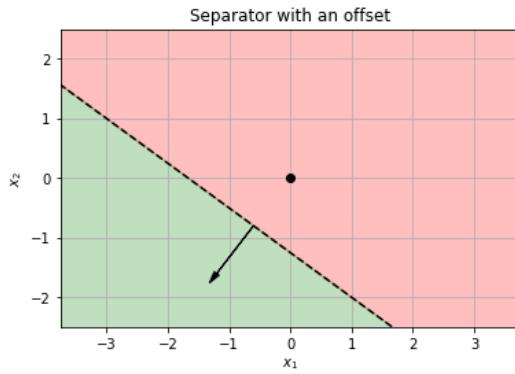
- $\theta$  gives us the **orientation** of our line.
- $\theta_0$  **shifts** that line around in **space**.

## 4.2.9 How does the offset affect our classifier?

So, how exactly does our offset  $\theta_0$  affect our **classifier**? Well, we mark our classifier with our **normal vector** and the **boundary**.

Our **normal vector** is entirely captured by  $\theta$ : it's unchanged by  $\theta_0$ .

What about our **boundary**? We have its **orientation**, but we don't know where it has **shifted** to.



Note that the origin has been marked.

Well, let's use our equation: if your formula is positive, you get +1. If your formula is negative, you get -1.

The **boundary** line is between positive and negative: it's at **zero**.

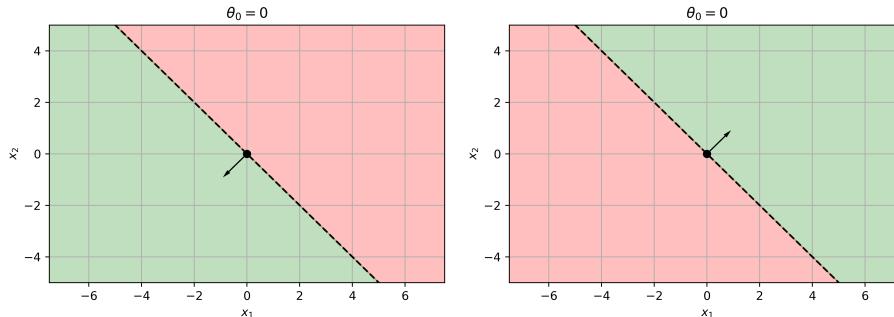
$$\theta^T x + \theta_0 = 0 \iff \theta^T x = -\theta_0 \quad (4.7)$$

We'll break the effects of  $\theta_0$  into three cases:

For each, we'll show a boundary, and a **flipped** version of that boundary.

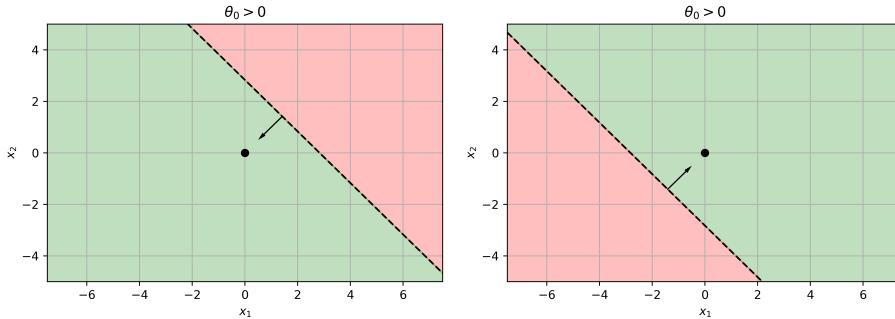
Note: the below statements are true no matter what  $\theta$  we choose!

- If  $\theta_0 = 0$ , then  $x = (0, 0)$  is **on the line**.
  - Without an **offset**, our line goes through the **origin**.



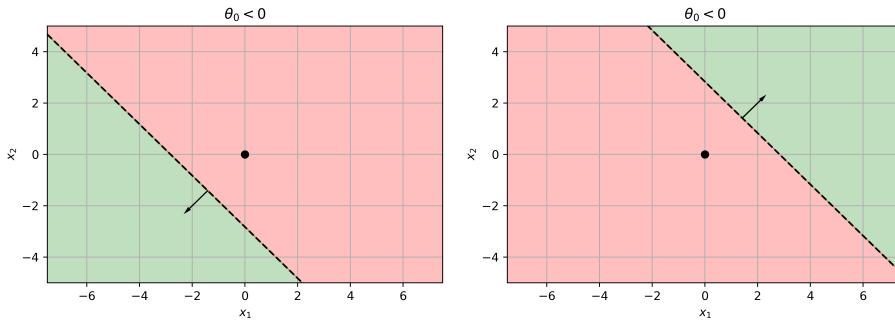
The boundary is on the origin.

- If  $\theta_0 > 0$ , then the **origin** is in the **positive** region.
  - That means the positive region is "larger": some space that was negative, is positive.
  - The boundary line has moved in the  $-\theta$  direction.



If we have a **positive** constant, it's "easier" to get a positive **result**: more positive space.

- If  $\theta_0 < 0$ , then the **origin** is in the **negative** region.
  - That means the positive region is "smaller": some space that was positive, is negative.
  - The boundary line has moved in the  $+\theta$  direction.



If we have a **negative** constant, it's "harder" to get a positive **result**: more negative space.

This can be a bit confusing, so we'll summarize:

### Concept 170

The **sign** of our  $\theta_0$  and the **direction** we move away from the origin are **opposite**.

If  $\theta_0 > 0$  (positive), our boundary moves in the  $-\theta$  **direction**.

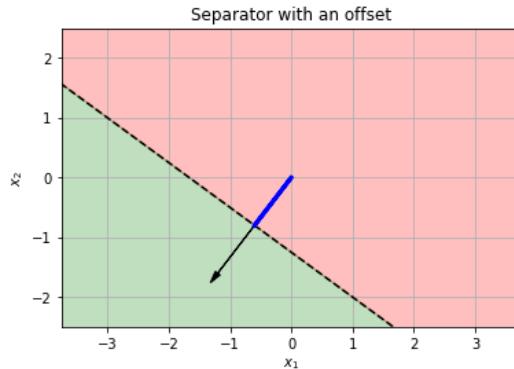
If  $\theta_0 < 0$  (negative), our boundary moves in the  $+\theta$  **direction**.

This gives us a general idea of how the offset affects it, but what is the **exact** effect of  $\theta_0$  on the line?

We'll focus on one point on the line: the **closest point to the origin**. We want to look at this **point** because it's **unique**.

Points that aren't unique are hard to keep track of!

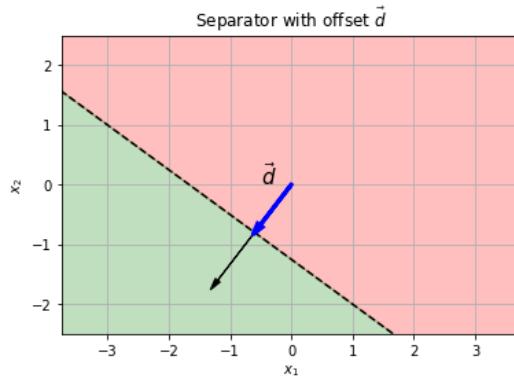
### 4.2.10 Distance from the Origin to the Plane



Notice that the **shortest** path from the origin to the line is **parallel** to  $\theta$ !

So, we can think of our **line** as having been **pushed** in the  $\theta$  direction. This **matches** what we did for 1-D separators:  $x_1 > 3$  was moved in the  $x_1$  direction.

So, we'll take the closest point on the line,  $\vec{d}$ . The **magnitude**  $d$  will give us the **distance** that the separator has been **shifted**.



Since  $\vec{d}$  is in the direction of  $\theta$ , the direction can be captured by the unit vector  $\hat{\theta}$ . Let's take a look at that:

$$\theta = \|\theta\| \hat{\theta} \quad (4.8)$$

Remember, a vector is direction (unit vector) times magnitude (scalar).

$$\vec{d} = d \hat{\theta} \quad (4.9)$$

They're in the same **direction**, so they have the same **unit vector**  $\hat{\theta}$ .

$\vec{d}$  is on the **line**, so it satisfies:

We'll use  $\theta \cdot \vec{d}$  instead of  $\theta^\top \vec{d}$  here.

$$\theta \cdot \vec{d} + \theta_0 = 0 \quad (4.10)$$

We can plug our equations 4.8 and 4.9, where we've separated magnitude from unit vector:

$$\underbrace{(\|\theta\|\hat{\theta})}_{\theta} \cdot \underbrace{(\vec{d}\hat{\theta})}_{\vec{d}} + \theta_0 \quad (4.11)$$

We can move the scalars  $\|\theta\|$  and  $d$  out of the way of the dot product:

$$\|\theta\|d (\hat{\theta} \cdot \hat{\theta}) + \theta_0 \quad (4.12)$$

We know that  $\hat{u} \cdot \hat{u} = 1$ :

$$\|\theta\|d + \theta_0 = 0 \quad (4.13)$$

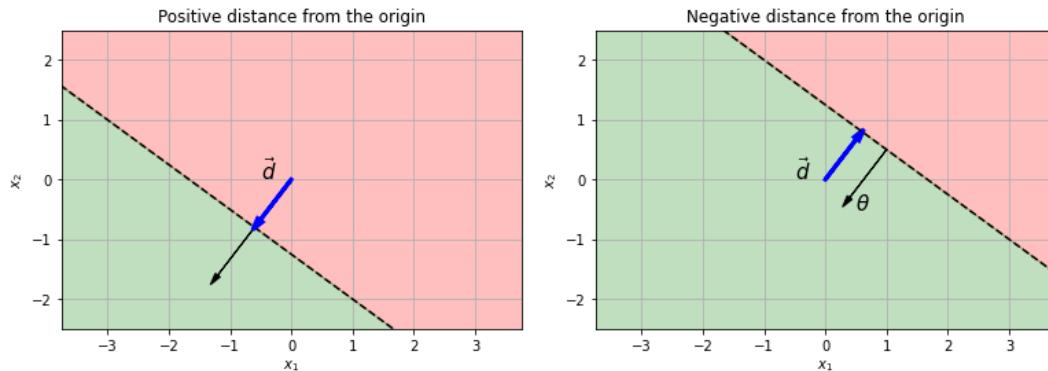
And now, we just solve for  $d$ :

### Concept 171

The **distance**  $d$  from the **origin** to our **linear separator** is

$$d = \frac{-\theta_0}{\|\theta\|} \quad (4.14)$$

A "negative" distance means  $\vec{d}$  (the vector from the origin to the line) is pointed in the opposite direction of  $\theta$ .



Notice, again, that this agrees with our **earlier** thought: the sign of  $\theta_0$  is the opposite ( $-1$ ) of the  $\theta$  direction we move in.

#### 4.2.11 Extending to higher dimensions

We've now fully conquered the 2D problem! Now, we can move up in **dimensions**.

In terms of equations, the answer is simple, just like it is for regression: just add more terms

to  $\theta$ .

### Key Equation 172

A general d-dimensional **linear separator** can do **binary classification** using the hypothesis

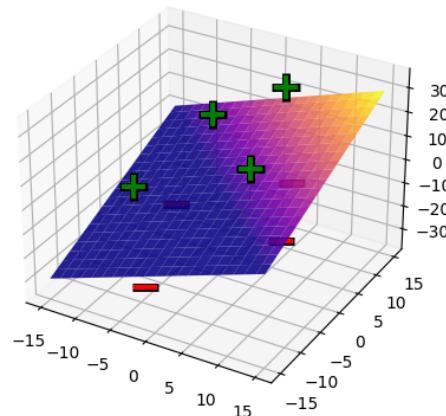
$$h(x; \theta) = \text{sign}(\theta^T x + \theta_0) = \begin{cases} +1 & \text{if } \theta^T x + \theta_0 > 0 \\ -1 & \text{otherwise} \end{cases}$$

Where

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{bmatrix} \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

What about how it looks? Well, if we have 3 input variables, our line turns into a **plane**:

Classification in 3D



Notice the green + signs are "above" the plane, while the red - signs are "below" the plane.

Just like with regression, this is when we introduce the **hyperplane**:

**Concept 173**

Our  $n$ -dimensional **linear separator** solution to the **binary classification** problem **splits** our space into two **halves**: a positive and a negative half.

The **surface** that **splits** space like this is a  $(n - 1)$ -dimensional **hyperplane**.

The hyperplane is **oriented**: there is a **normal** vector  $\theta$  which defines the **orientation** of the hyperplane, and which side is **positive**.

It also has an **offset** term  $+\theta_0$ , that slides it in the  $-\theta$  direction **away** from the origin.

- $\theta_0$  can be any real number, but the shift will always be in the opposite direction.

For any dimensional input, we can use hyperplanes as separators.

#### 4.2.12 IMPORTANT: A difference between regression and classification

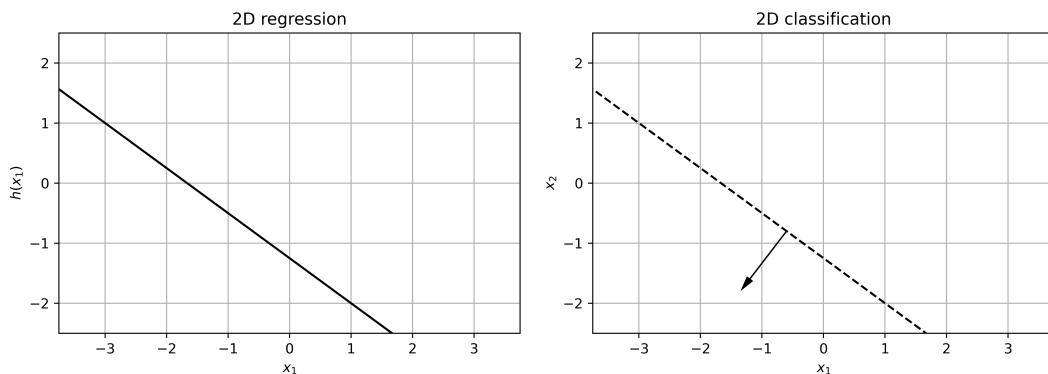
Here is an important misconception that comes up between regression and classification.

Both functions use the equation

$$\theta^T x + \theta_0 \quad (4.15)$$

So, one might think of them as interchangeable.

However, they are **not**. Why is that?



These two plots look almost the same, but represent completely different things!

Notice that these two plots are **both** plotted in 2-D, and both have a **line** plotted. But, they **aren't** as **similar** as they look.

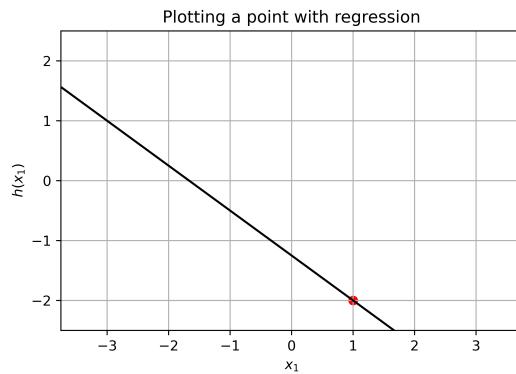
Notice, for example, that the regression plot has **only**  $x_1$ , while the classification plot has  $x_1$  **and**  $x_2$ .

The reason why? The **output**.

- In **regression**, the output is a **real number**: every point on that line represents an input  $x_1$ , and an output  $h(x_1)$ .
  - This plot can only contain **one** input variable: the **second** axis is reserved for the **output!**
- In **classification**, the output is **binary**. So, that line represents only the **values** where the output is  $h(x) = 0$ .
  - This plot can contain **two** input variables:  $x_1$  and  $x_2$ . Rather than **displaying** the output, we only show one **slice** of the output: the  $h(x) = 0$  slice.

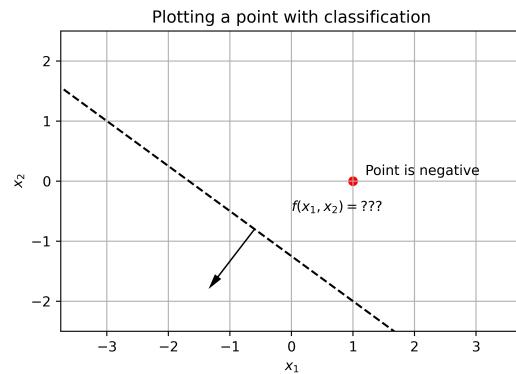
If we think in terms of  $f(x) = \theta^T x + \theta_0$ , we can compare them directly.

The regression plot shows the exact value on the y-axis. If we want to know what  $f(x_1 = 1)$  looks like, we can check the plot: we just get  $f(1) = -2$ .



We have one input, and we get the exact value of our output. We only plot values on the line.

But the classification plot **doesn't!** We aren't given the value of  $\theta^T x + \theta_0$  at  $x = (1, 0)$ : we just know that it's **negative**.



We have two inputs, and we **don't** get the exact output. We can plot inputs anywhere in space.

If we wanted to know the exact value of our 2-D classification, we would need to view it as a plane in 3-D space.

This is the trade-off between these two plots: one gives more information about the **output**, and the other allows for more inputs in a **lower-dimensional** visualization.

### Clarification 174

**Regression** and **classification** plots that look the same, have **different functions**:

When looking at the output of  $f(x) = \theta^T x + \theta_0$ ,

- A **regression** plot gives the **exact numeric**  $f(x)$ .
- A **classification** plot only shows where  $f(x) = 0$ , and the sign of  $f(x)$  elsewhere.

In short:

Regression adds the whole **y number line**, classification only shows  $y = 0$ . This saves **one dimension of space**.

~~~~~

When plotting d inputs,

- A **regression** plot uses a **$d+1$** dimensions (d -dim hyperplane) to plot: +1 dimension for the **output axis**.
 - **Example:** You have one input dimension, $d = 1$. You need to plot the **input and output**: you'll be plotting it on a 2D plane.
 - However, on that 2D plane, you'll plot a **line**: despite existing in 2D space, a line is a 1-d hyperplane.
- A **classification** plot only needs **d** dimensions (($d-1$)-dim hyperplane): we don't need an output axis, because we only plot **$y = 0$** .
 - **Example:** You have three input dimensions, $d = 3$. You'll be plotting every combination of **three inputs** that gives **$y = 0$** : you'll be plotting it in 3D space.
 - However, in 3D space, you'll plot a **plane**: despite existing in 3D space, a plane is 2-d hyperplane.

Notation 175

A $(d - 1)$ -dimensional hyperplane is a d -dimensional **separator**.

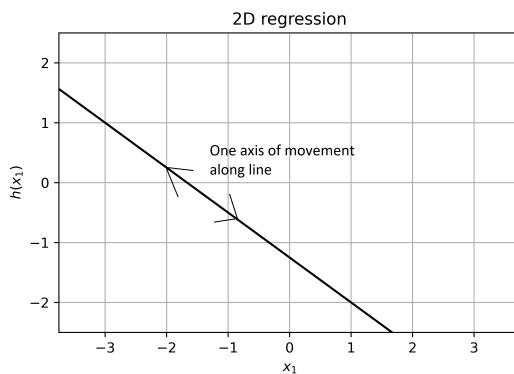
That's because it's **splitting** the d -dimensional space in half.

- **Example:** A **line** (1-dim) splits the **plane** (2-dim) in half.
- So, it separates the two halves of 2-d space.

Why do we need $d + 1$ dimensions to plot a d -dimensional **hyperplane**? Because the hyperplane doesn't fill the whole space.

Here's an example: a **line** in 2-D space is a 1-D **hyperplane**: we have only **one axis** we can move on the line.

It's a 1-d object "embedded" in 2-d space.



Our plot is 2-D, but we can only move along one axis on our line!

- Notice that our line does not fill up the whole space: that's why it's a lower dimension.
- You need 2 dimensions to see **where** the line is, but the line itself is only 1-d.

Because of these differences, θ also acts differently:

Clarification 176

θ appears differently in 2-D regression and classification:

- In **2-D regression**, θ is the **slope** of the line

$$h(x) = \theta x + \theta_0$$

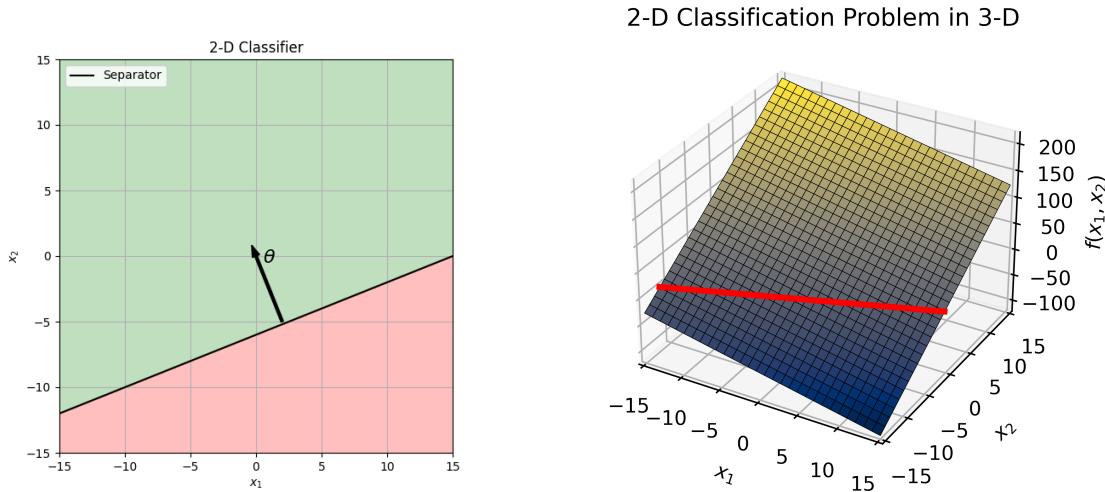
- In **2-D classification**, θ is the **normal vector** of the line

$$\theta = \theta^T x + \theta_0$$

4.2.13 3d plot of 2d separator

For additional understanding, you might view the full output of $\theta^T x + \theta_0$, before we simplify the output to $\{-1, +1\}$.

The below plot "reveals" the 3rd dimension (output of $h(x)$) that the classification plot usually hides.



We can **compare** what we usually see (left plot) to an **alternate** version that shows the 3rd dimension (right plot). These are the **same classifier**!

We mentioned before that, if we wanted to show the exact value of $f(x)$ for our 2-D classifier, we'd need a 3-D plot (just like for regression).

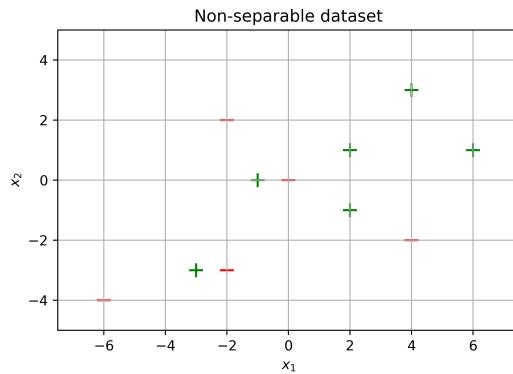
So here, we've done **exactly** that: the **height** is the output of $h(x)$.

But, because we don't **care** about the **exact** output in classification, we usually only graph the plane containing the **red line**: where $f(x_1, x_2) = 0$.

This shows how we're taking a 2D slice ($f(x) = 0$) out of a 3-D plot (full hyperplane), to **save** on one dimension of **plotting**.

4.2.14 Separable vs Non-separable data

One more consideration: **not all** data can be correctly **divided** by a linear separator!



There's no line we could draw through this data to **separate** the points from each other.

If we can, we call it **linearly separable**.

Definition 177

A **dataset** is **linearly separable** if you can **perfectly** classify it with a **linear classifier**.

A couple common reasons for data to not be linearly separable:

- A positive and negative data point have the exact **same position** in input space.
- Two points on either **side** of a point with opposite classification: $+ - +$ or $- + -$, for example.

Very often, real-world datasets **can't** linearly separated, because of **complexities** in the real world, or random **noise**.

But, sometimes, we can **almost** linearly separate it: we get very high **accuracy**. In those cases, it may be **fine** to use a linear separator: we might risk **overfitting** if we use a more complex model.

- Still, if a dataset is not **linearly separable**, or at least **high-accuracy** with a linear separator, that could mean we need a **richer** hypothesis class.
- We'll get into ways to make a **richer** class in the **next** chapter: **feature transformations**.

What is "high enough accuracy"? Depends on what you need it for!

Remember: a "richer" or more "expressive" hypothesis class is one that can create more hypotheses that our current one can't!

4.3 Linear Logistic Classifiers

4.3.1 The problem

Now, our goal is to create a **good model** for our problem, **binary classification**.

To do this, we can **try** using our 0-1 loss \mathcal{L} :

$$J(\theta, \theta_0) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(\text{sign}(\theta^T x^{(i)} + \theta_0), y^{(i)}) \quad (4.16)$$

The **first** thing to note is that there isn't an easy **analytical** solution, no simple **equation**: $\text{sign}(u)$ isn't a function that we can explicitly **solve**, like we could for **linear regression**.

So, we refer to our other approach, **gradient descent**.

First, we need to compute the **gradient**.

To be fair, this is true for most possible problems: most of them can't be solved analytically.

$$\nabla_{\theta} J = 0 \quad (4.17)$$

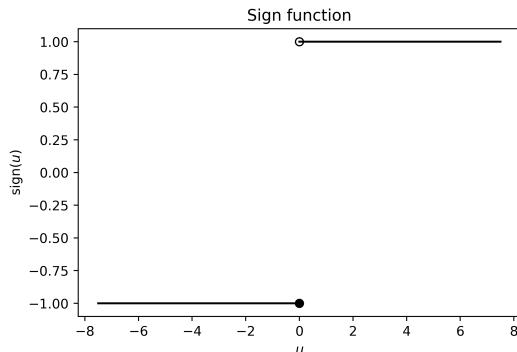
...Well that's not good.

Why not? Because we use our **gradient** to decide **how** to change θ .

4.3.2 The real problem: $\text{sign}(u)$ is flat

What's going on here? Let's look at the sign function:

If the gradient is 0, θ never changes, and we never **improve** θ at all!



Sign is a flat function! The slope is 0 everywhere, except $u = 0$, where it's **undefined**.

Well, that explains why we can't use the gradient: the function is **flat**.

- Another way to say this is that our function doesn't **tell** us when we're **closer** to being right.
- There's **no difference** between being **wrong** by 1 unit or being wrong by 10 units: you can't tell if you're getting **closer** to a correct answer.

- And the **gradient** doesn't tell you which way to move in **parameter space** to further improve.

In fact, the best way we know how to approach this kind of problem takes **exponential** time: it takes exponentially **longer** to solve based on our **number** of data points.

Remember, parameter space is what we move through as we change our parameter vector θ .

That's way too **slow**. So, we'll have to come up with a **better** function: something to **replace** $\text{sign}(u)$, that still serves the same role.

Concept 178

The **sign function** is difficult to optimize, because it isn't **smooth**: not only is the slope undefined at 0, it is 0 everywhere else.

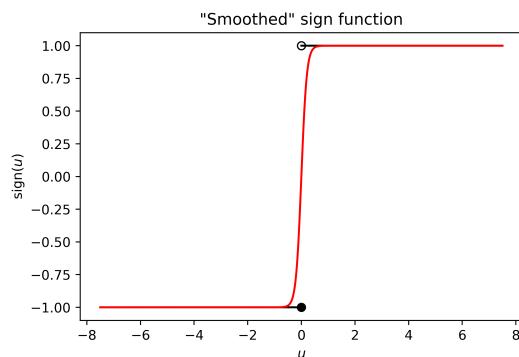
This causes two problems:

- We can't tell whether one **hypothesis** is **closer** to being **correct**, if it has gotten **better**, unless its accuracy has increased.
 - This makes it harder to **improve**.
- We can't indicate how **certain** we are in our answer: $\text{sign}(u)$ is **all-or-nothing**: we choose one class, with no information about how **confident** we are in our choice.
 - Knowing how **uncertain** we are can be **helpful**, both for **improving** our machine and also **judging** the choices our machine makes.

So, we need to explore a **new** approach: we'll **replace** $\text{sign}(u)$ with something else.

4.3.3 The sigmoid function

So, what do we **replace** sign with? We like the way sign **works** (choosing between two different classes based on a **threshold**), so maybe we want a **smoother** version of it.

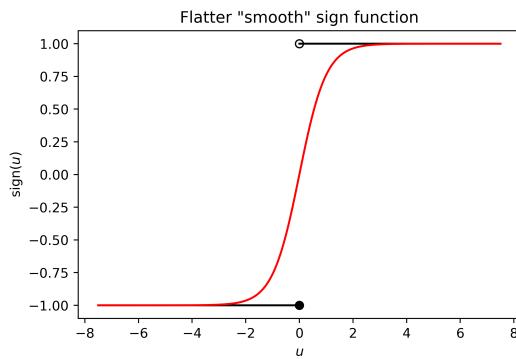


The red line shows a "smoother" sign function, that mostly behaves the same, while solving our problem.

This solves **one** of our two problems: the **gradient** is **nonzero**.

We could also make it less steep:

It's hard to see visually, but the function is **smooth**, and the slope is nonzero **everywhere**!



So, we need a **function** that accomplishes this. It turns out there are **several** that work: $\tanh u$, for example.

For our purposes, we'll use the following function:

Definition 179

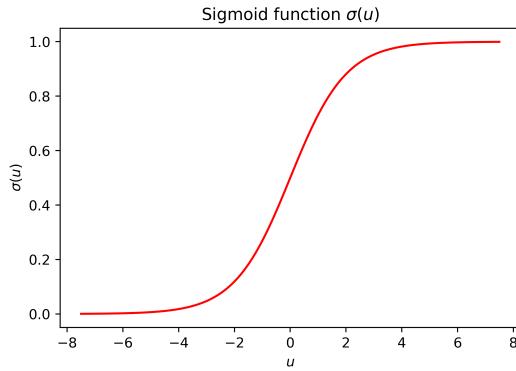
The **sigmoid** function

$$\sigma(u) = \frac{1}{1 + e^{-u}}$$

...is a **nonlinear** function that we use to **compute** the output of our **classification** problem.

- It is also called the **logistic** function.

The function looks like this:



4.3.4 Sigmoid as a probability

Something you may **notice** is that $\sigma(x)$ is always between 0 and 1. But before, $\text{sign}(x)$ was **always** between -1 and +1. Why would we use *this* function?

Because going between 0 and 1 has a different advantage: we can interpret it as a **probability**.

- Your **value** of $\sigma(u)$ can be stated as, "what does the machine think is the **probability** we **classify** this data point as +1".
- And, on the **flip** side, $1 - \sigma(u)$ is the "**probability** we **classify** as -1".

This solves the second problem we mentioned **earlier**: we can indicate how **confident** the machine is in its answer!

Concept 180

The output of the **sigmoid function** $\sigma(u(x))$ gives the **probability** that the data point x is classified **positively**.

$$\sigma(u) = P\{x \text{ is classified } +1\}$$

$$1 - \sigma(u) = P\{x \text{ is classified } -1\}$$

Note that this works because $\sigma(u) \in (0, 1)$.

4.3.5 Logistic Regression

So, we've seen the benefits of switching from $\text{sign}(u)$ to $\sigma(u)$. So we'll do that:

We're using $u(x) = \theta^T x + \theta_0$

Key Equation 181

Logistic Regression is a **modification** of **linear regression**.

$$h(x; \theta) = \sigma(\theta^T x + \theta_0)$$

where

$$\sigma(u) = \frac{1}{1 + e^{-u}}$$

It outputs the **probability** of a **positive** classification.

If we **plug** this in, we get this slightly ugly expression:

$$h(x; \theta) = \frac{1}{1 + e^{-(\theta^T x + \theta_0)}}$$

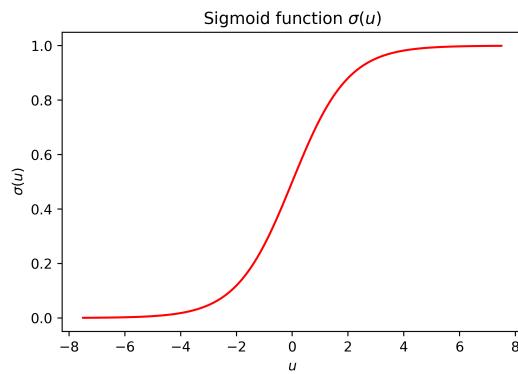
We have a problem, though: **logistic regression** is a... **regression** function. It takes in a real **vector**, and outputs a real **number**: $\mathbb{R}^d \rightarrow \mathbb{R}$.

We can't use this to do **classification**, where want $\mathbb{R}^d \rightarrow \{-1, +1\}$!

4.3.6 Prediction Threshold

When we were just using $u(x) = \theta^T x + \theta_0$, we classified data points by saying whether $u(x) > 0$. Our boundary was $u(x) = 0$.

What happens to σ if $u = 0$?

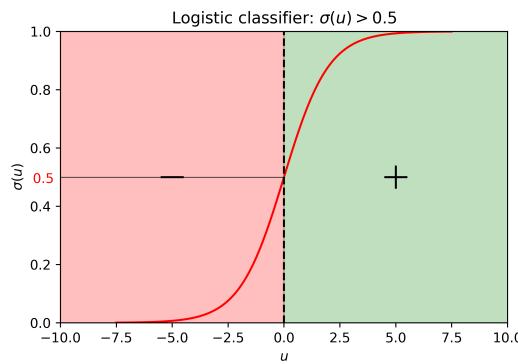


If we go to $u = 0$ on the x-axis, we find $\sigma = .5$ on the y-axis.

We get $\sigma(u = 0) = 0.5$. So, we could use that as our classification: $\sigma(u) > 0.5$.

$$u > 0 \iff \sigma(u) > 0.5 \quad (4.18)$$

If we want to plot the positive and negative regions:



But, we don't necessarily always want to use $\sigma = 0.5$ as our boundary:

Example: Imagine if you wanted to **classify** whether someone needs a **test** for a disease. Classify -1 if we test them, $+1$ if we don't.

Let's say you got $\sigma(u) = 0.6$, so you're only 60% sure they **don't** need it. You'd classify that as $\sigma(u) > 0.5$: they're assigned "**no test**".

If the disease is life-threatening, and the test is cheap, then a 40% chance could justify getting the test.

- Whether or not they **should** get that test isn't usually decided by whether the chance is greater than 50%: that's a pretty **arbitrary** number.
- In real life, the **certainty** you want depends on the situation.

We call the **boundary** between positive and negative the **prediction threshold**.

How expensive is the test? How bad is it, if we don't catch the disease now? Etc.

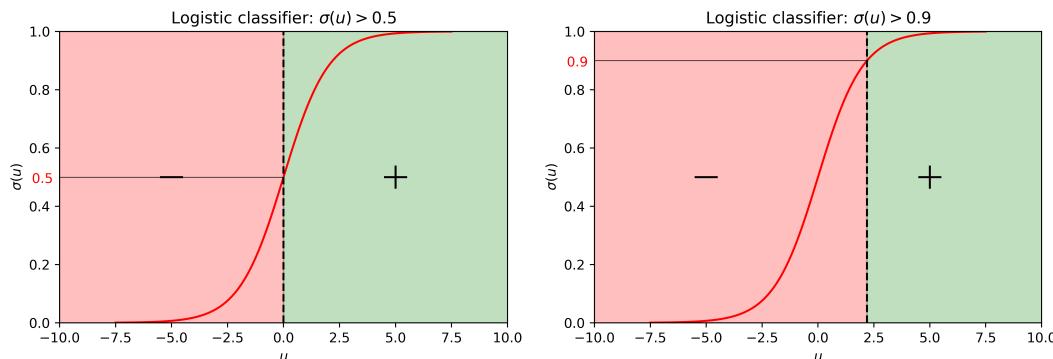
Definition 182

The **prediction threshold** σ_{thresh} is the value where you go from **negative** classification to **positive**.

Our **default** value is a threshold of 0.5, but our threshold can be **anywhere** in the range

$$0 < \sigma_{\text{thresh}} < 1$$

Example: If $\sigma_{\text{thresh}} = 0.9$, we would see:



We switch from a 0.5 threshold (left) to a 0.9 threshold (right).

In this example, more things will be negatively classified.

4.3.7 Linear Logistic Classifier

This finally gives us our **linear logistic classifier** (LLC)

Key Equation 183

The **linear logistic classifier** is a **binary** classifier of the form

$$h(x; \theta) = \begin{cases} +1 & \text{if } \sigma(u(x)) > \sigma_{\text{thresh}} \\ -1 & \text{otherwise} \end{cases}$$

where

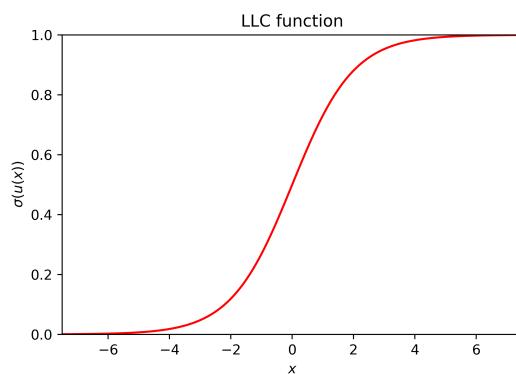
$$u = \theta^T x + \theta_0 \quad \sigma(u) = \frac{1}{1 + e^{-u}}$$

We call it linear because of the linear inner function $u(x)$, and logistic because of the outer function $\sigma(u)$.

4.3.8 Modifying our sigmoid

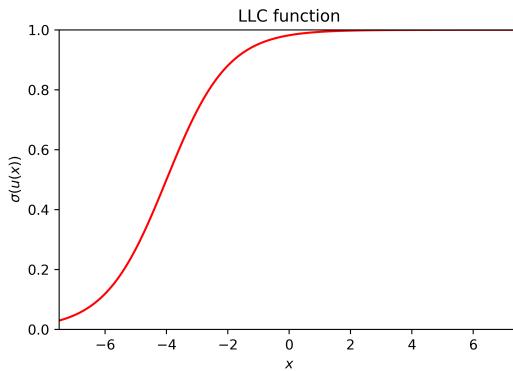
What happens when you modify the **parameters** of an LLC? Let's find out.

We'll use a 1-D input: our variables will be θ (scalar) and θ_0 : $\theta x + \theta_0$



Our baseline LLC: $u(x) = 1x + 0$

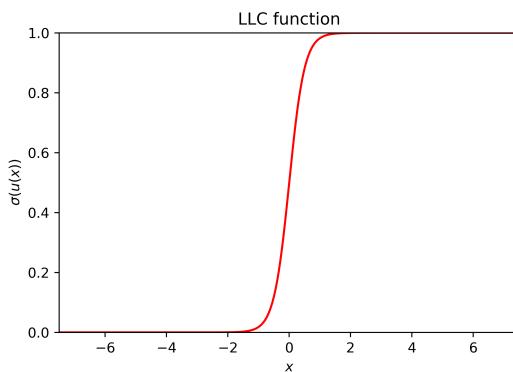
What if we shift by increasing θ_0 ?



Our shifted LLC: $u(x) = 1x + 4$. θ_0 shifts us along the x-axis!

Just like in linear regression, it **shifts** us in the **opposite** direction: if θ_0 is **positive**, we shift in the **negative** direction, and vice versa.

What if we increase the magnitude of θ !



Our new LLC: $u(x) = 4x$. Increasing θ makes our function steeper!

Making the magnitude of θ larger makes our function **change** faster.

- This makes some sense: if θ (linear slope of $u(x)$) makes $u(x)$ **change** faster, it will make $σ(u)$ change faster **too**.

You can combine these changes as well: you can shift your LLC with θ_0 , and also make it steeper/less steep by changing magnitude of θ .

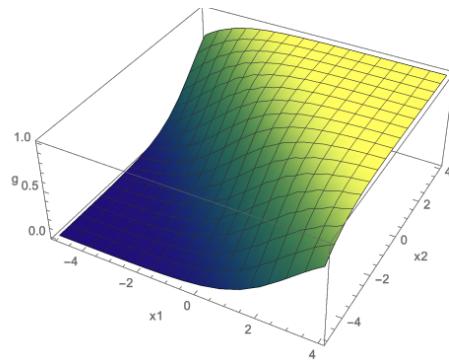
Concept 184

When working with **sigmoids**, you can **transform** them using your **parameters**:

- A higher **magnitude** $\|\theta\|$ makes the slope **steeper**, and answers more **confident**.
- **Increasing** θ_0 **shifts** the sigmoid in the $-\theta$ **direction**, and vice versa.

4.3.9 Viewing our sigmoid in 3D

Let's quickly take a look at a sigmoid in 3D, with two inputs:



As you can see, you get mostly the same shape: if you look at it from the side, it's exactly the same, in fact! Just stretched out into 3D.

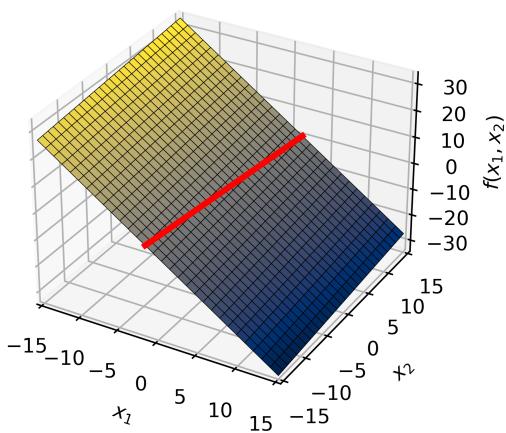
4.3.10 LLCs and LCs have the same boundary

One more important thing to note: noticed that we set $\sigma_{\text{thresh}} = 0.5$, because that was when $u(x) = 0$.

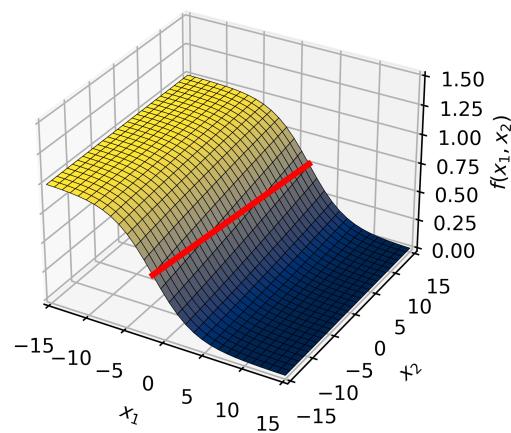
$$u = 0 \iff \sigma(u) = 0.5 \quad (4.19)$$

This means that, if our threshold is 0.5, then the boundary of our LLC should look exactly the same as if it were LC: the only difference is the values that we *can't* see:

2-D Classification Problem in 3-D



2-D Classification Problem in 3-D



Despite having different shapes in 3D, they both create 2-D **linear** classifiers: on the left, $u(x) = 0$, and on the right, $\sigma(u) = .5$.

One way to think about this difference is that while one may be logistic, they are both

linear: they both create the same **linear separator**.

The main benefit of switching to LLC is that $\sigma(u)$ has a useful **gradient**, while $\text{sign}(u)$ does **not**, so we can do **gradient descent**.

- Even if we adjust our threshold σ_{thresh} , that will simply shift the linear classifier.

The probability interpretation is also more appropriate: we usually aren't fully confident in our answers.

Concept 185

LLCs (Linear Logistic Classifiers) and **LCs** (Linear Classifiers) both create a **linear hyperplane separator** in $d - 1$ dimensional **space**.

If the **threshold value** is 0.5, then they have the **exact same** separator.

- This is because the **same set of inputs** x that cause $u(x) = 0$, will also cause $\sigma(u(x)) = 0.5$.
- So, if $u(x) = 0$ (LC) creates a hyperplane, then $\sigma(u(x)) = 0.5$ (LLC) will, too.

4.3.11 Learning LLCs: Loss Functions

Now that we have fully **built up** LLCs, we can start trying to **train** our own.

In order to do that, we need a way to **evaluate** our hypotheses: a **loss function**.

Earlier in the chapter, we tried **0-1 Loss**:

$$\mathcal{L}_{01}(\mathbf{h}(\mathbf{x}; \Theta), y) = \begin{cases} 0 & \text{if } y = \mathbf{h}(\mathbf{x}; \Theta) \\ 1 & \text{otherwise} \end{cases}$$

But, this **loss** function has the same problem our **sign** function did: it isn't **smooth**!

- It's a **discrete** function based on our **discrete classes**: so, it won't have a smooth **gradient** we can do **descent** on.

For our **sign** function, we switched to the **sigmoid** function, which measures in terms of **probabilities**: this gave us some **smoothness** to our classification.

Could we do the same here?

4.3.12 Building our new loss function

So, the **output** of our sigmoid $\sigma(u)$ is a **probability**: it tells us, "how **likely** do we think a point is to be in class +1?"

We want a loss function

$$\mathcal{L}(g, y) \tag{4.20}$$

That considers two facts: the **correct** answer y , and how likely we **expected** +1 to be, $g = \sigma(u)$.

Notation 186

For our **loss function**, rather than using $\hat{y} \in \{-1, +1\}$, we switch to **probabilities**: $y \in \{0, 1\}$.

$$\hat{y} \in \{-1, +1\} \rightarrow y \in \{0, 1\}$$

That way, $\sigma(u)$ and y **match**:

$$y \in \{0, 1\} \quad g \in (0, 1)$$

Both represent "the chance that $\hat{y} = +1$ ". $\sigma(u)$ makes a prediction, while y uses the known result:

- If $\hat{y} = +1$, there's a **100% chance** that $\hat{y} = +1$.
 - So, the "true" output is $y = 100\% = 1$
- If $\hat{y} = -1$, there's a **0% chance** that $\hat{y} = +1$.
 - So, the "true" output is $y = 0\% = 0$

In this problem, it's easier to think in terms of "goodness" of the result. We'll use a "goodness" function $G(g, y)$.

- We want the "true" and "predicted" probabilities to be as close as possible.
 - If the correct answer is $y = 1$, then we want $g = \sigma(u)$ to be **high**.
 - If the correct answer is $y = 0$, we want $g = \sigma(u)$ to be **low**.

To get the loss function, we just take the negative of it later.

We represent this as

$$G(g, y) = \begin{cases} g & \text{if } y = 1 \\ 1 - g & \text{else } (y = 0) \end{cases} \quad (4.21)$$

Remark (Optional) 187

This loss function can be interpreted as, "if we had a g chance of picking +1, how often would we have been right?"

- If $y = 1$, then we want +1. The chance of choosing +1 is g .
- If $y = 0$, then we want -1. The chance of choosing -1 is $1 - g$.

0-1 loss works the same way for a non-random model (one that picks the larger odds every time): what's percentage of the data we get right.

As we mentioned, we'll need to take the negative of this later to get the "loss" function.

4.3.13 Loss Function for Multiple Data Points

Now, how do we consider **multiple** data points? Well, let's think in terms of **probability**: guessing each point is a separate **event**.

We *could* add or **average** our guesses. But, since we're working with **probabilities**, there's a natural way to **combine** them: multiple events **occurring** at the same time.

Before, we asked, "how likely were we to be **right**?" for **one** data point. We could **extend** this question to, "how likely are we to get **every** question right?"

Well, each question we get right is an **independent** event E_i . If we want two independent events to **both** happen, we have to **multiply** their probabilities.

Key Equation 188

The probability of two independent events A and B happening at the same time is

$$P\{A \text{ and } B\} = P\{A\} * P\{B\}$$

So if we want **all** of them, we just multiply:

$$P\{E_{\text{all}}\} = P\{E_1\} * P\{E_2\} * \dots * P\{E_n\} \quad (4.22)$$

Written using pi notation, and also $g^{(i)}$ for multiple data points: _____

$$P\{E_{\text{all}}\} = \prod_{i=1}^n P\{E_i\} = \prod_{i=1}^n \begin{cases} g^{(i)} & \text{if } y^{(i)} = 1 \\ 1 - g^{(i)} & \text{if } y^{(i)} = 0 \end{cases} \quad (4.23)$$

Pi notation is described in the prerequisites chapter! The short version: instead of adding terms with \sum , you multiply with \prod .

4.3.14 Simplifying our expression - Piecewise

Our piecewise function is a bit **annoying**, though: is there a way to **simplify** it so that it doesn't have to be **piecewise**?

Our goal is to **combine** our two piecewise cases into a **single** equation. That means one of them needs to **cancel out** whenever the other is true.

Well, let's see what we have to **work** with.

Our **two** cases happen when $y = 0$ or $y = 1$: these are **nice** numbers! Why? Because of the **exponent** rules for these two:

- $c^0 = 1$: an exponent of 0 outputs 1: a factor of 1 in a product might as well **not be there**. It has been effectively **cancelled** out.
- $c^1 = c$: an **exponent** of 1 leaves the factor **unaffected**.

So, let's consider the **first** case, g . we can use $\textcolor{red}{g}^y$: if $y = 1$, it's **unaffected**. If $y = 0$, the term is **removed**.

We want the **opposite** for $1-g$. We can **swap** 1 and 0 by doing $1-y$. This gives us $(1-\textcolor{red}{g})^{1-y}$.

For one data point:

$$P\{E\} = \underbrace{\textcolor{blue}{g}^y}_{y=1} \underbrace{(1-\textcolor{red}{g})^{1-y}}_{y=0} \quad (4.24)$$

We've gotten rid of the piecewise function! Let's add back in the product:

$$P\{E_{all}\} = \prod_{i=1}^n P\{E_i\} = \prod_{i=1}^n \textcolor{red}{g}^{(i)} \textcolor{blue}{y}^{(i)} (1-\textcolor{red}{g}^{(i)})^{1-\textcolor{blue}{y}^{(i)}} \quad (4.25)$$

Looks pretty ugly, but we'll work on that.

4.3.15 Getting rid of the product

Our exponents look pretty **ugly**. Can we do something about that?

- More important than ugliness: **products** are also pretty unpleasant: we can't use **linearity!**

Linearity uses **addition** between variables. What sort of **function** could change a **product** into a **sum**?

Linearity makes lots of problems easy to work with, so we try to keep it.

Well, we could **list** out different basic functions, to see which ones connect sums and products. It turns out, one **interesting** function is

$$\overbrace{\log ab}^{\text{product}} = \overbrace{\log a + \log b}^{\text{sum}} \quad (4.26)$$

Aha! If we take the **log** of our function, we can turn a **product** into the **sum**!

$$\overbrace{\log \left(\prod_{i=1}^n p_i \right)}^{\text{product}} = \overbrace{\sum_{i=1}^n \log(p_i)}^{\text{sum}} \quad (4.27)$$

Plugging in $p_i = P\{E_i\}$:

$$\sum_{i=1}^n \log \left(g^{(i)y^{(i)}} (1 - g^{(i)})^{1-y^{(i)}} \right) \quad (4.28)$$

The below equation looks complicated, but all we've done is swap the product for a sum!

We can also separate our two **factors**, g^y and $(1 - g)^{1-y}$.

$$\sum_{i=1}^n \left(\log(g^{(i)y^{(i)}}) + \log((1 - g^{(i)})^{1-y^{(i)}}) \right) \quad (4.29)$$

And finally, we can remove the **exponents**:

$$\sum_{i=1}^n \left(y^{(i)} \log g^{(i)} + (1 - y^{(i)}) \log(1 - g^{(i)}) \right) \quad (4.30)$$

Concept 189

Our **negative log likelihood** (NLL) comes from a couple steps:

- Use $y \in \{0, 1\}$ instead of $y \in \{-1, +1\}$ so that y and g have **matching** outcomes.
- Get the **chance** the model is right on every **guess**: a **product**.
- Use **exponents** to convert the **piecewise** expression into a single **equation**.
- Take the **log** of our expression to switch from a **product** to a **sum**.
- Take the **negative** to get the **loss** rather than the "goodness" of our function.

4.3.16 Negative Log Likelihood

Remember, at the **beginning**, we said that we need to take the **negative**: our function represents how **good** our function is, but we want the **loss**.

With this, our function is in its final form:

Key Equation 190

We can get the loss of our **linear logistic classifier (LLC)** using the **negative log likelihood (NLL)** loss function

$$\mathcal{L}_{\text{nll}}(g^{(i)}, y^{(i)}) = - \left(y^{(i)} \log g^{(i)} + (1 - y^{(i)}) \log (1 - g^{(i)}) \right)$$

Or,

$$- \left((\text{answer}) \log(\text{guess}) + (1 - \text{answer}) \log(1 - \text{guess}) \right)$$

Our total loss is

$$\sum_{i=1}^n \mathcal{L}_{\text{nll}}(g^{(i)}, y^{(i)}) \quad (4.31)$$

Finally, we add our **regularizer**:

$$J_{\text{lr}}(\theta, \theta_0; \mathcal{D}) = \frac{1}{n} \sum_{i=1}^n \left(\mathcal{L}_{\text{nll}}(g^{(i)}, y^{(i)}) \right) + \lambda \|\theta\|^2 \quad (4.32)$$

Key Equation 191

The full **objective function** for **LLC** is given as

$$J_{\text{lr}}(\theta, \theta_0; \mathcal{D}) = \frac{1}{n} \sum_{i=1}^n \left(\mathcal{L}_{\text{nll}}(\sigma(\theta^T x + \theta_0), y^{(i)}) \right) + \lambda \|\theta\|^2$$

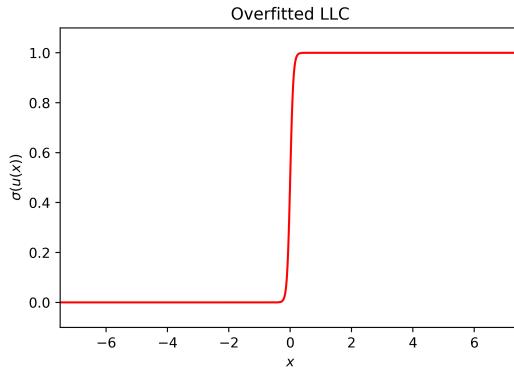
Using our **loss** function \mathcal{L}_{nll} , and our **logistic** function $\sigma(u)$.

4.3.17 LLCs and overfitting

In chapter 2, we reduced **overfitting** by limiting the **magnitude** of θ using

$$R(\theta) = \lambda \|\theta\|^2 \quad (4.33)$$

In this chapter, it's more clear why reducing **magnitude** reduces **overfitting**. Let's see what happens when θ is very **large**:

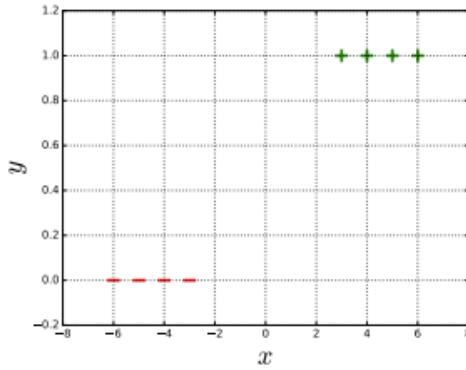


Our "squished" LLC: $u(x) = 20x$.

This function starts looking more and more like the **sign** function. This means we very, **very quickly** go from **confident** in one answer, to confident in another.

- If we go from $x = -0.5$ to $x = +0.5$, we go from "incredibly sure of -1 " to "incredibly sure of $+1$ ".
- That's a small change in input, for a big chance in confidence!

This sort of certainty isn't always appropriate: consider the below example.



In this case, you definitely want to separate the left and right side. Our $\theta = 20$ separator above, would work just fine.

- But, is it **appropriate**? Let's try to use our model to predict new data.
- If you give that model $x = 0.5$, then our model believes $\hat{y} = +1$ with .99995 confidence.
- When our closest data point to 0 is at ± 3 , that's unreasonably high certainty.

In other words, the model could create a very sensitive boundary, without having enough data to justify it: this could be a sign of **overfitting**.

Why would we train such an extreme model? Is this even a problem we have to worry about?

It's more serious than you'd expect: we can see why, by consider how our loss function works: here's an excerpt from earlier.

- We want the "true" and "predicted" probabilities to be as close as possible.
 - If the correct answer is $y = 1$, then we want $g = \sigma(u)$ to be **high**.
 - If the correct answer is $y = 0$, we want $g = \sigma(u)$ to be **low**.

The short version: we want to maximize our confidence in the correct answer. This is directly coded into our loss function, and all of the variations.

We just discussed that increasing θ will increase your confidence in the answer. So, this loss function encourages our model to make θ larger and larger!

So, we need to prevent θ from **growing** to an unreasonably high value. Thus, we **penalize** a large $\|\theta\|$.

This means we're **penalizing** the machine's **overconfidence** in its answer, so that it **generalizes** better.

Concept 192

In **classification**, the **regularizer** follows the form

$$R(\theta) = \lambda \|\theta\|^2$$

Regularization in this form reduces **overfitting** to our data by

- Making the **transition** between classifications less **sharp**, when it shouldn't be so **certain** of the boundary.
- It also prevents our model from becoming **overly confident** in its answer.

4.4 Gradient Descent for Logistic Regression

4.4.1 Summary

Now, we have developed all the tool we need to do binary classification with LLC:

- A **linear** model that lets us combine our **input** variables into a single, predictive **number**:

$$u(x) = \theta^T x + \theta_0 \quad (4.34)$$

- A **logistic** model that turns this **number** into a **probability** of a classification,

$$\sigma(u) = \frac{1}{1 + e^{-u}} \quad (4.35)$$

- A **threshold value** we use to determine how to use this **probability** to **classify**:

$$h(x; \theta) = \begin{cases} +1 & \text{if } \sigma(u(x)) > \sigma_{\text{thresh}} \\ 0 & \text{otherwise} \end{cases} \quad (4.36)$$

- A **loss function** NLL we use to evaluate the **quality** of our **classifications**:

$$\mathcal{L}_{\text{nll}}(g^{(i)}, y^{(i)}) = - \left(y^{(i)} \log g^{(i)} + (1 - y^{(i)}) \log (1 - g^{(i)}) \right)$$

- And an **objective function** we can **optimize** and reduce our **loss**:

$$J_{\text{lr}}(\theta, \theta_0; \mathcal{D}) = \lambda \|\theta\|^2 + \frac{1}{n} \sum_{i=1}^n \mathcal{L}_{\text{nll}}(g^{(i)}, y^{(i)}) \quad (4.37)$$

We have everything we need to do optimization.

4.4.2 The problem: Gradient Descent

We want to do **gradient descent** to minimize J_{lr}

$$J_{\text{lr}}(\theta, \theta_0) = R(\theta) + \frac{1}{n} \sum_{i=1}^n \mathcal{L}_{\text{nll}}(g^{(i)}, y^{(i)}) \quad (4.38)$$

We want repeatedly **adjust** our model $\Theta = (\theta, \theta_0)$ to improve J_{lr} . To do that, we want the gradients for θ and θ_0 . Let's start with θ .

$$\nabla_{\theta} J_{\text{lr}} = \frac{\partial J_{\text{lr}}}{\partial \theta} \quad (4.39)$$

First, J_{lr} has **two** terms, so we'll separate them.

$$\nabla_{\theta} J_{lr} = \frac{\partial R}{\partial \theta} + \frac{1}{n} \sum_{i=1}^n \frac{\partial \mathcal{L}_{NLL}}{\partial \theta}(g^{(i)}, y^{(i)}) \quad (4.40)$$

The regularization term is pretty easy, because we did it last chapter:

$$\frac{\partial R}{\partial \theta} = 2\lambda\theta \quad (4.41)$$

But what about our first term?

4.4.3 Getting the gradient: Chain Rule

Now, we just need to do

$$\frac{\partial \mathcal{L}_{NLL}}{\partial \theta}(g, y) \quad (4.42)$$

With our \mathcal{L}_{NLL} term, we run into an issue: how do we take the **derivative**? The function is very, very deeply **nested**. In our case:

x **affects** $u(x)$. $u(x)$ **affects** $\sigma(u)$. $\sigma(u) = g$ **affects** $\mathcal{L}_{NLL}(g, y)$, which finally **affects** $J(\theta, \theta_0)$.

How do we represent this **chain** of functions? With the **chain rule**:

This next line is a generic chain rule: not specific to our problem.

$$\frac{\partial A}{\partial C} = \frac{\partial A}{\partial B} \cdot \frac{\partial B}{\partial C} \quad (4.43)$$

So, we'll build up a **chain rule** for our needs. We'll use $g = \sigma(u)$.

$$\frac{\partial \mathcal{L}_{NLL}}{\partial \theta} = \frac{\partial \mathcal{L}_{NLL}}{\partial \sigma} \cdot \frac{\partial \sigma}{\partial \theta} \quad (4.44)$$

Sigma contains u , so we'll add that to the chain:

$$\frac{\partial \mathcal{L}_{NLL}}{\partial \theta} = \frac{\partial \mathcal{L}_{NLL}}{\partial \sigma} \cdot \frac{\partial \sigma}{\partial u} \cdot \frac{\partial u}{\partial \theta} \quad (4.45)$$

This is our full **chain rule**!

Key Equation 193

The **gradient** of **NLL** can be calculated using the **chain rule**:

$$\frac{\partial \mathcal{L}_{NLL}}{\partial \theta} = \frac{\partial \mathcal{L}_{NLL}}{\partial \sigma} \cdot \frac{\partial \sigma}{\partial u} \cdot \frac{\partial u}{\partial \theta} \quad (4.46)$$

4.4.4 Getting our individual derivatives

We can take the derivative of each of these objects. First, let's look at \mathcal{L}_{NLL}

$$\mathcal{L}_{\text{NLL}}(\sigma, y) = - \left(y \log \sigma + (1 - y) \log (1 - \sigma) \right)$$

And we'll use $\frac{d}{dx} \log(x) = \frac{1}{x}$

$$\boxed{\frac{\partial \mathcal{L}_{\text{NLL}}}{\partial \sigma} = - \left(\frac{y}{\sigma} - \frac{1-y}{1-\sigma} \right)} \quad (4.47)$$

Now, we look at $\sigma(u)$:

$$\sigma(u) = \frac{1}{1 + e^{-u}} \quad (4.48)$$

If we take the derivative, we can get:

$$\frac{\partial \sigma}{\partial u} = \frac{-e^{-u}}{(1 + e^{-u})^2} \quad (4.49)$$

Which we can rewrite, conveniently, as

Try this yourself if you're curious!

$$\boxed{\frac{\partial \sigma}{\partial u} = \sigma(1 - \sigma)} \quad (4.50)$$

Finally, our last derivative:

$$u = \theta^T x + \theta_0 \quad (4.51)$$

$$\boxed{\frac{\partial u}{\partial \theta} = x} \quad (4.52)$$

4.4.5 Simplifying our chain rule

So, now, we can put together our chain rule:

$$\frac{\partial \mathcal{L}_{\text{NLL}}}{\partial \theta} = \frac{\partial \mathcal{L}_{\text{NLL}}}{\partial \sigma} \cdot \frac{\partial \sigma}{\partial u} \cdot \frac{\partial u}{\partial \theta} \quad (4.53)$$

Plug in the derivatives:

$$\frac{\partial \mathcal{L}_{\text{NLL}}}{\partial \theta} = - \left(\frac{y}{\sigma} - \frac{1-y}{1-\sigma} \right) \cdot \sigma(1 - \sigma) \cdot x \quad (4.54)$$

Simplify:

$$\frac{\partial \mathcal{L}_{\text{NLL}}}{\partial \theta} = ((1 - \mathbf{y})\sigma - \mathbf{y}(1 - \sigma)) \cdot \mathbf{x} \quad (4.55)$$

And finally, we sum the terms. We can do the θ_0 gradient at the same time: the only difference is that $\frac{\partial u}{\partial \theta_0} = 1$, instead of x .

Key Equation 194

The **gradients** of NLL for gradient descent are

$$\nabla_{\theta} \mathcal{L}_{\text{NLL}} = (\sigma - \mathbf{y})\mathbf{x}$$

$$\frac{\partial \mathcal{L}_{\text{NLL}}}{\partial \theta_0} = (\sigma - \mathbf{y})$$

We can plug this into J_{lr} :

$$\nabla_{\theta} J_{\text{lr}} = \frac{1}{n} \sum_{i=1}^n \left((\mathbf{g}^{(i)} - \mathbf{y}^{(i)}) \mathbf{x}^{(i)} \right) + 2\lambda\theta \quad (4.56)$$

One comment we didn't make: remember that $R(\theta)$ won't show up in the θ_0 derivative!

$$\frac{\partial J_{\text{lr}}}{\partial \theta_0} = \frac{1}{n} \sum_{i=1}^n (\mathbf{g}^{(i)} - \mathbf{y}^{(i)}) \quad (4.57)$$

We can use this to do **gradient descent**!

$$\theta_{\text{new}} = \theta_{\text{old}} - \eta \nabla_{\theta} J_{\text{lr}}(\theta_{\text{old}}) \quad (4.58)$$

In $\theta^{(t)}$ notation:

$$\theta^{(t)} = \theta^{(t-1)} - \eta \left(\nabla_{\theta} J_{\text{lr}}(\theta^{(t-1)}) \right) \quad (4.59)$$

$$\theta_0^{(t)} = \theta_0^{(t-1)} - \eta \left(\frac{\partial J_{\text{lr}}(\theta^{(t-1)})}{\partial \theta_0} \right) \quad (4.60)$$

- This chain-rule approach will return when we do Neural Networks in a later chapter.

4.5 Handling Multiple Classes

Now, we have developed a **binary** classifier, using logistic regression. But, many (almost all) problems have **more than two classes!**

Example: Different animals, genres of movies, sub-types of disease, etc.

4.5.1 Approaches to multi-class classification

So, we need to a way to do **multi-classing**. Consider two main approaches:

- Train many binary classifiers on different **classes** and **combine** them into a single model.
 - There are several ways to **combine** these **classifiers**. We won't go over them here, but some **names**: OVO (one-versus-one), OVA (one-versus-all).
- Make **one** classifier that handles the multi-class problem by itself.
 - This model will be a **modified** version of **logistic regression**, using a variant of NLL.

The **latter** approach is what we will use in this **next** section.

4.5.2 Extending our Approach: One-Hot Encoding

Rather than being **restricted** to classes 0 and 1, we'll have **k distinct classes**. Our **hypothesis** will be

$$h : \mathbb{R}^d \rightarrow \{C_1, C_2, C_3, \dots, C_k\}$$

Where C_i is the i^{th} class. Meaning, we want to **output** one of those k **classes**.

Because we'll be using our computer to do **math** to get the **answer**, we need to represent this with **numbers**. Before, we would simply **label** with 0 or 1.

- We could return $\{1, 2, 3, 4, 5, \dots, k\}$ for each **label**. But this is **not a good idea**: it implies that there's a natural **order** to the classes, which isn't necessarily true.
- If we don't **actually** think C_1 is **closer** to C_2 than to C_5 , we probably shouldn't represent them with numbers that are **closer** to each other.

Instead, each class needs to be a **separate** variable. We can store them in a **vector**:

$$\begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_k \end{bmatrix} \quad (4.61)$$

So, our **label** will be

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix} \quad (4.62)$$

In binary classification, we used 0 or 1 to indicate whether we fit into one **class**. So, that's how we'll do each class: 0 if our data point is **not** in this class, 1 if it **is**.

This approach is called **one-hot encoding**.

Definition 195

One-hot encoding is a way to represent **discrete** information about a data point.

Our k classes are stored in a length- k column **vector**. For **each** variable in the vector,

- The value is **0** if our data point is **not in that class**.
- The value is **1** if our data point is **in that class**.

In one-hot encoding, items are **never** labelled as being in **two** classes at the **same time**.

Example: Suppose that we want to classify **furniture** as table, bed, couch, or chair.

$$\begin{bmatrix} \text{table} \\ \text{bed} \\ \text{couch} \\ \text{chair} \end{bmatrix} \quad (4.63)$$

For each class:

$$y_{\text{chair}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad y_{\text{table}} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad y_{\text{couch}} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad y_{\text{bed}} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad (4.64)$$

4.5.3 Probabilities in multi-class

So, we now know our **problem**: we're taking in a data point $x \in \mathbb{R}^d$, and **outputting** one of the classes as a **one-hot vector**.

So, now that we know what sorts of data we're **expecting**, we need to decide on the formats of our **answer**.

We'll be returning a vector of length- k : **one** for each **class**. When we were doing **binary** classification, we estimated the **probability** of the positive class.

So, it should make sense to do the same **here**: for each class, we'll return the estimated **probability** of our data point being in that class.

$$g = \begin{bmatrix} P\{x \text{ in } C_1\} \\ P\{x \text{ in } C_2\} \\ \vdots \\ P\{x \text{ in } C_k\} \end{bmatrix} = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_k \end{bmatrix} \quad (4.65)$$

We need one **additional** rule: the probabilities need to **add up to one**: we should assume our point ends up in some class or **another**.

$$g_1 + g_2 + \dots + g_k = \sum_i g_i = 1 \quad (4.66)$$

Concept 196

The different terms of our **multi-class** guess g_i represent the **probability** of our data point being in class C_i .

Because we should assume our data point is in **some** class, all of these probabilities have to **add** to 1.

Let's be careful, though: this is only true for probabilities within a single data point.

Example: Suppose you have two animals (data points).

- It's impossible for the first animal to be **both** 90% cat and 90% dog.
- *But*, there's no issue with the first animal being 90% cat and the second animal being 90% dog.

Clarification 197

It's only true that all of the probabilities for the **same data point** need to add to 1.

If you have $P\{\text{class 1}\}$ for one data point and $P\{\text{class 2}\}$ for another data point, those **aren't related**.

So, we want to scale our values so they add to 1: this is called **normalization**. How do we do that?

Well, let's say each class gets a **value** of r_i , before being **normalized**. For now, let's ignore how we got r_i , just know that we have it.

To make the total 1, we'll **scale** our terms by a factor C :

$$C(r_1 + r_2 + \dots + r_k) = C \left(\sum_{i=1}^k r_i \right) = 1 \quad (4.67)$$

We can get our factor C just by dividing:

$$C = \frac{1}{\sum r_i} \quad (4.68)$$

We've got our desired g_i now!

$$g = \begin{bmatrix} r_1 / \sum r_i \\ r_2 / \sum r_i \\ \vdots \\ r_k / \sum r_i \end{bmatrix} \quad (4.69)$$

4.5.4 Turning sigmoid multi-class

Now, we just need to compute r_i terms to plug in. To do that, we'll see how we did it using sigmoid:

$$g = \sigma(u) = \frac{1}{1 + e^{-u}} \quad (4.70)$$

This function is 0 to 1, which is good for being a probability.

Just for our convenience, we'll switch to positive exponents: all we have to do is multiply by e^u/e^u .

Negative numbers are easy to mess up in algebra.

$$g = \frac{e^u}{e^u + 1} \quad (4.71)$$

We'll think of **binary** classification as a **special case** of **multi-class** classification. The above probability could be thought of as g_1 : the probability for our first class.

Concept 198

Binary classification is a **special** case of **multi-class** classification with only **two** classes.

So, we can use it to figure out the **general** case.

So, what was our **second** probability, $1 - g$? This will be our second class, g_2 .

$$g_2 = 1 - g = \frac{1}{1 + e^u} \quad (4.72)$$

This follows an $r_i / (\sum r_i)$ format! The numerators (1 and e^u) add to **equal** the denominator ($1 + e^u$).

$$g = \begin{bmatrix} e^u / (e^u + 1) \\ 1 / (e^u + 1) \end{bmatrix} = \begin{bmatrix} r_1 / (r_1 + r_2) \\ r_2 / (r_1 + r_2) \end{bmatrix} \quad (4.73)$$

How do we **extend** this to **more** classes? Well, 1 and e^u are **different** functions: this is a problem. We want to be able to **generalize** to many r_i .

How do they make them **equivalent**? We could say $1 = e^0$. So, we could treat both terms as **exponentials**!

$$g_1 = \frac{e^u}{e^u + e^0} \quad g_2 = \frac{e^0}{e^u + e^0} \quad (4.74)$$

What if we want more classes? We just need more exponentials! They'll fit into the pattern from e^u and e^0 :

$$g_i = \frac{r_i}{\sum r_j} = \frac{e^{u_i}}{\sum e^{u_j}} \quad (4.75)$$

Now, we have a template for expanding into higher dimensions!

4.5.5 Our Linear Classifiers

What are each of those u_i terms? When we were doing **binary classification**, we used a **linear regression** function to help generate the probability:

$$u(x) = \theta^T x + \theta_0 \quad (4.76)$$

Remember that $u(x)$ is not a probability yet: we use a sigmoid to turn it *into* a probability.

Now, we want multiple probabilities. So, we create multiple different functions u_i : k different linear regression models (θ, θ_0) . We'll represent each vector as $\theta_{(i)}$.

$$\theta_{(1)} = \begin{bmatrix} \theta_{1(1)} \\ \theta_{2(1)} \\ \vdots \\ \theta_{d(1)} \end{bmatrix} \quad \theta_{(2)} = \begin{bmatrix} \theta_{1(2)} \\ \theta_{2(2)} \\ \vdots \\ \theta_{d(2)} \end{bmatrix} \quad \theta_{(k)} = \begin{bmatrix} \theta_{1(k)} \\ \theta_{2(k)} \\ \vdots \\ \theta_{d(k)} \end{bmatrix} \quad (4.77)$$

Each of these models could be seen as a "different perspective" of our data point: what about that data point is **prioritized** (large θ_i magnitudes), and how do we **bias** the result (θ_0)?

This "perspective" we call $\theta_{(i)}$ will tell us if our data point is "closer" to the **class** it represents

$$u_1(x) = \theta_{(1)}^T x + \theta_{0(1)} \quad u_2(x) = \theta_{(2)}^T x + \theta_{0(2)} \quad u_k(x) = \theta_{(k)}^T x + \theta_{0(k)} \quad (4.78)$$

These equations allow us to directly compute each u_i .

- Intuitively, if $u_i(x)$ is **larger than** $u_j(x)$, then our data point x is **more similar to** class i than class j .

In the last section, we emphasized that we can only use $\sum p_i = 1$ for the probabilities of a **single** data point. Based on this, we'll focus on only one data point.

Clarification 199

In this section, x represents only **one data point** $x^{(i)}$.

Softmax treats each data point **individually**, so it's easier to not group them together.

Having all these separate equations for θ_i is tedious. Instead, we can combine them all into a $(d \times k)$ **matrix**.

$$\theta = \begin{bmatrix} \theta_{(1)} & \theta_{(2)} & \cdots & \theta_{(k)} \end{bmatrix} = \begin{bmatrix} \theta_{1(1)} & \theta_{1(2)} & \cdots & \theta_{1(k)} \\ \theta_{2(1)} & \theta_{2(2)} & \cdots & \theta_{2(k)} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{d(1)} & \theta_{d(2)} & \cdots & \theta_{d(k)} \end{bmatrix} \quad (4.79)$$

k classes, so we need k classifiers. We'll stack them side-by-side like how we stacked multiple data points to create X.

And our constants, θ_0 , in a $(k \times 1)$ matrix:

$$\theta_0 = \begin{bmatrix} \theta_{0(1)} \\ \theta_{0(2)} \\ \vdots \\ \theta_{0(k)} \end{bmatrix} \quad (4.80)$$

Concept 200

We can combine **multiple classifiers** $\Theta_{(i)} = (\theta_{(i)}, \theta_{0(i)})$ into large **matrices** θ and θ_0 to compute **multiple** outputs u_i at the **same** time.

This will put all of our terms into a $(1 \times k)$ vector u .

$$u(x) = \theta^T x + \theta_0 = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_k \end{bmatrix} \quad (4.81)$$

Which creates a deceptively simple formula: this is one of the perks of matrix multiplication!

4.5.6 Softmax

We now have all the pieces we need.

- Our **linear regression** for each class:

$$\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_k \end{bmatrix} = \theta^T x + \theta_0 \quad (4.82)$$

- The **exponential** terms, to get **logistic** behavior

$$r_i = e^{u_i} \quad (4.83)$$

- The **averaging** to get the probabilities to add to 1:

$$g = \begin{bmatrix} r_1 / \sum r_i \\ r_2 / \sum r_i \\ \vdots \\ r_k / \sum r_i \end{bmatrix} \quad (4.84)$$

And so, our multiclass function is...

Definition 201

The **softmax function** allows us to calculate the probability of a point being in each class:

$$g = \begin{bmatrix} e^{u_1} / \sum e^{u_i} \\ e^{u_2} / \sum e^{u_i} \\ \vdots \\ e^{u_k} / \sum e^{u_i} \end{bmatrix}$$

Where

$$u_i(x) = \theta_{(i)}^T x + \theta_{0(i)} \quad (4.85)$$

If we are forced to make a **choice**, we choose the class with the **highest probability**: we return a **one-hot encoding**.

4.5.7 NLLM

One loose end left to tie up: our **loss function**. We need to evaluate our hypothesis, and be able to improve it.

For **binary classification**, we did **NLL**:

$$\mathcal{L}_{\text{nll}}(g, y) = - \left(y \log g + (1 - y) \log (1 - g) \right)$$

How do we make this work in **general**? Well, we want to make our two terms have a **similar** form, so we can generalize to more classes.

- g and $1 - g$ are both **probabilities**: we can think of them as g_1 and g_2 , respectively.
 - If $g = g_1$, then we would expect $y = y_1$.
 - Similarly, $1 - g = g_2$ pairs with $1 - y = y_2$.

$$\mathcal{L}_{\text{nll}}(g, y) = - \left(y_1 \log g_1 + (y_2) \log (g_2) \right)$$

They have the **same** format now! Much tidier. And it tracks: when one **label** is correct, the other term is $y_j = 0$, and **vanishes**.

Does this **generalize** well? It turns out it does: with **one-hot encoding**, the correct label is **always** $y_j = 1$, and the incorrect labels are **all** $y_j = 0$.

So, we'll write it out:

Key Equation 202

The **loss** function for **multi-class** classification, **Negative Log Likelihood Multiclass (NLLM)**, is written as:

$$\mathcal{L}_{\text{NLLM}}(\mathbf{g}, \mathbf{y}) = - \sum_{j=1}^k y_j \log(g_j)$$

Because of **one-hot encoding** for y , all terms except one have $y_j = 0$, and thus **vanish**.

Using all of these functions, we can finally do gradient descent on our multi-class classifier. However, we won't go through that work in these notes.

4.5.8 A side comment: Sigmoid vs. Softmax

Let's jump back to softmax real quick and clarify something.

Usually, we expect to use **softmax** if we have **more than 2** classes, because that's what we built it for.

- However, this isn't always the case – there's another aspect of softmax we haven't focused on:
- **Softmax** represents k different classes/events. These classes are assumed to be **mutually exclusive**: you can't be in multiple at the same time.

In other words, the classes of softmax are **disjoint**.

Definition 203

If two events are **disjoint**, they **can't** happen at the **same time**.

If n events are **disjoint**, only **one** of them can happen at a time.

Example: We usually wouldn't classify an animal as both a cat and a dog: it's either one or the other.

When events are disjoint, their probabilities are separate:

Concept 204

If two events are **disjoint**, then they have **separate** probabilities: there's no overlap. Since $P\{A \cap B\} = 0$, we can say:

$$P\{A \cup B\} = P\{A\} + P\{B\}$$

If we have **every** event and they're all **disjoint**, then their probabilities sum to 1.

$$\sum_i p_i = 1 \quad (4.86)$$

Example: If the weather options are rain, cloudy, and sunny, and you have to only choose one, you should expect that:

$$P\{\text{Rain}\} + P\{\text{Cloudy}\} + P\{\text{Sunny}\} = 1 \quad (4.87)$$

~~~~~  
But this only makes sense **if** events can't happen at the same time.

- In some situations, multiple classes/events can happen at the same time! They might even be **independent**.
- **Example:** There might be  $k$  different people we could find in an **image**. But, there can be multiple people in the **same** image!

So, it doesn't always make sense to assume that each event is **mutually exclusive**.

- If our events are not mutually exclusive, then we **shouldn't use softmax**.

The solution: for each class, find the **probability** for that class, vs. the **absence** of that class.

- This brings us back to binary classification: we just use **one sigmoid per class**.

**Clarification 205**

**Softmax** is used when each of our  $k$  classes is **disjoint** (mutually exclusive).

- However, if they aren't, then we **can't use softmax**.

~~~~~  
If our classes are independent, we can use k **sigmoid** functions: one for each of our k classes. We're using **binary classification** on each class separately.

- The i^{th} sigmoid tells us how likely the **data point** is to be in the i^{th} class.

Example: We might have an algorithm figuring out which **products** a customer might want. They might want **multiple**, so we can't treat them as disjoint.

In this case, each product is a class, and we determine the result based on the matching sigmoid

What happens if the events aren't disjoint, but they also aren't independent? You need a more complex model.

4.6 Prediction Accuracy and Validation

We've been working in **probabilities**, but in the end, the goal is usually to make a **decision** or **prediction**: which class do we pick?

In general, we just pick the class we predict with the **highest** probability.

And in practice, we often don't care about how close we were to right - we just care about how often we **were** right.

So, we use **accuracy**.

Definition 206

The **accuracy** A of our model is the **percentage** of the time we get the **right** answer.

We can write this as

$$A(h; D) = 1 - \frac{1}{n} \sum_{i=1}^n \mathcal{L}(h(x^{(i)}), y^{(i)})$$

Where \mathcal{L} is 0-1 loss (**counting** the number of **wrong** answers)

Or, "one minus how often we get the answer **wrong**".

4.7 Terms

- Class
- Classification
- Label
- Binary Classification
- 0-1 Loss
- Linear Classifier
- Separator
- Orientation
- Boundary
- Normal Vector
- Dot Product (Conceptual)
- Linear Separator
- Sign Function
- Hyperplane
- Separability
- Non-separate data
- Sigmoid Function
- Logistic Regression
- Prediction Threshold
- Linear Logistic Classifier (LLC)
- Negative Log Likelihood (NLL)
- Multi-class Classification
- One-Hot Encoding
- Normalization
- Softmax Function
- Negative Log Likelihood Multi-Class (NLLM)
- Accuracy

CHAPTER 5

Feature Representation

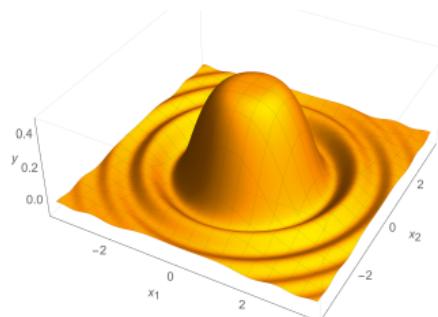
What's still missing?

Last chapter, we used our linear regression model to do classification: we created a "hyperplane" to **separate** the data that we placed in each class.

We also mentioned that regularization can increase **structural error**, by limiting what possible θ models we're allowed to use.

But, what if our linear model is already **too limited**? What if we need a more complicated model? This is true in a lot of real-world problems, like vibration:

Our goal was to decrease estimation error, but that's beside the point right now.



This wave doesn't seem particularly friendly to a planar approximation.

These kinds of situations are called, appropriately, **non-linear**.

Concept 207

Non-linear behavior cannot be accurately represented by any **linear** model.

In order to create an accurate model, we have to use some **nonlinear** operation.

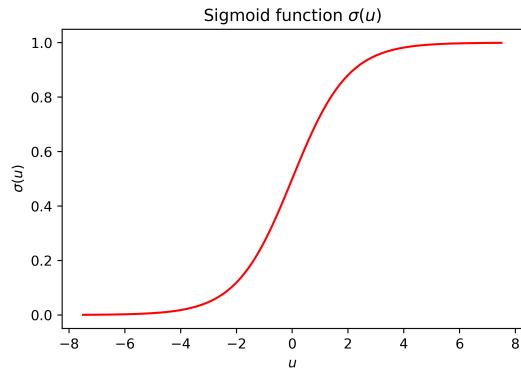
If we could create effective, non-linear models, we might even be able to deal with data that was previously "linearly inseparable".

~~~~~

## Possible Solutions: Polynomials

Let's try to think of ways to approach this problem. We'll start with a 1-D input, for simplicity.

Upon hearing "non-linear", we might remember the function we introduced last chapter: the **sigmoid**.



Your friendly neighborhood sigmoid.

Can we use this to create a new model class? For now, unfortunately not: remember that we used this in the last chapter, and we still got a **linear** separator. The reasons were discussed there.

Instead, we can get inspiration from our example of "structural error". For now, let's focus on **regression** (though classification isn't too different):

We'll show ways we can use this kind of approach, when we discuss Neural Networks.

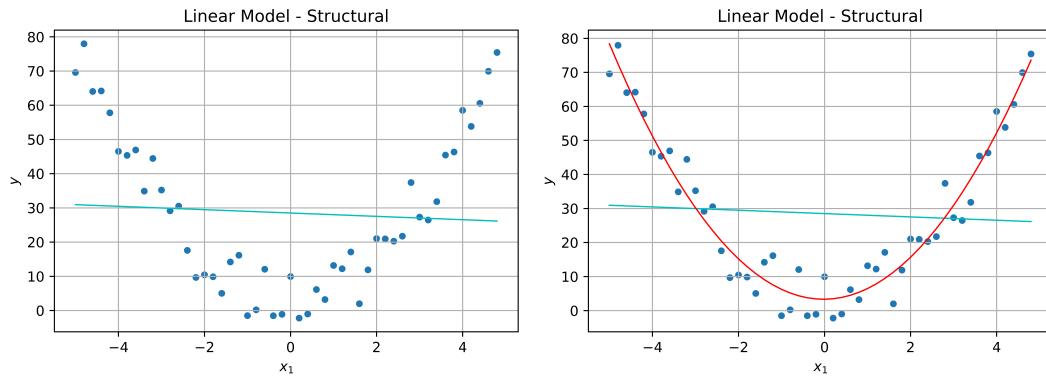


Figure 5.1: A linear regression can't represent this dataset. However, a parabola can!

We're still using our input variable  $x$ , but this time, we've "transformed" it: we have squared  $x$ , giving us a model of the form

Remember that  $x$  is 1-D right now!

$$h(x) = Ax^2 + Bx + C \quad (5.1)$$

It should be clear that this model is more **expressive** than the one before: it can create every model that our linear approach could (just by setting  $A = 0$ ), and it can create new models in a parabola shape.

### Concept 208

We can make our **linear** model more **expressive** by add a squared term, and turning it into a **parabolic** function.

Reminder: "expressiveness" or "richness" of a hypothesis class is how many models it can represent: a more expressive model can handle more different situations.

This concept can be extended even further, to any **polynomial**.

This is called a **polynomial transformation** of our input data.



## Transformation

How do we *generalize* this concept? Well, we have a set of constant parameters  $A, B, C$ . These are similar to our constants  $\theta_i$ . Let's change our notation:

$$h(x) = \theta_2 x^2 + \theta_1 x + \theta_0 \quad (5.2)$$

Now, we've got something more familiar. We could imagine extending this to any number of terms  $\theta_i x^i$ : if we needed a cubic function, for example, we could include  $\theta_3 x^3$ .

This is starting to look pretty similar to our previous model: in fact, we could even separate out  $\theta$  as a parameter:

Notice that  $\theta_0$  corresponds to  $x^0 = 1$ .

$$h(x) = \underbrace{\sum_{i=1}^k \theta_i x^i}_{\text{Polynomial sum}} = \underbrace{\begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \end{bmatrix} \cdot \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^k \end{bmatrix}}_{\text{Store as vectors}} = \underbrace{\theta^T}_{\text{Simplify}} \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^k \end{bmatrix} \quad (5.3)$$

This really *is* starting to look like our linear transformation  $\theta^T x$ . That's helpful: we might be able to use the techniques we developed before.

In fact, we can argue that they're **equivalent**: we've just changed what our input vector is.

Consider our new input  $\phi(x)$ :

$$\phi(x) = \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^k \end{bmatrix} \quad h(x) = \theta^T \underbrace{\phi(x)}_{\text{New input}} \quad (5.4)$$

Compare the structure of  $\theta^T x$  versus  $\theta^T \phi(x)$ : you've replaced  $x$  with  $\phi(x)$ .

This is called **transforming** our input. However, polynomial is only one of our transformations!

### Definition 209

A **transformation**  $\phi(x)$  takes our input vector  $x$  and converts it into a **new** vector.

This transformation can be used to:

- Allow our model to handle new, more **complex** situations
  - **Example:** Polynomial transformations
- **Pre-process** our data to make certain **patterns** more obvious, and easy for our model to detect.
  - **Example:** Radial transformations (to be discussed later!)
- Convert our data into a **usable** format (if the original format doesn't fit into our equations)
  - **Example:** One-hot encoding

**Example:** Taking our input  $x$  and converting it into a polynomial is a **transformation** of our input.

This chapter will focus on these kinds of transformations.



## Features

One benefit of only changing the input is that everything else we know about the model is still true: we can continue to use our linear representation.

- We will be able to optimize a "linear" model  $\theta$ , over data that has been made **nonlinear**.

These transformations can be complex, especially for **multi-dimensional** inputs. In this first case, we only combined one input ( $x = x_1$ ) with **itself**. But, often, we can combine multiple  $x_i$  together!

So, we need to be careful of our input variables  $x_i$ .

We'll cover this multi-dimensional polynomial transformation later in the chapter.

- We sometimes call a single input variable, a single "**feature**". However, we need to be careful: this word can have multiple meanings.

### Clarification 210

We often use the word **feature** in related (but not identical) contexts:

- A **feature** can be one unprocessed **aspect** of the **data**:
  - **Example:** Whether or not something is a cat or a dog, or the height of a patient.
- A **feature** can also be one mathematical **variable** in our **transformed input**.
  - $x_i$  is a **feature** of the **data**.
  - $\phi(x)_j$  (one variable in  $\phi(x)$ ) is a **feature** of the **transformed data**.

Just like how we have an **input space** and a **hypothesis space**, we call the collection of possible values for our features the **feature space**.

Combined, this is why we called this technique the **feature transformation**:

- We apply a transform to the **features**  $x$  of our data, and create a new list of **features**  $\phi(x)$ .

Since these transforms only apply to our features, they don't affect the rest of our model. So, we can still use **linear** tools:

**Definition 211**

Feature transformation allows us to do linear regression or classification on a list of features we have non-linearly transformed:

$$h(x) = \theta^T \phi(x)$$

- The  $\theta^T u$  operation is still linear ( $u = \phi(x)$ ).
- All non-linearity is stored in  $\phi(x)$ .

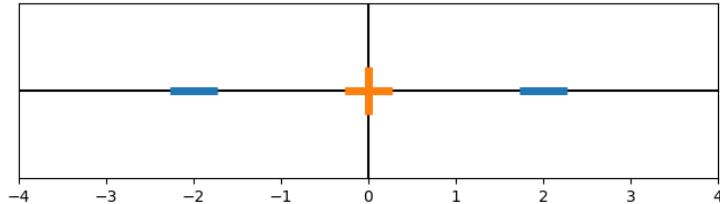


## 5.1 Gaining intuition about feature transformations

Now that we understand the general idea of feature transformations, we can begin work with them, particularly for **classification**.

Our goal is often to take data that linear models couldn't handle, and make them **more accurate**.

So, we'll consider maybe the simplest (solvable) case of a nonlinear data set in 1-D:



The y-axis doesn't exist, it just has vertical height to make it easier to view on a page.

Figure 5.2: In this state, there's no 0-d plane (point) that would **separate** these data points.

This is where our transform comes in: we can't separate using just  $x$ . So, we'll introduce a second variable:  $x^2$ .

- We want a function that lets us classify some data points as positive, and some as negative.
- For the dataset in this image,  $-x^2 + 2 > 0$  gives us 100% accuracy. Let's see it in action:

We're replacing  
 $\theta_0 + \theta_1 x > 0$  with  
 $\theta_2 x^2 + \theta_1 x + \theta_0 > 0$ .

$$\begin{array}{lll} x = 2 & \parallel & y = -(2)^2 + 2 = -2 \\ x = 0 & \parallel & y = -(0)^2 + 2 = 2 \\ x = -2 & \parallel & y = -(-2)^2 + 2 = -2 \end{array} \quad \begin{array}{ll} y < 0 \implies \text{Negative} \\ y > 0 \implies \text{Positive} \\ y < 0 \implies \text{Negative} \end{array}$$

How do we visualize this? It turns out, there are different perspectives:

**Clarification 212**

There are **two** different ways we can **graph** a transformation:

We transform the **hyperplane**:

- **Example:** If our model is  $f(x) = -x^2 + 2$ , we just graph  $y = f(x)$  as our separator in 2D space.
  - This is the approach we used to start the chapter: we wanted a line that **fit** to our data.
  - In practice, this bends our **hyperplane** into a curve: at the start of the chapter, we transformed a line into a parabola.

Or, we transform the **data**:

- **Example:** We plot each data point in 2D as  $[x, -x^2 + 2]$ .
  - This model allows us to keep a "**linear** separator": we "shift" the data nonlinearly, **then** linearly separate it.

These models are mathematically **equivalent**, and we'll switch the approach we're using based on which is easier/more useful to graph.

See our plot examples for each below.

Note that our nonlinear transformation "adds" dimensions: we had a 1D problem, and we used a second dimension to separate it.

It may seem concerning to transform the **data**, rather than the **model**. The data is what we're using to make decisions, after all.

However, keep in mind that:

- Our model already was a sort of **transformation**: even the linear model  $\theta$  "transforms" each data point  $x$  into  $y = \theta^T x$ .
- Usually, we try to preserve the **original structure** of the data, so we don't lose information: we just add more.
  - For example,  $[1, x, x^2]$  still contains the information  $x$ : we just append 1 and  $x^2$ .

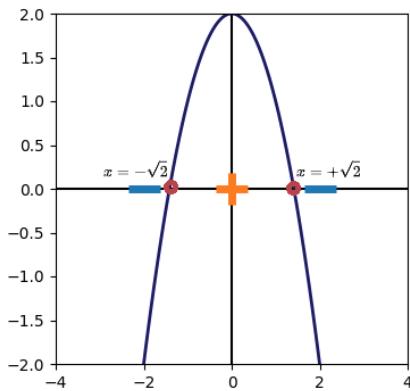
**Example:** Let's show both of these in action, using the 1-D dataset we showed above.



### 5.1.1 Transforming our separator

First, we transform our linear separator as desired: graphing  $-x^2 + 2 = f(x)$ .

- Our separator points are still on the  $x$ -axis: they "separate" our data wherever  $f(x) = 0$ .



In this version, we've taken our hyperplane separator and transformed it nonlinearly.

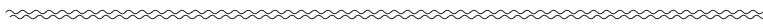
To correctly classify, we assign  $-x^2 + 2 > 0$  as positive.

In this version, we preserve the structure of the data, making it easier to see the original shape.

However, it's not as easy to think about the direction and orientation of the "plane" now that it's been deformed into a parabola.

- For example, we don't really have a good "normal" vector, even if we know which side is positive.

This is why, to keep our model "linear", we can transform the **data**, instead of the separator. We'll do that next.

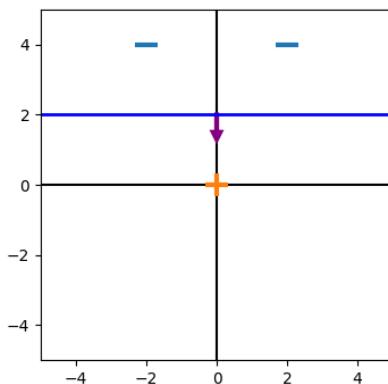


### 5.1.2 Transforming our data

In this case, every data point gets plotted on  $[x, x^2]$ . Our hyperplane is given by

$$-x^2 + 2 = \underbrace{\begin{bmatrix} 0 \\ -1 \end{bmatrix}}_{\theta^T}^T \begin{bmatrix} x \\ x^2 \end{bmatrix} + 2 = \theta^T \phi(x) + \theta_0 \quad (5.5)$$

Thus, we get a  $\theta$  plane pointing downward, with an offset of 2.



This time, we've transformed our data: the math is totally the same, but now we can identify our separator more easily.

Note that our transformation makes the data linearly separable!

#### Concept 213

Features transformations allow us to **non-linearly** transform our data, in order to make that data **linearly separable**, or at least, more **accurate** with a linear separator.

Often, we do this by transforming into a **higher dimensional** space.



### 5.1.3 Positive vs. Negative

While these perspectives are helpful, they can become too complicated with more dimensions/higher-dimensional transformations.

In an effort to simplify, we might ask ourselves, "what do we really want to know"? In the end, all we typically care about is classification: which data points are positive or negative?

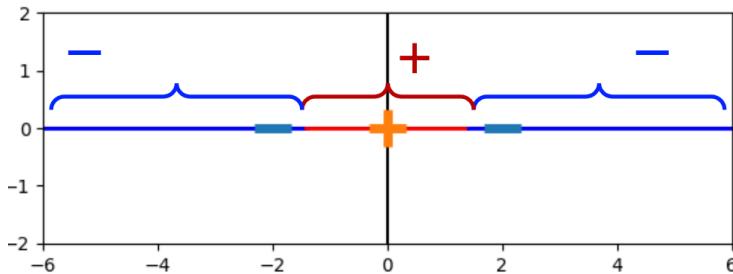
So, we'll create a third representation to correspond to that.

#### Concept 214

A third, **simplified** representation of our transformation doesn't show how it affects our data points or classifier. Instead, we just show the **result**: which regions are classified as positive, and which are classified as negative?

This allows us to see which points get **classified** in which way, without considering the high-dimensional details of the model itself.

**Example:** We can graph this for our sample data:



This way, we can stay in a 1-D space, while showing the information we need!

Note that the points where we switch between positive and negative,  $\pm\sqrt{2}$ , are the points corresponding to  $-x^2 + 2 = 0$ : they're the only part of the separator surface visible in our 1D plot.

They match our nonlinear hyperplane separator from section 5.1.1

## 5.2 Systematic feature construction

Now that we've established feature transformations, let's consider a couple options for how we'd want to do it, and how we can generalize to higher dimensions.

Here, we'll present two common ways to construct features, in a way that's consistent across problems, or "systematic".

We could also call this "problem independent": it works regardless of what kind of problem you have. Though, that doesn't mean problem type won't affect performance.

### 5.2.1 Polynomial Basis

At the start of this chapter, we introduced the idea of polynomial transformations.

If a linear function isn't "expressive" enough to solve a problem, then we can create a more complex model, based on how many  $x^i$  we include. This can be written as:

$$h(x) = \sum_{i=1}^k \theta_i x^i = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \end{bmatrix} \cdot \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^k \end{bmatrix} = \theta_0 + \theta_1 x^1 + \theta_2 x^2 + \cdots + \theta_k x^k \quad (5.6)$$

Another word for these  $x_i$  terms might be a "polynomial **basis**".

- Why call it a basis? Well, we can use our  $x_i$  terms to create a polynomial, using

$$\sum_i \theta_i x^i \quad (5.7)$$

- Using this procedure, we can combine **basic** elements  $x_i$  to create any **polynomial**.

$$\{1, x_1, x_2, \dots, x_k\} \quad (5.8)$$

- This is what defines it as a **basis**: the ability to combine these terms, make any polynomial we want.

#### 5.2.1.1 Order

An important question to ask is, "how many terms do we include"?

We categorize our polynomials based on the highest exponent included: this is called the **order**.

**Definition 215**

**Order**  $k$ , also known as **degree**, is the **largest** exponent allowed in our **polynomial**.

Every higher exponential  $x^j$  can be thought of as having a coefficient  $\theta_k = 0$ : as far as we're concerned, it **doesn't exist**.

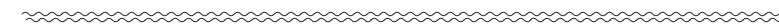
**Example:** We can compare different orders, by looking at the feature vector they create.

Here's a table of the first few:

| Order | $d = 1$                     | Example                     |
|-------|-----------------------------|-----------------------------|
| 0     | [1]                         | 3.5                         |
| 1     | $[1, x]^T$                  | $2.5x - 1$                  |
| 2     | $[1, x, x^2]^T$             | $4.1x^2 - 10x + 1$          |
| 3     | $[1, x, x^2, x^3]^T$        | $x^3 + 8x^2 + x - \sqrt{2}$ |
| :     | :                           | :                           |
| $k$   | $[1, x, x^2, \dots, x^k]^T$ | $\sum_{j=1}^k \theta_j x^j$ |
| :     | :                           | :                           |

Note that, while we chose every coefficient to be nonzero here, they don't have to be!  $-x^2 + 2$  from before is a valid second-order polynomial.

The order we choose is an important design choice.



### 5.2.1.2 Overfitting with order

It's difficult to know how many terms to include in our polynomial, but we run into two problems if our order is **too high**:

- It becomes time-consuming to calculate, with little benefit
- We start overfitting more and more.

The first part makes sense: with more terms, we have to do more multiplications, more additions, etc.

**Concept 216**

More **complex models** tend to be more **expensive** to train, and slower to use. This is a trade-off for more **accuracy**.

Usually, there's a point where cost **outweights** benefits. A problem is rarely perfectly solved, even by an excellent model, so you can't just continue until it's "perfect".

But what about the second part? Why do we increase overfitting?

With a higher order, our polynomial becomes more complex: it can take on more shapes, which are increasingly complex and perfectly fit to the data.

This can cause our data to overlook obvious patterns, and instead create a very precise shape that is paying attention to the noise in our model.

### Concept 217

**High-order polynomials** are very vulnerable to **overfitting**.

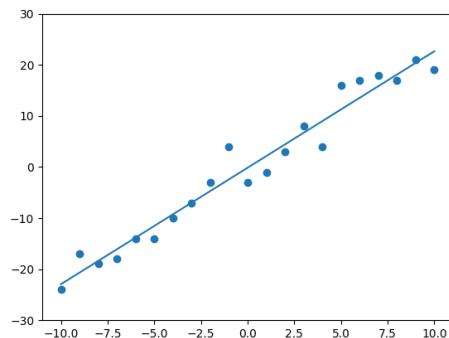
Because they can take on so many different, **complex** functions, they can very very closely **match** the original data set.

This can cause the model to "learn" noise, and **miss** broader and simpler patterns that actually exist. It may fail to learn something broad and useful, while **memorizing** the dataset with its expressiveness.

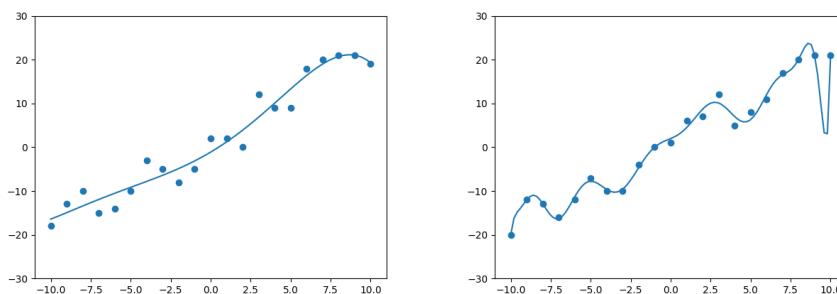
Let's see this in action: we'll generate some data based on  $2x + 1$ , while applying some random noise to it. We'll see the optimized linear regression model for each.

Rather than transform the data, we'll transform the separator: this really highlights the overfitting effect.

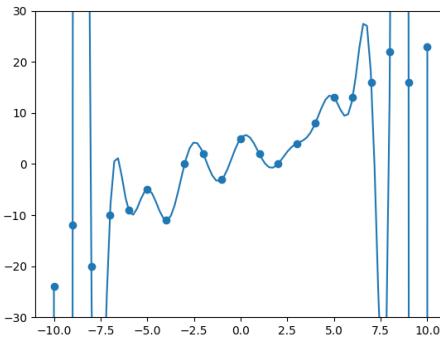
For ease, we'll exclude regularization: it does help mitigate this problem, but it doesn't totally solve it.



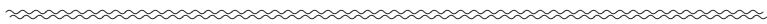
Here's the 1st order solution: in this case, correct for the underlying distribution. It fits our data fine.



5th and 15th order. The left model looks suspicious, and the right is way overfit. It's very unlikely that we know such an intricate pattern, from so little data.



20th order. We have one order for each data point: now, our model is capable of doing regression going through every single data point: as overfit as physically possible, perfectly matching the data.



### 5.2.1.3 Higher dimensions

Until now, we've only been focusing on the 1-D case of data. Let's change that. Let's consider a 2D dataset  $[x_1, x_2]^T$ .

We start with our typical 2D model:

$$\theta^T x + \theta_0 = \theta_1 x_1 + \theta_2 x_2 + \theta_0 \quad (5.9)$$

Is polynomial basis, with "order 1": the largest exponent is 1. This is still a "linear" model.



If we want to move up to order 2, we increase the **largest exponent**, adding  $x_1^2$  and  $x_2^2$  to the basis.

However, this doesn't take full advantage of the expressiveness of our model: this only creates parabolas aligned with the  $x_1$  and  $x_2$  axes. How do we create other options?

Well, we created these options by multiplying  $x_1$  with another  $x_1$ . It seems like we could logically expand to multiplying  $x_1$  by  $x_2$ .

#### Definition 218

For **higher dimension**  $d > 1$  **polynomials**, we allow for multiplication **between variables**  $x_i$  and  $x_j$ .

The **order** of the polynomial is the maximum number of times you can **multiply variables** together.

For order  $k$ , the **sum of exponents** must be **less than or equal to** the order.

So, for  $d = 2$ , order=2, we get the basis:

$$\begin{bmatrix} 1 & \textcolor{red}{x}_1 & \textcolor{red}{x}_1^2 & \textcolor{blue}{x}_2 & \textcolor{blue}{x}_2 x_1 & \textcolor{blue}{x}_2^2 \end{bmatrix}^\top \quad (5.10)$$

For  $d = 2$ , order=3, it starts getting a bit messy: we have 10 different basis terms.

You don't need to memorize these.

$$\begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & x_2 & x_2 x_1 & x_2 x_1^2 & x_2^2 & x^2 x_1 & x^3 \end{bmatrix}^\top \quad (5.11)$$

~~~~~

5.2.1.4 Total number of features

How many do we have in general? Well, every term results from multiplying variables **at most** k times. Or, the exponents **add up** to at most k .

If we count 1 as a factor, we can say that the exponents always add up to k (since 1^j has no effect). So, we have $d + 1$ different numbers, which add up to exactly k .

We have d variables, so $d + 1$ if we include 1 as a variable.

For example, here's how this would look for $d = 2$, $k = 2$ (5.10 above). All you need to notice is that the exponents add to 2.

$$\begin{bmatrix} 1^2 x_2^0 x_1^0 & 1^1 x_2^0 x_1^1 & 1^0 x_2^0 x_1^2 & 1^1 x_2^1 x_1^0 & 1^0 x_2^1 x_1^1 & 1^0 x_2^2 x_1^0 \end{bmatrix}^\top \quad (5.12)$$

This is a well-known problem in combinatorics: how many ways are there to add up $d + 1$ numbers to total k ? The solution to this problem gives us:

$$\binom{(d+1)+k-1}{k} = \binom{d+k}{k} = \frac{(d+k)!}{d!k!} \quad (5.13)$$

Explaining the math here is beside the point of this course. If you're curious, search up "stars and bars math", or visit [here](#).

5.2.1.5 Summary of Polynomial Basis

Definition 219

The **polynomial basis** of order k and dimension d includes **every feature**

$$\prod_{i=1}^d x_i^{c_i}$$

Where all of the integer exponents $c_i \geq 0$ add up to **at most** k .

Creating features such as:

$$x_1^k, x_1 x_2, x_2 x_3^3 x_6, 1, \dots$$

We can represent this in a table:

This table is different from the one we saw earlier!

Order	$d = 1$	in general ($d > 1$)
0	[1]	[1]
1	$[1, x]^T$	$[1, x_1, \dots, x_d]^T$
2	$[1, x, x^2]^T$	$[1, x_1, \dots, x_d, x_1^2, x_1x_2, \dots]^T$
3	$[1, x, x^2, x^3]^T$	$[1, x_1, \dots, x_1^3, x_1x_2, \dots, x_1x_2x_3, \dots]^T$
\vdots	\vdots	\vdots

5.2.1.6 Polynomial Basis in Action

Below, we'll show examples of how polynomial basis being used, to demonstrate just how useful it is.

But how do we use feature transformations? The best part: remember that we can view it as a linear separator? We can train it just the same way!

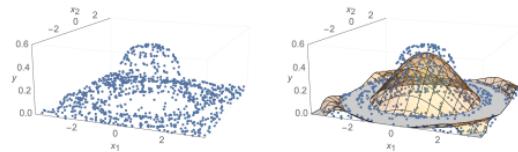
Concept 220

Feature transformations don't change how we train our model. We can still treat our model as a **linear** vector θ , even if our data has been **non-linearly** transformed.

So, after we transform our data, we can use normal techniques (OLS, gradient descent, SGD) to fit our model.

In this situation, the benefits of regularization become more clear: by preventing θ from becoming too large, we discourage a surface that is too "extreme", with larger changes across its surface.

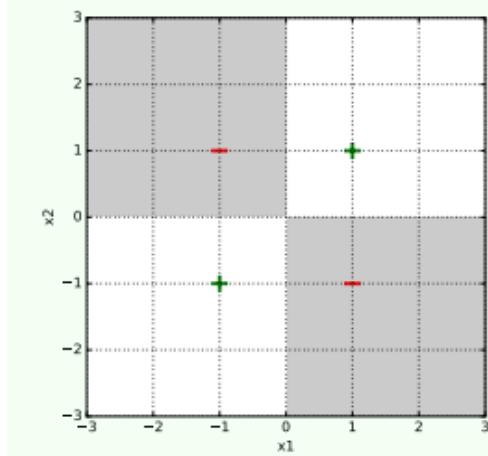
We'll train our model for various situations, to see what it can do. Different problems require different orders k , still.



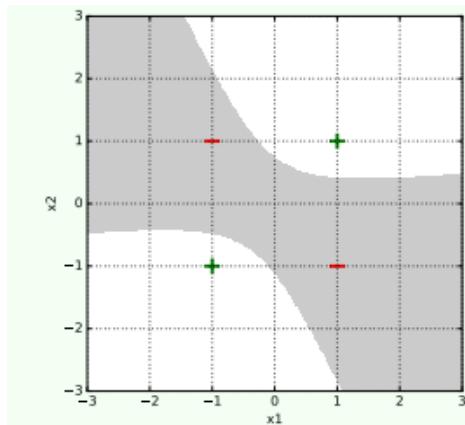
We start with the waveform we showed at the start of the chapter: on the left, we see a bunch of datapoints we want to take a regression over. With $k = 8$, we get a pretty good result.

For 2D separators, it's easier to show only the +/- classification, rather than the transformed data/boundary. That means, these below graphs are hiding the numeric outputs.

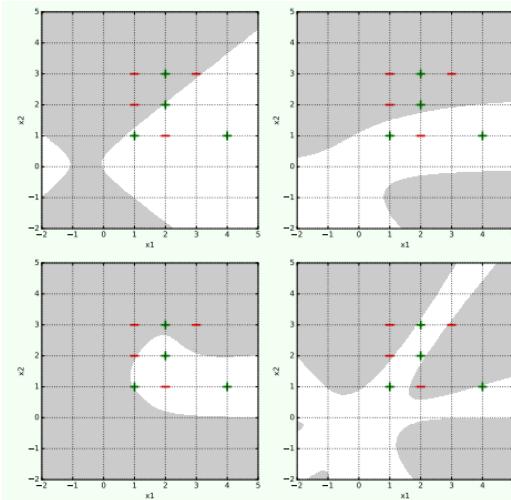
Light indicates "positive" for the model, dark indicates "negative" for the model.



This is the classic "xor" problem: a typical case of "linearly unseparable". With $k = 2$, we can classify it well with the chosen model $4x_1x_2 = 0$.



This time we use gradient descent and a random initialization to get a less rigid, but still effective classification.



This dataset is pretty brutal: we try with $k = 2, 3, 4$, and finally 5. The shapes we get are... complex, to say the least. But, we successfully solve it with $k = 5$.



5.2.2 Radial Basis

Finally, we consider an alternative way to create a feature space.

- With the "polynomial basis" approach, we **combined features** to create more complex surfaces to **fit** the structure of the data.
- This "radial basis" approach, on the other hand, **combines data points** to **learn** more about the structure of the data.

What do we mean by that? Well, let's consider what we mean by "structure": when we're judging data, what sorts of patterns are we looking for?

Often, we're looking to see, "what data is near/similar to other data?" Similar data is more likely to behave similarly, after all.

So, it might be useful to include distance between data points as a feature: how do we implement this? Well, let's do this one-by-one: we'll create a feature for the distance to a single data point p .

We'll come back to these ideas when we talk about clustering!

We start with squared distance, for smoothness reasons.

$$\|p - x\|^2 \quad (5.14)$$

This feature would *grow* as data points get further apart, though. We want to see what data is *close*: the opposite.

We could use a function like $\frac{1}{u}$. However, this would explode to infinity as distances shrink: not good.

e^{-u} is a better fit: it approaches 1 when $u = 0$, and, relatedly, it tends to drop off more smoothly and gradually than $1/u$.

Finally, we add a coefficient β to the exponent to give us more control: it will tell us how quickly our function decays with distance.

The word "decay" is used commonly for exponential decrease.

Definition 221

We define the **radial basis function**

$$f_p(x) = e^{-\beta \|p-x\|^2}$$

As a **feature** in the RBF feature transformation.

This transformation takes a data point p and provides a feature $f_p(x)$ that represents "**closeness**" of x to p .

Note some useful properties of this transform:

- For small distances, this feature creates a **connection** between p and our data point: representing some local "structure" of **closeness**.
- If points are far away, this effect gradually **vanishes**: points which are **far** away have very little to do with each other.
- β controls what is considered "close" and "far":
 - if β is large, points have to be very close for an effect.
 - if β is small, we have a larger "neighborhood" of points with a relevant $f_p(x)$.

Definition 222

The **radial basis functions (RBF)** transform takes each of the data points in the input, and uses it to create a set of **radial basis function** features.

Collectively, they make the **feature space**:

$$\phi(x) = [f_{x^{(1)}}(x), f_{x^{(2)}}(x), \dots, f_{x^{(n)}}(x)]^T$$

Where:

$$f_p(x) = e^{-\beta ||p-x||^2}$$

This transform allows us to represent "closeness" within our dataset. With it, we can compare new data points to some "reference" points $x^{(i)}$.

It's often used to allow us to represent our dataset in a way that is approximate, but still useful.

This general idea is useful for problems like:

- Function approximation,
- Optimization,
- Reducing noise in signals

This approach is not limited to the "squared distance" idea of closeness, either: if you can come up with another way to define distance, you can use the same approach.

Reminder that "noise" just refers to anything undesired in the signal. Usually added by randomness or the environment.

These ways to define distance are called "distance metrics".

5.3 Hand-constructing features for real domains

So far, we've focused on transformations that handle two of our main problems (which have a lot of overlap): _____

- Allow our model to handle new, more **complex** situations (more **expressiveness**)
- **Pre-process** our data to make it **easier** for our model to find **patterns**.

Borrowed from the transformation definition above.

Now, we'd like to turn our attention to the last of the three:

- Convert our data into a **usable** format (if, say, the original format doesn't fit into our equations)

One challenge with our models are they rely on computation and calculation. This usually require our input to be something like a **number**.

But we don't always receive data in this way: words, brands, colors, and odd others data types, are often presented instead. Frequently, we even need to **adjust** our numerical inputs.

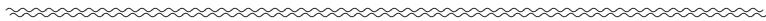
The **transformations** in this section address these kinds of problems. We take data that is informative, but not currently usable, and converting them to something we can **compute** with.

Concept 223

Often, we have to **convert** between data types, in order to do the machine learning work we want to do.

This requires using the appropriate **transforms** to get data we can work with, without **losing** important **information**.

As mentioned above, we have to be **careful**: if we use the wrong data type, we can **lose** crucial information that our model could have made use of.



5.3.1 Discrete Features

One of the most common issues with data types is figuring out what to do with **discrete** features: ones that are broken up into categories.

These categories may or may not have an order, or some other important information. We need to use the right data type to keep as much information as possible. This will allow our model to more easily discover patterns.

We'll make the following assumption:

Clarification 224

In this textbook/course, we assume that all **input vectors** x should be **real-valued vectors** (or: $x \in \mathbb{R}^d$)

And now, we go through some common examples of feature transformations:

5.3.1.1 Numeric

First, we consider the case where our pre-processing **feature** is "almost" in a number format: each class could reasonably correspond to a **number**.

Definition 225

In the **numeric** transformation, we convert each of our k classes into a **number**.

- This approach is only appropriate if each class is roughly "numeric": it fits appropriately into the **real numbers**.
 - We have a clear **ordering**, and
 - The numbers have the **structure** of real numbers: **distance** between points, or the idea of **adding/multiplying**, makes sense.

Example: There are many ways to do this. Here, we evenly distribute values evenly between 0 and 1: _____

$$\left[\frac{1}{k} \quad \frac{2}{k} \quad \dots \quad \frac{k-1}{k} \quad 1 \right], \quad \text{Class } i \longrightarrow \frac{i}{k} \quad (5.15)$$

Remember: which way you transform should reflect the nature of your data!

5.3.1.2 Thermometer Code

Next, we'll relax how number-like our feature is. This time, we don't need our data to behave like a number, but it does have an **ordering**. _____

Some examples:

- Results of an opinion poll:
 - "Strongly Agree", "Agree", "Neutral", "Disagree", "Strongly Disagree"
- Education level:
 - "Below High School", "High School Degree", "Associates Degree", "Bachelors" "Advanced Degree"
- Ranking of athletes

By "relax", we mean we'll remove some requirements for our feature, like being able to add them together.

In this case, we can't just use numbers $\{1, 2, 3, \dots\}$. Why not?

Because that implies that there's a specific "scaling" between points: Is the #1 athlete twice as good as the #2 athlete? Maybe, but that's not what the ranking tells us!

Concept 226

Data that is **ordered** but not **numerical** cannot be represented with **a single real number**.

Otherwise, we might consider one element to be a certain amount "larger" or "smaller" than another, when that's not what **ordering** means.

Example: Suppose we assign $\{1, 2, 3\}$ for $\{\text{Disagree}, \text{Neutral}, \text{Agree}\}$. The person who writes 'agree' is doesn't "agree three times as much" as the person who writes 'disagree'!

But, we still want to keep that ordering: counting up from one element to the next. How do we create an order without creating an exact, numeric difference?

Just now, we tried to count by increasing a single variable. But, there's another way to count: counting up using multiple different variables!

This approach is more similar to counting on your fingers.

$$\begin{array}{llll} \text{Class 1} \rightarrow \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} & \text{Class 2} \rightarrow \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} & \text{Class 3} \rightarrow \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} & \text{Class 4} \rightarrow \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \end{array}$$

This version allows us to avoid the problems we had before: this doesn't behave the same way as a **numerical** value.

To better understand what's going on, here's another way to frame it:

Class(x) is just shorthand for, "which class is x in?"

$$\phi(x) = \begin{cases} \text{Class}(x) \geq 4 \\ \text{Class}(x) \geq 3 \\ \text{Class}(x) \geq 2 \\ \text{Class}(x) \geq 1 \end{cases} \quad (5.16)$$

Example: Suppose x is in class 3. The bottom three statements are all true, the top one is false: so we get $[0, 1, 1, 1]^T$.

This helps us understand why this encoding is so useful:

- We aren't directly "adding" variables to each other: they stay separated by **index**.
- When using a linear model $\theta^T \phi(x)$, each class matches a different θ_i .
- Despite not behaving like numbers, "higher" classes in the order still keep track of all of the classes "below" them.

θ_i scales the i^{th} variable. So, each class can be scaled differently!

- **Example:** Class 2-4 all share the feature $\text{Class}(x) \geq 2$ (equivalent to $\text{Class}(x) > 1$).

This technique is called **thermometer encoding**.

Definition 227

Thermometer encoding is a **feature transform** where we take each class and turn it into a feature vector $\phi(x)$ where

$$\text{Class 1} \rightarrow \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad \text{Class 2} \rightarrow \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 1 \end{bmatrix}, \quad \text{Class 3} \rightarrow \begin{bmatrix} 0 \\ \vdots \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \text{Class k} \rightarrow \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

- The **length of the vector** is the **number of classes** k we have.
 - The i^{th} class has i **ones**.
-
- This transformation is only appropriate if the data
 - Is **ordered**,
 - But not **real number-compatible**: we can't add the values, or compare the "amount" of each feature.

Example: We reuse our example from earlier:

$$\phi(x_{\text{Class 1}}) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad \phi(x_{\text{Class 2}}) = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}, \quad \phi(x_{\text{Class 3}}) = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \phi(x_{\text{Class 4}}) = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

5.3.1.3 One-hot Code

We introduced this technique in the **previous** chapter:

When there's no clear way to **simplify** our data, we accept the current discrete classes, and **convert** them to a number-like form that implies no order.

- Examples:
 - Colors: {Red, Orange, Yellow, Green, Blue, Purple}

- Animals: {Dog, Cat, Bird, Spider, Fish, Scorpion}
- Companies: {Walmart, Costco, McDonald's, Twitter}

We can't use thermometer code, because that suggests a natural **order**. And we definitely can't use real numbers.

Example: {Brown, Pink, Green} doesn't necessarily have an obvious order: you could force one, but there's no reason to.

But, we can use one idea from thermometer code: each class in a different variable.

$$\begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_k \end{bmatrix} \quad (5.17)$$

But in this case, we don't "build up" our vector: we replace $\text{Class}(x) \geq 4$ with $\text{Class}(x) = 4$.

$$\phi(x) = \begin{bmatrix} \text{Class}(x) = 4 \\ \text{Class}(x) = 3 \\ \text{Class}(x) = 2 \\ \text{Class}(x) = 1 \end{bmatrix} \quad (5.18)$$

This approach is called **one-hot encoding**.

Definition 228

One-hot encoding is a way to represent **discrete** information about a data point.

Our k classes are stored in a length- k column **vector**. For **each** variable in the vector,

- The value is **0** if our data point is **not in that class**.
- The value is **1** if our data point is **in that class**.

$$\text{Class 1} \rightarrow \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad \text{Class 2} \rightarrow \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad \text{Class 3} \rightarrow \begin{bmatrix} 0 \\ \vdots \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \text{Class } k \rightarrow \begin{bmatrix} 1 \\ \vdots \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

In one-hot encoding, items are **never** labelled as being in **two** classes at the **same time**.

- This transformation is only appropriate if the data is
 - Does not have another **structure** we can reduce it to: it's neither like a **real number** nor **ordered**
 - We don't have an **alternative** representation that contains more (accurate) information.

Example: Suppose that we want to classify **furniture** as table, bed, couch, or chair.

$$\begin{bmatrix} \text{table} \\ \text{bed} \\ \text{couch} \\ \text{chair} \end{bmatrix} \quad (5.19)$$

For each class:

$$y_{\text{chair}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad y_{\text{table}} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad y_{\text{couch}} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad y_{\text{bed}} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad (5.20)$$

5.3.1.4 One-hot versus Thermometer

One common question is, "why can't we use one-hot for ordered data? We could sort the indices so they're in order".

However, there's a problem with this logic: the computer **doesn't care** about the order of the variables in an array: it contains no information!

Why is that? If the vector has an order, shouldn't that **affect** the model?

Well, remember that our model is represented by

$$\theta^T x = \sum_i \theta_i x_i \quad (5.21)$$

The vector format $\theta^T x$ is just a way to **condense** our equation: addition ignores ordering of elements!

Concept 229

Order of elements in a vector **don't** affect the behavior of our model.

This is because a linear model is a **sum**, and sums are the same regardless of **order**.

If our model has the same math regardless of order, then it doesn't encode that ordering.

Example: We'll take a vector, and rearrange it.

Despite shuffling, these two equations are equivalent!

$$\theta^T \phi(x) = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 & \theta_4 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \longrightarrow (\theta^T)^* (\phi(x))^* = \begin{bmatrix} \theta_3 & \theta_1 & \theta_4 & \theta_2 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

The math is the same, despite changing order: our model knows nothing about ordering.

Clarification 230

One hot encoding **cannot** encode information about ordering.

Thermometer encoding is required to **represent ordered objects**.

Why is thermometer encoding able to represent ordering? Let's try shuffling it, too.

$$\theta^T \phi(x) = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 & \theta_4 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad (5.22)$$

$$(\theta^T)^* (\phi(x))^* = \begin{bmatrix} \theta_3 & \theta_1 & \theta_4 & \theta_2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} \quad (5.23)$$

Even though we've changed the order, we still know this is the **third** in the order, because we have **three 1's**!

Concept 231

Even though the **order of elements** in a vector **doesn't matter**, we can retrieve the order of **thermometer coding** based on the **number of 1's in the vector**.



5.3.1.5 Factored Code

Now, we move away from number-like properties. Instead, what other **patterns** of our feature could be useful?

Sometimes, a single feature will contain **multiple** pieces of information. Separating those pieces (or **factors**) from each other can make it easier for our machine to understand.

- A **car** is often described by the "make" (brand) and "model" (which exact type of car by that brand).
 - These could be broken into two **features**: "make" is one feature, "model" is another.
 - **Example:** "Nissan Altima" becomes "Make: Nissan" and "Model: Altima".
- Most **blood types** are in the following categories: {A+, A-, B+, B-, AB+, AB-, O+, O-}.
 - You could factor this based on the letter, and positive/negative: {A, B, AB, O} and {+,-}.
 - Since "O" means we contain neither A nor B, we could factor the first feature further: {A, not A}, {B, not B}
 - Example: Using the first factoring, A- becomes [A, -]. Using the second it becomes [A, not B, -].
- **Addresses** have many parts: street number, zip code, state, etc.
 - Each of these can be given their own factor.

Definition 232

Factored code is a **feature transformation** where we take one **discrete class** and break it up into other discrete classifications, called **factors**.

Class m and n \rightarrow Class m, Class n

- This transformation is only appropriate if
 - We have some feature(s) that can be **broken up** into **simpler**, meaningful parts.

Often, we apply **other** feature transformations after factored coding.

Note the final comment: often, we turn a discrete class into multiple new discrete classes.

But, we still need to convert these into a usable, numeric-vector form!

Example: We can re-use our blood type example from above.

$$\phi(x) = \begin{bmatrix} x \text{ contains A} \\ x \text{ contains B} \\ x \text{ is +} \end{bmatrix} \quad (5.24)$$

Each of these are binary features. For example:

$$\phi(AB-) = \begin{bmatrix} \text{True} \\ \text{True} \\ \text{False} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \quad (5.25)$$

5.3.1.6 Binary Code

One possible way to encode data is to **compress** data using a **binary code**.

This might be tempting, because k values can be represented by $\log_2(k)$ values.

Example: Suppose you have the one-hot code for 6, and want to represent it with binary:

$$\text{Class 6} \xrightarrow{\text{One-hot}} [0 \ 1 \ 0 \ 0 \ 0 \ 0]^T \xrightarrow{\text{Binary}} [1 \ 1 \ 0]^T \quad (5.26)$$

Please do not do this

Concept 233

Using **binary code** to compress your features is usually a **bad idea**.

This forces your model to spend resources learning how to **decode** the binary code, before it can do the task you want it to!

5.3.2 Text

Just now, we showed different ways to transform **discrete** features.

Another very common data type we work with is **language**: bodies of text, online articles, corpora, etc.

Later in this course, we will discuss more powerful ways to analyze text, such as **sequential models**, and **transformers**.

Obligatory chatgpt reference.

There's a very simple encoding that we'll focus on here: the **bag of words** approach.

This approach is meant to be as simple as possible: for each word, we ask ourselves, "if this word in the text?", and answer yes (1) or no (0) for every single word.

Definition 234

The **bag of words** feature transformation takes a body of text, and creates a **feature** for every **word**: is that word in the text, or not?

$$\phi(x) = \begin{bmatrix} \text{Word 1 in } x \\ \text{Word 2 in } x \\ \vdots \\ \text{Word k in } x \end{bmatrix} \quad (5.27)$$

This approach is used for **bodies of text**.

Example: Consider the following sentence: "She read a book."

With the words: {She, he, a, read, tired, water, book}

We get:

$$\phi(\text{"She read a book."}) = [1 \ 0 \ 1 \ 1 \ 0 \ 0 \ 1] \quad (5.28)$$

A couple weaknesses to this approach:

- Ignores the order of words and syntax of the sentence.
- Doesn't encode meaning directly.

- Duplicate words are only included once.
- It doesn't create much structure for our model to use.

But, it's very easy to implement.

5.3.3 Numeric values

Now, on to the (typically) more manageable data type:

Concept 235

Typically, if your feature is **already a numeric value**, then we usually want to **keep it as a data value**.

Example: Heart rate, stock price, distance, reaction time, etc.

However, this may not be true if there is some difference between different ranges of numbers:

- Being below or above the age of 18 (or 21) for legal reasons
- Temperature above or below boiling
- Different age ranges of children might need different range sizes: the difference between ages 1-2 is very different from ages 7-8.

Concept 236

Sometimes, if there are distinct **breakpoints**/boundaries between different values of a numerical feature, we might use **discrete** features to represent those.

5.3.3.1 Standardizing Values

We still aren't done, if our data is numeric. We likely want to **scale** our features, so that they all tend to be in similar ranges.

Why is that? If some features are much **larger** than others, then they will have a much larger impact on the answer.

For example, suppose we have $x_1 = 4000$, $x_2 = 7$:

$$h(x) = \theta^T x = 4000\theta_1 + 7\theta_2 \quad (5.29)$$

The first term is going to have a way bigger impact on $h(x)$. If we change x_1 by 10%, that's going to be bigger than if we changed x_2 by 100%!

$4000 * 10\% = 400$
$7 * 100\% = 7$

Concept 237

If one **feature** is much **larger** than **another** feature, it will tend to have a much **larger** effect on the result.

This is often a bad thing: just because one feature is **larger**, doesn't mean it's more **important**!

Example: Income might be in the range of tens of thousands (10,000-100,000), while age is a two-digit number(20-100). Income will be weighed more heavily.

How do we solve that problem? We need to do two things:

- **Shift** the data so that our range is not too high/low. Our goal is to have it centered on 0.
 - We want it centered on 0 so we can distinguish between the above-average and below-average data points.
 - We do this by **subtracting the mean**, or the **average** of all of our data points.

You could try to solve this by scaling down θ .

But, we're already using regularization to bias against large θ : that will affect small variables (big θ_i) more than larger ones (small θ_i).

$$\phi_1(x) = x - \bar{x} \quad (5.30)$$

- **Scale** the **range** of possible values, so they all vary by roughly the same amount.

- So, if one variable tends to **vary** by a **larger** amount, it doesn't have a bigger impact on the result.

$$\phi(x_i) = \frac{x_i - \bar{x}_i}{\sigma_i} \quad (5.31)$$

Where σ is the **standard deviation**, a measure of how much our data varies.

If you are interested, we define **standard deviation** below.

Note that each feature has its own σ_i : we have to compute this equation for each feature.

Definition 238

To make sure that all of our data is **on the same size scale**, we **normalize/standardize** our dataset using the operation

$$\phi(x_i) = \frac{x_i - \bar{x}_i}{\sigma_i}$$

For every variable x_i in a data point x .

- \bar{x}_i is the **mean** of x_i
- σ_i is the **standard deviation** of x_i

This results in a dataset which has

- A mean \bar{x}_i of **0**
- A standard deviation σ_i of **1**

So, all of our features have the same **average**, and **vary** by the same amount.

This prevents some features getting prioritized because they're on different size scales.

Example: Suppose we have 1-D data $x = [1, 2, 3, 4, 5, 6]$

The mean is

$$\bar{x} = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5 \quad (5.32)$$

And the standard deviation is

$$\sigma = \sqrt{\frac{2.5^2 + 1.5^2 + .5^2 + .5^2 + 1.5^2 + 2.5^2}{6}} = \sqrt{\frac{35}{12}} \approx 1.7078 \quad (5.33)$$

5.3.3.2 Variance and Standard Deviation (Optional)

This section describes the origin of σ above. Feel free to skip if you're familiar.

In order to scale our data, we need a measure of how much our data **varies**. So, if our data varies by more, we can scale it down, and vice versa.

We can measure this using the **variance**.

Definition 239

We can measure how spread out/varying our data with **variance**

$$\sigma^2 = \sum_i \frac{(x^{(i)} - \bar{x})^2}{n} \quad (5.34)$$

In other words, the **average squared distance** from the **mean**.

Why do we square the terms? Same reason we square our loss:

- We want only positive values, for distance.
- We don't want to use absolute value, for smoothness.

However, this is too large: we want something similar to "average distance from the mean".

This is the average **squared** distance.

We also get nicer statistical properties we won't discuss here.

So, we take a square root!

Definition 240

A more common way to measure how our data varies is using **standard deviation** σ

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_i \frac{(x - \bar{x})^2}{n}}$$

This term is **not** the average distance from the mean, but can be used for **scaling** our data in the same way.

This term allows us to scale our data appropriately. If our data varies by a larger amount, σ will be larger. So, $\frac{1}{\sigma}$ will cancel that variance out.

5.4 Terms

- Non-linear
- Transformation
- Feature
- Feature Transformation
- Polynomial Basis
- Order/Degree of a polynomial
- Radial Basis
- Discrete Feature
- Numeric transformation
- Thermometer Code
- One-hot Code
- Factored Code
- Binary Code
- Bag of Words
- Standardization
- Normalization
- Standard Deviation

CHAPTER 6

Neural Networks 1 - Neurons, Layers, and Networks

The tools we've developed so far are interesting, and **varied**. We've discussed:

- **Regression**: the problem of creating **real-number** outputs based on data.
- **Classification**: the problem of **sorting** data points into **categories**.
- Gradient **descent**: A technique for gradually **improving** your model using **calculus**.

These concepts are fascinating in their own right, and can be used to handle some **simple** problems. But, when they are **combined** together, we get something much more **powerful**: **neural networks**.

6.0.1 Machine Learning Applications

Neural networks in the modern area are used to tackle complex and challenging problems:

- Image labelling and generation
 - **Example**: Recognizing a picture of a dog. Or, creating a picture of a dog when prompted.
- Physics simulation
 - **Example**: Simulating water flow realistically, or special-effects smoke for a movie.
- Financial prediction

- **Example:** Predicting how the **market** moves over time, and what the best **financial** choices in the present are.
- Text processing and generation
 - **Example:** Creating machines that can understand human text **prompts**, and writing useful **explanations** for humans.
- Data analysis
 - **Example:** **Compressing** data, or processing it to discover the **important** information.

As you can see, **neural networks** are used in a wide array of very **difficult** problems. No wonder it's become so popular!

6.0.2 Neural Network Perspectives: The brain

So, what *is* a neural network? There are several perspectives we can take.

First, the **name** comes from the fact that NNs are inspired by the **brain**:

- We call the basic unit of a neural network, a **neuron**.
- This gives us some general idea of the **structure** of a neural network:
 - Just like in the **brain**, we take many individual units, called **neurons**, which we connect together to do more **complicated** tasks. That combined structure is a **neural network**.

Concept 241

Neural networks are inspired by the brain and its **neurons**, in an effort to do better, **human-like** computation.

Based on this, neural networks are **built** out of simple **units** called **neurons**, connected to each other.

Funny enough, as effective as neural networks are, we now think they don't work very much like the human brain! But we keep the terminology.

6.0.3 Neural Network Perspectives: Classification and Regression

In this class, we **won't** focus on the brain analogy, though it did inspire the model.

Instead, we will mostly think of **neural networks** in terms of what they're able to do, and how they work.

- Our biggest problem so far is the "nonlinear task": tasks that can't be solved by our **linear** regression/classification models.
- In short: some problems require solutions outside of our **hypothesis space**.

Before, we used **feature representation** to solve this problem. Through the polynomial basis, we found a **richer** hypothesis class.

In this chapter, we present a different (but related) way of creating a richer hypothesis class. In this case, rather than transforming the input, we use a different **structure** for the model.

- By combining lots of simple **units** ("neurons"), we can get a very **complex** model for solving our problems.

With such a **rich** hypothesis class, combined with the power of **gradient descent**, we can create a model that can do **classification** or **regression** for much more difficult problems.

Concept 242

Neural Networks make up a very **rich** hypothesis class, by combining many simple **units**.

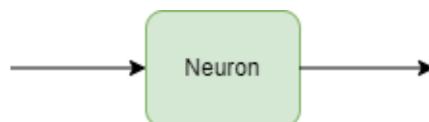
With this **hypothesis class**, we can handle **regression** or **classification** for very challenging **problems**.

Reminder: "richness" or "expressiveness" of a hypothesis reflect how wide our options are. Neural networks give us many possibilities for models. With more options, we can handle more problems!

6.0.4 Building up a basic neural network

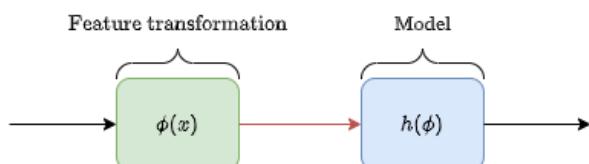
Let's make sense of what we said above, and **visualize** what a neural network might look like.

We start with one function: a **neuron**. This function could be, for example, one we've used before: our logistic **classifier**, or linear **regression**. We'll ignore the details for now.



One neuron might not be very powerful, or **expressive**. It's useful, but limited. We've seen its weaknesses.

We could try to use **feature transformations** to help us. But, let's think in a more **general** way: a transformation is just another **function** we apply to our input!



This gives us an **idea**: rather than trying to think of a single, more **complex** model, we could combine **multiple** simple models!

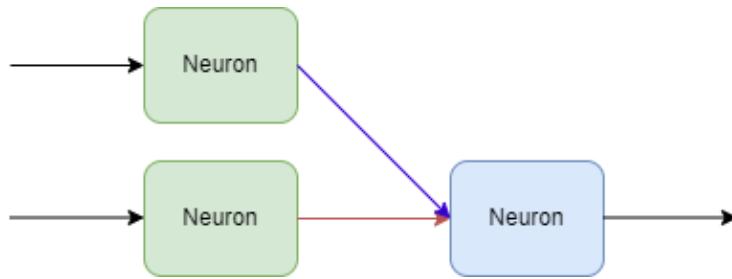
Last Updated: 09/03/24 03:53:41

Feature transformations, like polynomial or radial basis, are a bit more **complex** than what we'd usually put in a **neuron**. But, it gives us the right inspiration.



We could repeatedly add more neurons in **series**: each one being the input to another. And we'll do that later!

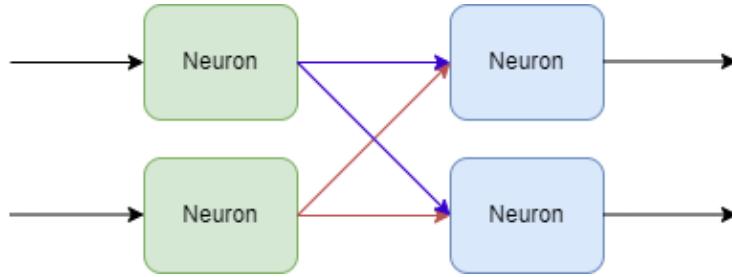
But, there's another type of **complexity** we haven't explored: we could have two neurons in **parallel**.



This parallel/series vocabulary is borrowed from circuits. We'll just use it for demonstration: you don't need to remember it.

Now, we have **two** neurons feeding into one output neuron! This already looks like a more **complicated** model.

We can go even further: what if we have two outputs as well?



Because we had two **inputs**, we had to add two new **links** when we added the output neuron. This is getting difficult to **view**!

We'll stop here for now, but you can imagine repeatedly **adding** more neurons in **parallel** (with the same inputs/outputs) or in **series** (as an input or output).

- And with each addition, the function gets more and more **complex**: you can create a **richer** hypothesis class!

We'll explore how to do this **systematically** later in the chapter.

By "systematic", we just mean "in a way that we can develop with simple instructions".

Definition 243

Neural Networks are a **class** of models that can be used to solve **classification**, **regression**, or other interesting problems.

They create very **rich** hypothesis classes by combining many **simple** models, called **neurons**, into a **complex** model.

- We do this combination **systematically**, so that it is easy to **analyze** and work with our model.

This creates a very **flexible** hypothesis, which can be **broken down** into its **simple** parts and what **connects** them.

6.0.5 Neural Network Perspectives: Predictions with Big Data

Our last major **perspective** on neural networks is one that you see in lots of modern **applications**. We won't work much with this perspective in this **class**, but our techniques **enable** it.

- Neural networks, because they can create such **sophisticated** models, can be used for problems in very **complex** domains: the kind of **applications** we discussed at the beginning of this chapter.
- These applications require a lot of **data** to build a good **model**, however. So, machine learning models often take **huge** amounts of data, with lots of energy and time to train them.
- But, once they are fully **trained**, they can give predictions very **quickly**, and often very **accurately**.

Concept 244

Neural networks can be seen as a way to make **predictions** based on huge amount of **data** for very **complex** problems.

6.1 Basic Element

Now, we have idea of what neural networks **are**. But, we have yet to handle the **details**:

- What **is** a neuron?
- How do we "systematically" **combine** our neurons?
- How do we **train** this, like we would a **simple** model?

We'll handle all of these steps and more - the above description was just to give a **high-level** view of what we want to **accomplish**.

Now, we go down to the **bottom** level, and think about just **one neuron**: what does it look like, and how does it work?

First, some terminology:

Notation 245

Neurons are also sometimes called **units** or **nodes**.

They are mostly **equivalent** names. They just reflect different **perspectives**.

6.1.1 What's in a neuron: The Linear Component

As we mentioned before, our goal is to combine **simple** units into a **bigger** one. So, we want a model that's **simple**.

Well, let's start with what we've done before: we've worked with the **linear** model

$$h(x) = \theta^T x + \theta_0 \quad (6.1)$$

This model has lots of nice properties:

- It limits itself to **addition** and **multiplication** (easy to compute)
- **Linearity** lets us prove some mathematical things, and use vector/**matrix** math
- The dot product between θ and x has a nice **geometric** interpretation.

This will make up the **first** part of our model.

Concept 246

Our **neuron** contains a **linear** function as its **first** component.

6.1.2 Weights and Biases

But, there's one minor **change**: before, we used θ because it represented our **hypothesis**.

But, every neuron is going to have its own **values** for its **linear** model:

$$\overbrace{f_1(x)}^{\text{Neuron 1}} = Ax + B \quad \overbrace{f_2(x)}^{\text{Neuron 2}} = Cx + D \quad (6.2)$$

It wouldn't make much **sense** to call both A and C by the name θ .

We could use some clever **notation**, but why treat them as **hypotheses**? They are each only a **part** of our hypothesis Θ .

So, instead of thinking of each as a "hypothesis", let's switch perspectives.

Each value θ_k **scales** how much x_k affects the **output**: if we're doing

$$g(x) = 100x_1 + 2x_2 \quad (6.3)$$

Then, changing x_1 will have a much **bigger** effect on $g(x)$. Another way to say this is it **weights** more heavily: it matters **more**.

Because of that, we call the number we scale x_1 by a **weight**.

Notation 247

A **weight** w_k tells you how heavily a **variable** x_k **weights** into the output.

w_k is **equivalent** to θ_k : it's a **scalar** $w_k \in \mathbb{R}$.

$$(\theta_1 x_1 + \theta_2 x_2) \iff (w_1 x_1 + w_2 x_2)$$

We can combine it into a vector $w \in \mathbb{R}^m$.

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix} \quad \theta^T x \iff w^T x$$

What about our other term, θ_0 ? We call it an **offset**: it's the value we **shift** our linear model away from **origin**.

Remember that $a \iff b$ means a and b are equivalent!

We'll use the same notation:

Notation 248

An **offset** w_0 tells you how far we **shift** $h(x)$ away from the origin.

w_0 is **equivalent** to θ_0 : it's a **scalar** $w_0 \in \mathbb{R}$

$$((\theta^T x) + \theta_0) \iff ((w^T x) + w_0)$$

We also sometimes call this the **threshold** or the **bias**.

Alternate notation: we might call this variable b , for bias.

This gives us our linear model using our new notation:

Definition 249

The **linear component** for a neuron is given by

$$z(x) = w^T x + w_0$$

where $w \in \mathbb{R}^m$ and $w_0 \in \mathbb{R}$.

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix}$$

6.1.3 Linear Diagram

Now, we want to be able to depict our **linear** subunit. Let's do it piece-by-piece.

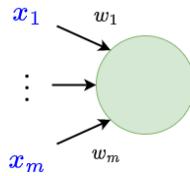
First, we have our vector $x = [x_1, x_2, \dots, x_m]^T$:

x_1

\vdots

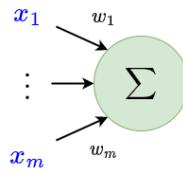
x_m

Now, we want to **multiply** each term x_i by its corresponding **weight** w_i . We'll combine them into a **function**:



The circle represents our function.

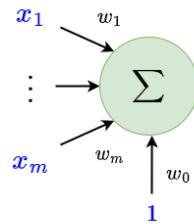
How are we combining them? Well, we're adding them together.



Note that we use the \sum symbol, because we're **adding** after we **multiply**. In fact, we can write this as

$$\mathbf{w}^T \mathbf{x} = \sum_{i=1}^m w_i x_i \quad (6.4)$$

We'll include the bias term as well: remember that we can represent w_0 as $1 * w_0$ to match with the other terms.

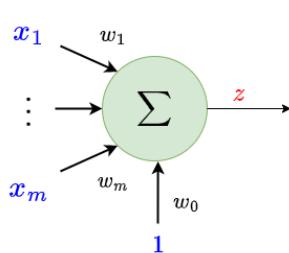


The blue "1" term is **multipled** by w_0 , just like how x_k gets multiplied by w_k .

We have our full function! All we need to do is include our output, z :

Notation 250

We can depict our linear function $z = \mathbf{w}^T \mathbf{x} + w_0$ as



Thus, z is a function of x :

$$z(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0 \quad (6.5)$$

Which, in \sum notation, we could write as

$$z(\mathbf{x}) = \left(\sum_{i=1}^m w_i x_i \right) + w_0 \quad (6.6)$$

6.1.4 Adding nonlinearity

We'll continue building our neuron based on what we've done **before**. When doing linear regression, that linear unit was all we had.

But, once we do classification, we found that it was helpful to have a second, **non-linear** component: we used **sigmoid** $\sigma(u)$.

- We might not necessarily want the **same** nonlinear function, so instead, we'll just generalize: we have *some* second component, which is allowed to be **nonlinear**.

We call this component our **activation** function. Why do we call it that? It comes from the historical **inspiration** of neurons in the brain.

- Biological neurons only "fire" (give an output) above a certain threshold of **input**: that's when they **activate**.

You might remember that we had a problem with the logistic linear model still behaving linear, despite having a nonlinear function.

We'll show how we fix this later on.

Some activation functions reflect this, but they don't have to.

Definition 251

Our **neuron** contains a potentially **nonlinear** function f , called an **activation function**, as its **second** component.

We note this as

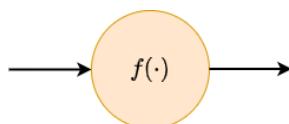
$$a = f(z)$$

Where z is the **output** of the **linear** component, and a is the **output** of the **activation** component.

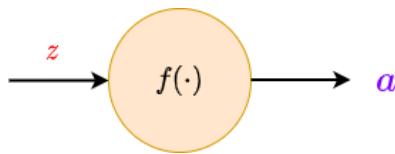
Note that z and a are **real numbers**: we have $f : \mathbb{R} \rightarrow \mathbb{R}$

6.1.5 Nonlinear Diagram

We'll depict a function f .



It takes in our **linear** output, z , and outputs our **neuron** output, a .



Note some vocabulary used for z :

Notation 252

z , the **output** of our **linear** function, is called the **pre-activation**.

This is because it is the result **before** we run the **activation** function.

And for a :

Notation 253

a , the **output** of our **activation** function, is called the **activation**.

6.1.6 Putting it together

So now, our neuron is complete.

Definition 254

Our **neuron** is made of

- A **linear** component that takes the neuron's input x , and applies a linear function

$$z = w^T x + w_0$$

- A (potentially nonlinear) **activation** component that takes the pre-activation z and applies an **activation function** f :

$$a = f(z)$$

When we **compose** them together, we get

$$a = f(z) = f(w^T x + w_0)$$

When we say "compose", we mean **function composition**: combining $f(x)$ and $g(x)$ into $f(g(x))$.

Definition 255

Our **neuron** has several important intermediate outputs:

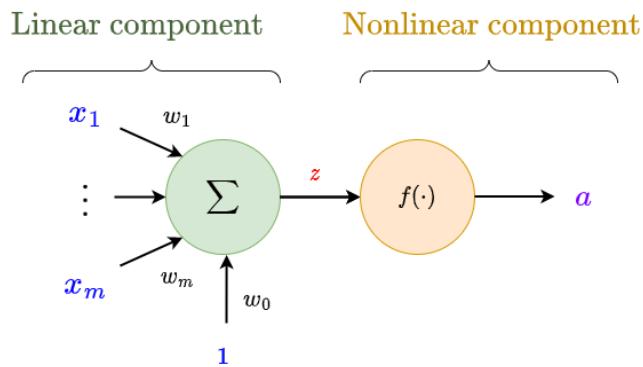
- The **pre-activation z** is the **output** of the **linear** function.
 - It is also the **input** of the **activation function f** .
- The **activation a** is the **output** of the **activation function**.
 - It is also the **output** of the **neuron**.

We can also use \sum notation to get:

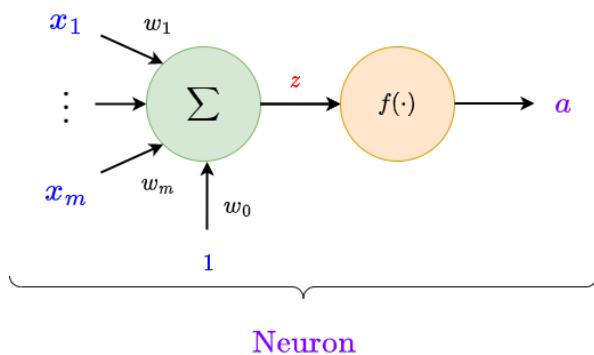
$$a = f(z) = f\left(\left(\sum_{i=1}^m w_i x_i\right) + w_0\right)$$

6.1.7 Neuron Diagram

Finally, we can **compose** our neuron into one big **diagram**:

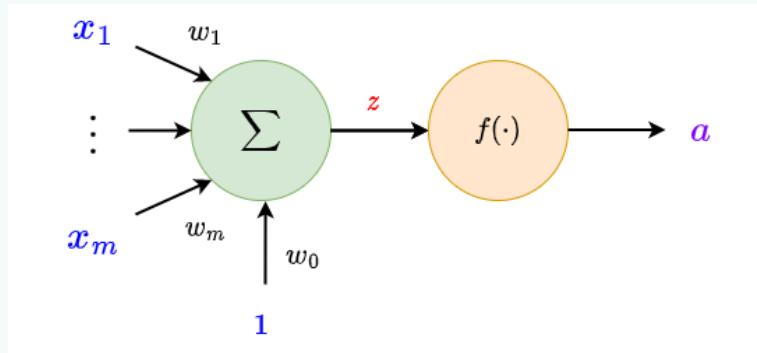


From here on out, we'll treat this as a **single** object:



Notation 256

We can depict our **neuron** $f(w^T x + w_0)$ as



- x is our **input** (neuron input, linear input)
- z is our **pre-activation** (linear output, activation input)
- a is our **activation** (neuron output, activation output)

This neuron will be the **basic unit** we work with for the rest of this **chapter** - it's one of the most **important** objects in all of machine learning.

6.1.8 Our Loss Function

One more detail: we will want to **train** these neurons. In order to be able to **measure** their performance, we'll need a **loss** function.

This **isn't** any different from usual: we just need a **function** of the form

$$\mathcal{L}(g, y) \tag{6.7}$$

In **regression**, we wrote our loss as

$$\mathcal{L} \left(h(x; \Theta), y \right)$$

The right term, $y^{(i)}$, is unchanged: we still need to compare against the **correct** answer.

The main change is we aren't using Θ notation: we'll **replace** it with (w, w_0)

$$\mathcal{L} \left(h(x; (w, w_0)), y \right)$$

And finally, we get the loss for multiple data points: _____

We skip doing $1/n$ averaging because we often use this for SGD: we plan to take small steps as we go, rather than adding up our steps all at once.

$$\sum_i \mathcal{L} \left(h(\mathbf{x}^{(i)}; (\mathbf{w}, w_0)), \mathbf{y}^{(i)} \right)$$

And with this, not only is our neuron **complete**, but we have everything we need to **work** with it.

Concept 257

For a **complete neuron**, we need to specify

- Our **weights** and **offset**
- Our **activation** function
- Our **loss** function

From here, we could do **stochastic gradient descent** as we usually do, to **optimize** this neuron's **performance**.

6.1.9 Example: Linear Regression

Let's go through some **examples**. We mentioned in the **beginning** of this chapter that our neuron could be most of the simple **models** we've worked with.

So, let's give that a go: the most simple version that's useful. We'll start by doing **linear regression**.

$$h(x) = \theta^T x + \theta_0$$

This model is exclusively **linear**: we just have to replace θ with w .

$$z(x) = w^T x + w_0$$

So, our linear component is **done**: $(\theta, \theta_0) = (w, w_0)$.

What about our **activation** function?

- Well, activation allows for **nonlinear** functions. But, we don't **want** to make it nonlinear.
- In fact, we've already got what we **want**: we don't want the **activation** to do anything **at all**.

So, we'll use **this** function:

Concept 258

The **identity function** $f(z)$ is a function that has no **effect** on your **input**.

$$f(z) = z$$

By "having no effect", we mean that the input is **unchanged**: this is true even if your input is **another function**:

$$f(g(x)) = g(x) \quad (6.8)$$

So, the **identity** function is our activation function: it keeps our **linearity**.

We call it the "identity" because the input's identity is unchanged!

Concept 259

Linear Regression can be represented with a **single neuron** where

- We keep our **linear component**, but set $(\theta, \theta_0) = (w, w_0)$.

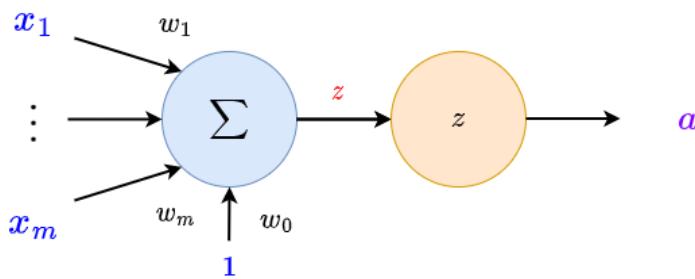
$$z(x) = w^T x + w_0$$

- Our **activation function** is the **identity** function,

$$f(z) = z$$

- Our **loss function** is **quadratic loss**.

$$\mathcal{L}(a, y) = (a - y)^2$$



6.1.10 Example: Linear Logistic Classifiers

Now, we do the same for LLCs: it's already broken up into **two** parts in our **classification** chapter.

First, the **linear** component. This is the same as linear regression:

$$\textcolor{red}{z} = \theta^T x + \theta_0 \quad (6.9)$$

And then, the **logistic** component:

$$\sigma(\textcolor{red}{z}) = \frac{1}{1 + e^{-\textcolor{red}{z}}} \quad (6.10)$$

This second part is **nonlinear**: it's our **activation** function!

Concept 260

A **Linear Logistic Classifier** can be represented with a **single neuron** where

- We keep our **linear component**, but set $(\theta, \theta_0) = (w, w_0)$.

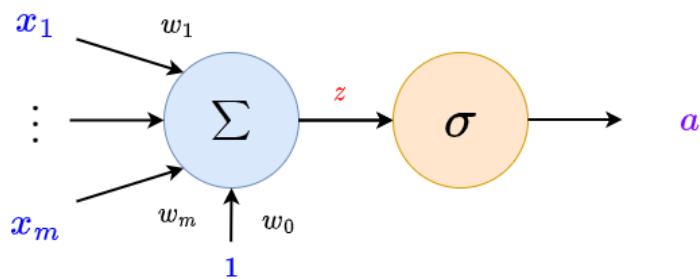
$$\textcolor{red}{z}(x) = w^T x + w_0$$

- Our **activation function** is the **sigmoid** function,

$$f(\textcolor{red}{z}) = \sigma(\textcolor{red}{z}) = \frac{1}{1 + e^{-\textcolor{red}{z}}}$$

- Our **loss function** is **negative-log likelihood** (NLL)

$$\mathcal{L}_{\text{nll}}(\textcolor{violet}{a}, \textcolor{blue}{y}^{(i)}) = - \left(\textcolor{blue}{y}^{(i)} \log \textcolor{violet}{a} + (1 - \textcolor{blue}{y}^{(i)}) \log (1 - \textcolor{violet}{a}) \right)$$



6.2 Networks

Now, we have fully developed the individual **neuron**.

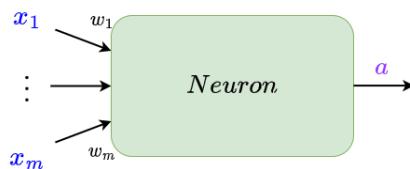
We can even do **gradient descent** on it: just like when we were doing LLCs, we can use the **chain rule**.

So, we return to the idea from the beginning of this chapter: combining multiple neurons into a **network**.

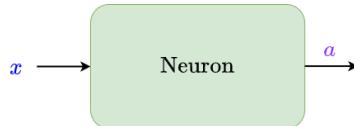
We'll get into this more, later in the chapter.

6.2.1 Abstraction

For this next section, we'll **simplify** the above diagram to this:



In fact, for more **simplicity**, we'll draw **one** arrow to represent the whole vector x . However, nothing about the **actual** math has changed.



This is also called **abstraction** - we need it a lot in this chapter.

Definition 261

Abstraction is a way to view your system more **broadly**: removing excess details, to make it **easier** to work with.

Abstraction takes a **complicated** system, and focuses on only the **important** details. Everything else is **excluded** from the model.

Often, this **simplified** view boils a system down to its the **inputs** and **outputs**: the "interface".

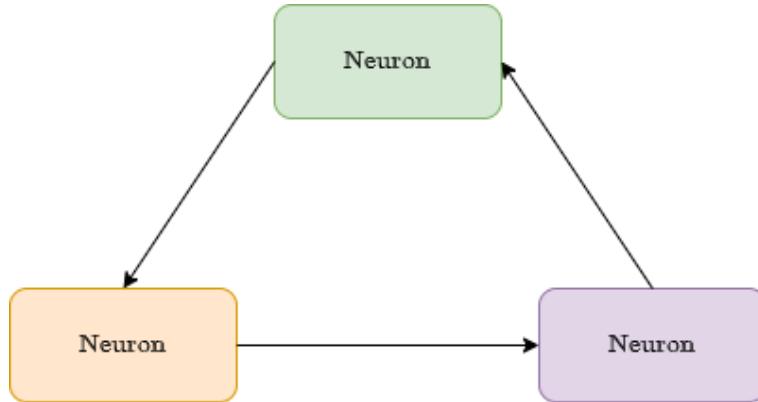
Example: Rather than thinking about all of the **mechanics** of how a car works, you might **abstract** it down to the pedals, the steering wheel, and how that causes the car to move.

6.2.2 Some limitations: acyclic networks

We won't allow for just **any** kind of network: we can create ones that might be unhelpful, or just very **difficult** to **analyze**.

For now, we can get interesting and **useful** behavior while keeping it **systematic**. We'll define this "system" later.

We'll assume our networks are **acyclic**: they do not create closed **loops**, where something can affect its own input.



This is a cyclic network: this is messy and we won't worry about this for now.

This means information only **flows** in one direction, "forward": it never flows "backwards".

Concept 262

For simple **neural networks**, we assume that they are **acyclic**: there are no **cycles**, or loops.

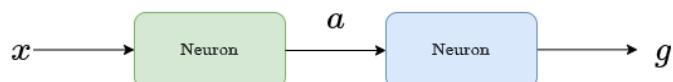
This means that **no neuron** has an output that affects its **input**, directly or indirectly.

We call these **feed-forward** networks: information can only go "forward", not "backward".

We'll show how to build up the rest of what we need.

6.2.3 How to build networks

Suppose we have two neurons in **series**, our **simplest** arrangement:



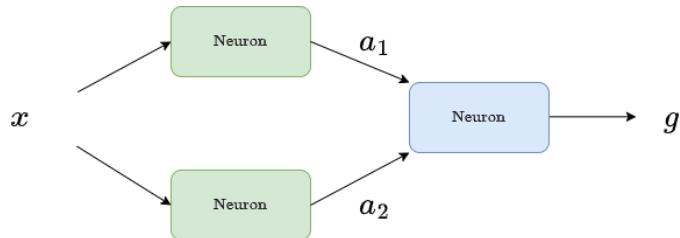
Our first neuron takes in a whole **vector** of values, $x = [x_1, x_2, \dots, x_m]^T$. But, it only **outputs** a single value, a .

- That means the second neuron only receives **one** value.

Remember that while we only see one arrow from x , each data point x_i is included.

But, just like our first neuron, it's capable of handling a full **vector**. We can add more values!

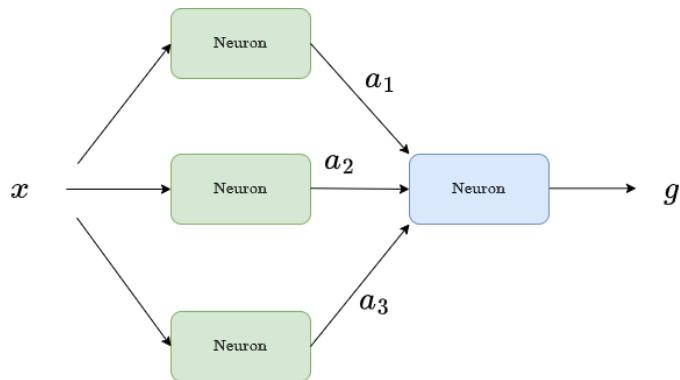
Let's add **another** neuron.



Our rightmost neuron now has **2 inputs**, which can be stored in a vector,

$$\mathbf{A} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \quad (6.11)$$

We could increase the **length** of this vector by adding more **neurons**.



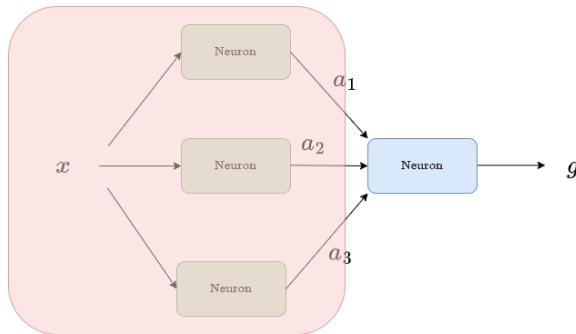
$$\mathbf{A} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \quad (6.12)$$

For our **rightmost** neuron, this is effectively the **same** as x : an **input vector**.

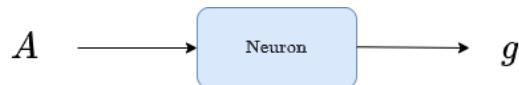
6.2.4 Layers

This gives us an idea for how to **build** our network: using multiple neurons in **parallel**, we can output a new vector \mathbf{A} !

This is useful, because it means we can **simplify**: from the rightmost neuron's perspective, it just sees that **vector** as an input.



We can take this entire layer...



And just reduce it down to the vector A .

Because it's so useful, we'll give this set of neurons a name: a **layer**.

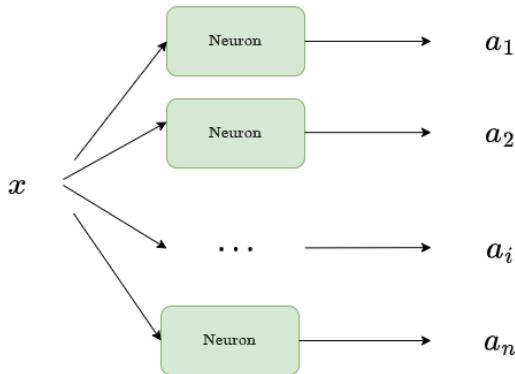
Definition 263

A **layer** is a set of **neurons** that are in "parallel":

- They all have **inputs** from the same **previous layer**
 - This **previous layer** could also be the **original input** x .
- They all have **outputs** to the same **next layer**
 - This **next layer** could also be the **final output** of the neural network.
- And none of the neurons in the same layer are directly **connected** to each other.

This **layering** structure allows us to simplify our **analysis**: anything that comes after the layer only has to work with a **single vector**.

A layer in general might look like this:



A general layer in a neural network.

6.2.5 The Basic Structure of a Neural Network

We could pick many structures for neural networks, but for simplicity, this will define our **template** for this chapter.

Definition 264

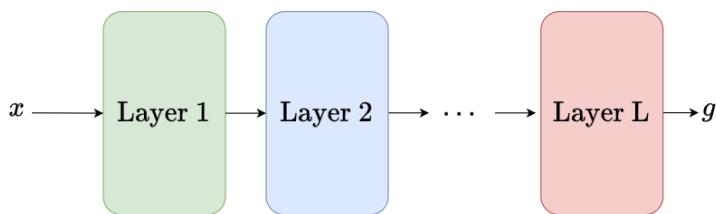
We structure our **neural networks** as a series of **layers**, where each layer is the **input** to the next layer.

This means that **layers** are a basic unit of a neural network, one level above a **neuron**.

In short, we have:

- A **neuron**, made of a linear and an activation component
- A **layer**, made of many **neurons** in parallel
- A **neural** network, made of many **layers** in series

Our goal is some kind of structure that looks something like this:

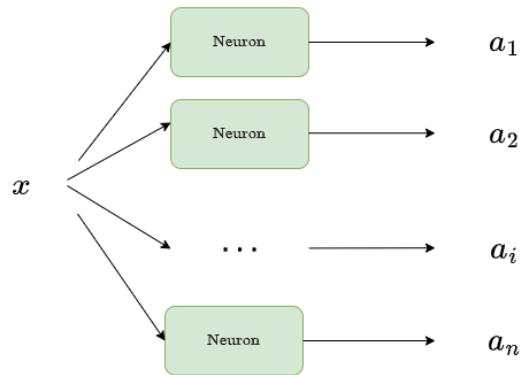


A neural network.

We now have a high-level view of our entire neural network, so now we dig into the details of a single layer.

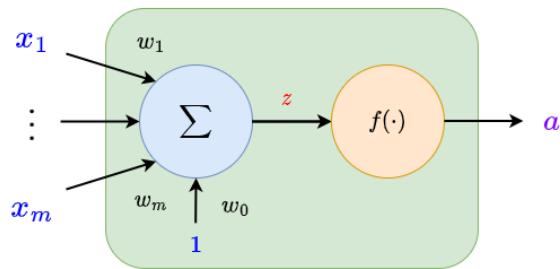
6.2.6 Single Layer: Visualizing our Components

Now, rather than analyzing a single neuron, we will analyze a single layer.



Our first layer.

In order to **analyze** this layer, we have to open back up the **abstraction**:

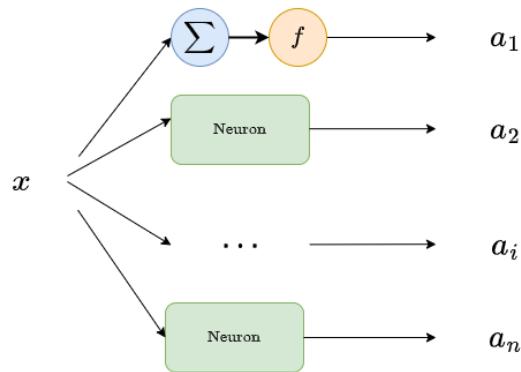


Each of those neurons looks like this.

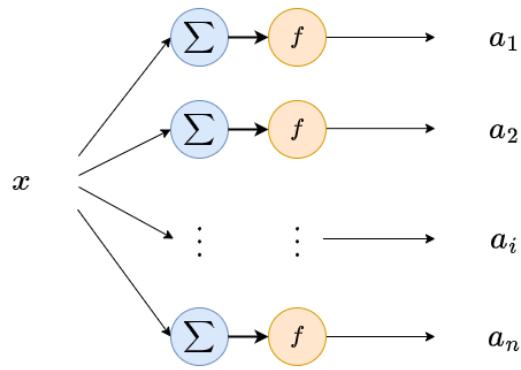
There are two important pieces of **information** we're hiding:

- We have two components inside of our neuron.
- We have many inputs x_i for one neuron.

The first piece of information is easier to visualize: we just replace each neuron with the two components.



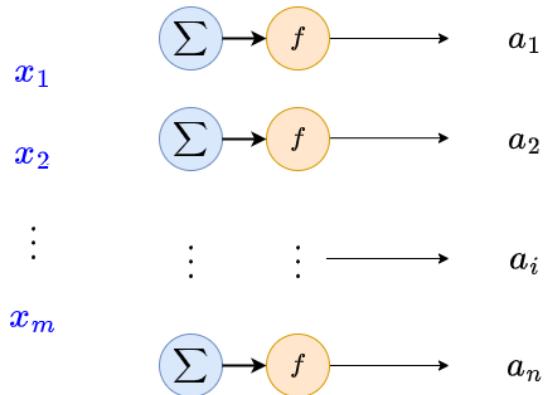
Replacing one neuron...



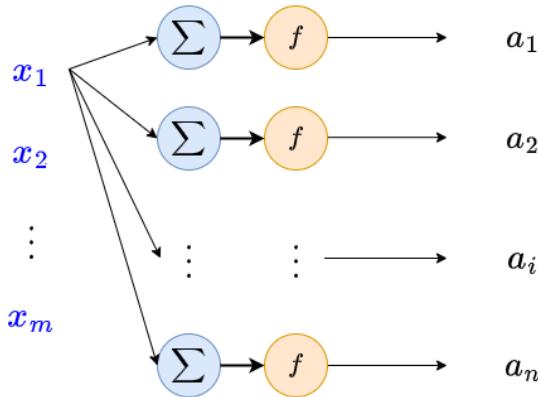
Replacing all neurons!

6.2.7 Single Layer: Visualizing our Inputs

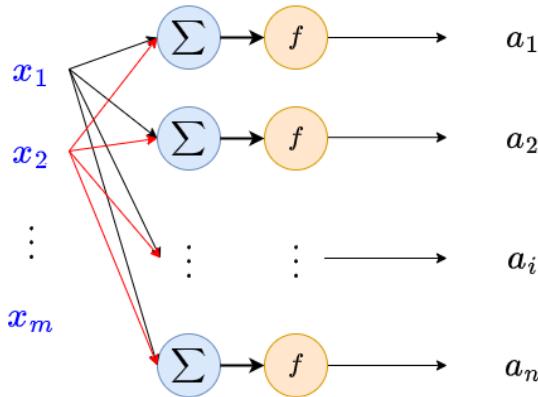
The second piece of information is much more difficult: we show all of the x_i outputs.



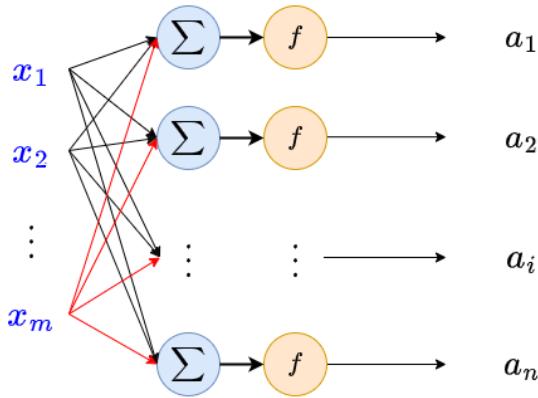
Now we have to draw the arrow for each input.



Every neuron receives the first input.



Every neuron receives the second input, too. This is getting messy...



The completed version: this is hard to look at.

Don't worry if this looks **confusing!** It's natural for it to be **hard** to read: the only thing you need to know is that we pair **every** input with **every** neuron.

This is our **final** view of this layer: because each of our m inputs has to go to every of n neurons, we end up with mn different **weights**.

This is a ton of **information**, and its only one layer! This shows how **complex** a neural network can be, just by **combining** simple neurons.

Note that this is a **fully connected** network: not all networks are FC.

Definition 265

A layer is **fully connected** if every neuron has the **same input vector**.

In other words, every neuron in our layer is connected to every input value.

Example: If one of our neurons **ignored** x_1 , but the others did **not**, the layer would not be **fully connected**.

6.2.8 Dimensions of a layer

Now that we've seen the **full** view, we can **analyze** it. Our goal is to create a more **useful** and **accurate** simplification.

Our first point: note that the input and output have a **different** dimensions!

Clarification 266

A **layer** can have a different **input** and **output** dimension. In fact, they are completely **separate** variables.

This is because **every** input variable is allowed to be applied to the **same** neuron:

Example: You can have one neuron of the form

$$z = \begin{bmatrix} w_1 & w_2 & w_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + w_0$$

In this case, our neuron has **one** output variable $f(z)$, but **three** inputs x_1, x_2, x_3 . Input dimension 3, output dimension 1.

Thus, our output dimension has been separated from our input dimension. Instead, it is the number of neurons.

So, in general, we can say:

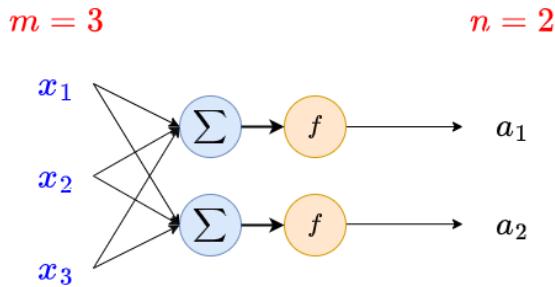
Notation 267

A **layer** has two associated **dimensions**: the **input** dimension m and the **output** dimension n .

- The **input** dimension m is based on the vector output from the **previous layer**:
 $x \in \mathbb{R}^m$
- The **output** dimension n is equal to the **number of neurons** in the **current** layer:
 $A \in \mathbb{R}^n$

These dimensions can be any pair of numbers: the value of m doesn't affect the value of n .

Example: Suppose you have an **input** vector $x = [x_1, x_2, x_3]$ and two **neurons**. The dimensions are $m = 3$, and $n = 2$.



The input dimension and output dimensions are **separate**.

6.2.9 The known objects of our layer

So, we know we have two objects so far:

- Our **input** vector $x \in \mathbb{R}^m$
- Our **output** vector $A \in \mathbb{R}^n$

Where they each take the form

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \quad A = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \quad (6.13)$$

But, there are a couple other things we haven't **generalized** for our entire **layer**:

- Our weights

- Our offsets
- Our preactivation

6.2.10 The other variables of our layer: weights and offsets

First, our **weights**: each neuron has its own vector of weights $w \in \mathbb{R}^m$.

The dimension needs to match x so we can compute $w^T x$.

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix} \quad (6.14)$$

To distinguish them from each other, we'll represent the i^{th} neuron's weights as \vec{w}_i .

$$\vec{w}_i = \begin{bmatrix} w_{1i} \\ w_{2i} \\ \vdots \\ w_{mi} \end{bmatrix} \quad (6.15)$$

Each weight needs to be used to **compute** a_i , but having so many objects is annoying.

Remember that, when we had **multiple** data points $x^{(i)}$, we worked with them at the **same time** by stacking them in a **matrix**. Let's do the same here:

$$W = \underbrace{\begin{bmatrix} \vec{w}_1 & \vec{w}_2 & \cdots & \vec{w}_n \end{bmatrix}}_{\text{Each neuron has a weight vector}} \quad (6.16)$$

If we expand it out, we get a full matrix...

$$W = \underbrace{\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m1} & \cdots & w_{mn} \end{bmatrix}}_{\text{n neurons}} \Bigg\} \text{m inputs} \quad (6.17)$$

This is our **weight matrix** W : it's an $(m \times n)$ matrix. It contains all of our mn weights, sorted by

- **Input variable** (row)
- **Neuron** (column)

We can do this for our **offsets** too: thankfully, there is only **one** offset per neuron, so we can write:

This is our offset vector, with the shape $(n \times 1)$.

Notation 268

We can store our **weights** and **offsets** as **matrices**:

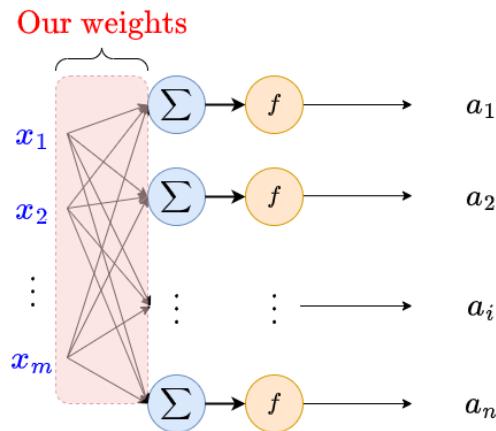
- **Weight** matrix W has the shape $(m \times n)$

$$W = \underbrace{\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{bmatrix}}_{m \text{ inputs}} \overbrace{\quad}^{n \text{ neurons}}$$

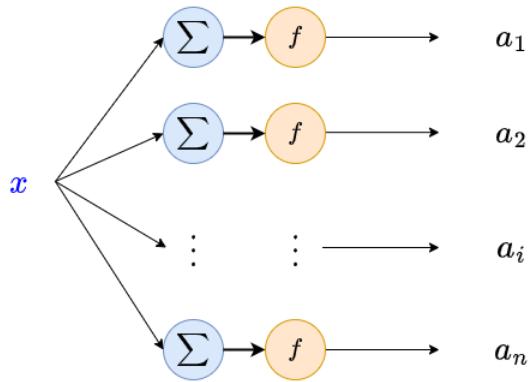
- **Offset** matrix W_0 has the shape $(n \times 1)$

$$W_0 = \underbrace{\begin{bmatrix} w_{01} \\ w_{02} \\ \vdots \\ w_{0n} \end{bmatrix}}_{\text{Each neuron has an offset}}$$

These matrices give us a tidy way to understand all of this mess:



Now that we understand it, we'll **hide** those weights again, for readability.



6.2.11 Pre-activation

Now, all that remains is the pre-activation z .

Before, we did

$$w^T x + w_0 = z \quad (6.18)$$

Because we so carefully kept our weights and offsets separate, we can still do this!

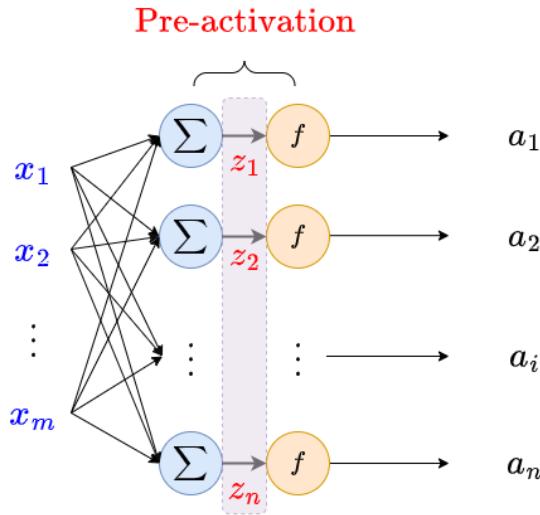
$$W^T x + W_0 = Z \quad (6.19)$$

You can check for yourself that this behaves the way you expect it to.

This pre-activation vector Z contains all of the outputs of our linear components:

$$Z = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} \quad (6.20)$$

On our diagram, we can see it here:



This section is what Z details with.

And we can connect this to our activation: each a_i is the result of running our function f on z_i :

$$A = f(Z) = \begin{bmatrix} f(z_1) \\ f(z_2) \\ \vdots \\ f(z_n) \end{bmatrix} \quad (6.21)$$

Because we run the function on each element in Z, we call this an **element-wise** use of our function.

6.2.12 Summary of a layer

So, we can now break our our layer up into pieces:

Notation 269

Our **layer** is a **function** that takes in $x \in \mathbb{R}^m$, and returns $A \in \mathbb{R}^n$.

It is defined by:

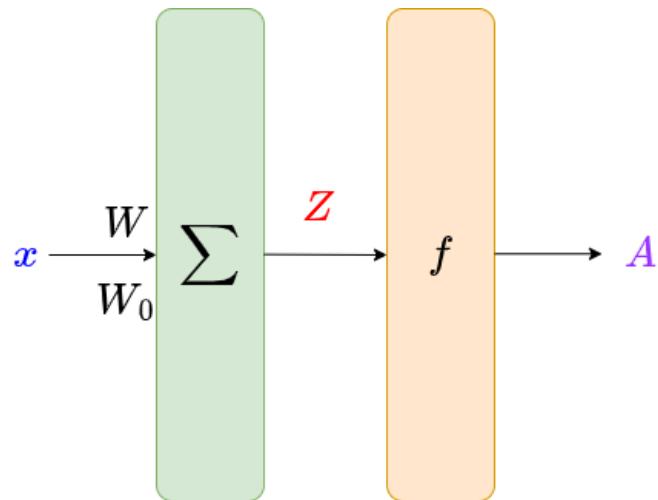
- **Dimensions:** m for **input**, n for **output** (number of neurons)

And our different **matrices**:

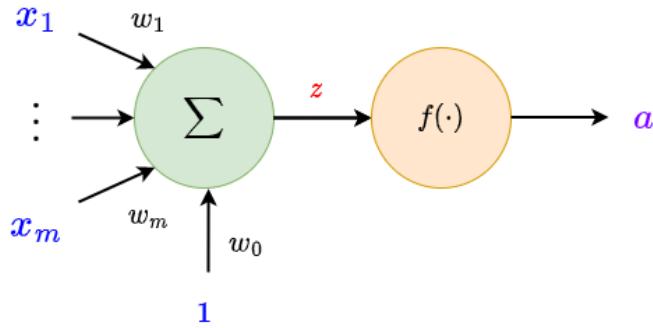
- **Input:** a **column vector** X in the shape $(m \times 1)$
- **Weights:** a **matrix** W in the shape $(m \times n)$
- **Offset:** a **column vector** W_0 in the shape $(n \times 1)$
- **Pre-activation:** a **column vector** Z in the shape $(n \times 1)$
- **Activation:** a **column vector** A in the shape $(n \times 1)$

We've now accomplished our goal: **simplify** the layer into its **base** components, without losing any crucial **information**.

We've can represent an entire layer like this:



Note how similar this looks to a **single** neuron: this works because the neurons in a **layer** are in **parallel**!



The math is very similar as well:

Definition 270

Our **layer** can be represented by

- A **linear** component that takes in x , and outputs **pre-activation** Z :

$$Z = W^T x + W_0$$

- A (potentially nonlinear) **activation** component that takes in Z , and outputs **activation** A :

$$A = f(Z)$$

When we **compose** them together, we get

$$A = f(Z) = f(W^T x + W_0)$$

6.2.13 The weakness of a single layer

What can we do with a single layer? Well, our LLC model gives us an example: it has the **nonlinear** sigmoid activation, but acts as a **linear** separator.

Why is that? Why is the separator still linear, if the **activation** isn't?

Well, let's take the **linear** separator created by the pre-activation:

$$z = w^T x + w_0 = 0 \quad (6.22)$$

This is our **boundary** for just a linear function. But adding the nonlinear activation should make it more **complex**, right?

Well, it turns out, we can represent our **activation** boundary with a **linear** boundary.

Example: Continue our LLC example. If $z = 0$, then $\sigma(z) = \sigma(0)$. Our boundary is

$$\sigma(z) = \sigma(0) = \frac{1}{2} \quad (6.23)$$

Wait. But that means that $\sigma(z) = .5$ is the same as $z = 0$: the same inputs x cause both of them, so they have the same boundary!

$$\text{Linear boundary } z = 0 \iff f(z) = \frac{1}{2} \quad (6.24)$$

Summary:

- $\sigma(z) = .5$ is the **same** as $z = 0$.
- $z = 0$ is **linear**.
- Thus, our sigmoid boundary is **linear**.

We can apply this to other activation functions. In general, any constant boundary for most $f(z)$ is equivalent to some linear boundary $z = C$:

Assuming that f is invertible, which it often is.

$$z = C \iff f(z) = f(C) \quad (6.25)$$

Since $z = C$ is linear, we know that our activation separator $f(x) = f(C)$ is linear too.

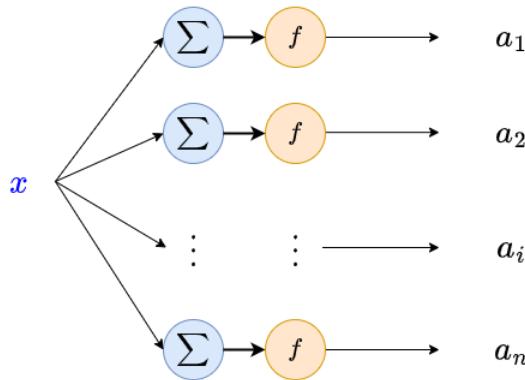
Concept 271

A single neuron creates a **linear separator**, even if it has a **nonlinear** activation.

This is because any **boundary** for $f(z)$ we can create, can be represented by some **linear** boundary in z .

There are exceptions, but this is true for most useful activation functions.

It turns out, adding more neurons **within** the layer doesn't change much: because they act in **parallel**, each neuron acts separately, and the things we said above are still **true** for each output a_i .



Each of these neurons has the same input, x .

So, in order to create nonlinear behavior, we need at least two layers of neurons in **series**.

So, we'll start **stacking** layers on each other: each layer **feeds** into the next one.

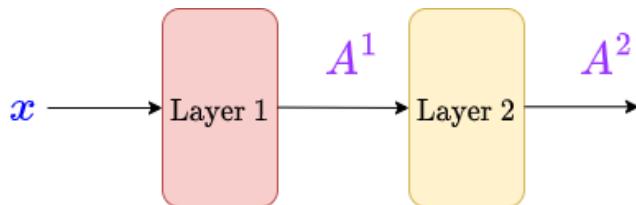
Concept 272

A **single layer** of neurons has **linear** behavior.

We need **multiple** layers to get a nonlinear **neural network**.

6.2.14 Adding a second layer

So, let's add one more **layer**. We'll label layers by using a **superscript**: W^1 is the set of **weights** for the **first** layer, for example.



We have two separate outputs: A^1 and A^2 .

Clarification 273

Superscripts in our notation indicate the **layer** that our value is associated with.

They do **not** represent exponentiation!

Example: Z^3 would be the **pre-activation** for layer 3: it is **not** Z "cubed".

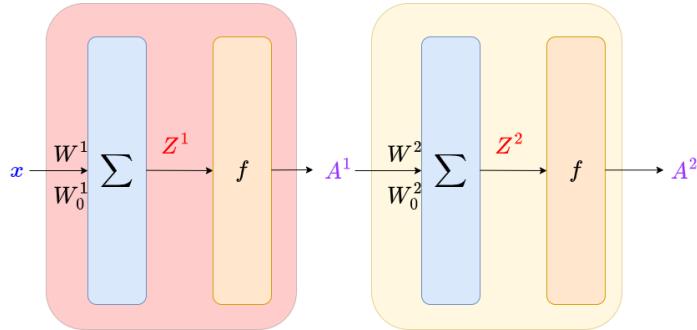
What can we learn from this?

- The **output** of layer 1, A^1 , is the **input** to layer 2.

- Thus, the output dimension n^1 of layer 1 must **match** the input m^2 of layer 2:

$$n^1 = m^2 \quad (6.26)$$

Let's break these into their components again.



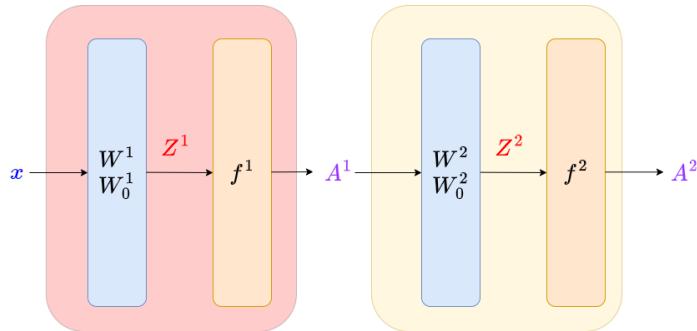
We have two separate outputs: A^1 and A^2 .

To distinguish between the linear functions in each layer, we'll just notate them using the weights and offsets.

$$\begin{matrix} W^1 \\ W_0^1 \end{matrix} \leftrightarrow \sum$$

These two are equivalent (if in the same layer)! We'll use the notation on the left, so that you know which layer our unit is in.

And this gives us:



Now, we can make our functions. For layer one:

$$A^1 = f(Z^1) = f((W^1)^T x + W_0^1) \quad (6.27)$$

And layer two:

$$A^2 = f(Z^2) = f((W^2)^T A^1 + W_0^2) \quad (6.28)$$

We can use this to build our **general** pattern.

6.2.15 Many Layers

We are finally ready to build our **complete** neural network. We'll just retrace the steps of the 2-layer case.

Notation 274

The total **number** of **layers** in our **neural network** is notated as L .

Typically we notate an **arbitrary** layer as ℓ (or l).

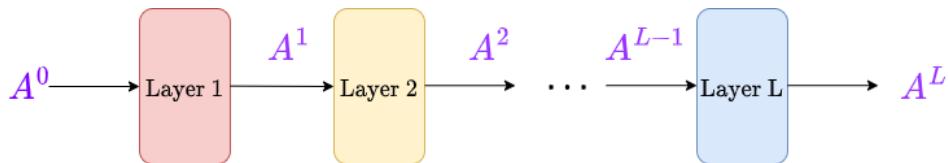
Since x is, for all purposes, **equivalent** to a vector A , we will call it A^0 .

Notation 275

Our **neural network**'s input x is used in the **same** way as every term A^ℓ .

So, we will **represent** it as

$$x = A^0$$



Again, we see that the **output** of layer ℓ is the **input** of layer $\ell + 1$.

Concept 276

Each layer **feeds** into the next layer.

A^ℓ is the **output** of layer ℓ , and the **input** of layer $\ell + 1$.

This means that the **output** dimension must **match** the next **input** dimension.

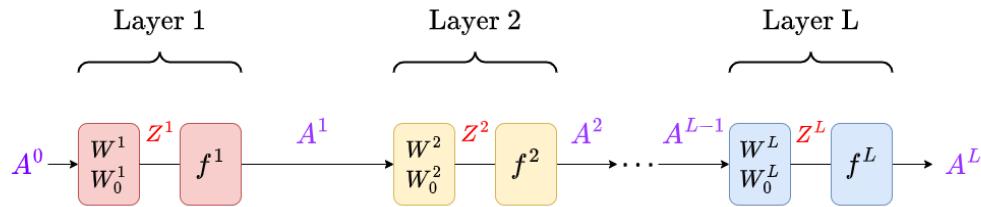
$$\underbrace{n^\ell}_{\text{Output}} = \underbrace{m^{\ell+1}}_{\text{Output}}$$

And the **dimension** of A^ℓ is $(n^\ell \times 1) = (m^{\ell+1} \times 1)$.

6.2.16 Our Complete Neural Network

We can break our layers into components, so we can see the functions involved.

With this, we build our final neural network:



With this, we can see how each layer is **related** to each other: as we **mentioned**, the **output** of one layer is the **input** of the next layer.

Here is the computation we do for layer ℓ :

Key Equation 277

The calculations done by layer ℓ are broken into two parts:

- Input $A^{\ell-1}$ turns into pre-activation Z^ℓ

$$Z^\ell = (\mathbf{W}^\ell)^T A^{\ell-1} + \mathbf{W}_0^\ell$$

- Pre-activation Z^ℓ turns into A^ℓ

$$A^\ell = f(Z^\ell)$$

Which combine into:

$$A^\ell = f(Z^\ell) = f\left((\mathbf{W}^\ell)^T A^{\ell-1} + \mathbf{W}_0^\ell\right)$$

6.2.16.1 Hidden Layers and the "First Layer"

Now that we have a full network, we introduce some useful vocab.

Definition 278

A **hidden layer** is any functional layer except for the **output** (last) layer.

It is called a "**hidden**" layer because, if you're viewing the whole neural network based on

- **Input** x (first input)
- **Output** A^L (final output)

You can't see the output of the **hidden layers** from outside the network.

Based on this definition, the **number of hidden layers** in a network is the layer count, minus one: $L - 1$.

Note that there's one point of confusion: online, you may see that the hidden layer is "any layer other than the **input** (first) or **output** (last) layer".

This is because, often, we consider the input itself to be a separate "**input layer**".

Despite this fact, when someone counts the number of layers in a neural network, they're usually only counting the hidden and output layers: we **don't count** the input layer.

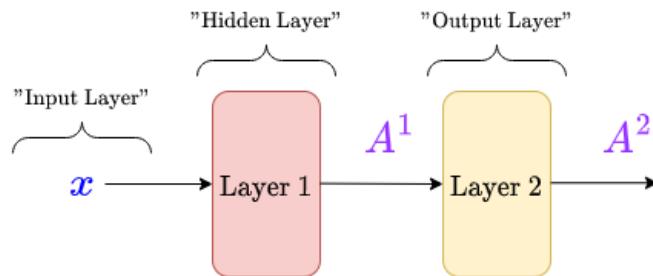
It confused me, too.

Definition 279

The **input layer** is a layer that brings the **input** into the network. It applies **no functions** to the data.

Because the input layer has **no effect** on our data (it just moves it), we **don't count the input layer** when we're saying how **many layers** a network has.

Example: Consider the following network from earlier:



In this network, x is passed into the network by the **input layer**. This layer is **before** layer 1 (you could think of it as "Layer 0").

Despite having the input layer, plus layer 1 and 2, we count only

- **Two** layers in our network:

- One hidden layer: Layer 1.
- One output layer: Layer 2.

6.3 Choices of activation function

Our linear model is entirely **defined** by its input: the number of **weights** in a neuron is just the number of **inputs** m .

But our **activation** function is up to us to decide: what works best?

6.3.1 Trying out linear activation

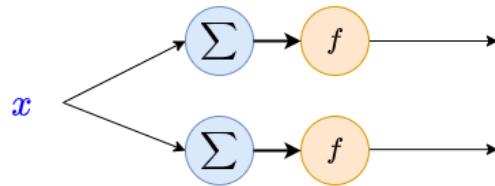
The simplest assumption would be to just use the **identity** function

$$f(z) = z \quad (6.29)$$

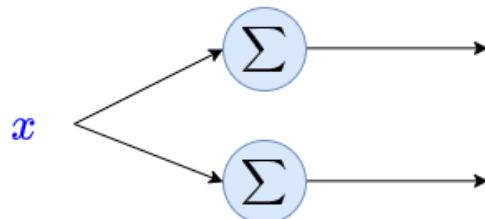
We might hope that we can combine a bunch of simple, **linear** models, and get a more sophisticated model. Why bother having a **nonlinear** activation at all?

Well, it turns out, combining **multiple** linear layers doesn't make our model any stronger. Let's try an example: we'll take a network with 2 layers, two neurons each.

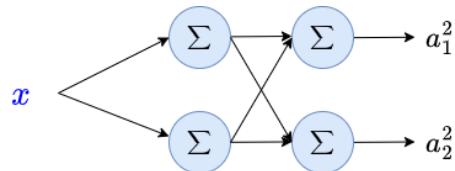
Let's look at layer 1:



Since the activation function has **no effect** on our result, we can **omit** it:

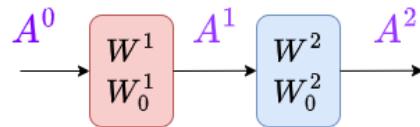


And now, we can show our **full** network:



6.3.2 Linear Layers: An example

We'll assume **two** inputs $A_0 = [x_1, x_2]^T$. For our sanity, we'll lump all of the weights in each layer: _____



In each layer, we're "combining" the linear component with the linear activation:
linear * linear=linear.

We'll leave out W_0 terms to make it more readable, but the same will apply.

Layer 1:

$$A^1 = (\textcolor{red}{W}^1)^T \textcolor{blue}{A}_0 \quad (6.30)$$

Layer 2:

$$A^2 = \overbrace{(\textcolor{blue}{W}^2)^T (\textcolor{red}{W}^1)^T}^{\text{Weight matrices}} \textcolor{blue}{A}_0 \quad (6.31)$$

The full function for this equation is two matrices, **multiplied** by our input vector.

Let's take an arbitrary example:

$$\textcolor{red}{W}^1 = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \quad \textcolor{blue}{W}^2 = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} \quad (6.32)$$

Our equation becomes:

$$A^2 = \overbrace{\begin{bmatrix} 5 & 7 \\ 6 & 8 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}}^{\text{Transposed matrices}} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (6.33)$$

We created this function by applying two matrices separately. But, can't we **combine** them?

$$A^2 = \begin{bmatrix} 19 & 43 \\ 22 & 50 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (6.34)$$

Wait, but this looks like a **one-layer** network with those weights! The second layer is **pointless**, we could have represented it with a single layer...

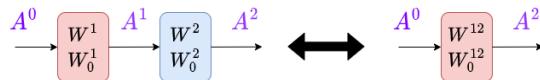
$$(\textcolor{blue}{W}^{12})^T = \begin{bmatrix} 19 & 43 \\ 22 & 50 \end{bmatrix} \quad (6.35)$$

6.3.3 The problem with linear networks

In fact, this is true in general: we can always take our **two** linear layers and combine them into **one**.

$$(\mathbf{W}^2)^T (\mathbf{W}^1)^T = \mathbf{W}^{12} \quad (6.36)$$

Our network is **equivalent** to the supposedly "simpler" one-layer network.



What if we have more layers? Well, we can just combine them one-by-one. At the end, we're just left with one layer:

$$(\mathbf{W}^L)^T (\mathbf{W}^{L-1})^T \dots (\mathbf{W}^2)^T (\mathbf{W}^1)^T = \mathbf{W} \quad (6.37)$$

And so, we can't just use linear layers: we **need** a **nonlinear** activation function.

Concept 280

Having multiple consecutive **linear layers** (i.e. layers with linear **activation** functions) is **equivalent** to having one linear layer in its place.

This means that we do not expand our **hypothesis** class by using more linear layers: we have to use **nonlinear** activation functions.

This problem is even worse than it seems: let's see why. Since we can **combine** **n** linear layers together into one, what happens if we only have **one** linear layer?

Suppose layer ℓ is linear. The next layer contains a **linear** component and a non-linear **activation** component. We'll focus on just the linear part.

Activation comes after this step, so we would just use $f(z^{\ell+1})$.

$$z^{\ell+1} = (\mathbf{W}^{\ell+1})^T \mathbf{x}^{\ell+1} = (\mathbf{W}^{\ell+1})^T (\mathbf{W}^\ell)^T \mathbf{x}^\ell \quad (6.38)$$

Wait: we have **two** consecutive **linear** components. We can combine layer ℓ with the linear component of the next layer!

$$(\mathbf{W}^{\ell+1})^T (\mathbf{W}^\ell)^T \mathbf{x}^\ell = \mathbf{W} \mathbf{x}^\ell \quad (6.39)$$

Now, we've removed layer ℓ entirely: it makes no difference to have even just one **hidden** linear layer!

Concept 281

Even having one hidden **linear layer** is **redundant**: it's **equivalent** to not having that layer at all.

Since this requires **more computation** for no benefit, we **almost never** make linear hidden layers.

So, linear models are out. What if we use something **nonlinear**?

$$A^2 = f((W^2)^T A^1) \quad (6.40)$$

We get something that doesn't seem to **simplify**:

This is ugly, but we don't have to worry about the details.

$$A^2 = f((W^2)^T \boxed{f((W^1)^T x)}) \quad (6.41)$$

If we choose our function right (and avoid linearity), this cannot be simplified to a single layer! That means, this function is **different** (and likely more **complex**) than a one-layer model.

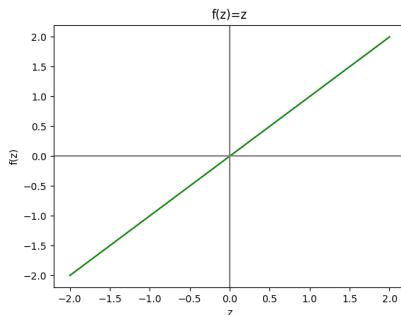
And this kind of **expressiveness** is exactly what we're looking for.

6.3.4 Example of Activation Functions

So, let's look at some possible **activation** functions:

- **Identity** function z :

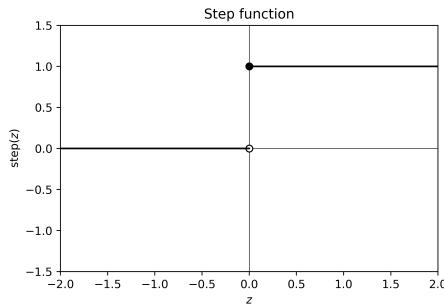
$$f(z) = z \quad (6.42)$$



- This function is called an **identity** function because it "preserves the identity" of the input: the output is the same.

- This is an example of a **linear** function.
 - * As we described in the last section, linear activation can't make our model more **expressive**.
 - * So, we **almost never** use it (or any other **linear** function) as an activation for a **hidden** layer.
- We mainly use this as an **output** activation function: it allows our final output to be any real number.
 - * This is a good activation function for a **regression** model, which returns a **real** number.
 - * It's a simple function, that can return **any** real number. By contrast, sigmoid and ReLU both have **limited** output ranges.
- **Step** function $\text{step}(z)$:

$$\text{step}(z) = \begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases} \quad (6.43)$$

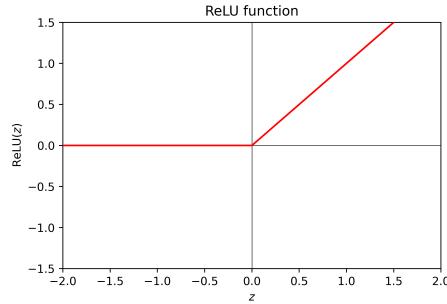


- This function is basically a **sign** function, but uses $\{0, 1\}$ instead of $\{-1, +1\}$.
- Step functions were a common early choice, but because they have a **zero** gradient, we can't use **gradient descent**, and so we basically **never** use them.

- **Rectified Linear Unit** $\text{ReLU}(z)$:

Same reason we replaced the sign function with sigmoid, when we were doing linear logistic classifiers.

$$\text{ReLU}(z) = \max(0, z) = \begin{cases} z & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases} \quad (6.44)$$

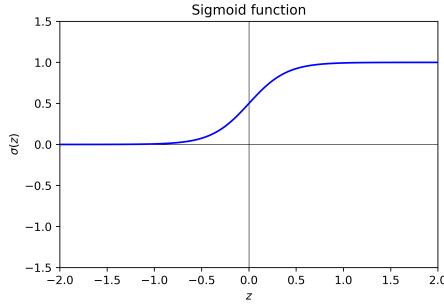


- This is a very **common** choice for activation function, even though the derivative is undefined at 0.
- We specifically use it for internal ("**hidden**") layers: layers that are neither the **first** nor **last** layer.

- **Sigmoid** function $\sigma(z)$:

They're "hidden" because they aren't visible to the input or output.

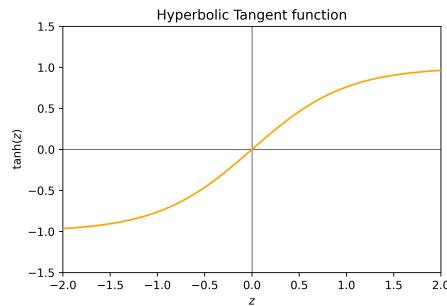
$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (6.45)$$



- This is the **activation** function for our **LLC** neuron from before.
- Just like LLC, it's useful for the **output neuron** in **binary classification**.
- Can be interpreted as the **probability** of a positive (+1) binary classification.
- We can also use this for multiclass when classes are **NOT** disjoint: we use one sigmoid per class.
 - * Each sigmoid tells us how likely the data point is to be in that class.

- **Hyperbolic Tangent** $\tanh(z)$:

$$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \quad (6.46)$$



- This function looks similar to sigmoid over a different **range**.
- Unfortunately, it will not get much use in this class.
- **Softmax** function $\text{softmax}(z)$:

$$\text{softmax}(z) = \begin{bmatrix} \exp(z_1) / \sum_i \exp(z_i) \\ \vdots \\ \exp(z_n) / \sum_i \exp(z_i) \end{bmatrix} \quad (6.47)$$

- Behaves like a disjoint, **multi-class** version of **sigmoid**.
- Appropriately, we use it as the **output neuron** for **multi-class** classification.
- Can be interpreted as the **probability** of our k possible classifications.
 - * "Disjoint" probability: each option is separate. Sum of the rows adds up to 1.

Concept 282

For the different **activation functions**:

- $f(z) = z$ isn't used for **hidden** layers, but we can use it for regression **output**.
- $\text{sign}(z)$ is **rarely** used.
- $\text{ReLU}(z)$ is often used for "**hidden**" layers.
- $\sigma(z)$ is often used as the **output** for **binary classification**.
- $\text{softmax}(z)$ is often used as the **output** for **multi-class classification**

$\tanh(z)$ is useful, but not a focus of this class.

Remember this caveat, though:

Clarification 283

Multi-class depends on whether a **data point** can be in **multiple classes at the same time**.

- $\text{softmax}(z)$ assumes our classes are **disjoint**: you can only be in **one** class.
 - This is usually what people mean by **multi-class**.
- $\sigma(z)$ can be used when classes are **not disjoint**: you can be in **multiple** classes.
 - You can think of this as **binary classification** for each class.

When using sigmoids, we need **one** sigmoid for each **class**.

Example: We can compare use cases for each of these:

- Softmax could be used to answer, "which word is the next one in the sentence?"
 - Every word in a sentence is only followed by one word: they're mutually exclusive.
- Sigmoids could be used to answer, "what genre of book is that?"
 - A book is often in more than one genre.

6.4 Loss functions and activation functions

As we can see above, your **activation** function depends on what kind of **problem** you're dealing with.

The same is true for our **loss** function: we used **different** loss functions for classification and regression.

Classification can be further broken up into **binary** versus **multiclass** classification.

To summarize our findings, we'll **sort** this information:

Concept 284

Each of our **tasks** requires a different **loss** and output **activation** function.

We emphasize that we specifically mean the **output** activation function: the activation function used in **hidden layers** doesn't have to match the loss function.

task	f^L	Loss	
Regression	Linear z	Squared	$(g - y)^2$
Binary Class	Sigmoid $\sigma(z)$	NLL	$y \log g + (1 - y) \log(1 - g)$
Multi-Class	Softmax $\text{softmax}(z)$	NLLM	$\sum_j y_j \log(g_j)$

Special Case: If we allow **multiple** classes at the **same** time (non-disjoint), we use **binary** classification for each of them, rather than multi-class.

Example: An example for each type:

- **Regression:** Predicting the amount of rainfall in centimeters tomorrow.
- **Binary Classification:** Will the stock market go up or down tomorrow?
- **Multi-Class:** What species of tree is this?
- **Multiple Binary:** What are the themes in this movie?

6.4.1 Other Considerations

You might consider using other functions, based on the needs of a more specialized task. We'll ignore those cases, for the most part.

But, if you want to try a new function, the **data type** is the most important for whether we

can use it.

Concept 285

If you want to use a new **activation** or **loss** function, you have to pay attention to the **input/output** type.

Example: $\tanh(z)$ outputs over the range $(-1, 1)$. We could use it, if that was the range we wanted.

Be careful, though:

Clarification 286

It's important to stress that while our **output activation** depends on the task, **hidden layers** don't have to.

Hidden layers can use one of several **different** activation functions, regardless of the **task**.

However, some activation functions tend to be **better** for making a model than others.

Example: Often, we use ReLU for hidden layers, but it's rarely used as an output activation function.

We also might use **sigmoid** as a hidden layer for a regression model, even though regression most commonly uses a **linear** output.

Terms

- Neuron (Unit, Node)
- Neural Network
- Series and Parallel
- Linear Component
- Weight w
- Offset (Bias, Threshold) w_0
- Activation Function f
- Pre-activation z
- Activation a
- Identity Function
- Acyclic Networks
- Feed-forward Networks
- Layer
- Fully Connected
- Input dimension m
- Output dimension n
- Weight Matrix
- Offset Matrix
- Layer Notation A^ℓ
- Step function
- ReLU function
- Sigmoid function
- Hyperbolic tangent function
- Softmax function

CHAPTER 6

Neural Networks 2 - Back-Propagation and Training

6.5 Error back-propagation

We have a complete neural network: a **model** we can use to make predictions or calculations.

Now, our mission is to **improve** this neural network: even if our hypothesis class is good, we still have to **find** the hypotheses that are useful for our problem.

As usual, we will start out with **randomized** values for our weights and biases: this **initial** neural network will not be useful for anything in particular, but that's why we need to improve it.

For such a complex problem, we definitely can't find an explicit solution, like we did for ridge regression. Instead, we will have to rely on **gradient descent**.

Concept 287

Neural networks are typically optimized using **gradient descent**.

We randomize them because otherwise, if our initialization is $w_i = 0$, we get

$$w^T x + w_0 = 0$$

no matter what input x we have.

6.5.1 Review: Gradient Descent

What does it really mean to do gradient descent on our **network**? Let's remind ourselves of how gradient descent works, and then **build** up to a network.

Concept 288

Gradient descent works based on the following reasoning:

- We have a function we want to **minimize**: our loss function \mathcal{L} , which tells us how **badly** we're doing.
- We want to perform "less badly". Our main tool for **improving** \mathcal{L} is to alter θ and θ_0 .
 - These are our **parameters**: we're adjusting our model.
- The **gradient** is our main tool: $\frac{\partial \mathcal{L}}{\partial \theta}$ tells you the direction to **change** θ in order to **increase** \mathcal{L} .
- We want to **change** θ to **decrease** \mathcal{L} . Thus, we move in the direction of

$$\Delta\theta = -\eta \frac{\partial \mathcal{L}}{\partial \theta}$$
- We take steps $\Delta\theta$ (and $\Delta\theta_0$) until we are satisfied with \mathcal{L} , or it **stops** improving.

Remember that η is our **step size**: we can take bigger or smaller steps in each direction.

6.5.2 Review: Gradient Descent with LLCs

Let's start with a familiar example: **LLCs**.

Our LLC model uses the following equations:

We'll use w instead of θ .

$$z(x) = w^T x + w_0 \quad g(z) = \sigma(z) = \frac{1}{1 + e^{-z}} \quad (6.1)$$

$$\mathcal{L}(g, y) = y \log(g) + (1 - y) \log(1 - g) \quad (6.2)$$

Our goal is to minimize \mathcal{L} by adjusting θ and θ_0 .

So, we want

$$\frac{\partial \mathcal{L}}{\partial w} \quad \text{and} \quad \frac{\partial \mathcal{L}}{\partial w_0} \quad (6.3)$$

We did this by using the **chain rule**:

We'll focus on w , but the same goes for w_0 .

$$\frac{\partial \mathcal{L}}{\partial w} = \overbrace{\frac{\partial \mathcal{L}}{\partial g}}^{\mathcal{L}(g)} \cdot \overbrace{\frac{\partial g}{\partial w}}^{\sigma'(z)} \quad (6.4)$$

We can break it up further using **repeated** chain rules:

$$\frac{\partial \mathcal{L}}{\partial w} = \overbrace{\frac{\partial \mathcal{L}}{\partial g}}^{\mathcal{L}(g)} \cdot \underbrace{\frac{\partial g}{\partial z}}_{g(z)} \cdot \frac{\partial z}{\partial w} \quad (6.5)$$

Plugging in our derivatives, we get:

$$\frac{\partial \mathcal{L}}{\partial w} = -\left(\frac{y}{\sigma} - \frac{1-y}{1-\sigma}\right) \cdot \underbrace{\frac{\partial g}{\partial z}}_{\sigma(1-\sigma)} \cdot \underbrace{\frac{\partial z}{\partial w}}_x \quad (6.6)$$

Concept 289

The **chain rule** allows us to take the gradient of **nested functions**, where each function is the **input** to the next one.

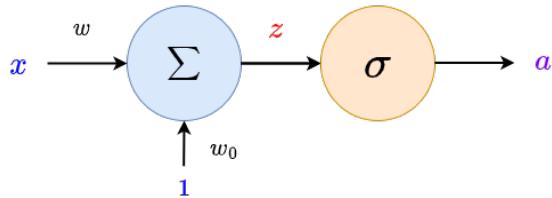
Another way to say this is that one function **feeds into** the next.

If you aren't familiar with "nested" functions, consider this example:

If you have functions $f(x)$ and $g(x)$, then $g(f(x))$ is the **nested** combination, where the output of f is the input of g .

6.5.3 Review: LLC as Neuron

Remember that we can represent our LLC as a **neuron**: this could give us the first idea for how to train our **neural network**!



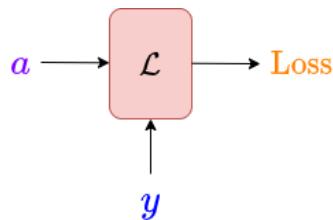
As usual, our first unit \sum is our **linear** component. The output is z , nothing different from before with LLC.

The **output** of σ , which we wrote before as g , is now a .

Remember that x is a whole vector of values, which we've condensed into one variable.

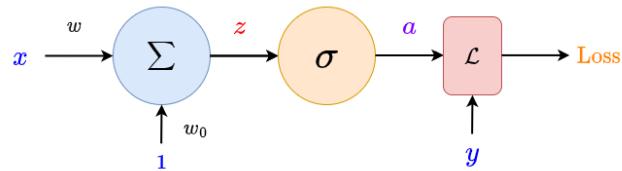
Something we neglected before: this diagram is **missing** the **loss function**. Let's create a small unit for that.

$\mathcal{L}(a, y)$ has **two** inputs: our predicted value a , and the correct value y .



We have two inputs to our loss function.

We **combine** these into a single unit to get:



Our full unit!

6.5.4 LLC Forward-Pass

Now, we can do gradient descent like before. We want to get the effect our **weight** has on our **loss**.

But, this time, we'll pair it with a **visual** that is helpful for understanding how we **train** neural networks.

First, one important consideration:

As we saw above, the **gradient** we get might rely on z , a , or $\mathcal{L}(a, y)$. So, before we do anything, we have to **compute** these values.

Each step **depends** on the last: this is what the **forward** arrows represent. We call this a **forward pass** on our neural network.

Definition 290

A **forward pass** of a neural network is the process of sending information "**forward**" through the neural network, starting from the **input**.

This means the **input** is fed into the **first** layer, and that output is fed into the **next** layer, and so on, until we reach our **final** result and **loss**.

Example: If we had

- $f(x) = x + 2$

- $g(f) = 3f$
- $h(g) = \sin(g)$

Then, a forward pass with the input $x = 10$ would have us go function-by-function:

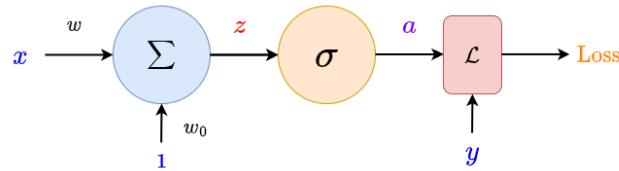
- $f(10) = 10 + 2$
- $g(f) = 3 \cdot 12$
- $h(g) = \sin(36)$

So, by "forward", we mean that we apply each function, one after another.

In our case, this means computing z , a , and $\mathcal{L}(a, y)$.

6.5.5 LLC Back-propagation

Now that we have all of our values, we can get our gradient. Let's **visualize** this process.



We want to link \mathcal{L} to w . In order to do that, we need to **connect** each thing in between.

- This lets us **combine** lots of simple **links** to get our more complicated result.

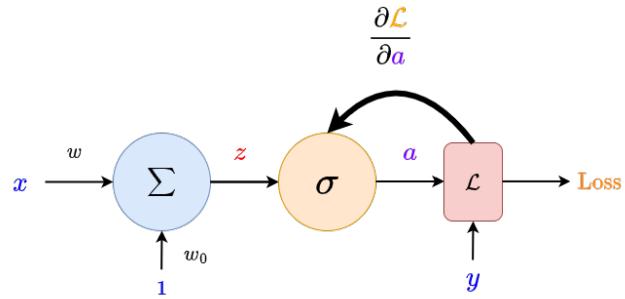
We can also call this "chaining together" lots of derivatives.

Loss \mathcal{L} is what we really care about. So, what is the loss directly **connected** to? The **activation**, a .

- Our loss function $\mathcal{L}(a, y)$ contains information about how \mathcal{L} is linked to a .

$$\overbrace{\frac{\partial \mathcal{L}}{\partial a}}^{\text{Loss unit}} \quad (6.7)$$

We send this information backwards, so it can be used later.

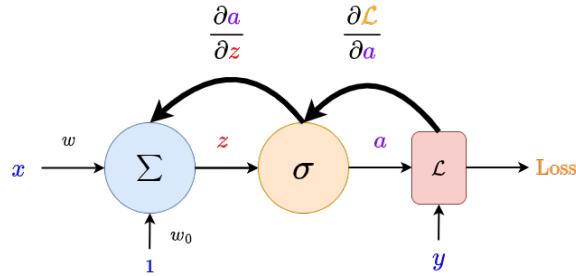


Now, we're on the $\sigma(z)$ unit.

- The $\sigma(z)$ unit contains information about how a is linked to z .
- We've connected \mathcal{L} to a , and a to z . We chain them together, connecting \mathcal{L} to z .

$$\overbrace{\frac{\partial \mathcal{L}}{\partial a}}^{\text{Loss unit}} \cdot \overbrace{\frac{\partial a}{\partial z}}^{\text{Activation function}} \quad (6.8)$$

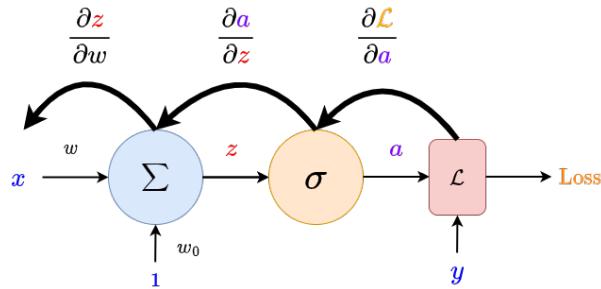
We haven't reached w yet, so we send this information further back.



Finally, we reach \sum .

- The \sum unit contains information about how z is linked to w .
- Finally, we have a chain of links, that allows us to connect \mathcal{L} to w .

This last derivative uses x , because $w^T x + w_0 = z$.



And, we built our chain rule! This contains the **information** of the derivatives from **every** unit.

$$\frac{\partial \mathcal{L}}{\partial w} = \underbrace{\frac{\partial \mathcal{L}}{\partial a}}_{\text{Loss unit}} \cdot \underbrace{\frac{\partial a}{\partial z}}_{\text{Activation}} \cdot \underbrace{\frac{\partial z}{\partial w}}_{\text{Linear subunit}} \quad (6.9)$$

Moving backwards like this is called **back-propagation**.

Definition 291

Back-propagation is the process of moving "backwards" through your network, starting at the **loss** and moving back layer-by-layer, and gathering terms in your **chain rule**.

We call it "propagation" because we send backwards the **terms** of our chain rule about later derivatives.

An **earlier** unit (closer to the "left") has all of the **derivatives** that come after (to the "right" of) it, along with its own term.

6.5.6 Summary of neural network gradient descent: a high-level view

So, with just this, we have built up the basic idea of how we **train** our model: now that we have the gradient, we can do **gradient descent** like we normally do!

This summary covers some things we haven't fully discussed. We'll continue digging into the topic!

Concept 292

We can do **gradient descent** on a **neural network** using the ideas we've built up:

- Do a **forward pass**, where we compute the value of each **unit** in our model, passing the information **forward** - each layer's **output** is the next layer's **input**.
 - We finish by getting the **loss**.

- Do **back-propagation**: build up a **chain rule**, starting at the **loss** function, and get each unit's **derivative** in **reverse order**.
 - **Reverse** order: if you have 3 layers, you want to get the 3rd layer's **derivatives**, then the 2nd layer, then the 1st.
 - **Each weight** vector has its own **gradient**: we'll deal with this later, but we need to calculate one for each of them.

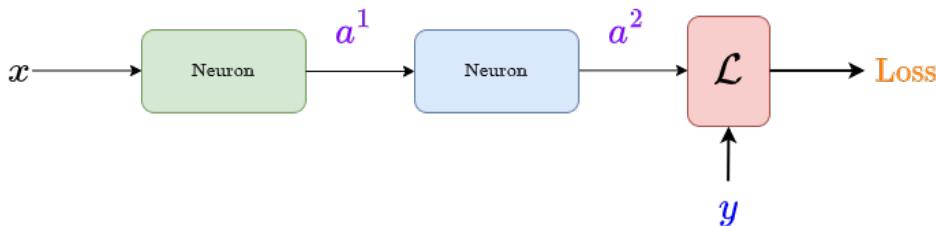
- Use your chain rule to get the **gradient** $\frac{\partial \mathcal{L}}{\partial w}$ for your **weight** vector(s). Take a **gradient descent** step.
- **Repeat** until satisfied, or your model **converges**.

6.5.7 A two-neuron network: starting backprop

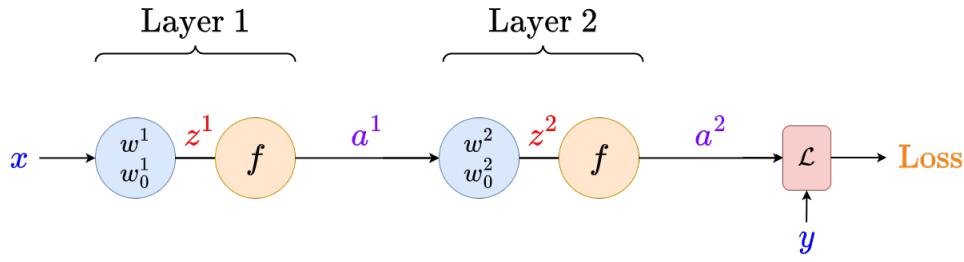
Above, we mention "each layer": we'll now transition to a **two-neuron** system, so we have "two layers". Then, we'll build up to many layers.

Remember, though, that the **ideas** represented here are just extensions of what we did **above**.

Let's get a look at our **two-neuron** system, now with our **loss** unit:



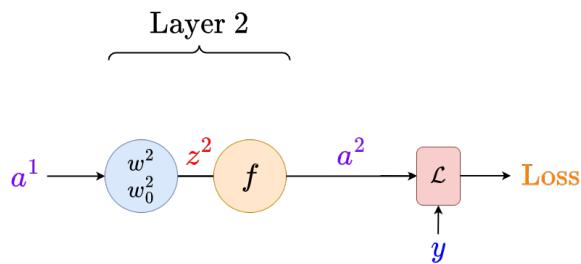
And unpack it:



We want to do **back-propagation** like we did before. This time, we have **two** different layers of weights: w^1 and w^2 . Does this cause any problems?

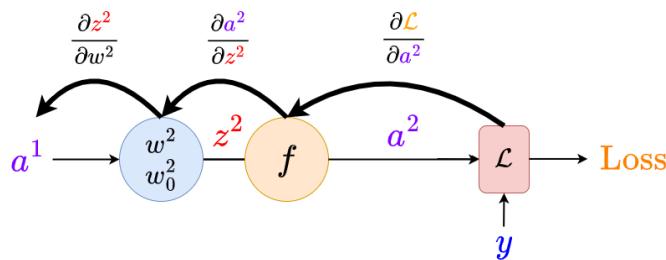
It turns out, it doesn't! We mentioned in the first part of chapter 7 that we can treat the **output** of the **first** layer a^1 as the same as if it were an **input** x .

This is one of the biggest benefits of neural network layers!



Now, we can do backprop safely.

"Backprop" is a common shortening of "back-propagation".



We can get:

$$\frac{\partial \mathcal{L}}{\partial w^2} = \overbrace{\frac{\partial \mathcal{L}}{\partial a^2}}^{\text{Loss unit}} \cdot \overbrace{\frac{\partial a^2}{\partial z^2}}^{\text{Activation}} \cdot \overbrace{\frac{\partial z^2}{\partial w^2}}^{\text{Linear}} \quad (6.10)$$

The same format as for our **one-neuron** system! We now have a gradient we can update for our **second** weight vector.

But what about our **first** weight vector?

6.5.8 Continuing backprop: One more problem

We need to continue further to reach our **earlier** weights: this is why we have to work **backward**.

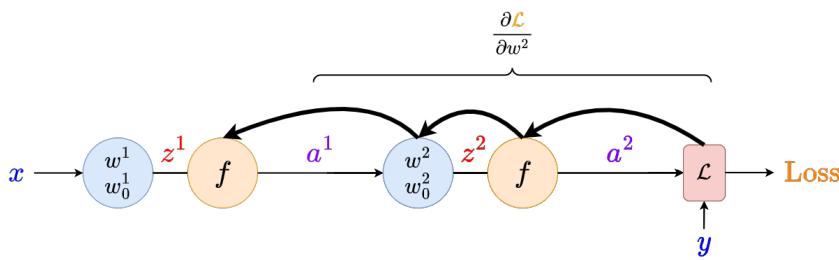
Concept 293

We work **backward** in **back-propagation** because every layer after the **current** one **affects** the gradient.

Our current layer **feeds** into the next layer, which feeds into the layer after that, and so on. So this layer affects **every** later layer, which then affect the loss.

So, to see the effect on the **output**, we have to **start** from the **loss**, and get every layer **between** it and our weight vector.

Remember that when we say "f feeds into g", we mean that the output of f is the input to g.



We have one problem, though:

We just gathered the derivative $\partial \mathcal{L} / \partial w^2$. If we wanted to continue the chain rule, we would expect to add more terms, like:

$$\frac{\partial w^2}{\partial a^1} \quad (6.11)$$

The problem is, what is w^2 ? It's a vector of constants.

Since our current derivative includes w^2 , we would continue it with a w^2 in the "top" of a derivative,

$$\frac{\partial \mathcal{L}}{\partial w^2} \frac{\partial w^2}{\partial r}$$

We're not sure what "r" is yet.

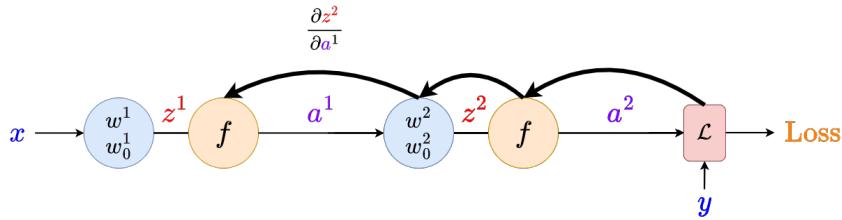
$$w^2 = \begin{bmatrix} w_1^2 \\ w_2^2 \\ \vdots \\ w_n^2 \end{bmatrix}, \quad \text{Not a function of } a^1! \quad (6.12)$$

That derivative above is going to be **zero!** In other words, w^2 isn't really the **input** to z^2 : it's a **parameter**.

So, we can't end our derivative with w^2 . Instead, we have to use something else. z^2 's real input is a^1 , so let's go directly to that!

We were building our chain rule by combining inputs with outputs: that's what links two layers together.

So, it should make sense that using something like w (that doesn't link two layers) prevents us from making a longer chain rule.



Using this allows us to move from layer 2 to layer 1.

Now, we have our new chain rule:

$$\frac{\partial \mathcal{L}}{\partial a^1} = \underbrace{\frac{\partial \mathcal{L}}{\partial a^2}}_{\text{Other terms}} \cdot \underbrace{\frac{\partial a^2}{\partial z^2}}_{\text{Link Layers}} \cdot \underbrace{\frac{\partial z^2}{\partial a^1}}_{\text{Link Layers}} \quad (6.13)$$

Concept 294

For our **weight gradient** in layer l , we have to end our **chain rule** with

$$\frac{\partial z^l}{\partial w^l}$$

So we can get

$$\frac{\partial \mathcal{L}}{\partial w^l} = \underbrace{\frac{\partial \mathcal{L}}{\partial z^l}}_{\text{Other terms}} \cdot \underbrace{\frac{\partial z^l}{\partial w^l}}_{\text{Get weight grad}}$$

However, because w^l is not the **input** of layer l , we can't use it to find the gradient of **earlier layers**.

Instead, we use

$$\frac{\partial z^l}{\partial a^{l-1}} \quad (6.14)$$

To "link together" two different layers l and $l - 1$ in a **chain rule**.

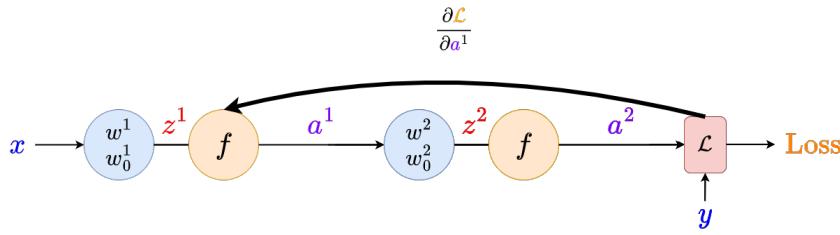
6.5.9 Finishing two-neuron backprop

Now that we have safely connected our layers, we can do the rest of our gradient. First, let's lump together everything we did before:

In this section, we compressed lots of derivatives into

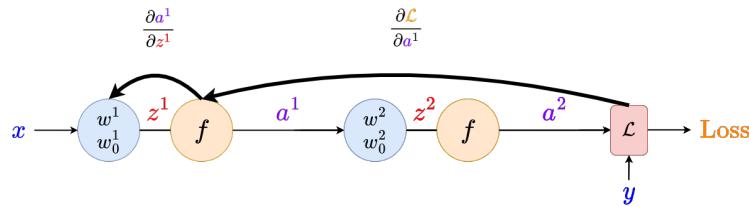
$$\frac{\partial \mathcal{L}}{\partial z^l}$$

Don't let this alarm you, this just hides our long chain of derivatives!

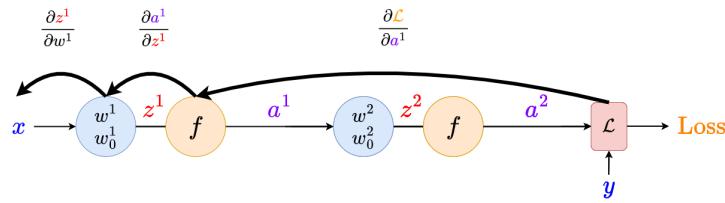


All the info we need is stored in this derivative: it can be written out using our friendly chain rule from earlier.

Now, we can add our remaining terms. It's the same as before: we want to look at the pre-activation



And finally, our input:



We can get our second chain rule

$$\frac{\partial \mathcal{L}}{\partial w^1} = \overbrace{\frac{\partial \mathcal{L}}{\partial a^1}}^{\text{Other layers}} \cdot \overbrace{\frac{\partial a^1}{\partial z^1} \cdot \frac{\partial z^1}{\partial w^1}}^{\text{Layer 1}} \quad (6.15)$$

Which, in reality, looks much bigger:

$$\frac{\partial \mathcal{L}}{\partial w^1} = \overbrace{\left(\frac{\partial \mathcal{L}}{\partial a^2} \right)}^{\text{Loss unit}} \cdot \overbrace{\left(\frac{\partial a^2}{\partial z^2} \cdot \frac{\partial z^2}{\partial a^1} \right)}^{\text{Layer 2}} \cdot \overbrace{\left(\frac{\partial a^1}{\partial z^1} \cdot \frac{\partial z^1}{\partial w^1} \right)}^{\text{Layer 1}} \quad (6.16)$$

We see a clear **pattern** here! In fact, this is the procedure we'll use for a neural network with **any** number of layers.

Concept 295

We can get all of our **weight gradients** by repeatedly appending to the **chain rule**.

If we want to get the **weight gradient** of layer ℓ , we **terminate** with

$$\text{Within layer} \quad \overbrace{\frac{\partial a^\ell}{\partial z^\ell}}^{\text{Get weight grad}} \quad . \quad \overbrace{\frac{\partial z^\ell}{\partial w^\ell}}$$

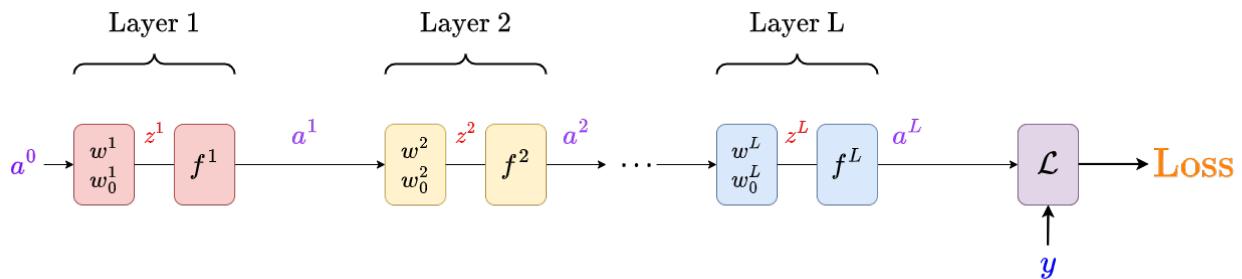
If we want to **extend** to the previous layer, we **instead** multiply by

$$\text{Within layer} \quad \overbrace{\frac{\partial a^\ell}{\partial z^\ell}}^{\text{Link layers}} \quad . \quad \overbrace{\frac{\partial z^\ell}{\partial a^{\ell-1}}}$$

6.5.10 Many layers: Doing back-propagation

Now, we'll consider the case of many possible layers.

To make it more readable, we'll use boxes instead of circles for units.



This may look intimidating, but we already have all the tools we need to handle this problem.

Our goal is to get a **gradient** for each of our **weight** vectors w^ℓ , so we can do gradient descent and **improve** our model.

According to our above analysis in Concept 9, we need only a few steps to get all of our gradients.

Concept 296

In order to do **back-propagation**, we have to build up our **chain rule** for each weight gradient.

- We start our chain rule with one term shared by every gradient:

$$\overbrace{\frac{\partial \mathcal{L}}{\partial a^L}}$$

Loss unit

Then, we follow these two steps until we run out of layers:

- We're at layer ℓ . We want to get the **weight gradient** for this layer. We get this by **multiplying** our chain rule by

$$\begin{array}{c} \text{Within layer} \quad \text{Get weight grad} \\ \overbrace{\frac{\partial a^\ell}{\partial z^\ell}} \quad \cdot \quad \overbrace{\frac{\partial z^\ell}{\partial w^\ell}} \end{array}$$

We **exclude** this term for any other gradients we want.

- If we aren't at layer 1, there's a previous layer we want to get the weight for. We reach layer $\ell - 1$ by multiplying our chain rule by

$$\begin{array}{c} \text{Within layer} \quad \text{Link layers} \\ \overbrace{\frac{\partial a^\ell}{\partial z^\ell}} \quad \cdot \quad \overbrace{\frac{\partial z^\ell}{\partial a^{\ell-1}}} \end{array}$$

Once we reach layer 1, we have **every single** weight vector we need! Repeat the process for w_0 gradients and then do **gradient descent**.

Let's get an idea of what this looks like in general:

$$\frac{\partial \mathcal{L}}{\partial w^\ell} = \overbrace{\left(\frac{\partial \mathcal{L}}{\partial a^L} \right)}^{\text{Loss unit}} \cdot \overbrace{\left(\frac{\partial a^L}{\partial z^L} \cdot \frac{\partial z^L}{\partial a^{L-1}} \right)}^{\text{Layer L}} \cdot \overbrace{\left(\frac{\partial a^{L-1}}{\partial z^{L-1}} \cdot \frac{\partial z^{L-1}}{\partial a^{L-2}} \right)}^{\text{Layer L-1}} \cdot \left(\dots \right) \cdot \overbrace{\left(\frac{\partial a^\ell}{\partial z^\ell} \cdot \frac{\partial z^\ell}{\partial w^\ell} \right)}^{\text{Layer } \ell} \quad (6.17)$$

That's pretty ugly. If we need to hide the complexity, we can:

Notation 297

If you need to do so for **ease**, you can **compress** your derivatives. For example, if we want to only have the last weight term **separate**, we can do:

$$\frac{\partial \mathcal{L}}{\partial w^\ell} = \overbrace{\frac{\partial \mathcal{L}}{\partial z^\ell}}^{\text{Other}} \cdot \overbrace{\frac{\partial z^\ell}{\partial w^\ell}}^{\text{Weight term}}$$

But we should also explore what each of these terms *are*.

6.5.11 What do these derivatives equal?

Let's look at each of these derivatives and see if we can't simplify them a bit.

First, every gradient needs

- The **loss derivative**:

$$\frac{\partial \mathcal{L}}{\partial a^L} \tag{6.18}$$

This **depends** on our loss function, so we're **stuck** with that one.

Next, within each layer, we have

- The **activation function** - between our activation a and preactivation z :

$$\frac{\partial a^\ell}{\partial z^\ell} \tag{6.19}$$

What does the function between these **look** like?

$$a = f(z) \tag{6.20}$$

Well, that's not super interesting: we **don't know** our function. But, at least we can **write** it using f : that way, we know that this term only depends on our **activation** function.

$$\frac{\partial a^\ell}{\partial z^\ell} = \overbrace{(f^\ell)'}^{\text{deriv of func for layer } \ell} \overbrace{(z^\ell)}^{\text{Deriv input}} \tag{6.21}$$

This expression is a bit visually clunky, but it works. Without the annotation:

z^ℓ is not being multiplied by $(f^\ell)'$, it's the input to that derivative.

$$\frac{\partial a^\ell}{\partial z^\ell} = (f^\ell)'(z^\ell) \tag{6.22}$$

Between layers, we have

- We can also think about the derivative of the **linear function** that **connects two layers**:

$$\frac{\partial z^\ell}{\partial a^{\ell-1}} \quad (6.23)$$

So, we want the function of these two:

Be careful not to get this mixed up with the last one!
They look similar, but one is within the layer, and the other is between layers.

$$z^\ell = w^\ell a^{\ell-1} + w_0^\ell \quad (6.24)$$

This one is pretty simple! We just take the derivative manually:

$$\frac{\partial z^\ell}{\partial a^{\ell-1}} = w^\ell \quad (6.25)$$

Finally, every gradient will end with...

- The derivative that directly connects to a **weight**, again using the **linear function**:

$$\frac{\partial z^\ell}{\partial w^\ell} \quad (6.26)$$

The linear function is the same:

$$z^\ell = w^\ell a^{\ell-1} + w_0^\ell \quad (6.27)$$

But with a different **variable**, the **derivative** comes out different:

$$\frac{\partial z^\ell}{\partial w^\ell} = a^{\ell-1} \quad (6.28)$$

Notation 298

Our **derivatives** for the **chain rule** in a **1-D neural network** take the form:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}^L}$$

$$\frac{\partial \mathbf{a}^\ell}{\partial \mathbf{z}^\ell} = (f^\ell)'(\mathbf{z}^\ell)$$

$$\frac{\partial \mathbf{z}^\ell}{\partial \mathbf{a}^{\ell-1}} = \mathbf{w}^\ell$$

$$\frac{\partial \mathbf{z}^\ell}{\partial w^\ell} = \mathbf{a}^{\ell-1}$$

Now, we can rewrite our generalized expression for gradient:

$$\frac{\partial \mathcal{L}}{\partial w^\ell} = \overbrace{\left(\frac{\partial \mathcal{L}}{\partial \mathbf{a}^L} \right)}^{\text{Loss unit}} \cdot \overbrace{\left((f^L)'(\mathbf{z}^L) \cdot \mathbf{w}^L \right)}^{\text{Layer L}} \cdot \overbrace{\left((f^{L-1})'(\mathbf{z}^{L-1}) \cdot \mathbf{w}^{L-1} \right)}^{\text{Layer L-1}} \cdot \left(\dots \right) \cdot \overbrace{\left((f^\ell)'(\mathbf{z}^\ell) \cdot \mathbf{a}^{\ell-1} \right)}^{\text{Layer } \ell} \quad (6.29)$$

Our expressions are more concrete now. It's still pretty visually messy, though.

6.5.12 Activation Derivatives

We weren't able to **simplify** our expressions above, partly because we didn't know which **loss** or **activation** function we were going to use.

So, here, we will look at the **common** choices for these functions, and **catalog** what their derivatives look like.

- **Step function** $\text{step}(z)$:

$$\frac{d}{dz} \text{step}(z) = 0 \quad (6.30)$$

This is part of why we don't use this function: it has no gradient. We can show this by looking piecewise:

$$\text{step}(z) = \begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases} \quad (6.31)$$

And take the derivative of each piece:

$$\frac{d}{dz} \text{ReLU}(z) = 0 = \begin{cases} 0 & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases} \quad (6.32)$$

- **Rectified Linear Unit** $\text{ReLU}(z)$:

$$\frac{d}{dz} \text{ReLU}(z) = \text{step}(z) \quad (6.33)$$

This one might be a bit surprising at first, but it makes sense if you **also** break it up into cases:

$$\text{ReLU}(z) = \max(0, z) = \begin{cases} z & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases} \quad (6.34)$$

And take the derivative of each piece:

$$\frac{d}{dz} \text{ReLU}(z) = \text{step}(z) = \begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases} \quad (6.35)$$

- **Sigmoid** function $\sigma(z)$:

$$\frac{d}{dz} \sigma(z) = \sigma(z)(1 - \sigma(z)) = \frac{e^{-z}}{(1 + e^{-z})^2} \quad (6.36)$$

This derivative is useful for simplifying NLL, and has a nice form.

As a reminder, the function looks like:

We can just compute the derivative with the single-variable chain rule.

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (6.37)$$

- **Identity** ("linear") function $f(z) = z$:

$$\frac{d}{dz} z = 1 \quad (6.38)$$

This one follows from the definition of the derivative.

We cannot rely on a linear activation function for our **hidden** layers, because a linear neural network is no more **expressive** than one layer.

But, we use it as the output activation for **regression**.

- **Softmax** function $\text{softmax}(z)$:

This function has a difficult derivative we won't go over here.

If you're curious, here's a [link](#).

- **Hyperbolic tangent** function $\tanh(z)$:

$$\frac{d}{dz} \tanh(z) = 1 - \tanh(z)^2 \quad (6.39)$$

This strange little expression is 1 minus the "hyperbolic secant" squared. We won't bother further with it.

Notation 299

For our various **activation** functions, we have the **derivatives**:

Step:

$$\frac{d}{dz} \text{step}(z) = 0$$

ReLU:

$$\frac{d}{dz} \text{ReLU}(z) = \text{step}(z)$$

Sigmoid:

$$\frac{d}{dz} \sigma(z) = \sigma(z)(1 - \sigma(z))$$

Identity/Linear:

$$\frac{d}{dz} z = 1$$

6.5.13 Loss derivatives

Now, we look at the loss derivatives.

- **Square loss** function $\mathcal{L}_{\text{sq}} = (a - y)^2$:

$$\frac{d}{da} \mathcal{L}_{\text{sq}} = 2(a - y) \quad (6.40)$$

Follows from chain rule+power rule, used for regression.

- **Linear loss** function $\mathcal{L}_{\text{lin}} = |a - y|$:

$$\frac{d}{da} \mathcal{L}_{\text{lin}} = \text{sign}(a - y) \quad (6.41)$$

This one can also be handled piecewise, like $\text{step}(z)$ and $\text{ReLU}(z)$:

$$|u| = \begin{cases} u & \text{if } z \geq 0 \\ -u & \text{if } z < 0 \end{cases} \quad (6.42)$$

We take the piecewise derivative:

$$\frac{d}{du}|u| = \text{sign}(u) = \begin{cases} 1 & \text{if } z \geq 0 \\ -1 & \text{if } z < 0 \end{cases} \quad (6.43)$$

- **NLL** (Negative-Log Likelihood) function $\mathcal{L}_{\text{NLL}} = -(y \log(a) + (1-y) \log(1-a))$

$$\frac{d}{da}\mathcal{L}_{\text{NLL}} = -\left(\frac{y}{a} - \frac{1-y}{1-a}\right) \quad (6.44)$$

- **NLLM** (Negative-Log Likelihood Multiclass) function $\mathcal{L}_{\text{NLL}} = -\sum_j y_j \log(a_j)$

Similar to softmax, we will omit this derivative.

Notation 300

For our various **loss** functions, we have the **derivatives**:

Square:

$$\frac{d}{da}\mathcal{L}_{\text{sq}} = 2(a - y)$$

Linear (Absolute):

$$\frac{d}{da}\mathcal{L}_{\text{lin}} = \text{sign}(a - y)$$

NLL (Negative-Log Likelihood):

$$\frac{d}{da}\mathcal{L}_{\text{NLL}} = -\left(\frac{y}{a} - \frac{1-y}{1-a}\right)$$

6.5.14 Many neurons per layer

Now, we just have left the elephant in the room: what do we do about the case where we have *big* layers? That is, what if we have **multiple** neurons per layer? This makes this more complex.

Well, the solution is the same as earlier in the course: we introduce **matrices**.

But this time, with a twist: we have to do serious **matrix** calculus: a difficult topic indeed.

To handle this, we will go in somewhat **reversed** order, but one that better fits our needs.

- We begin by considering how the chain rule looks when we switch to matrix form.
- We give a general idea of what matrix derivatives look like.
- We list some of the results that matrix calculus gives us, for particular derivatives.
- We actually reason about how matrix calculus *works*.

The last of these is by far the **hardest**, and warrants its own section. Nevertheless, even without it, you can more or less get the idea of what we need - hence why we're going in reversed order.

6.5.15 The chain rule: Matrix form

Let's start with the first: the punchline, how does the chain rule and our gradient descent **change** when we add **matrices**?

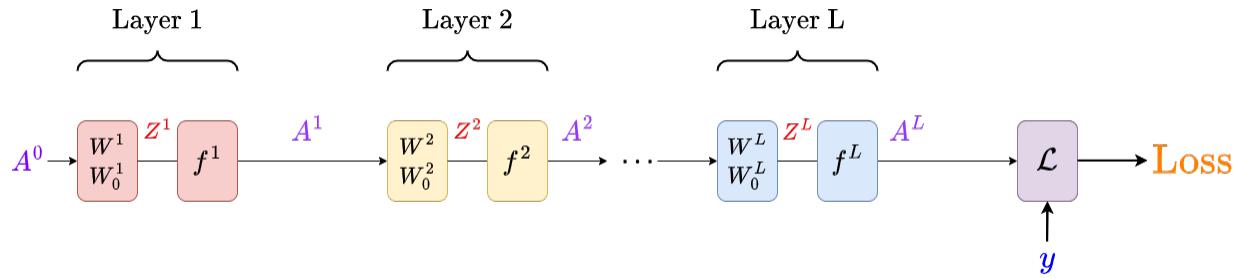
It turns out, not much: by using **layers** in the last section, we were able to create a pretty powerful and mathematically **tidy** object.

- With layers, each layer feeds into the **next**, with no other interaction. And neurons **within** the same layer do **not** directly **interact** with each other, which simplifies our math greatly.
 - Basically, we have a bunch of functions (neurons) that, within a layer, have **nothing** to do with each other, and only **output** to the **next** layer of similar functions.
- So, we can often **oversimplify** our model by thinking of each **layer** as like a "big" function, taking in a vector of size m^l and outputting a vector of size n^l .

Our main concern is making sure we have agreement of **dimensions!**

So, here's how our model looks now:

In fact, if you just rearranging your matrices and transposing them can be a helpful way to debug. Be careful, though!



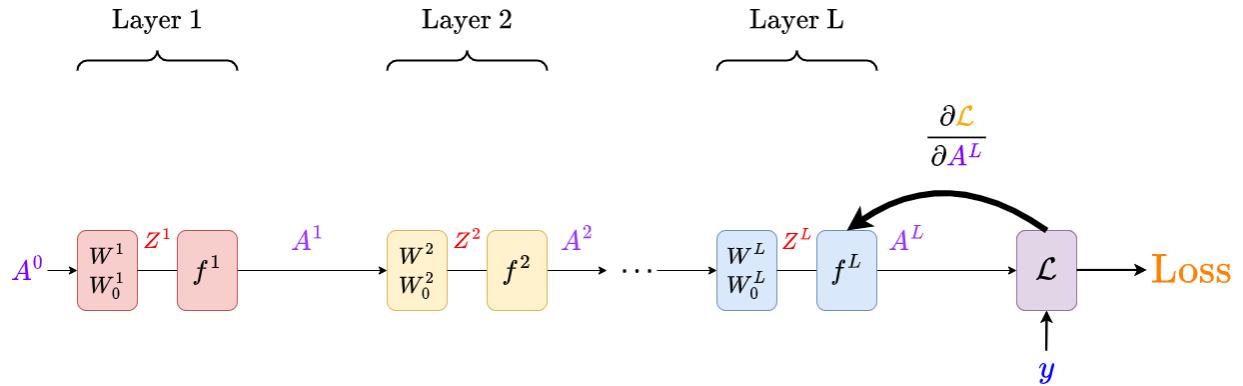
Pretty much the same! Only major difference: swapped scalars for vectors, and vectors for matrices (represented by switching to uppercase)

And, we do backprop the same way, too.

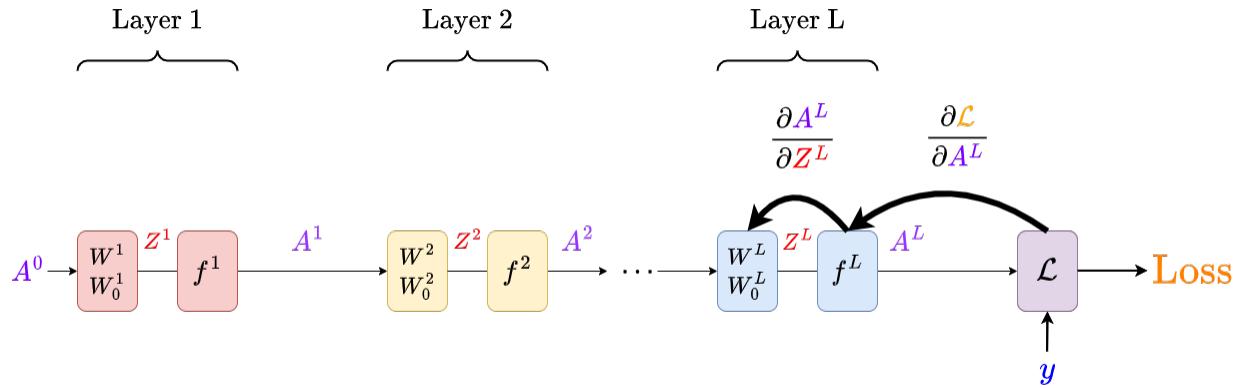
Here, we're not going to explain much as we go: all we're doing is getting the **derivatives** we need for our **chain rule!**

As we go **backwards**, we can build the gradient for each **weight** we come across, in the way we described above.

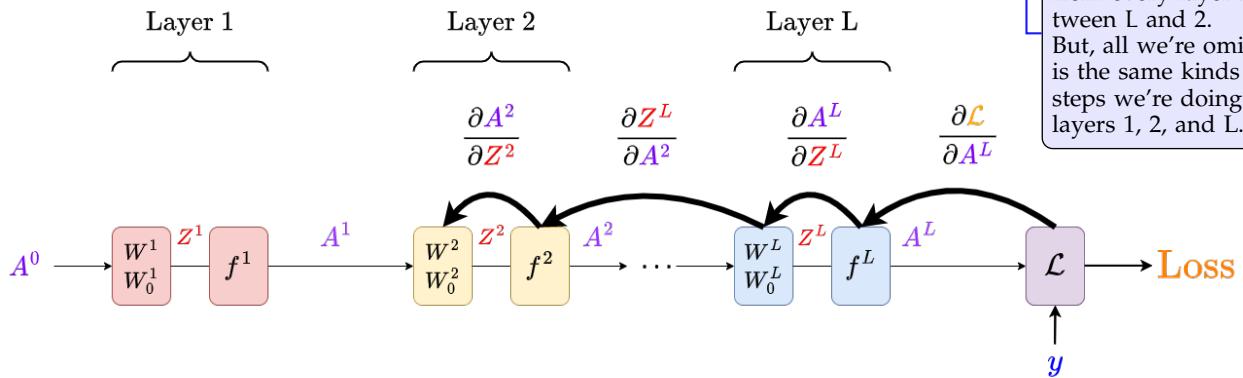
As always, we start from the loss function:



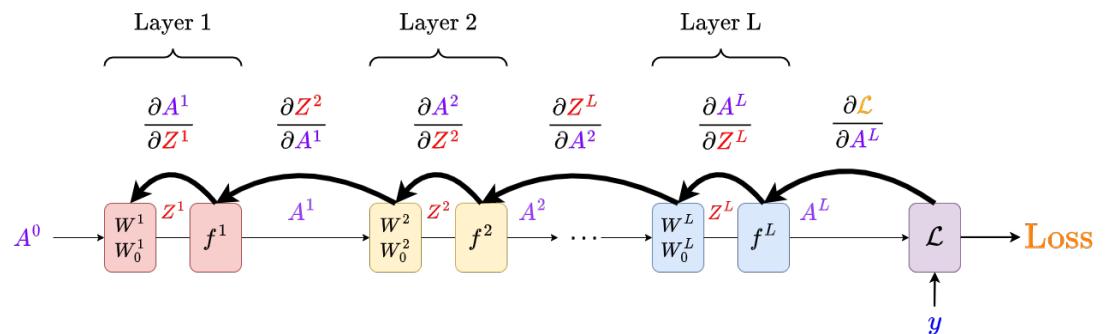
Take another step:



We'll pick up the pace: we'll jump to layer 2 and get its gradient.



Now, we finally get to layer 1!



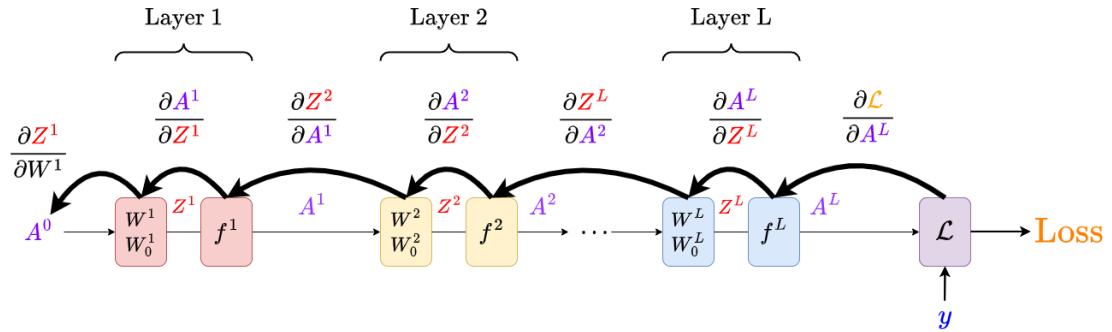
We finish off by getting what we're after: the gradient for W^1 .

Notation 301

We depict neural network gradient descent using the below diagram (outside the box):

The **right-facing straight** arrows come **first**: they're part of the **forward pass**, where we get all of our values.

The **left-facing curved** arrows come **after**: they represent the **back-propagation** of the gradient.



And, with this, we can rewrite our general equation for neural network gradients.

6.5.16 How the Chain Rule changes in Matrix form

As we discussed before, we can't just add onto our weight gradient to reach another layer: the final term

$$\frac{\partial Z^\ell}{\partial W^\ell} \quad (6.45)$$

Ends our chain rule when we add it: W^ℓ isn't part of the input or output, so it doesn't connect to the previous layer.

So, for this section, we'll add it **separately** at the end of our chain rule:

$$\frac{\partial \mathcal{L}}{\partial W^\ell} = \underbrace{\frac{\partial \mathcal{L}}{\partial Z^\ell}}_{\text{Weight link}} \cdot \underbrace{\left(\frac{\partial \mathcal{L}}{\partial Z^\ell} \right)^\top}_{\text{Other layers}}$$

That way, we can add onto $\partial \mathcal{L} / \partial Z^\ell$ without worrying about the weight derivative.

Notice two minor changes caused by the switch to matrices:

- The order has to be **reversed**.

- We also have to do some weird **transposing**.

Both of these mostly boil down to trying to be careful about **shape**/dimension agreement.

Notation 302

The **gradient** $\nabla_{W^\ell} \mathcal{L}$ for a neural network is given as:

$$\frac{\partial \mathcal{L}}{\partial W^\ell} = \overbrace{\frac{\partial \mathcal{L}}{\partial Z^\ell}}^{\text{Weight link}} \cdot \overbrace{\left(\frac{\partial \mathcal{L}}{\partial Z^\ell} \right)^T}^{\text{Other layers}}$$

There are also deeper interpretations, but they aren't worth digging into for now.

We get our remaining terms $\partial \mathcal{L} / \partial Z^\ell$ by our usual chain rule:

$$\frac{\partial \mathcal{L}}{\partial Z^\ell} = \overbrace{\left(\frac{\partial A^\ell}{\partial Z^\ell} \right)}^{\text{Layer } \ell} \cdot \left(\dots \right) \cdot \overbrace{\left(\frac{\partial Z^{L-1}}{\partial A^{L-2}} \cdot \frac{\partial A^{L-1}}{\partial Z^{L-1}} \right)}^{\text{Layer } L-1} \cdot \overbrace{\left(\frac{\partial Z^L}{\partial A^{L-1}} \cdot \frac{\partial A^L}{\partial Z^L} \right)}^{\text{Layer } L} \cdot \overbrace{\left(\frac{\partial \mathcal{L}}{\partial A^L} \right)}^{\text{Loss unit}}$$

This is likely our most important equation in this chapter!

6.5.17 Relevant Derivatives

If you aren't interesting in understanding matrix derivatives, here we provide the general format of each of the derivatives we care about.

Notation 303

Here, we give useful **derivatives** for **neural network gradient descent**.

Loss is not given, so we can't compute it, as before:

$$\overbrace{\frac{\partial \mathcal{L}}{\partial \mathbf{A}^L}}^{(n^L \times 1)}$$

We get the same result for each of these terms as we did before, except in matrix form.

$$\overbrace{\frac{\partial \mathbf{Z}^\ell}{\partial W^\ell}}^{(m^\ell \times 1)} = \mathbf{A}^{\ell-1}$$

$$\overbrace{\frac{\partial \mathbf{Z}^\ell}{\partial \mathbf{A}^{\ell-1}}}^{(m^\ell \times n^\ell)} = W^\ell$$

The last one is actually pretty different from before:

$$\overbrace{\frac{\partial \mathbf{A}^\ell}{\partial \mathbf{Z}^\ell}}^{(n^\ell \times n^\ell)} = \begin{bmatrix} f'(z_1^\ell) & 0 & 0 & \cdots & 0 \\ 0 & f'(z_2^\ell) & 0 & \cdots & 0 \\ 0 & 0 & f'(z_3^\ell) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & f'(z_r^\ell) \end{bmatrix}$$

Where r is the length of Z^ℓ .

- In short, we only have the z_i derivative on the i^{th} diagonal
- Why? Check the **matrix derivative explanatory notes**.

Example: Suppose you have the activation $f(z) = z^2$.

Your pre-activation might be

$$z^\ell = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad (6.46)$$

The output would be

For $\partial \mathbf{Z}^\ell / \partial W^\ell$, check section A.9.2.

For $\partial \mathbf{Z}^\ell / \partial \mathbf{A}^{\ell-1}$, check section A.9.3.

For $\partial \mathbf{A}^\ell / \partial \mathbf{Z}^\ell$, check section A.9.4.

$$\mathbf{a}^{\ell} = f(\mathbf{z}^{\ell}) = \begin{bmatrix} 1 \\ 2^2 \\ 3^2 \end{bmatrix} \quad (6.47)$$

But the derivative would be:

$$f(z) = 2z \quad (6.48)$$

Which, gives our matrix derivative as:

$$\frac{\partial \mathbf{a}^{\ell}}{\partial \mathbf{z}^{\ell}} = \begin{bmatrix} f'(1) & 0 & 0 \\ 0 & f'(2) & 0 \\ 0 & 0 & f'(3) \end{bmatrix} = \begin{bmatrix} 2 \cdot 1 & 0 & 0 \\ 0 & 2 \cdot 2 & 0 \\ 0 & 0 & 2 \cdot 3 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 6 \end{bmatrix}$$

If you want to be able to **derive** some of the derivatives, without reading the matrix derivative section, just use this formula for vector derivatives:

If you have time, do read – you won't understand what you're doing otherwise!

$$\frac{\partial \mathbf{w}}{\partial \mathbf{v}} = \left\{ \begin{array}{cccc} \frac{\partial w_1}{\partial v_1} & \frac{\partial w_1}{\partial v_2} & \dots & \frac{\partial w_1}{\partial v_m} \\ \frac{\partial w_2}{\partial v_1} & \frac{\partial w_2}{\partial v_2} & \dots & \frac{\partial w_2}{\partial v_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial w_n}{\partial v_1} & \frac{\partial w_n}{\partial v_2} & \dots & \frac{\partial w_n}{\partial v_m} \end{array} \right\} \quad (6.49)$$

Column j matches w_j

Row i matches v_i

We can use this for scalars as well: we just treat them as a vector of length 1.

With some cleverness, you can derive the Scalar/Matrix and Matrix/Scalar derivatives as well.

This is contained in the matrix derivatives chapter.

Clarification 304

Note that we have chosen a **convention** for how our matrices work: plenty of other resources use a transposed version of matrix derivatives.

This alternate version means the exact **same** thing as our version. Our choice is called the **denominator layout notation** for matrix derivatives.

6.6 Training

6.6.1 Comments

A few important side notes on training. First, on derivatives:

Concept 305

Sometimes, depending on your **loss** and **activation** function, it may be easier to directly compute

$$\frac{\partial \mathcal{L}}{\partial Z^L}$$

Than it is to find

$$\partial \mathcal{L} / \partial A^L \text{ and } \partial A^L / \partial Z^L$$

So, our algorithm may change slightly.

Another thought: initialization.

Concept 306

We typically try to pick a **random initialization**. This does two things:

- Allows us to avoid weird **numerical** and **symmetry** issues that happen when we start with $W_{ij} = 0$.
- We can hopefully find different **local minima** if we run our algorithm multiple times.
 - This is also helped by picking **random data points** in **SGD** (our typical algorithm).

Here, we choose our **initialization** from a **Gaussian** distribution, if you know what that is.

6.6.2 Pseudocode

Our training algorithm for backprop can follow smoothly from what we've laid out.

Here, we'll use the @ symbol to indicate matrix multiplication, following numpy conventions.

If you do not know a gaussian distribution, that shouldn't be a problem. It is also known as a "normal" distribution.

SGD-NEURAL-NET($\mathcal{D}_n, T, L, (m^1, \dots, m^L), (f^1, \dots, f^L), \text{Loss}$)

```

1  for every layer:
2      Randomly initialize
3          the weights in every layer
4          the biases in every layer
5
6  While termination condition not met:
7      Get random data point i
8      Keep track of time t
9
10     Do forward pass
11         for every layer:
12             Use previous layer's output: get pre-activation
13             Use pre-activation: get new output, activation
14
15     Get loss: forward pass complete
16
17     Do back-propagation
18         for every layer in reversed order:
19             If final layer: #Loss function
20                 Get  $\partial \mathcal{L} / \partial A^L$ 
21
22             Else:
23                 Get  $\partial \mathcal{L} / \partial A^\ell$ : #Link two layers
24                  $(\partial Z^{\ell+1} / \partial A^\ell) @ (\partial \mathcal{L} / \partial Z^{\ell+1})$ 
25
26                 Get  $\partial \mathcal{L} / \partial Z^\ell$ : #Within layer
27                  $(\partial A^\ell / \partial Z^\ell) @ (\partial \mathcal{L} / \partial A^\ell)$ 
28
29         Compute weight gradients:
30             Get  $\partial \mathcal{L} / \partial W^\ell$ : #Weights
31                  $\partial Z^\ell / \partial W^\ell = A^{\ell-1}$ 
32                  $(\partial Z^\ell / \partial W^\ell) @ (\partial \mathcal{L} / \partial Z^\ell)$ 
33
34             Get  $\partial \mathcal{L} / \partial W_0^\ell$ : #Biases
35                  $\partial \mathcal{L} / \partial W_0^\ell = (\partial \mathcal{L} / \partial Z^\ell)$ 
36
37     Follow Stochastic Gradient Descend (SGD): #Take step
38         Update weights:
39              $W^\ell = W^\ell - (\eta(t) * (\partial \mathcal{L} / \partial W^\ell))$ 
40
41         Update biases:
42              $W_0^\ell = W_0^\ell - (\eta(t) * (\partial \mathcal{L} / \partial W_0^\ell))$ 
43
44 Return final neural network with weights and biases  Last Updated: 09/03/24 03:53:41

```

6.7 Optimizing neural network parameters

We now understand both how neural networks work, and how to **train** them. We can use gradient descent to **optimize** their parameters.

But, we can do **better** than a simple SGD approach with step size $\eta(t)$. We'll try out some **modifications** that can speed up our training, and make better models.

6.7.1 Mini-batch

6.7.1.1 Review: Gradient Descent Notation

Let's review some gradient descent notation. We want to **optimize** our objective function J using W .

We do this using the gradient. This gradient depends on our current weights at time t , W_t .

$$\overbrace{\nabla_W J}^{\text{General Gradient}} \longrightarrow \overbrace{\nabla_W J(W_t)}^{\text{Gradient at time } t} \quad (6.50)$$

Our update rule is:

$$W_{\text{new}} = W_{\text{old}} - \eta \overbrace{(\nabla_W J(W_{\text{old}}))}^{\text{Gradient}} \quad (6.51)$$

Or, using timestep t :

$$W_{t+1} = W_t - \eta \overbrace{(\nabla_W J(W_t))}^{\text{Gradient}} \quad (6.52)$$

What is our objective function J ? Without regularization, it's based on our **loss** function. We can get loss for each of our data points: _____

$$\mathcal{L}^{(i)} = \overbrace{\mathcal{L}(g^{(i)}, y^{(i)})}^{\text{Loss for data point } i} \quad (6.53)$$

We won't define J here, because it is slightly different for SGD and BGD. We'll get to that below.

Our guess $g^{(i)}$ depends on both our current data point $x^{(i)}$, and the current weights W_t :

$$\mathcal{L}^{(i)}(W_t) = \mathcal{L}(\overbrace{h(x^{(i)}; W_t)}^{g^{(i)}}, y^{(i)}) \quad (6.54)$$

6.7.1.2 Review: BGD vs. SGD

Let's review our two main types of gradient descent, using the equation

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta \overbrace{\nabla_{\mathbf{W}} J(\mathbf{W}_t)}^{\text{Gradient}} \quad (6.55)$$

First, we have **batch gradient descent**, where we use our **whole** training set each time we take a step.

Definition 307

Batch Gradient Descent (BGD) is a form of gradient descent where we get the **gradient** of our loss function using **all of our training data**.

$$\nabla_{\mathbf{W}} J(\mathbf{W}_t) = \sum_{i=1}^n \overbrace{\nabla_{\mathbf{W}} (\mathcal{L}^{(i)}(\mathbf{W}_t))}^{\text{Each data point}}$$

We get the gradient for each data point, and then **add** all of those gradients up. We use this **combined gradient** to take **one step**.

We **repeat** this process every time we want to take a new step.

Then, we have **stochastic gradient descent**, where we use only **one** data point for each step we take.

Definition 308

Stochastic Gradient Descent (SGD) is a form of gradient descent where we get the **gradient** of our loss function using **one data point at a time**.

$$\nabla_{\mathbf{W}} J(\mathbf{W}_t) = \overbrace{\nabla_{\mathbf{W}} (\mathcal{L}^{(i)}(\mathbf{W}_t))}^{\text{One data point}}$$

We **randomly** choose one data point $(x^{(i)}, y^{(i)})$ and get the **gradient**. Based on this one gradient, we take our **step**.

For each step, we choose a new **random** data point.

These two approaches have tradeoffs:

Concept 309

There are **tradeoffs** between **SGD** and **BGD**:

- Each step is **faster** in **SGD**: we only use one data point.
 - Meanwhile, **BGD** is **slower**: each step uses all of our data.
 - **SGD** could improve a lot with only a **small subset** of a data.
- Because **BGD** uses all our data, its gradient is much more **accurate**.
 - **SGD** often uses **smaller** steps: the gradient is less accurate, with less data.
 - This is worse if the data is **noisy**: each SGD step becomes less effective.
- **SGD randomly** chooses data points: this random noise makes it harder to over-fit.
 - **BGD** uses all of the data, so we don't reduce overfitting.

6.7.1.3 Mini-batch

Rather than picking one or the other, one might think, "why do we have to pick **every** data point or **one** data point? Couldn't we pick only a **few**?"

This is the premise of **mini-batch**: instead of making a batch out of the entire training set, we **randomly** select a few data points, and use that as our batch.

Definition 310

Mini-batch is a way to **compromise** between SGD and BGD.

To create a mini-batch, we **randomly** select K data points from our training data.

We treat this mini-batch the same way we would a regular **batch**: get the **gradient** of each data point, **add** those gradients, and take one step of gradient descent.

$$\nabla_{\mathbf{W}} J(\mathbf{W}_t) = \underbrace{\sum_{i=1}^K}_{K \text{ data points in a mini-batch}} \nabla_{\mathbf{W}} (\mathcal{L}^{(i)}(\mathbf{W}_t))$$

We gather a **new** mini-batch for each step we want to take.

Mini-batch is the **default** used in most modern packages: it gives us more **control** over our algorithm, and can often find the **best** of both worlds.

We do have to be careful to randomly select data in an efficient way, though. Packages usually take care of this.

Concept 311

Mini-batch has a lot of benefits of both SGD and BGD:

- Steps are **faster** than BGD: we only need to get the gradient for K points.
 - The **speed** no longer depends on the total training data size (more data, more gradients): instead, it depends on our **batch size** K.
- Steps are more **accurate** than SGD: with more data, we have a better **gradient**.
 - This means we can take **bigger** steps.
- Our batches are **random**, like SGD: we reduce overfitting and escape local minima.

One more important benefit:

- If we find that a particular problem is better suited for something closer to BGD or SGD, we can **adjust** our batch size K.
 - This gives us more **control** over our learning algorithm.

6.7.2 Adaptive Step Size - Challenges

We'll stop discussing mini-batches, and the SGD vs. BGD problem. Instead, let's improve our **step size**.

Step size η is a difficult problem:

- If η is **small**, then our training can take a long **time**.
- If η is too **large**, we might **diverge**: our answer gets way too large.
- A **large** step size might also cause **oscillation**: most of our step is wasted going back and forth, so we go **slowly** again.

SGD and mini-batch have a step size-related problem, too:

- In order to **converge** according to our theorems (see chapter 3), the step size $\eta(t)$ has to be **decreasing** in a certain way.

Check chapter 3 for the exact requirements of the theorem.

In Appendix B, we discuss some common techniques:

- Momentum
- Adadelta
- Adam

6.7.3 Vanishing/Exploding Gradient

Now, neural networks have one more **problem**, that we've ignored so far: **deep** neural networks can cause a problem called "**exploding/vanishing gradient**".

Here's an example: suppose you have a long chain rule, with 8 terms. Our chain rule gets **longer** with more layers, because each layer needs its own derivatives.

By "deep", we just mean "many layers".

$$\frac{\partial A}{\partial H} = \frac{\partial A}{\partial B} \cdot \frac{\partial B}{\partial C} \cdot \left(\dots \right) \cdot \frac{\partial G}{\partial H} \quad (6.56)$$

This chain rule gets **longer** as we move "**backwards**" through our network, so the chain rule is longest for the "**early**" layers: $\ell = 1, 2$, and so on.

Suppose all of our derivatives are roughly $.1$. What happens when we multiply them **together**?

$$\frac{\partial A}{\partial H} = .1 \cdot .1 \cdot \left(\dots \right) \cdot .1 = 10^{-8} \quad (6.57)$$

The derivative becomes really, really **tiny**! This is the case of the **vanishing** gradient: if our gradients are less than one, then as we append more layers, they multiply to get smaller and smaller.

- This is a problem: if our gradients in our earlier layers become too **small**, we'll never make any progress! They'll hardly change.

Definition 312

Vanishing gradient occurs when a deep neural network ends up with **very small gradients** in the **earlier** layers.

This happens because a deeper neural network has a **longer chain rule**: if all of the terms are **less than one**, they'll multiply into a very small value, "**vanishing**".

This means that our gradient descent will have **almost no effect** on these earlier weights, **slowing down** our algorithm considerably.

What if the gradients are larger than 1 ? Let's say our derivatives are 10 each.

$$\frac{\partial A}{\partial H} = 10 \cdot 10 \cdot \left(\dots \right) \cdot 10 = 10^8 \quad (6.58)$$

Now, the early derivatives are becoming **huge**! This is the case of **exploding** gradient: if our gradients are greater than one, then as we add layers, they multiply to get bigger.

- This is also a problem: we don't want to take **huge** steps, or we will **diverge**, or **oscillate**, and jump huge distances across the **hypothesis space**.

Definition 313

Exploding gradient occur when a deep neural network ends up with **very large gradients** in the **earlier** layers.

This happens because a deeper neural network has a **longer chain rule**: if all of the terms are much **greater than one**, they'll multiply into a very large value, "**exploding**".

This means that our gradient descent will take **huge steps** in the hypothesis space. This can cause us to **diverge**, miss local minima, or **oscillate**.

So, to avoid this, we can't just blindly multiply our gradients and keep a fixed step size.

The solution? Each **weight** gets its own step size η .

Concept 314

In order to avoid **vanishing/exploding** gradient problems, we give each **weight** in our network its own **step size** η .

This allows us to **adjust** the step size for some weights more than others: if our gradient is too large or small, we can fix it.

6.8 Regularization

Something we haven't discussed in a while, that we might investigate, is **regularization**: techniques against overfitting.

- We teach our model using "training data", which, by chance, may not perfectly reflect the **true** distribution.

We want to apply this to our modern, deep neural networks, with their huge number of **parameters**, and huge amount of **data**. And yet...

These modern neural nets **don't** tend to have as much problem with **overfitting**, and we're not sure **why**!

Regardless, we do still have *some* methods for regularization: this will be our focus for the rest of this chapter.

6.8.1 Methods related to ridge regression

We'll start with methods that we can bring back from classic ridge regression.

6.8.1.1 Early Stopping

One component built into our learning algorithm for regression is **early stopping**: we check if the model is still making **progress**. If it isn't, then we **stop** training.

- The longer we train our model, the more time it has to "**memorize**" the exact structure of our training data: including **noise**.

We would typically either measure the size of the **gradient**, or the change in the **loss**. If either was small, then we might be in a local minimum: we've finished training.

So, we do the same here: after a period of training (over the whole dataset, called an **epoch**), we measure the **loss** on a validation set.

- If the loss stops decreasing, or begins **increasing**, our model is probably **overfitting**. We **stop early**.

Then, you return the weights with the lowest error.

Definition 315

An **epoch** is the time frame during which your model sees your whole **training data** set, **once**.

- Note that sometimes, "epoch" just refers to how long you train before you check your loss.
- In this case, it might be smaller than the whole dataset.

Definition 316

With **early stopping**, you evaluate your model using your **validation data**, computing the loss.

- If the loss has decreased from the last epoch, you **continue** training.
- If the loss has stopped decreasing, or is increasing, you **stop** training.

This continues until you've either run out of **epochs**, or you've met your **termination** condition, and stopped early.

6.8.1.2 Weight Decay

We can also use the same kind of regularization term that we used for linear regression: penalizing the **squared magnitude**.

- Starting with our loss function:

$$J_{\text{loss}} = \sum_{i=1}^n \mathcal{L}(\text{NN}(x^{(i)}), y^{(i)}; W) \quad (6.59)$$

- And we penalize based on the square magnitude of our weights:

$$J = \lambda \|W\|^2 + J_{\text{loss}} \quad (6.60)$$

If we take the gradient, we get:

$$\nabla_W J = 2\lambda W + \nabla_W J_{\text{loss}} \quad (6.61)$$

Let's see how the regularization affects our step:

$$W_t = W_{t-1} - \eta \left(2\lambda \|W_{t-1}\| + \nabla_W J_{\text{loss}} \right) \quad (6.62)$$

It directly subtracts from our weight, **decaying** it.

$$W_t = W_{t-1} \left(1 - 2\lambda\eta \right) - \eta \nabla_W J_{\text{loss}} \quad (6.63)$$

Thus, we call it **weight decay**.

Concept 317

When we apply **square magnitude** regularization to the weights of a neural network, we call it **weight decay**.

$$J_{\text{reg}} = J_{\text{loss}} + \lambda \|W\|^2$$

That's because, when you take the gradient, it directly subtracts from weight W , causing it to **decay** by a factor of $(1 - 2\lambda\eta)$.

$$W_t = W_{t-1} \left(1 - 2\lambda\eta \right) - \eta \nabla_W J_{\text{loss}}$$

6.8.1.3 Perturbation

One last way to reduce overfitting is to add some **random noise** to our data: each variable has a small, random number added to it.

This value is typically **zero-mean** and **normally distributed**:

- Zero-mean: it has 0 effect, on average, so it doesn't bias the data high or low.
- Normally distributed: the noise is **symmetric**: +2 and -2 are equally likely.

How large the noise is, depends on the chosen **variance**.

This reduces overfitting, because if the data is slightly different each time you see it, it's harder to perfectly "memorize" the exact shape and structure.

The "normal distribution" contains more information than that, but the symmetry is important.

Definition 318

Perturbation is a technique where you slightly modify your system.

- In our case, we're **randomly** adding small amount of **noise** to our input data.

This makes it more difficult for the model to **overfit**, because the patterns aren't always exactly the same.

- Only the "general", high-level patterns are preserved, each time you view the dataset.

Of course, if you perturb your data too strongly, you can miss real patterns. Your perturbations shouldn't be too large.

6.8.2 Dropout

We also have **structural** ways of dealing with overfitting. We discussed perturbing the **dataset**, but we could, instead, perturb the **model** itself!

We do this by randomly **removing** some weights from the neural network, and training.

- Each weight has a probability p of being "turned off": the **activation** is set to zero.

$$a_j^\ell = 0 \quad (6.64)$$

- Thus, that neuron's output is **ignored** by the next layer, and receives no training.
- At the next step, we remove a **different** random selection of weights.

Because our model keeps changing slightly, it's harder to exactly **overfit** to the data.

This particular approach also addresses a **different** kind of overfitting:

- One model might heavily "rely" on a **small** number of neurons to make decisions.
- This makes our model less flexible, uses the weights less efficiently.

To solve this, we prevent the neural network from using some of these weights, **randomly**.

Thus, the whole network "**shares**" some responsibility for getting the right answer.

Definition 319

Dropout is a process where, at each training interval, you **randomly** "drop out", or de-activate, some of the weights in the network.

- Each neuron has probability p of being turned off.
- These neurons have their **activation** set to zero: $a_j^\ell = 0$.

This process is designed to reduce **overfitting**. As the network randomly changes, it's harder for it to perfectly match the data structure.

This process is also designed to create "collective responsibility" for your neurons. It prevents your network from relying on a small number of neurons to solve problems.

It generally improves **robustness** against random variations in the data.

Clarification 320

When a network using dropout is finished training, we multiply all the weights by p . Why?

- Because during training, only p fraction of the neurons were active.
- We want to replicate that average activity level, even when we use all of our neurons.

This approach has, in recent years, become somewhat less popular, for a couple reasons:

- Very **large** networks don't struggle as much with overfitting.
- CNNs tend not to benefit from this procedure, because of **weight-sharing**: the same weights are re-used in multiple places.
- Like most ML techniques, their usefulness depends on the situation.

We'll discuss CNNs in our next chapter.

It still finds use in some smaller models, RNNs, etc.

In many places, it has been replaced by **batch normalization**.

6.8.3 Batch Normalization

6.8.3.1 Covariate Shift

Our last approach related to regularization was designed to handle a new type of problem we call **covariate shift**:

When you run **gradient descent** on a neural network, you're adjusting the weights of all of our layers, at the **same time**.

Let's focus on layer 1 and layer 2. By updating layer 1, we've changed the outputs it creates: the same $x^{(i)}$ now creates a different output, going from a_{old}^1 to a_{new}^1 .

But, this output is the **input** of layer 2.

- This means that layer 2 is now receiving **different inputs** than it was before.
- This is a problem: layer 2 just received training based on the **old** inputs a_{old}^1 !

This makes life a lot harder for our layer 2: not only is it learning to make better predictions, it also has to adjust for the change in layer 1. This is a form of **covariate shift**.

Definition 321

Covariate shift occurs when the distribution of input variables **changes** over time.

- This can cause our original model to become inaccurate, or "outdated": it was trained on **different** data.

Internal covariate shift occurs because of changes to the network itself.

- The distribution of inputs to **later layers** changes, because **earlier layers** have changed through training.

Example: If the weights in layer 1 all get smaller (as a side effect of correcting them), layer 2 may have to make all of its weights bigger to compensate.

- Our expectation is that this extra work would slow down training.

6.8.3.2 Layer Normalization

So, we want to counter the problem of how the input to layer 2 **changes** based on layer 1's **learning**.

- But, at the same time, we don't want to **undo** the work that we did **training** layer 1.
- So, we want to preserve the information in layer 1, while making it easier to use.

In our example, we mentioned that the scale of the inputs might get larger: they all get bigger or smaller, at the same time.

But, we often want to know what makes these inputs **different** from each other, so we can compare them: it's not helpful if all of them become larger/smaller.

So, we'll **standardize** each of our mini-batches of data, between each layer: _____

This is exactly the same as when we standardized in the Feature Transformation chapter.

- **Subtracting the mean:** we take the mean of the mini-batch input, and subtract it from each data point. So, our standardized data is always centered at 0.
- **Dividing by standard deviation:** we compute how "spread out" the data is, and scale by that factor. Our standardized data is always the same amount "spread out".

We repeat this process in between each layer of our network. Each layer receives a set of inputs with mean 0, standard deviation 1, no matter how the earlier layers change. _____

For now, we'll apply this to the pre-activation Z.

Below, we exclude the ℓ superscript in z^ℓ , μ^ℓ , σ^ℓ , and n^ℓ for readability.

Notation 322

We'll represent the element in the i^{th} row, j^{th} column, as Z_{ij} .

Key Equation 323

*Here, we focus on a single element: dimension i of the j^{th} data point in our batch, z_{ij} .

When we **standardize**, we first compute the **mean** and **standard deviation**:

$$\mu_i = \frac{1}{K} \sum_{j=1}^K Z_{ij} \quad \sigma_i^2 = \frac{1}{K} \sum_{j=1}^K (Z_{ij} - \mu_i)^2$$

Once we've done that, we can properly standardize, creating data with mean 0, and sd 1.

- We include a very small ϵ term to avoid dividing by zero.

$$\bar{Z}_{ij} = \frac{Z_{ij} - \mu_i}{\sigma_i + \epsilon}$$

Concept 324

In order to deal with **internal covariate shift**, we'll take each mini-batch of k **pre-activations** to each **layer** of our neural network, and **standardize**/normalize it.

- Each **dimension** of the input is standardized **separately**.

After **standardizing**, this data is sent forward through the network.

- This is equivalent to using a "standardizing function" **after** the weight function $Z = W^T A$, and **before** the activation function $f(Z)$ (now $f(\bar{Z})$).
- We could also standardize **after** activation, but it's unclear which approach is better.

We can insert our new module into our existing model:



Normalization/standardization can be treated just like any other module.

Note that this preserves the **information** in our input data:

- If one data point has one feature much larger than another, you'll still see that: the gap will just be shifted over to zero, and normalized.

Example: Suppose you have some data: [1, 2, 100, 3, 4, 5]. If you standardize, you get

$$[-0.458, -0.433, 2.04, -0.408, -0.383, -0.358,] \quad (6.65)$$

The larger data point still stands out above the rest!

We need to be careful of dimensions:

Clarification 325

Normalization relies on the distribution (mean, s.d.) of our **mini-batch**.

But that means we can't just compute one data point at a time: we need to include the whole mini-batch of k **at the same time**.

So, we have to change the dimensions of Z^ℓ .

k is our **batch size**, while n is the number of **dimensions**.

- Z^ℓ without norm: $(n^\ell \times 1)$
- Z^ℓ with norm: $(n^\ell \times k)$

We use Z_{ij}^ℓ to indicate the i^{th} dimension of the j^{th} data point.

6.8.3.3 Post-Normalization: Choose Mean and S.D.

Now, we've adjusted it so that our distribution doesn't "drift", based on our training.

But, now, we've **restricted** our model:

- We don't necessarily want our mean and standard deviation to be 0 and 1: it would be better to be able to **control** it.

To accomplish this, we'll **scale** and **shift** our input. Thus, we're choosing our mean/s.d. in a deliberate way.

- Each dimension needs its own mean and standard deviation.
- We have n dimensions, and we need one variable to handle mean (or s.d.) for each: we'll need an $(n \times 1)$ vector.

Concept 326

By doing **normalization**, we've transformed Z into \bar{Z} .

- This "resets" our mean and standard deviation to 0 and 1.

However, we want to be able to **control** our mean and standard deviation. To do so, we introduce two new parameters:

- G : An $(n \times 1)$ vector that **scales** each of our dimensions \bar{Z}_i , giving our **standard deviations**.
- B : An $(n \times 1)$ vector that **shifts** each of our dimensions \bar{Z}_i , giving our **means**.

Thus, we get the true output of **batch normalization**:

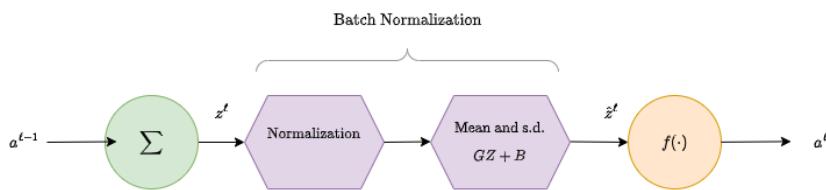
$$\hat{Z}_{ik} = G_i * \bar{Z}_{ij} + B_i$$

Example: Here's a sample example using a vector \bar{z}_j : only considering one, post-normalization data point j .

$$\begin{bmatrix} \hat{Z}_{1j} \\ \hat{Z}_{2j} \\ \vdots \\ \hat{Z}_{kj} \end{bmatrix} = \begin{bmatrix} G_1 \\ G_2 \\ \vdots \\ G_k \end{bmatrix} * \begin{bmatrix} \bar{Z}_{1j} \\ \bar{Z}_{2j} \\ \vdots \\ \bar{Z}_{kj} \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_k \end{bmatrix} \quad (6.66)$$

Where $*$ indicates element-wise multiplication.

If we include this in our neuron graph, we now have two new modules:



6.8.3.4 Full definition

Definition 327

Batch Normalization is a process where we

- Standardize the pre-activation for each layer using the mean μ_i and standard deviation σ_i (for the i^{th} dimension). Select $\epsilon: 0 < \epsilon \ll 1$.

$$\bar{Z}_{ij} = \frac{Z_{ij} - \mu_i}{\sigma_i + \epsilon}$$

- Choose the new mean and standard deviation for the pre-activation using $(n \times 1)$ vectors G and B

$$\hat{Z}_{ik} = G_i * \bar{Z}_{ij} + B_i$$

Concept 328

Batch Normalization is meant to accomplish the following:

- Remove possible **internal covariance shift**: training earlier layers may change the scale of inputs to later layers.
 - This could make training more difficult.
- Allow our model to **select** a particular mean and s.d. for its pre-activation values, rather than arriving at them by chance.

It also tends to have a regularizing effect, and, in some learning algorithms, has replaced dropout.

6.8.3.5 Effectively Perturbs Data

We're not actually sure why normalization helps our models train. We originally designed it for **internal covariate shift**, but we're not sure if that's really **why** it works.

One explanation might be that, due to random sampling, each mini-batch ends up slightly **modified** by our normalization.

- Since each batch is likely to have a slightly different mean/standard deviation, each one ends up differently "perturbed" by normalization.

6.8.3.6 Applying batch normalization to backprop

We defer discussion of backprop to Appendix B.

6.9 Terms

Section 7-1

- Neuron (Unit, Node)
- Neural Network
- Series and Parallel
- Linear Component
- Weight w
- Offset (Bias, Threshold) w_0
- Activation Function f
- Pre-activation z
- Activation a
- Identity Function
- Acyclic Networks
- Feed-forward Networks
- Layer
- Fully Connected
- Input dimension m
- Output dimension n
- Weight Matrix
- Offset Matrix
- Layer Notation A^ℓ
- Step function
- ReLU function
- Sigmoid function
- Hyperbolic tangent function
- Softmax function

Section 7-2

- Forward pass
- Back-Propagation
- Weight gradient
- Matrix Derivative
- Partial Derivative
- Multivariable Chain Rule
- Total Derivative
- Size of a matrix
- Planar Approximation
- Scalar/scalar derivative
- Vector/scalar derivative
- Scalar/vector derivative
- Vector/vector derivative
- Mini-batch
- Vanishing/Exploding Gradient
- Momentum (Optional)
- Adadelta (Optional)
- Adagrad (Optional)
- Adam (Optional)
- Normalization
- Early stopping (Review)
- Weight Decay
- Perturbation
- Dropout
- Covariate Shift
- Internal Covariate Shift
- Batch Normalization

- Multivariable Chain Rule (Review)

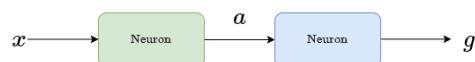
CHAPTER 7

Convolutional Neural Networks

7.0.1 Fully Connected Networks

Up to this point, we've focus on "**fully connected**" neural networks.

- "Connected" refers to the "connection" between neurons in **adjacent** layers: one neuron provides the input for another.



These two neurons are connected.

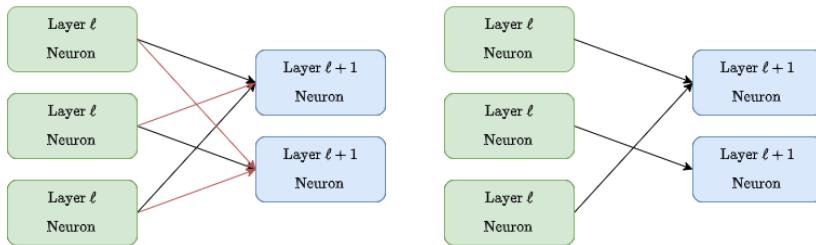
Thus, "fully connected" means that every possible connection between pairs of neurons **exists**.

Definition 329

A **fully connected** (FC) layer is one where every **input** neuron is connected to every **output** neuron.

The network layer only needs to be missing **one** connection between neurons to not be fully connected.

Example: We compare two networks:



The left network is fully connected: the right is not, having removed the red arrows.

7.0.2 The drawbacks of fully-connected networks

The "fully connected" approach includes a **weight** $w_{a,b}$ for every pair of input neuron a , and output neuron b .

- Each of these weights determines the **relationship** between our two neurons.
- In an FC settings, we're allowing for every possible pattern between pairs.

This is a very, very flexible model: any combination of patterns is possible.

Concept 330

Fully-connected networks are very useful when we **know very little** about how to **predict** our result.

By including so many possible connections and patterns, we're open to lots of **different models** we could try.

- This is especially helpful if we expect these relationships to be complex.

With a non-FC model, on the other hand, some connections have been severed. With this model, we're creating making some **assumptions** about which patterns **don't** exist.

- **Example:** If you think fact A is irrelevant for computing fact B, you wouldn't include it in the equation.
- This is similar to how you want to exclude inputs that won't help you predict your output.

Concept 331

Removing a connection in a neural network is equivalent to saying, "I don't think this variable a **should** affect this other variable b ".

This highlights a major **drawback** of fully connected networks: sometimes, it's inappropriate to allow for every possible connection.

Having connections we don't need can cause plenty of problems:

Concept 332

Fully-connected networks come with some problems:

- Having many parameters can risk **overfitting**,
- Our model takes **more time** to converge
 - Both because it has to train **more weights**,
 - And because our model can get "distracted" by dead-end possibilities, that a simpler model wouldn't consider.
- It's often difficult to interpret **how** our neural network comes to the conclusions it does.

In this chapter, we'll introduce some more specific problems, and one model type that allows us to overcome these problems: **Convolutional Neural Networks**.

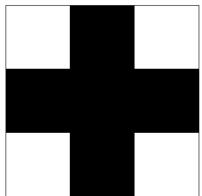
7.0.3 Intro to Image Processing

An excellent example for how FC neural networks can fail, is **image processing**.

Example: Facial recognition, self-driving vehicles, classifying the object in a picture

Let's give it a try: suppose we have an image. For simplicity, it's black and white. We'll need to represent the **brightness** of each pixel with a number.

- We'll use the most common range of values: the integers [0, 255].



$$\implies \begin{bmatrix} 255 & 0 & 255 \\ 0 & 0 & 0 \\ 255 & 0 & 255 \end{bmatrix}$$

Our machine stores the "picture" on the right.

Our neural network takes a single $d \times 1$ vector as its input. But right now, we have an $(r \times k)$ matrix. How do we solve that? With **flattening**.

Definition 333

Flattening is the process of taking a **matrix** of inputs, and transforming it into a single **vector**.

We usually do this by **concatenating** (combining consecutively) each row/column, in order.

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \rightarrow \begin{bmatrix} a \\ c \\ b \\ d \end{bmatrix}$$

If the input is an $(r \times k)$ matrix, the result is an $(rk \times 1)$ vector.

Example: We can apply this to our data above.

We'll represent it with the transpose to save space.

$$\begin{bmatrix} 255 & 0 & 255 \\ 0 & 0 & 0 \\ 255 & 0 & 255 \end{bmatrix} \rightarrow \begin{bmatrix} 255 & 0 & 255 & 0 & 0 & 0 & 255 & 0 & 255 \end{bmatrix}^T \quad (7.1)$$

And so transform from (3×3) to (9×1) .

Looking at the image, or even the **matrix**, it's relatively easy to see the "**cross**" pattern.

But, when we **flatten** it, those patterns immediately stop being obvious.

- We've lost information above which pixels are "**beside**" one another, for example.

Based on this, we'll find **two** main problems, that our CNNs will hopefully solve:

Remember that we only took the row vector for visualization: they're stacked vertically based on column!

7.0.4 Spatial Locality

First: as we just mentioned, we need an idea of which pixels are close **horizontally**.

- In fact, our network doesn't even care which pixels are **above** each other **vertically**:

Concept 334

Review from the Feature Representation chapter

The **order** we choose for elements in a vector **doesn't** affect the behavior of our model, so long as we **consistently** use that order.

This is because a linear model is a **sum**:

$$w^T a = \sum_i w_i a_i$$

And sums are the same, regardless of **order**.

$$a + b = b + a \implies w_1 a_1 + w_2 a_2 = w_2 a_2 + w_1 a_1$$

~~~~~

- We emphasize that the order does need to be **consistent** between data points.

**Example:** Suppose  $x_1$  represents height and  $x_2$  represents weight.

- We could do either [weight, height] or [height, weight]: it doesn't matter.

In other words, we could **shuffle** the order of our pixels, and as long as we shuffled it the **same way** for all of our training data, it wouldn't matter to our **model**.

BUT, we have to be **consistent** with which order we pick: otherwise, someone is measured as 180 feet tall, instead of 180 pounds.

- This was fine for the above example, but it doesn't make sense for **image processing**, where each dimension is just a **pixel**:

Human vision works differently: we look for shapes, which are often made up of pixels **near** each other: edges, points, corners, curves.

- In other words, we want to encode **local** information, across the physical **space** of our image. Thus, we call it **spatial locality**.

### Definition 335

**Spatial locality** is the knowledge of which objects are **close** in **space** to each other.

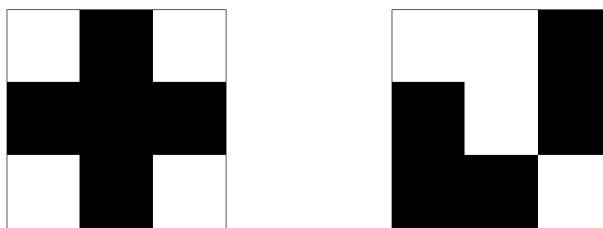
In an **image**, we might think of which pixels are "close" in that image.

- Which pixels are next to, or on top of each other? How far apart? How are they "arranged"?
- ~~~~~
- This is a property we want to build into the structure of our CNNs.

When we use "space" here, it's **different** from "latent space", or "input space".

Instead, we're talking about physical space: how physically **close** real things are, measured in **meters**, or in this case, **pixels**.

**Example:** If told that the white was blank space, and the black represented "objects", a human would have a concrete understanding of how these two images might be different:



The left image contains **one** object, while the right image contains **two** objects.

We figure this out based on which pixels are **toucning** or not: a spatial property.

- The neural network would struggle to encode anything like that.

### 7.0.5 Translation Invariance

A second problem is that, if the same **pattern** occupies different pixels, then it's completely new to the model.

- Example:** Suppose you have a cat on the left side of an image. You **move** it to the right side of the image.
- A person would consider that image "**almost the same**".
- But our FC NN does not: the cat is occupying a completely **different set of pixels**, which have a completely separate set of weights attached.

So, our NN can't find structures that are **similar** across different parts of the input.

Instead, we want a different behavior: we want our model to treat our input as the **same** (invariant), even if we move, or **translate** it.

- Thus, we're looking for **translation invariance**.

Not language translation: "translation" as in "moving around in space".

### Definition 336

**Translation invariance** is the property of treating patterns as the **same** even if we **translate** them in space.

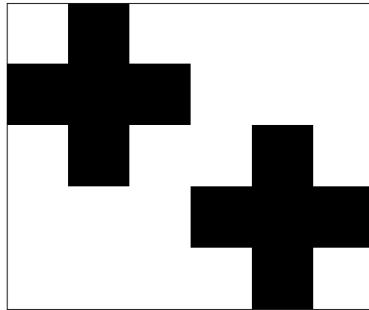
In an **image**, we might want to recognize the same **pattern** in two different **positions** on the image.

- In other words, the pattern has "translated" from one of those positions, to the other.



- This is a property we want to build into our CNN.

**Example:** In the following image, you would probably recognize "two crosses".



We have two of the **same** object: just **translated** over.

But, because the top left pixels have separate weights from the bottom right pixels, the NN will react differently to each.

Now that we've defined our problem, we can come up with a solution: **filters**.

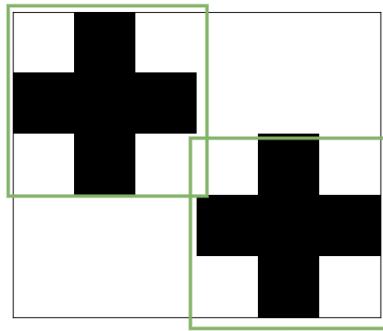
## 7.1 Filters

### 7.1.1 Motivating the Filter

So, we want a technique that handles both of these problems.

First, **translation invariance**: we want a calculation that can find the same pattern, in multiple locations.

- So, we'll apply the same calculation repeatedly, in **multiple positions** on our image.
- We'll **move** across our image, shifting to a new position each time we **scan** for that pattern.



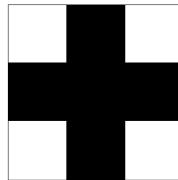
As we "scan over" our image, we'll hopefully find both of our crosses separately.

Next, **spatial locality**: we want this calculation to encode **spatial** information.

- As we "scan" across our image, each computation will look for a particular "shape", or "**pattern**" for our pixels.
- This pattern will be based on the **relative location** of each pixel.

So, we're looking for a tool that repeatedly shifts (or **translates**) across our image, and looks for a spatial **pattern** in the image.

- **Example:** Above, we would be looking for the "3x3 cross" pattern, and shift across rows/columns.



This is the shape we are looking for at each position.

The tool in question is "looking" for a pattern. Another way to see it, is that it's **filtering** out everything that doesn't match that pattern.

- Thus, we call it a **filter**.

#### Concept 337

**Filters** handle both the problems of **spatial locality** and **translation invariance** at the same time.

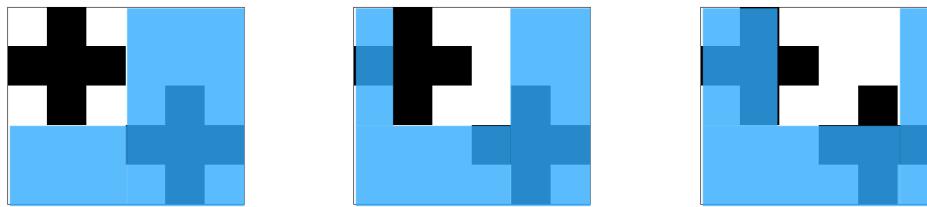
Notably, all of this works better if we keep our data in the **matrix** format, not the **flattened** one.

### 7.1.2 Windowing

We still need to figure out **how** we're going to find these patterns.

We've already established that our algorithm will look at a **local** region of the image, and search for the pattern.

To make life easier, we'll cut out a **piece** of the image, and only compare that to the pattern.



If we're viewing the top-left corner, we're ignoring everything else (blue-shaded). Then, we'll check the next position.

This region is the only part we "see", so we call it a **window**.

#### Definition 338

The **window**  $v$  is the region of our image that we are, at a given moment, looking for our **pattern** in.

This is the region we are applying our **filter** to.

The window has the **same dimensions** as our filter, so we can compare them directly.

- 
- As we continue filtering, we'll repeatedly move our filter, shifting it to every valid position on the image.

### 7.1.3 1-D case

To get going, we'll start with a 1D example.

- To make the math easier, we'll replace 0 and 255 with +1 and -1.

Suppose we're looking for "bright spots": pixels that are much brighter than their surroundings.

This isn't just a simplification: when processing sound data, it'll be in a 1D form.

We've decided to make dark pixels +1, and bright pixels -1. Which convention we choose isn't important: it's just more easily visible.

$$\begin{bmatrix} +1 & -1 & +1 \end{bmatrix}$$

So, we're looking for something like this.

How do we find "bright spots", like this? Well, we want to find regions which are **similar** to our pattern.

- Our sequence is a vector, so we want to **get the similarity between two vectors**.
- We have a tool for this! The **dot product**  $a \cdot b$ .

#### Concept 339

*Review from the Classification chapter*

You can use the **dot product** between non-unit vectors to measure their "similarity" **scaled by their magnitude**.

If two vectors are more **similar**, they have a **larger** dot product.

- If  $\text{angle} < 90^\circ$  they are "similar":  $\vec{a} \cdot \vec{b} > 0$
- If  $\text{angle} > 90^\circ$  they are "different":  $\vec{a} \cdot \vec{b} < 0$
- If they are **perpendicular** ( $\text{angle}=90^\circ$ ) to each other,  $\vec{a} \cdot \vec{b} = 0$

So: as an approximation, the higher the dot product, the more similar they are!

Now, we know what to do: we'll get the **dot product** between our window, and the **filter**, to see how similar they are.

- If they're similar enough, then we found the pattern!

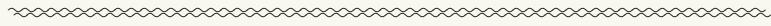
This works extra well for something like an image, where the pixels have a restricted "range" of values: it's not as easy to get an extra-high dot product just because the magnitude are too large.

**Concept 340**

To determine whether the window contains our pattern, we take the **dot product** between our **window v** and our **filter f**.

$$v \cdot f$$

The **higher** the dot product, the **more likely** that we have our pattern.

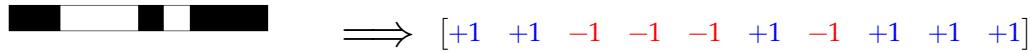


- There's no exactly dot product value to be "sure" you've found your pattern: you have to choose your threshold based on context.

We'll show our example below.

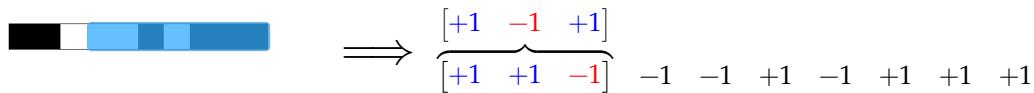
### 7.1.4 1D Example

So, suppose we have our input image:



$$\xrightarrow{\quad} [+1 \quad +1 \quad -1 \quad -1 \quad -1 \quad +1 \quad -1 \quad +1 \quad +1 \quad +1]$$

Our filter is size 3 (3 elements), so we'll grab a window of 3 elements.



$$\xrightarrow{\quad} \underbrace{[+1 \quad -1 \quad +1]}_{[+1 \quad +1 \quad -1]} \quad -1 \quad -1 \quad +1 \quad -1 \quad +1 \quad +1 \quad +1$$

We ignore everything after the first three elements.

We then compute the result:

$$\underbrace{[+1 \quad +1 \quad -1]}_v \xrightarrow{\quad} \overbrace{\begin{bmatrix} +1 \\ -1 \\ +1 \end{bmatrix}}^f \cdot \overbrace{\begin{bmatrix} +1 \\ +1 \\ -1 \end{bmatrix}}^v = +1 - 1 - 1 = -1 \quad (7.2)$$

This is our first filtering; we get -1. This is the first element of our output:

$$y = x * f = [-1 \quad ? \quad ? \quad ? \quad ? \quad ? \quad ?] \quad (7.3)$$

We'll repeat for the rest of our 1d signal.



$$[-1 \quad +1 \quad \dots \quad ?] \xrightarrow{\quad} [-1 \quad +1 \quad -1 \quad \dots \quad ?] \xrightarrow{\quad} [-1 \quad +1 \quad -1 \quad +1 \quad \dots \quad ?]$$

This is **convolution**.

### 7.1.5 Convolution

Convolution simply applies our filter, at each position:

#### Concept 341

When filtering (doing **convolution**), the **output** at the  $i^{\text{th}}$  index is given by having **shifted** your window over from 0,  $(i - 1)$  times.

- The **indices** for our output usually start from index 1.

**Example:** The last example above, ending at index 3, outputs +1 after shifting right 3 times.

The result is a new vector:

$$y = x * f = [-1 \quad +1 \quad -1 \quad +1 \quad -3 \quad \textcolor{orange}{+3} \quad -1 \quad +1] \quad (7.4)$$

The pixel we have labelled in orange corresponds to the "bright spot" in our sequence:



As we hoped, the "matching" pattern is the highest positive magnitude!

With this, we've fully demonstrated 1-d **convolution**.

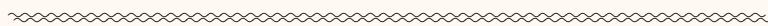
#### Definition 342

**Convolution**  $x * f$  is the process of searching through a **signal**  $x$  for a particular **pattern**, using a **filter**  $f$ .

- The filter **matches** the pattern we're looking for.

The convolution process follows the following steps:

- Taking a **window**  $v$  in the same shape as the filter, **isolating** a section of your signal
- Applying a **dot product-like** operation between your filter  $f$  and your window  $v$ .
- **Sliding** your window, and **repeating**, until every output is computed.



- The  $(i + 1)^{\text{th}}$  index is given for the dot product computed by shifting over  $i$  times from index 1.

We call this a "**dot product-like operation**" to prepare us for **higher-dimensional** equiva-

lents.

We can even write this in formula terms. To represent **windowing**, we'll use python slices, with the same conventions.

### Key Equation 343

If we have **signal**  $x$ , **filter**  $f$  of **size**  $k$ , we can create a **window**  $v_i$  by...

Starting at the **leftmost** pixel, and shifting right by  $i$  units:

$$v_{i+1} = x[i : i + k] = \begin{bmatrix} x_{i+1} \\ x_{i+2} \\ \vdots \\ x_{i+k} \end{bmatrix}$$

- Note the subscript  $v_{i+1}$ : we start from  $i = 0$ , and thus  $v_1$ .

This is used to create our **convolution**  $y = x * f$ :

$$y_i = f \cdot v_i$$

You might see a different version of indexing in some situations: \_\_\_\_\_

The fact that  $x$  is 1-indexed, but python slicing is 0-indexed, is the reason why  $x[i : i + k]$  starts at  $x_{i+1}$ .

### Clarification 344

Above, we used  $i$  to give us the leftmost slice of our input.

Like in the official notes!

We did this because we assumed the **leftmost** pixel would be assigned  $i = 0$ .

- However, in some cases, the **middle** pixel is assigned  $i = 0$ : the pixels indices go equally positive or negative.

In which case, we would need to **replace** our slicing procedure above:

$$x[i : i + k] \rightarrow x[(i - \lfloor k/2 \rfloor) : (i + \lfloor k/2 \rfloor)]$$

- We use the floor operator  $\lfloor x \rfloor$  so that we index correctly, by integers.

One more thing: we need to be careful when we say we're doing "Convolution".

### Clarification 345

In other fields, convolution requires **reversing the order** of your filter, before you apply it to your input.

However, this is typically **not** the case in machine learning.

### 7.1.6 Convolution Output Size

Something you might notice is that our output is **smaller** than our input was.

How much shorter? 2 elements: in general, the output of a convolution is  **$k - 1$**  elements **shorter** than the input.

Why is this? We can see why, by focusing on the **leftmost** element of our filter: we can only shift it until our vector ends.

Where  $k$  is, again, the size of our filter.

$$\begin{array}{ccccccccc} & & & & & \overbrace{\begin{bmatrix} +1 & -1 & +1 \end{bmatrix}}^k \\ +1 & +1 & -1 & -1 & -1 & +1 & -1 & \underbrace{\begin{bmatrix} +1 & +1 & +1 \end{bmatrix}}_{n-k} \end{array} \quad (7.5)$$

But, our leftmost element hasn't reached the end of the vector: if it did, then the rest of the vector would be **sticking out**, with nothing to multiply with:

$$\begin{array}{ccccccccc} & & & & & \overbrace{\begin{bmatrix} +1 & -1 & +1 \end{bmatrix}}^k \\ +1 & +1 & -1 & -1 & -1 & +1 & -1 & +1 & +1 & \underbrace{\begin{bmatrix} +1 & ? & ? \end{bmatrix}}_{n-k} \end{array} \quad (7.6)$$

When our leftmost position is as far right as we can go, there are  $k - 1$  positions remaining: the rest of the filter is "in the way".

#### Concept 346

For a length- $n$  input and a length- $k$  filter, **1d convolution** creates an output of size:

$$n - (k - 1)$$

### 7.1.7 Padding

We don't necessarily want to be shrinking the size of our output. How do we solve this?

Well, our equation above gives us two options: increase the input size, or decrease the filter size.

- Decreasing filter size is **restrictive**: the smaller the filter, the smaller the pattern we can search for.
- So, we'll just increase the size of our input.

We'll increase input size with **padding**: adding extra elements to the ends of our vector.

- Typically, we pad with 0's, to have the most neutral effect possible on our output.

**Definition 347**

**Padding** is a technique for increasing the size of the output of convolution.

To pad an input, you add filler values (usually 0's) to the **edges** of the input vector.

This allows the filter shift further in both directions.



A padding of  $p$  adds  $p$  values to **both sides** of our input vector, transforming our  $n$ -sized input into a  $n + 2p$  sized input. Thus, our output size is:

$$(n + 2p) - (k - 1)$$

Where  $k$  is our filter size.

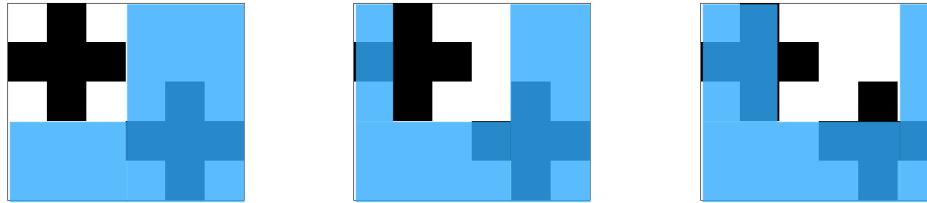
**Example:** Here's an example of **zero-padding** with  $p = 2$ :

$$\begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix} \Rightarrow \begin{bmatrix} 0 & 0 & 1 & 2 & 3 & 4 & 0 & 0 \end{bmatrix} \quad (7.7)$$

Often, we select our padding size so the output size is the same as the input size:  $2p = k - 1$ .

### 7.1.8 2D Filter

Now, we want to extrapolate this idea to higher dimensions: in particular, 2D, but this approach will work for any dimension.



We're back to this example.

Let's review what we can already guess: first, we now have a 2-D pattern we're looking for, and a 2-D window cut out of our image.

$$\begin{bmatrix} -1 & +1 & -1 \\ +1 & +1 & +1 \\ -1 & +1 & -1 \end{bmatrix}$$

This is our filter.

$$\begin{bmatrix} +1 & -1 & -1 \\ +1 & +1 & -1 \\ +1 & -1 & -1 \end{bmatrix}$$

This is our window  $w_{0,1}$ : we've shifted over by one column.

How do we measure their similarity?

$$\begin{bmatrix} -1 & +1 & -1 \\ +1 & +1 & +1 \\ -1 & +1 & -1 \end{bmatrix} \text{ vs } \begin{bmatrix} +1 & -1 & -1 \\ +1 & +1 & -1 \\ +1 & -1 & -1 \end{bmatrix} \quad (7.8)$$

We measured similarity between vectors using the **dot product**.

We can break our matrices up into vectors, by treating them as vectors of vectors.

$$\begin{bmatrix} [-1] \\ [+1] \\ [-1] \end{bmatrix} \quad \begin{bmatrix} [+1] \\ [+1] \\ [+1] \end{bmatrix} \quad \begin{bmatrix} [-1] \\ [+1] \\ [-1] \end{bmatrix} \quad \text{vs} \quad \begin{bmatrix} [+1] \\ [+1] \\ [+1] \end{bmatrix} \quad \begin{bmatrix} [-1] \\ [+1] \\ [-1] \end{bmatrix} \quad \begin{bmatrix} [-1] \\ [-1] \\ [-1] \end{bmatrix} \quad (7.9)$$

So, we can compare the similarity between the first vector in our **filter**, and the first vector in the **window** using the **dot product**.

### Concept 348

If the  $j^{\text{th}}$  vector that makes up **matrix A** is similar to the  $j^{\text{th}}$  vector that makes up **matrix B**, then A and B are **similar**.

$$\vec{d} \approx \vec{d}$$

$$\vec{b} \approx \vec{e} \implies [\vec{a} \quad \vec{b} \quad \vec{c}] \approx [\vec{d} \quad \vec{e} \quad \vec{f}]$$

$$\vec{c} \approx \vec{f}$$

- We'll repeat this process for each column.

$$\underbrace{\begin{bmatrix} [-1] \\ [+1] \\ [-1] \end{bmatrix} \cdot \begin{bmatrix} [+1] \\ [+1] \\ [+1] \end{bmatrix}}_{\text{Col 1}} + \underbrace{\begin{bmatrix} [+1] \\ [+1] \\ [+1] \end{bmatrix} \cdot \begin{bmatrix} [-1] \\ [+1] \\ [-1] \end{bmatrix}}_{\text{Col 2}} + \underbrace{\begin{bmatrix} [-1] \\ [+1] \\ [-1] \end{bmatrix} \cdot \begin{bmatrix} [-1] \\ [-1] \\ [-1] \end{bmatrix}}_{\text{Col 3}} \quad (7.10)$$

So, we've matched the  $j^{\text{th}}$  column of our window with the  $j^{\text{th}}$  column of our filter.

- And a dot product matches the  $i^{\text{th}}$  row of that window vector, with the  $i^{\text{th}}$  row of the filter vector.

That means, we're multiplying **element-wise** across our matrix: the  $(i, j)$  element of  $f$  is multiplied by the  $(i, j)$  element of  $w$ .

### Definition 349

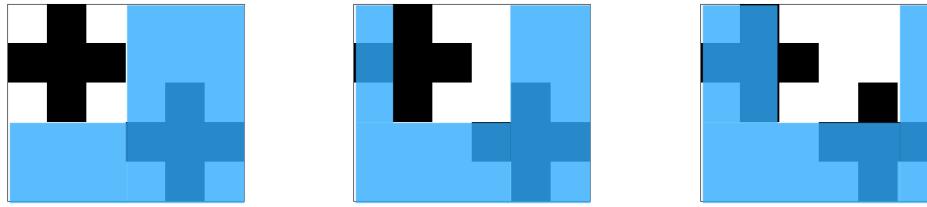
We introduce a **dot product generalization**, for a **matrix** (2-tensor):

- We compute it by multiplying **element-wise**, then **summing** all the elements.

$$A \cdot B = \sum_i \sum_j A_{ij} B_{ij}$$

This operation measures **similarity** between our two vectors.

This allows us to do higher-dimensional convolution.



We're back to this example.

### 7.1.9 2-D convolution

We'll need to shift around the image, in the same way that we did for the 1-d case.

- Before, we shifted over our **window** (size s) across our **input** (size k) by every possible position: this created  $n - (k - 1)$  outputs.

The process is the same here: we just need to shift along two axes.

- We'll need to consider every combination of shifting i rows down, and j rows right.
- In python, this is equivalent to a double for-loop: "for i in m, for j in n".

#### Concept 350

For **2-D convolution**, we need to **shift** our window along two axes.

- So, we have one window for each **combination** of shifting i rows down, j columns right.

If we have an **input** with an axis of length **n**, and a **filter** of size **k**, that output axis has **length**

$$(n + 2p) - (k - 1)$$

- k is typically the same on both of the 2d axes: it's usually **square**.

**Remark (Optional)**

Convolution was originally designed based on the way human eyes work: we use it to look for edges, and other distinct features in our vision.

### 7.1.10 Dot Product Generalization

Later, we'll need to generalize this to higher dimensions: we'll review the higher-dimensional version of a matrix, the **tensor**:

**Definition 351**

*Review from the Matrix Derivatives Chapter:*

An **array** of objects is an **ordered sequence** of them, stored together.

- The most typical example is a **vector**: an ordered sequence of **scalars**.
- A **matrix** can be thought of as a **vector** of **vectors**. For example: it could be a row vector, where every column is a column vector.

Thus, a vector is a 1-d array, and a matrix is a 2-d array.

We can extend this to any number of dimensions. We call this kind of generalization a **tensor**.

**Definition 352**

*Review from the Matrix Derivatives Chapter:*

In machine learning, we think of a **tensor** as a "**multidimensional array**" of numbers.

- Each "dimension" is what we have been calling an "**axis**".
- A tensor with  $c$  axes is called a **c-Tensor**.

**Example:** If we stacked a bunch of matrices in a box in 3-d, that would be a 3-tensor.

To get element-wise multiplication, we'll need a way to index into tensors: we'll use numpy notation.

**Notation 353**

We want to **index** into a tensor  $T$ , with  $n$  axes ("dimensions")

We'll use indices  $i_1, i_2, i_3, \dots, i_n$  to get an element:

$$T[i_1, i_2, i_3, \dots, i_n]$$

Finally, we can show the dot-product generalization for tensors: \_\_\_\_\_

**Definition 354**

The **dot product generalization** for an arbitrary **n-Tensor**

- We compute it by multiplying **element-wise**, then **summing** all the elements.

$$A \cdot B = \sum_{i_1, i_2, i_3, \dots, i_n} A[i_1, i_2, i_3, \dots, i_n] \cdot B[i_1, i_2, i_3, \dots, i_n]$$

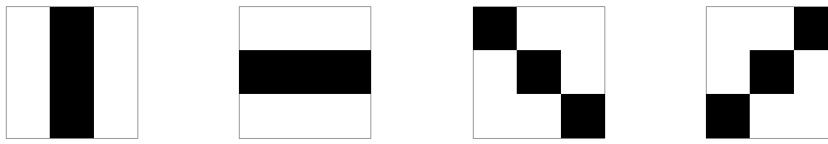
This is where python slicing really shines: it makes it easier to talk about grabbing an element from an unknown tensor.

### 7.1.11 Filter Banks

As we've shown, one filter produces one 2-d output: telling us where it "finds" or does not find the given pattern.

But, typically, when doing complex image analysis, we don't just want **one** filter. There are lots of different patterns we might be looking for.

- **Example:** Rather than programming every larger shape **directly**, it might be easier to look for smaller edges.
- You'll need a **different** filter for a vertical edge, or a horizontal edge, or a diagonal edge.



All four of these might be useful for the same image.

In practice, you almost always want to look for more than one pattern at the same time, in an image.

We'll store all of these filters together. Suppose we have  $m$  of these filters: each filter has size  $k$ .

- Each is a 2d matrix, so we'll **stack** them in the third dimension.
- This creates a **tensor** in the shape  $(k \times k \times m)$ .

This collection is called a **filter bank**.

#### Definition 355

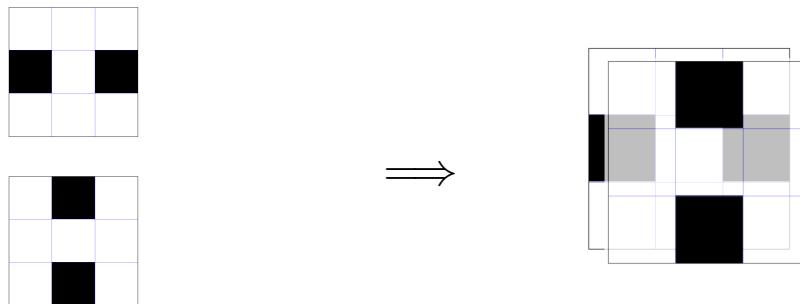
A **filter bank** is a collection of all of our  $(k \times k)$  filters stacked into a **3-tensor**.

- Thus, if we have  $m$  of these filters, the **shape** is  $(k \times k \times m)$ .

These filters are all applied to our image in **parallel**: meaning, each is applied to the original image, and each creates a separate output.

**Example:** We might, for example, combine the two following filters:

It's difficult to visualize a 3d thing like this, so if this looks strange, don't worry.



Now, we have a very simple filter bank.

#### Clarification 356

This  $(k \times k \times m)$  object could be a **filter bank**, but it could **also** be a single **3-tensor filter**, for a 3-tensor input.

Why would our **input** be a 3-tensor? We'll see why in a bit.

So, we'll use each of these filters, and **convolve** them with the input. Each creates a separate output stored in a separate **channel**.

#### Concept 357

A **channel** is the output of convolving **one filter** with our **image**.

In a 2d image problem, one channel is a **matrix**.

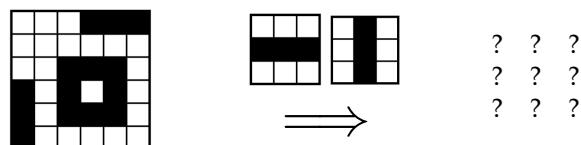
Each filter creates one channel. So, in order to depict all of our channels of output, we'll need another 3-tensor.

#### Concept 358

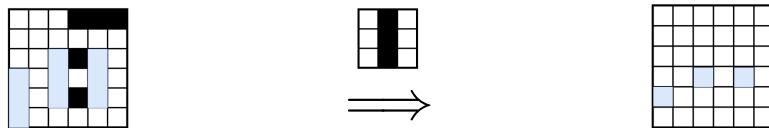
Suppose we have our 2d input.

If we have  $m$  filters in our **filter bank**, we end up with  $m$  **channels** in our output.

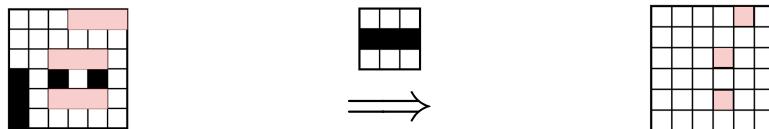
- Example:** Here, we'll apply two filters: one detecting vertical lines, one detecting horizontal lines. It'll create two channels of output.



We're applying a simple filter bank to the image on the left.

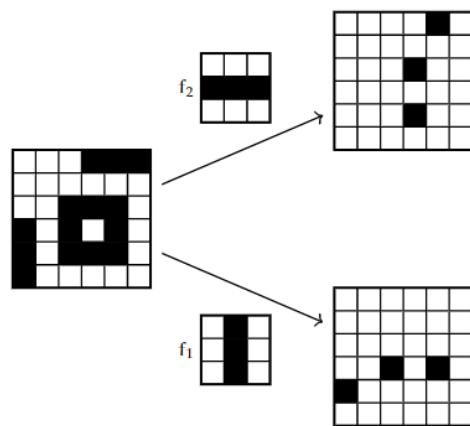


Our vertical detection.



Our horizontal detection.

Together, these create two channels:



### 7.1.12 Tensor Filters

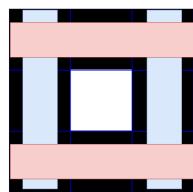
Now, we have two different channels in our output. What do we do with this result?

Each of our filters was designed to find a particular **pattern**: you could say it represents one "**perspective**" on the data.

- Our two filters above think about the data in terms of vertical lines, and horizontal lines.

We want to **combine** those perspectives to get useful information.

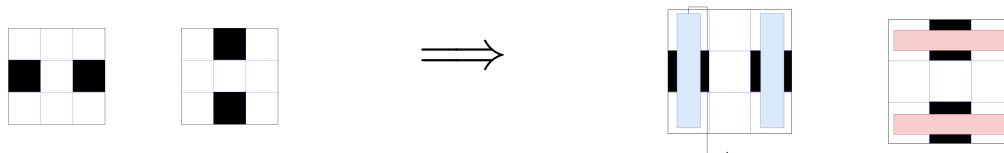
**Example:** A "square" is made out of two vertical lines, and two horizontal lines.



That means we want to find two **vertical** lines, and two **horizontal** lines: each on the opposite side of our center pixel.

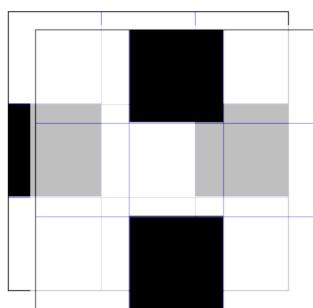
- This is a kind of **pattern** we could search for, but it's a pattern across **two channels**.
- That means that we need a filter occupying **multiple channels**.

Let's see the pattern we want to see on each channel:



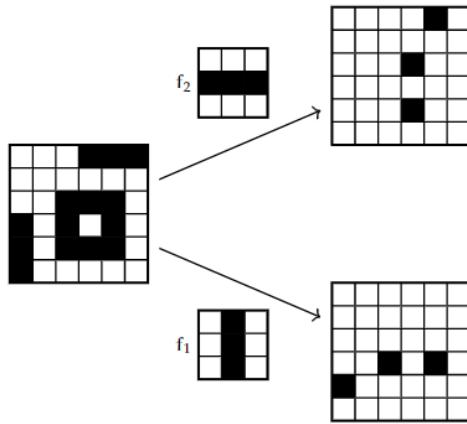
The right side shows what each pixel on the filter "represents".

We want both at the same time, so we create a **3d filter**:



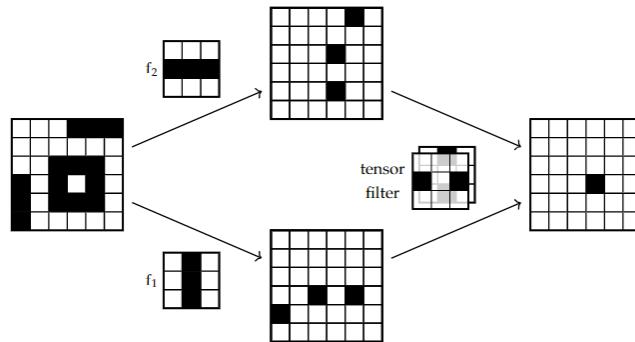
This looks like our 3d filter bank, of **2 filters**. But in this case, it's a **single** 3d filter.

Let's apply this trick to the two channels we designed earlier: we're going to find "donut" patterns in the original image.



Couldn't we have just used a donut-shaped filter in the first place, and then only need one filter?  
Yes, but this is useful for demonstrative purposes.

Here's our previous work, finding vertical and horizontal lines.



And now, we'll combine those lines to create a square.

Look at that: our result is, we **only** get an output where there's a hollow square in the **original** input!

We've found a more interesting pattern, using a **second layer** of convolution.

### 7.1.13 Tensor Filters: All channels

When we introduce a **3-tensor filter**, we could imagine not only moving across the rows and columns of the input, as we convolve, but also the **channels**.

- Thus, we would need to shift our filter along 3 axes.

However, in practice, we frequently **avoid** this: instead, our tensor filter tends to have the **same number of channels** as the input tensor.

- If they have the same number of channels, then there's no space to "shift" along the third axis.

- That means that the output of this filtering is a single matrix again!

### Concept 359

Typically, our **3-tensor filters** occupy **all channels of the input**.

- So, if the input is shape  $(a \times b \times c)$ , the filter has the shape  $(k \times k \times c)$ .

That means that our 3-tensor filter creates a **matrix** as its output, when we do convolution.

Technically, it's a tensor with the shape  $(m \times n \times 1)$ . The last dimension being 1 is why it's effectively a "matrix".

This allows us to add more tensor filters: our filter bank can contain multiple 3-tensors, and each one creates one channel of the output.

### Concept 360

Often, we use several **3-tensor** filters: each one occupies **all of the input channels** at the same time.

- And each one outputs a single **matrix**.

That means that we can stack these 3-tensors into a **filter bank**: this object is now a **4-tensor**.

- When we apply this **4-tensor filter bank** to our **3-tensor input**, we get a **3-tensor output**.

This means that our filter bank is a 4-tensor..... don't think too hard about it.

## 7.1.14 Convolution is Linear

You might have noticed that, above, we could have **replaced** all of our filters with a single, donut-shaped filter.

- We didn't do this, so we could **demonstrate** how convolution works conceptually.

This is possible because convolution, being entirely made out of **multiplication** and **addition**, is a **linear** operation.

So, two consecutive convolutions are "**compressible**" to one, just like linear layers.

**Concept 361**

Machine learning "convolution", or "cross-correlation", is a **linear** operation.

- Here, we'll summarize linearity as "**multiplying** our variables  $x$  by scalars (in this case, weights from  $f$ ), and **adding** the result together.

Summing Scalar  
Variables Product

$$v \cdot f = \sum_i \overbrace{v_i}^{\text{Scalar}} \overbrace{f_i}^{\text{Variable}}$$

In fact, as we'll see later when doing backprop, **convolution** between  $x$  and  $f$  can be represented using a particular **matrix multiplication**.

This isn't a problem in practice because we'll add ReLU and max-pool layers in between: both are **nonlinear**.

That said, a "convolutional layer" is not a "linear layer":

We haven't discussed max-pool yet.

**Clarification 362**

While convolution is **linear**, a **convolutional layer** is different from a **linear layer**.

Why is that?

Because convolution is a very **restricted** kind of linear:

- In **convolution**, we use the **same filter** for every dot product – every operation uses the same weights in  $f$ .
- In a **fully connected**, "**linear layer**", every input-output pair has a **separate, independent** weight, that can be tuned freely.



You could think that a FC/"linear" layer refers to the broadest, **least-restricted** kind of linear transformation:

- Every possible linear relationship is allowed.

You could also think of it as the "simplest" linear layer: it makes the least assumptions.

**7.1.15 RGB colors (Optional)**

One quick concern, that's more pragmatic: all of our images, so far, have been in black-and-white.

How do real pictures create color? Using the **RGB** system: each pixel has a certain bright-

ness of red, green, and blue.

- So, to represent the pixel output, you need **three** numbers: a brightness in the  $[0, 255]$  range for each color.

That means that our input isn't a 2d image: it's actually 3d.

### Concept 363

If we're using **RGB** color instead of black-and-white (BW), each pixel requires **3 values** to represent the **brightness** of each color.

Thus, if our BW image had the shape  $(m \times n)$ , our RGB image has the shape  $(m \times n \times 3)$ : we use a **3-tensor** to store the extra information about color.

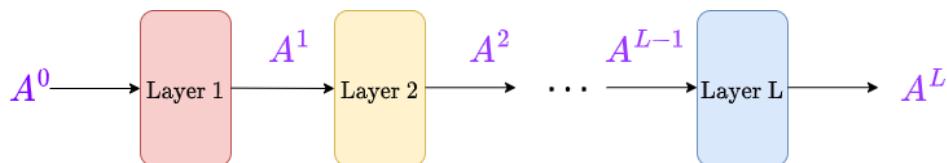
- Thus, our filters have to be 3d filters, as well.

### 7.1.16 Adding Convolution to our Neural Networks

So, we've developed a complete system for **convolution**, using **filter banks**. How do we apply this to **machine learning**?

Well, thankfully, our neural networks are very **modular**:

- Each FC layer is **self-contained**, and **abstracted**: when we depict it this way, we only care about the input and output dimensions.



By "self-contained" and "abstracted", we mean that we can hide the contents of the layer, while still having a useful representation.

We've broken our model into "modules" that we could swap out: this representation doesn't acknowledge the weights, or the structure.

So, it's not too different if, instead of a **fully-connected** layer, we were to have a **convolutional** layer.

#### Concept 364

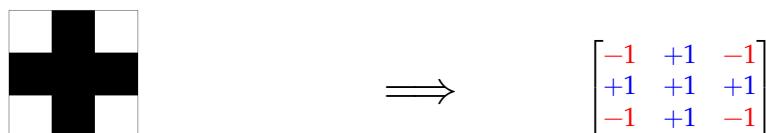
A **convolutional layer** can be inserted into a neural network by placing it **between** two other layers.

- You just need to make sure the input/output have the right **dimensions**.

Let's figure out how to **implement** that, while thinking in familiar NN terminology.

Convolution is based on your window and filter.

- Your **window** is simply given by your input tensor, and how far you've **shifted**.
- Your **filter** is what really defines your convolution: it chooses the **pattern** you're looking for.



So, we'll focus on the filter: this is how we determine the behavior of our convolutional layer.

- This is similar to how our **weight** matrix  $W$  is used to configure a layer of our FC NN.

- So, we say that our convolutional layer is defined by the **weights** in our filters.

Note that we say **filters**: we already established that we can use multiple filters. If we do, our output will have multiple channels.

### Definition 365

Our **convolutional layer** is entirely determined by the **weights** we choose for our **filter bank**.

- For **each filter** in our filter bank, we'll also include a single **offset**, or bias term.

~~~~~

Suppose that, in the 1d case, we have a window of size k , for our input x . Our window shifted by i , is labelled v_{i+1} .

$$v_{i+1} = x[i : i + k] = \begin{bmatrix} x_{i+1} \\ x_{i+2} \\ \vdots \\ x_{i+k} \end{bmatrix}$$

We've chosen weights for our filter f , with a bias term f_0 . The i^{th} element of our **output** for that filter is:

$$y_i = v_i \cdot f + f_0$$

Example: If we have a 2d filter of length k , then we need k^2 weights, and 1 bias. We have $k^2 + 1$ parameters.

$$f = \begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1n} \\ W_{21} & W_{22} & \cdots & W_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ W_{m1} & W_{m2} & \cdots & W_{mn} \end{bmatrix} \quad f_0 = W_0 \quad (7.11)$$

Each convolutional layer has one filter bank, typically.

We casually tossed in a **bias** term, similar to our neural network structure.

- But we should ask, what effect does that bias term have?

Concept 366

A **higher** filter output suggests a **higher chance** that our desired **pattern** is found at that position.

Thus, our **bias/offset** term can increase or decrease how "**sensitive**" we are to inputs similar to our pattern.

- If we increase the offset, we get a higher filter output: more inputs will appear as **positive**: possibly matching our pattern.

7.1.17 Training our Convolutional Layer

In the past, experts would **hand-craft** their filters, manually experimenting with them.

However, we've defined our convolutional layer in terms of **trainable** weights and biases.

- So, as long as we can take the **derivative**, we can use **gradient descent** to find filters which are more suitable for the task.

Concept 367

We can **train** the weights and biases used in our filter bank.

This requires doing **backpropagation** to find the gradient, but that's possible so long as we have well-defined **derivatives**.

We'll come back to how to compute these derivatives in the last section of this chapter.

Sometimes, these filters teach us interesting things about the **structure** of the data, based on which ones ended up being **useful**!

- They've even been found to sometimes recreate the types of successful designs made by humans.

7.1.18 Benefits of Convolution

We've already discussed some of the benefits of convolution:

Concept 368

Convolution provides **spatial locality** and **translation invariance**.

- **Spatial Locality:** our "filter" focuses on a **local** region of the image, and looks for a specific, **spatial** arrangement of pixels.
- **Translation Invariance:** we repeatedly apply the **same** filter as we **move** across the image. So, it will find our pattern and recognize it the same, no matter the position.

Of course, there's some caveats:

Clarification 369

Convolution doesn't perfectly provide translation invariance.

This is because of the **edges** of our image.

- If we don't use zero-padding, then information close to the edge of the image is scanned over **fewer** times.
- If we do use zero-padding, then the information close to the edge is **distorted** by the zeroes.

But there's one more surprising benefit.

The same filter is used, over and over again, as we move over the image.

- That means we repeatedly re-use the **same weights** for multiple different calculations.

This can be a bit confusing: the same weights will appear in different calculations, and thus different derivatives.

Definition 370

Weight sharing is a useful property of convolution, where the **same weights** are re-used for **multiple calculations**.

- In particular, the weights in a filter are used for many **dot products**, in the same convolution.

Having fewer weights allows our model to **train faster**, and possibly **overfit** less: it's a form of **regularization**.

Example: Let's compare two situations: in both, we have a (5×5) image, and we want a (5×5) output.

- FC Layer: We flatten our input and output to (25×1) .
 - To get every combination of input and output, we need $25 * 25$ weights.
 - 1 bias for each output: $25 * 1$ biases.
 - **Total:** 650 parameters.
- Conv. Layer: We keep our current shape and use a single filter.
 - We use a (3×3) filter, with one unit of padding ($p = 1$). That means 9 weights.
 - We have one bias: 1 bias term.
 - **Total:** 10 parameters.

In a way, weight-sharing makes our model more **efficient**. In exchange, it's less **flexible**: it makes some assumptions about how our data is structured.

7.1.19 Our NN dimensions

So, we are considering introducing a convolutional layer with layer ℓ .

We should be careful of how to notate our dimensions:

Notation 371

For a convolutional layer on layer ℓ :

- **Input length:** $n^{\ell-1}$
- **Input channels:** $m^{\ell-1}$
- **Filter size:** k^ℓ
- **Number of filters:** m^ℓ
- **Padding length:** p^ℓ

A few notes:

- The input parameters are $\ell - 1$, because they're the **output** of the previous layer $\ell - 1$.
- Notice that m is used for the filter count, and the channels of the input.
 - The input is a previous output, and the **output** has the same number of **channels** as the **filter bank**.

Now, we can use these to get the shapes of some objects:

Definition 372

For a convolutional layer on layer ℓ :

- **Input tensor shape:** $(n^{\ell-1} \times n^{\ell-1} \times m^{\ell-1})$
- **Filter shape:** $(k^\ell \times k^\ell)$
- **Filter bank shape:** $(k^\ell \times k^\ell \times m^\ell)$

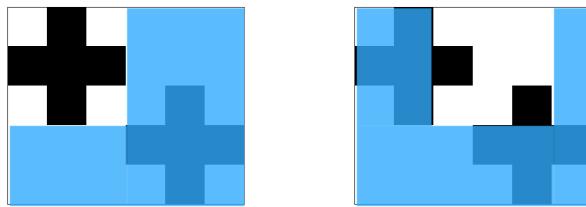
Again, we see that our input can be a **3-tensor**, if it has multiple channels.

7.1.20 Stride

There's one more thing we've skipped over: until now, we've assumed that, as we take a convolution, we move over, **one index** at a time.

But, this isn't required: we could move by **multiple** units.

This can either be due to filter banks, or the structure of the data (like in the RGB section).



This is the same as what we did before, except now, we've "skipped" one of our intermediate steps.

There are multiple reasons to do this, one of which involving "max pooling", which we discuss in the next section.

Definition 373

Stride is the distance you travel each time you **move indices** to take another **window** from your input.

Example: The stride we were using until now is 1. The stride we used in the above diagram is 2.

Naturally, this will shrink the size of your output:

Concept 374

Increasing **stride** s decreases the **size** of your output.

$$\text{New Size} = \left\lceil \frac{\text{Old Size}}{s} \right\rceil$$

We divide by s , because we're "skipping over" some windows: we are only taking a "fraction" of them.

Note the symbols on the side:

Notation 375

$\lceil x \rceil$ takes the "**ceiling**" of x : if x is between two integers, we round up.

We need this because the size of our output matrix is an **integer**.

- If we have a length-5 output with stride 1, but we take stride 2 instead, without rounding, we end up with size 2.5.
- So, instead we round.

7.1.21 Output shape

Now, we have all the tools we need, to compute the output shape, based on the input shape.

Three things can affect our input shape:

- Filter size: $n - (k - 1)$
- Padding: $n + 2p$
- Stride: $\lceil n/s \rceil$

Taking all of these variables together, we get this result (which is important, and worth saving!):

Key Equation 376

Suppose we apply **convolution** to a matrix, with

- **Input size** $n^{\ell-1}$
- **Filter size** k
- **Padding** p
- **Stride** s

The **output size** will be

$$n^\ell = \left\lceil \frac{n^{\ell-1} - (k^\ell - 1) + 2p^\ell}{s^\ell} \right\rceil$$

~~~~~

More commonly, you will see a (surprisingly) equivalent expression:

$$n^\ell = \left\lfloor \frac{n^{\ell-1} - k^\ell + 2p^\ell}{s^\ell} + 1 \right\rfloor$$

- Instead of the ceiling function  $\lceil x \rceil$ , which rounds up, we have the floor function  $\lfloor x \rfloor$ , which rounds down.

**Example:** Let's take an input tensor of shape  $(64 \times 64 \times 3)$ .

Our filter is size 2 ( $k = 2$ ), with stride 2 ( $s = 2$ ).

- It needs to have 3 channels, to match the input. Thus,  $(2 \times 2 \times 3)$ .

Using our equation for the size of our output, we get

$$\lceil(64 - (2 + 1) + 2 \cdot 0)/2\rceil = \lceil(63)/2\rceil = 32$$

So, our output dimensions are  $(32 \times 32 \times 1)$ .

This example is slightly different from the official notes: there, we were doing max-pool, so we keep all 3 channels.

## 7.2 Max-pooling

So, we've used our filters to find **basic** patterns, roughly matching the filter.

- But earlier, we showed an example where we used **two layers** of convolution, to create a more complex pattern from a simpler one.

Of course, in that case, the two layers were reducible to a **single** layer, because convolution is a **linear** operation.

But we could get a more "true" version of this idea, by introducing a new function: the **max-pool** operation.

### 7.2.1 Aggregating information

Let's clearly state our goal:

#### Concept 377

One goal of **multi-layer convolution** is to

- Find **local**, smaller patterns
- Combine them to create **bigger**, more complex patterns

With each layer, we find broader and broader patterns.

~~~~~  
However, we need a way to truly "aggregate" those patterns together.

This is the goal of our **max-pool** function.

- **Example:** We combine simple edges, into larger, **longer** edges, then into shapes like **squares**.
- Then, those combine into **windows** and doors and roofs, and finally, if they're arranged correctly, we use them to draw a "**house**".

We need a function that allows us to "**aggregate**" data this way.

~~~~~  
Here's one idea: what happens as we move to higher size scales, building up a more **complex** object?

- We tend to care **less** about the smaller, individual details.

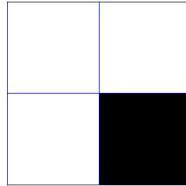
Rather than knowing **exactly** where a pattern is, we might simplify that to knowing **approximately** where that pattern is.

Our house picture doesn't stop being a house, if we slightly corrupt some of the edges, for example.

That way, we can gather information: "in this general **section** of the image, I found the pattern we're looking for!"

- How do we implement this?

It sounds like we want to replace "here's the exact pixel we found the pattern in", with "we found the pattern in this **general area**".



If black pixel indicates a "success" for finding the pattern, then this  $(2 \times 2)$  grid does contain our pattern!

#### Definition 378

The **region** we're applying our **maxpool** operation to is called our **receptive field**.

- This is similar to the **window** used by filters.

Typically, it's square grid:  $(k \times k)$ .

### 7.2.2 Deriving max-pool

In this image, how did we know that the pattern was here? We noticed, "**at least** one pixel in this grid detected the pattern".

- In other words, we don't care about pixels that **don't** detect the pattern.
- And we don't care if there are **multiple**.

We can get our desired behavior by taking the **max** of these pixels: the pixel with the greatest output, is the **most similar** to our pattern.

- So, if the "most similar" pixel doesn't match our pattern, then none of the others will, either.
- Naturally, this also ignores multiple instances of our pattern.

This process, of **pooling** together all the pixels in our receptive field, and taking their **maximum**, is called the **max-pool** operation.

We'll repeatedly apply this process, all over our image: that way, we can aggregate over each region.

**Definition 379**

The **max-pool** operation, applied to a **receptive field**, takes the **max** of all values over that region.

Similar to convolution, we repeatedly **shift** our receptive field, and compute the max again.

- Also similar to convolution: the **number of times** we've **shifted** over, is the **index** of the output.

**Example:** Here's a 2x2 max-pool operation, with a **stride** of 2:

$$\begin{bmatrix} 1 & 3 & -9 & -22 \\ 44 & -10 & -11 & -1 \\ 0 & 10 & 4 & -3 \\ 11 & 9 & 321 & 99 \end{bmatrix} \xrightarrow{\text{max-pool}} \begin{bmatrix} 44 & -1 \\ 11 & 321 \end{bmatrix} \quad (7.12)$$

**Concept 380**

**Max-pool** typically only uses a **2d matrix** for its **receptive field**: we apply the max-pool operation separately, for each channel of our input.

So, the number of channels is the same, before and after our input.

### 7.2.3 Max-pool stride

Notice that, in our example above, we've **shrunk** the image, while preserving some general data.

Max-pooling is typically designed to "gather" data across our image:

- If we apply it after a **convolutional layer**, it can help us figure out if the receptive field contains a **pattern**.

In other words, we're **searching** for our pattern over a **larger** region.

So, it might be natural to take **bigger steps** in between each max-pool, since each max-pool condenses a whole receptive field of information.

**Key Equation 381**

In order to get a simplified, "broader" view of the data, our **max-pool** often uses a larger **stride s**:

$$s > 1$$

This means we move our receptive field by a larger amount, between max-pools. This **shrinks** our output.

- If we apply this for multiple consecutive layers, we get a "**pyramid**" shape, where our output gradually shrinks in size.

With this approach, we can store the information we care about ("pattern found roughly here/not here"), without focusing on the exact, **pixel-perfect** detail.

- It's also useful for building larger objects: if two patterns (from two **filter channels**) are **roughly** nearby, it'll be easier to recognize a **larger shape**.

If we don't want to shrink our image as much, we can use **zero-padding**, usually on our convolution step.

Because often, different instances of a shape won't be exactly the same: this makes it easier to recognize them anyway.

That said, while we want a stride that's bigger than 1, we don't want it to be so big that we **skip** some of the input.

**Key Equation 382**

In order to avoid **skipping** portions of the input, we don't want our **stride s** to be larger than our filter of **size k**.

$$k \geq s$$

### 7.2.4 Clarifications on max-pooling

Our max-pool essentially chooses the "most likely to match" output, after the filtering. Based on that result, we can guess whether this region contains our pattern.

#### Clarification 383

Neither **convolution** nor **max-pooling** exactly tell us if we found a pattern **match** at a particular **index**.

Instead, they give us a **number**, based on a (generalized) **dot product**, that can be **interpreted** to check for a pattern match.

- Whether that number confirms our pattern, depends on how **high** it is.

Our **offset** can help with this problem: it helps us set a "**threshold**":

$$v_i \cdot f > -b \implies v \cdot f + b > 0$$

If we set  $b$  correctly, we could simply say, "the pattern appears if  $v \cdot f + b > 0$ ".

- This threshold, like our other parameters, will be **learned** by the neural network.

**Example:** Consider the following example:

$$\begin{bmatrix} +1 \\ +1 \\ +1 \end{bmatrix} \cdot \begin{bmatrix} +1 \\ +1 \\ -1 \end{bmatrix} = 1 + 1 - 1 = +1 \quad (7.13)$$

This output is **positive**, but the two patterns aren't visually the same. Whether they're **similar** enough depends on the context.

- If they're not similar enough to justify a positive output, we could use a **negative offset** to make our filter less sensitive.

A related comment: some pattern matches might be ambiguous.

#### Clarification 384

In this chapter, for our images, we've exclusively used simplified images, with only **extreme** brightnesses (-1 or +1).

However, in most real images, there's a **spectrum** of brightnesses.

This can make it **harder** to figure out whether you really find a pattern, in a particular place.

**Example:** Whether the following image contains our pattern is unclear. The left is our filter,

the right is our window.



The pixel in the center of the window is a bit brighter than the surroundings, but... not by very much.

This is another place where our bias can be useful to set a threshold.

### 7.2.5 Max-pool: A functional layer

Max-pool, as a specific variant of the max function, has **no parameters**: "compute the maximum input value" isn't a function that requires **adjustment**.

#### Concept 385

**max-pool** has no **parameters**: it always behaves the same way.

This also means that it doesn't need to be **trained**.

Because it has no weights, and behaves like a **function**, we can think of this a **pure functional layer**.

- Activation functions for linear layers, like **ReLU**, behave the same way: they require **no parameters**.

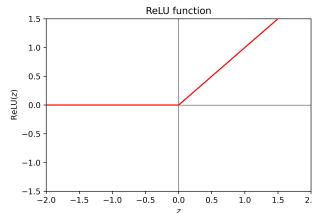
On that same note: while max-pool is *technically* nonlinear, we usually don't use it to provide nonlinearity to our model.

Instead, we accomplish this by applying ReLU **after** our convolution.

#### Concept 386

After convolution, we often apply a **ReLU function** to provide **nonlinearity** to our model.

Typically, we follow a sequence of **convolution**, **relu**, and then **max-pool**, before repeating.



A reminder of how the relu function appears.

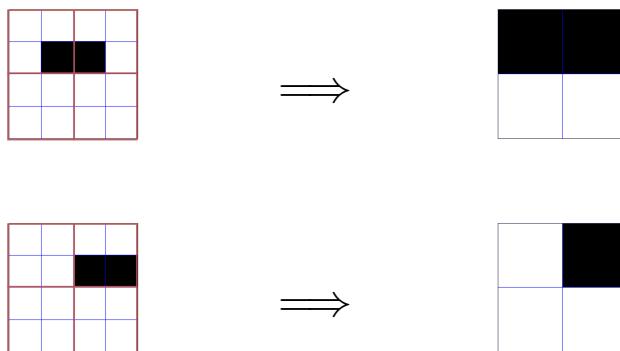
### 7.2.6 Max-pool: Some problems with "translation invariance".

There's one problem with having a larger stride:

- We skip some of the possible receptive fields.

This means that the same pattern could look different, based on where it's placed.

**Example:** We can get markedly different results of our max-pooling, just by shifting the input:



We shifted over the input, and lost one of our two black pixels: that doesn't seem very translation-invariant.

#### Concept 387

Using a stride  $s$  greater than one creates an output that isn't translation-invariant:

- if you shift part of the input slightly, it can alter the pattern recognition of the output.

This is notable for max-pool, which almost always uses  $s > 1$ .

Here, we provide a paper that provides a potential solution.

<https://arxiv.org/pdf/1904.11486.pdf>

- In short, the idea is to max-pool with stride 1, and then scale down our output by averaging over the results.

### 7.3 Typical architecture

Now that we've built all the pieces of our new neural net, we can get the general flow of a neural network that implements convolution: a **Convolutional Neural Network** (CNN).

First, let's consider what we need for each layer of convolution:

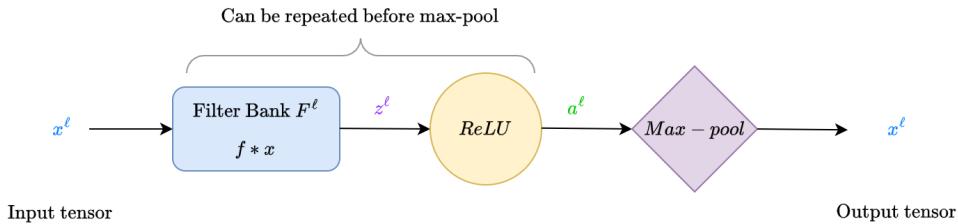
Concept copied from above, for convenience.

#### Concept 388

After convolution, we often apply a **ReLU function** to provide **nonlinearity** to our model.

We typically use layers of **convolution**, **relu**, and then **max-pool**, then repeat.

- Notably, we may do conv+relu multiple times before one max-pool.



This is our basic structure: often, we repeat this multiple times.

Once we reduced our input to a reasonably small size, we finish by using a **fully connected layer**.

The convolutional layers can be seen as "preparing" the data, so that our fully-connected network can more easily find the patterns it needs.

We could even model it as a very complex feature transformation!

#### Definition 389

A **Convolutional Neural Network** (CNN) is a neural network which uses **convolution** to transform data.

Most typically, we take the following structure:

- Several layers of **conv-relu-maxpool**, gradually shrinking the **output** size
- A **fully-connected** network, typically intended for classification or regression.

The model is **evaluated** based on the performance on the chosen classification/regression task.

- This can be viewed as several layers of convolution, applied before a regular neural network.

We can refer to our earlier comments for some benefits of CNNs:

### Concept 390

Convolution provides benefits through **spatial locality**, **translation invariance**, and **weight sharing**.

- **Spatial Locality**: our "filter" focuses on a **local** region of the image, and looks for a specific, **spatial** arrangement of pixels.
- **Translation Invariance**: we repeatedly apply the **same** filter as we **move** across the image. So, it will find our pattern and recognize it the same, no matter the position.
- **Weight sharing**: the same filter weights are **re-used** for many calculations. This can speed up training, and reduce overfitting.

As a result, CNNs tend to perform well for image-based problems.

One might ask: how many **layers** of convolution? How many **filters** per layer? What **size** should these filters have?

- These questions are good ones, but they're very **difficult** to answer: very few hard rules exist for how to design this kind of network.
- Often, designs are based on what has worked in the **past**, or some **intuition** about the data.

Once we've designed our network, we can begin training.

### Concept 391

**CNNs** can be trained just like normal neural networks: we train both the **fully connected** network, and the **filter weights** throughout the convolutional layers.

To do gradient descent, we measure the **performance** of our CNN on the classification/regression task in question.

We close out this section with an example of a "typical" CNN:

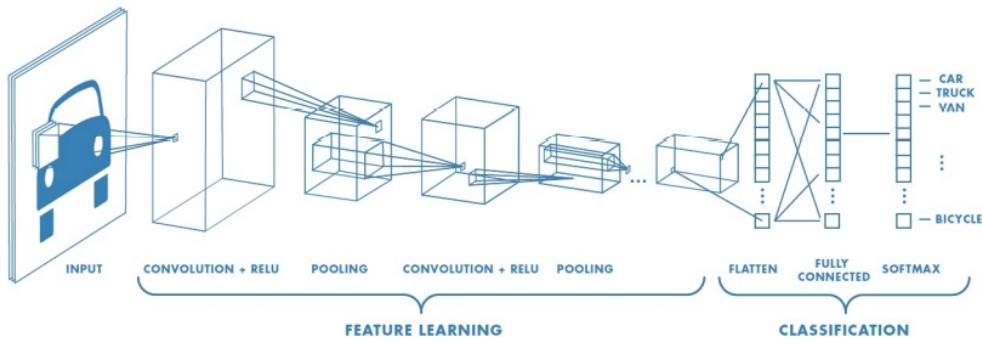


Figure source: <https://www.mathworks.com/solutions/deep-learning/convolutional-neural-network.html>

A few comments:

- Our input has **three** layers, to show the **RGB** color channels.
- In our convolution and pooling sections, we have **boxes**, not flat matrices: those represent **3-tensors**.
- Our last layer is **softmax**: our typical output for multi-class classification.
- We do end up **flattening** after our convolution is finished: that's necessary for feeding our FC network.

## 7.4 Backpropagation in a simple CNN

Now that we have a new type of neural network, it's only appropriate that we learn how to **train** it!

- Our filtering, ReLu, and maxpool functions are (mostly) **continuously differentiable**, so we can get useful derivative for **gradient descent**.

### 7.4.1 Our Simplest Example

We'll consider the simplest possible example, with a **1d** input.

- An **n**-length, one-channel, 1d input  $x$ : shape  $(n \times 1 \times 1)$ .

- We'll use zero-padding of length **p**, to create our input  $X = A^0$ .

Reminder that we're using  $A^\ell$  notation to indicate the  $\ell^{\text{th}}$  layer of our network.

With padding, our shape is  $((n+2p) \times 1 \times 1)$ .

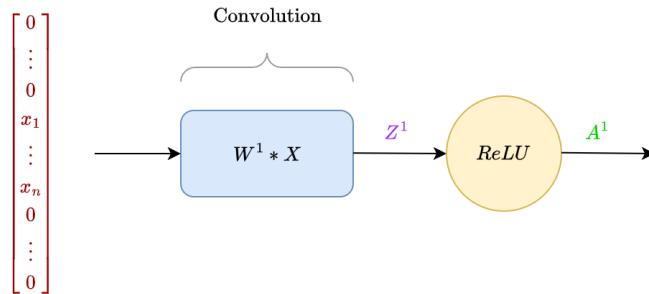
$$A^0 = X = \begin{bmatrix} 0 & \cdots & 0 & \overbrace{x_1 \cdots x_n}^x & 0 & \cdots & 0 \end{bmatrix}^T \quad (7.14)$$

- One layer of **conv-relu**

- Our convolution has one size-**k** filter, with weights  $W^1$ : shape  $(k \times 1 \times 1)$ .
- Stride **s = 1**.

$$Z^1 = \underbrace{W^1 * A^0}_{\text{Convolution}} \quad \rightarrow \quad A^1 = \text{ReLU}(Z^1) \quad (7.15)$$

We can visualize our results so far:



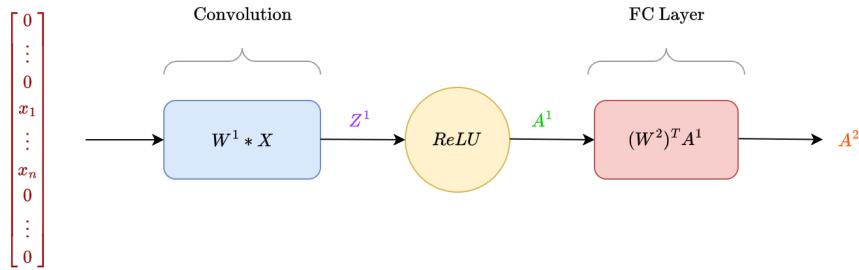
- A single FC layer for **regression**

- Using weights  $W^2$ , with no bias.

$$A^2 = (W^2)^T A^1 \quad (7.16)$$

- A loss function using **squared difference**.

$$\mathcal{L}(A^2, y) = (A^2 - y)^2 \quad (7.17)$$



### Notation 392

Reminder that, in this chapter,  $a * b$  refers to the machine learning convolution/[cross-correlation](#) between  $a$  and  $b$ .

- In other fields,  $a * b$  refers to the "true" convolution, where we flip the order of either  $a$  or  $b$  before using the same operation.

### 7.4.2 Chain rule to get full derivative

We know how to do gradient-descent on our **fully connected** layer already, and our **ReLU** layer is purely functional: no trainable weights.

So, all that's left is our **filter** derivative: our filter is parametrized by  $W^1$ .

$$\frac{\partial \mathcal{L}}{\partial W^1} \quad (7.18)$$

Looking at our above diagram, we can "move backwards" in steps, building up a **chain rule**, until we reach  $W^1$ .

$$\frac{\partial \mathcal{L}}{\partial A^2} \rightarrow \frac{\partial \mathcal{L}}{\partial A^2} \cdot \frac{\partial A^2}{\partial A^1} \rightarrow \frac{\partial \mathcal{L}}{\partial A^2} \cdot \frac{\partial A^2}{\partial A^1} \cdot \frac{\partial A^1}{\partial Z^1} \quad (7.19)$$

Finally, we get:

$$\frac{\partial \mathcal{L}}{\partial W^1} = \frac{\partial \mathcal{L}}{\partial A^2} \cdot \frac{\partial A^2}{\partial A^1} \cdot \frac{\partial A^1}{\partial Z^1} \cdot \frac{\partial Z^1}{\partial W^1} \quad (7.20)$$

### 7.4.3 Easy, Familiar Derivatives

We already know several of these terms:

$$\mathcal{L}(A^2, y) = (A^2 - y)^2 \implies \frac{\partial \mathcal{L}}{\partial A^2} = 2(A^2 - y) \quad (7.21)$$

Remember that  $A^2$  is not "A squared": it's A for layer 2.

$$A^2 = (W^2)^T A^1 \implies \frac{\partial A^2}{\partial A^1} = W^2 \quad (7.22)$$

### 7.4.4 ReLU Derivative

Our next derivative is ReLU, one of the tricky **functional layers**.

$$A^1 = \text{ReLU}(Z^1) \quad (7.23)$$

For a full dive explanation of this derivative, go to [Explanatory Notes – Matrix Derivatives, A.9.4](#).

For now, we'll take the result for granted.

**Concept 393**

*Review from Matrix Derivative Chapter:*

Each **activation** is only affected by the **pre-activation** in the **same neuron**.

So, if the **neurons** don't match, then our derivative is zero:

- $i$  is the neuron for pre-activation  $z_i$
- $j$  is the neuron for activation  $a_j$

$$\frac{\partial a_j}{\partial z_i} = 0 \quad \text{if } i \neq j$$

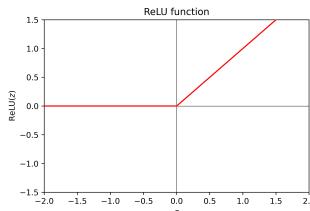
So, our only nonzero derivatives are

$$\frac{\partial a_i}{\partial z_i}$$

So, our result is a **diagonal** matrix: the off-diagonal elements are all zero.

$$\frac{\partial A_i}{\partial Z_j} = \begin{cases} f'(Z_i) & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \quad (7.24)$$

What about the **diagonal** elements? Well, that's based on the **ReLU** function.



Another picture of our ReLU function.

$$f(Z_i) = \begin{cases} Z_i & \text{if } Z_i > 0 \\ 0 & \text{otherwise} \end{cases} \implies f'(Z_i) = \begin{cases} 1 & \text{if } Z_i > 0 \\ 0 & \text{otherwise} \end{cases} \quad (7.25)$$

We get our final result by combining these two facts: the diagonal structure, with the ReLU derivative.

**Key Equation 394**

The **derivative** between the length- $m$  input  $Z$  and output  $A$  of the **ReLU** function is an  $(m \times m)$  **diagonal matrix**, whose diagonals are

$$\frac{\partial A_i}{\partial Z_i} = \begin{cases} 1 & \text{if } Z_i > 0 \\ 0 & \text{otherwise} \end{cases}$$


---

Being a diagonal matrix, the off-diagonal elements are all zero.

$$\frac{\partial A_i}{\partial Z_j} = \begin{cases} 1 & \text{if } Z_i > 0 \text{ and } i = j \\ 0 & \text{otherwise} \end{cases}$$

**Example:** One possible matrix might look like

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (7.26)$$

Where  $Z_3$  was negative, and thus it was in the 0, flat region of ReLU.

### 7.4.5 Filter Derivative

Lastly, we want to compute a **derivative**, based on the output of our **filter**.

$$Z^1 = W^1 * A^0 \implies \frac{\partial Z^1}{\partial W^1} = ??? \quad (7.27)$$

The problem is: we don't have a derivative for our **convolution**. Let's find an **expression** to get the derivative from.

The easiest way to do that is to look at our **equations** for convolution:

#### Key Equation 395

*Review from above, section 9.1.5*

If we have **signal**  $x$ , **filter**  $f$  of **size**  $k$ , we can create a **window**  $v_i$  by...

Starting at the **leftmost** pixel, and shifting right by  $i$  units:

$$v_{i+1} = x[i:i+k] = \begin{bmatrix} x_{i+1} \\ x_{i+2} \\ \vdots \\ x_{i+k} \end{bmatrix}$$

- Note the subscript  $v_{i+1}$ : we start from  $i = 0$ , and thus  $v_1$ .

This is used to create our **convolution**  $y = x * f$ :

$$y_i = f \cdot v_i$$

Let's try to create something differentiable out of our equation for elements of  $Z$ : \_\_\_\_\_

$$Z_i = W \cdot v_i = \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_k \end{bmatrix} \cdot \begin{bmatrix} X_i \\ X_{i+1} \\ \vdots \\ X_{i+k-1} \end{bmatrix} = \sum_{j=1}^k W_j X_{i+j-1} \quad (7.28)$$

We'll swap out  $v_{i+1}$  for  $v_i$ . This makes our slicing a little ugly, but we'll just omit that.

Now we have something differentiable: a sum of products! Let's find  $\frac{\partial Z_i^1}{\partial W_j^1}$ .

- If we're differentiating with  $W_j^1$ , we can ignore every term of the sum that doesn't include it.
- All that remains is the one term containing  $W_j$ .

$$W_j^1 X_{i+j-1} \quad (7.29)$$

Thus, we find:

$$Z_i^1 = \sum_{j=1}^k W_j^1 X_{i+j-1} \implies \frac{\partial Z_i^1}{\partial W_j^1} = X_{i+j-1} \quad (7.30)$$

### Key Equation 396

The derivative between the output of the **convolution** and its **weights** are given by a matrix, containing elements from the **input**:

$$\frac{\partial Z_i}{\partial W_j} = X_{i+j-1}$$

The matrix for  $\frac{\partial Z^1}{\partial W^1}$  has the shape  $(k \times n)$ .

**Example:** Suppose we had a simple example: 4 weights in  $W$ , 6 variables in  $X$ . With stride 1, that gives 3 outputs in  $Z$ .

$$\frac{\partial Z^1}{\partial W^1} = \begin{bmatrix} X_1 & X_2 & X_3 \\ X_2 & X_3 & X_4 \\ X_3 & X_4 & X_5 \\ X_4 & X_5 & X_6 \end{bmatrix} \quad (7.31)$$

With this last derivative, we can assemble our chain rule, and compute the gradient for  $\partial \mathcal{L} / \partial W^1$ .

We can confirm this with some shapes:

- Size of  $X$  is  $m$ :  $(m \times 1)$ .
- Size of  $Z^1$  and  $A^1$  is  $n$ :  $(n \times 1)$ .
- Size of  $A^2$  is 1: it's a scalar.
- Size of filter  $W^1$  is  $k$ :  $(k \times 1)$ .

Using our knowledge from the matrix derivatives chapter, we can confirm our shapes:

$$\underbrace{\frac{\partial \mathcal{L}}{\partial W^1}}_{(k \times 1)} = \underbrace{\frac{\partial Z^1}{\partial W^1}}_{(k \times n)} \cdot \underbrace{\frac{\partial A^1}{\partial Z^1}}_{(n \times n)} \cdot \underbrace{\frac{\partial A^2}{\partial A^1}}_{(n \times 1)} \cdot \underbrace{\frac{\partial \mathcal{L}}{\partial A^2}}_{(1 \times 1)} \quad (7.32)$$

### 7.4.6 Maxpool derivative

We didn't include a **maxpool** unit in our CNN. How do we compute the **derivative** of that?

Well, let's consider a simplified case: a 1d window of 2 elements:  $a_1$  and  $a_2$ .

$$\text{maxpool}(A) = \max \left( \begin{bmatrix} a_1 & a_2 \end{bmatrix} \right) = \begin{cases} a_1 & \text{if } a_1 \geq a_2 \\ a_2 & \text{if } a_1 < a_2 \end{cases} \quad (7.33)$$

- Notice that if  $a_1 = a_2$ , it **doesn't matter** which of the two you select.

We can just take the derivative from one of these inputs, let's say  $a_1$ .

$$\text{maxpool}(A) = \begin{cases} a_1 & \text{if } a_1 \geq a_2 \\ a_2 & \text{if } a_1 < a_2 \end{cases} \implies \frac{\partial \text{maxpool}(A)}{\partial a_1} = \begin{cases} 1 & \text{if } a_1 \geq a_2 \\ 0 & \text{if } a_1 < a_2 \end{cases} \quad (7.34)$$

This gives us something we can **generalize** to more  $a_i$  terms:

- If  $a_i$  is **biggest**, then it'll be the output of maxpool, and it'll have an effect – a **nonzero** derivative.
- If  $a_i$  is **not biggest**, then it's not included in the maxpool output, and it has **no effect** – a zero derivative.

**Example:** If you're taking the maximum, and the largest number is 1000, it doesn't matter if the second largest number is 999 or 2.

#### Key Equation 397

The **maxpool derivative** is only nonzero for its **maximum** value.

$$\frac{\partial \text{maxpool}(A)}{\partial a_i} = \begin{cases} 1 & \text{if } a_i = \text{maxpool}(A) \\ 0 & \text{otherwise} \end{cases}$$

When we take all of the  $a_i$  derivatives and combine them into a **vector**, we realize that we have a **one-hot vector**, telling us which output was the **maximum**.

**Example:** Suppose  $a_4$  was the largest out of 6 inputs in a column vector  $a$ .

$$\frac{\partial \text{maxpool}(A)}{\partial a} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad (7.35)$$

### 7.4.7 Maxpool derivative: somewhat similar to sign function (Optional)

Interestingly, when two values  $a_i$  and  $a_j$  are both max, maxpool behaves like ReLU.

If  $a_1 = a_2$ , and both are max, we have two cases:

- If  $a_1$  decreases, it has no effect on the output: **derivative 0**.
- If  $a_1$  increases, it increases the maxpool output directly: **derivative 1**.

So, the derivative is different moving left or right: it's **undefined**.

We'll ignore this edge case, just like we usually do for ReLU.



But what if we're not at the edge case, and  $a_i$  is our max value?

- Well, **decreasing** or **increasing**  $a_i$  will have a 1-1, linear effect, until we reach the **second-largest** term,  $a_j$ .
- Then once we're below  $a_j$ ,  $a_i$  it has no effect.

We still see that, for any one particular  $a_i$  term, maxpool behaves like a shifted ReLU, and its derivative like the step function.

#### Key Equation 398

The derivative of maxpool for a particular  $a_i$  term behaves like the **step function**.

- The transition from 0 to 1 occurs at the **highest  $a_j$  term, excluding  $a_i$** .

This is true regardless of whether  $a_i$  is currently the max.



By integrating, we see that maxpool, then, behaves like a **ReLU function**, shifted on the input/output dimensions.

## 7.5 Terms

- Connected
- Fully Connected
- Flattening
- Spatial Locality
- Translation Invariance
- Window
- Dot Product (Review)
- Filter
- Convolution
- Cross-Correlation
- Padding
- Dot Product Generalization
- Filtering
- Tensor (Review)
- Filter bank
- Channel
- 3-tensor Filter
- Linear Layer (Review)
- Convolutional Layer
- Weight sharing
- Stride
- Max-pooling
- Receptive Field
- Functional Layer
- Convolutional Neural Network

# CHAPTER 8

---

## Transformers

---

In this chapter, we want to focus on processing language. In particular:

**Definition 399**

**Natural Language Processing** (NLP) is a field of machine learning all about processing, understanding, and using **human language**.

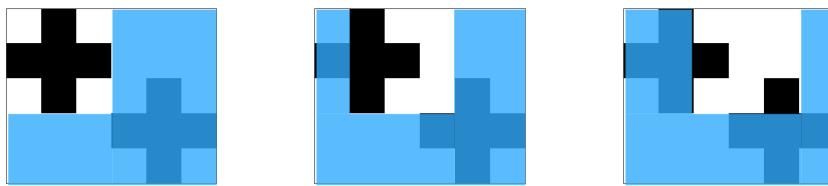
- **Example:** Chatbots, language translation, etc.

We'll start by considering a few candidate models for NLP, before moving to the state-of-the-art: **transformers**.

### 8.0.1 CNNs

In the previous chapter, we introduced the notion of a CNN:

- **Convolutional Neural Networks (CNNs)** view small regions of data, searching for patterns across the image.



In this example, we focus on a 3x3 segment of our data.

This kind of structure is useful for image processing: nearby pixels tend to be related to each other.

- By prioritizing "nearby" information, we can create models that easily find those **localized** patterns.
- We called this property **spatial locality**.

They might form a single line, or a corner, for example.

#### Concept 400

CNNs are designed to represent **locality**:

- In a CNN, **nearby** data is used to search for patterns.

This allows us to use smaller, **simpler** models:

- Rather than thinking about every possible connection between data, we only connect "nearby" data. Thus, we need **fewer** parameters.

## 8.0.2 The problem with locality

This presents one simple weakness, that we've ignored so far:

- If we focus on information that is **nearby**, we're missing out on information that's **far away**.
- We need a way to encode "distance" of information, that doesn't ignore the "distant" info.

#### Concept 401

If information is spread over **long distances**, our CNN model won't capture it.

- If a pattern is **too big** for our CNN filter, we'll have more trouble finding it.

This can become especially problematic for **language** processing.

**Example:** Consider the following sentence:

- The **sweater** that I found in the back of my old closet, which I hadn't opened since we moved into the house several years ago, **still fits me perfectly**.

Note that the beginning and the end of this sentence are linked as a single idea: "**The sweater still fits me perfectly**".

- But there's a **huge gap** between these phrases: it might be difficult to connect information over such a wide gap, while **ignoring** what's in-between.
- This also comes up in longer passages: in a paragraph, the first sentence might create **context** for the last sentence.

#### Concept 402

In language, words can be **far apart**, while still providing important **context** for the meaning of the text.

- Thus, language processing is difficult for models which focus too much on **locality**.

### 8.0.3 RNNs

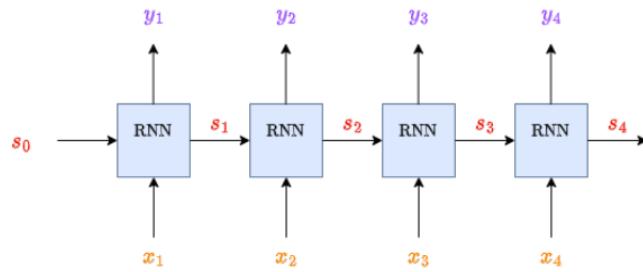
One useful observation might be that language tends to be *sequential*: words come in a very particular order.

#### Concept 403

In **image processing**, we see many pixels at the same time: the whole image is processed in **parallel**.

In **language processing**, we hear/read words one-by-one, in order: the data has a **sequential** structure.

Recurrent Neural Networks (RNNs) are, thus, a **sequential** model, designed for processing language.



Each  $x_t$  is one word in our sentence: we process the text, one word at a time. After every word, we update our memory ("state"  $s_t$ ).  $y_t$  is our output at time  $t$ .

By storing information about previous words (using a **state**), our model can "read" each word **in order**, while still remembering earlier parts of the text.

- While a CNN can only observe  $k$  consecutive pixels/words in a row, our RNN might be able to contain *some* information about words that are much further back in time.

How well does this work? RNNs have seen success in the past, but it struggles with **forgetting**: our RNN can only store so much information about words it's seen before.

- As a passage gets longer, our RNN is only paying attention to words it's seen **recently**.

Moreover, our RNN doesn't have any way to **choose** which words to **prioritize**: each new word will have to replace some information about older words.

- So, our RNN naturally prioritizes the most **recent words**. \_\_\_\_\_
- But the most recent word isn't always the most important one, as we saw above (in the sweater example)!

The more recent words haven't been replaced yet.

#### Concept 404

**RNNs** (Recurrent Neural Networks) tend to struggle with longer bodies of text:

- The **longer** we run our RNN, the less it usually remembers about the **distant past**.

Moreover, it prioritizes recent words, even when more distant words may be **more important**.

In the end, RNNs have, in most language applications, been replaced by transformers: a different model for language processing. \_\_\_\_\_

However, some transformer models have begun using the concepts of LSTMs, an RNN variant. We won't cover this topic here.

### 8.0.4 Transformers

One clever way to think about this problem is to recognize that our goal is to decide which words are **related** to each other, whether they're nearby or far apart.

- In other words, which words should we pay **attention** to, in order to understand the text we're reading?

This is exactly the problem that **transformer models** solve, using the appropriately named **attention mechanism**.

#### Clarification 405

In this chapter, we'll use **transformers** to **process language**, using the mechanism of **attention**.

- But the same tools can be applied to **many other problems**: image and audio processing, robotics, etc.

We'll develop this model in several steps:

- First (11.1), we'll convert words into vectors. One-hot encoding is too simple, so we'll use a different approach: **vector embeddings**.
- Next (11.2), we'll figure out which words in a passage are **relevant** (or connected) to each other, using a clever system called **attention**.
- Finally (11.3), we'll put together these ideas to create a complete model, known as a **transformer**.

## 8.1 Vector embeddings and tokens

### 8.1.1 One-hot encoding isn't enough

First, we want to turn words into something computable, like a **vector**.

The simplest approach would be **one-hot encoding**.

It's difficult to try to do math on the word "cheddar". It's not numerical.

- **Example:** Suppose that we want to classify **furniture** as table, bed, couch, or chair.

$$\begin{bmatrix} \text{table} \\ \text{bed} \\ \text{couch} \\ \text{chair} \end{bmatrix} \quad (8.1)$$

- For each class:

$$v_{\text{chair}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad v_{\text{table}} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad v_{\text{couch}} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad v_{\text{bed}} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad (8.2)$$

This approach is simple, but often, it's *too* simple.

#### Concept 406

**One-hot encoding** loses a lot of information about the objects it's representing.

- It's hard to say which words are "**similar**" to each other, for example.

**Example:** You probably associate the word "**sugar**" with "**sweet**", and "**salt**" with "**savory**".

- But, if you use one-hot encoding, all of these words are "**equally different**".

$$v_{\text{salt}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad v_{\text{savory}} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad v_{\text{sugar}} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad v_{\text{sweet}} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad (8.3)$$

You could **shuffle** the rows of one-hot vectors, and represent the same information.

So, we can't use the order of 1's and 0's to determine "**closeness**": the order can be **freely changed**.

In order to incorporate this information, we'll need a better way to represent words as vectors.

### 8.1.2 Word Embeddings: Similarity between words

Our new approach will convert each word  $w$  into a **vector**  $v_w$  of **length d**.

Unlike one-hot encoding, we don't require that  $d$  equals the size of our vocabulary.

Last Updated: 09/03/24 03:53:41

$$w \longrightarrow v_w \quad v_w \in \mathbb{R}^d \quad (8.4)$$

How do we want to convert words into vectors? Above, we mentioned that one-hot doesn't tell us how **similar** two words are.

#### Clarification 407

There are many ways for words to be **similar**: similar word length, similar choice of letters, etc.

But in our case, we're interested in **semantics**: the **meanings** of the words. We want to know which words have similar meanings.

- **Example:** We don't consider "sugar" and "sweet" to be similar because they both start with "s".
  - They're similar because of **meaning**: sugar tastes sweet. Sweet strawberries contain sugar.

#### Concept 408

We often want our **word embeddings**  $v_w$  to tell us which words are **semantically similar** to each other: which words have similar **meanings**.

$$v_a \text{ and } v_b \text{ are } \text{similar vectors} \iff a \text{ and } b \text{ are semantically similar words}$$

Our goal is to make this statement true. But we have a problem: these are *concepts*, rather than computable *numbers*.

- So, we'll have to turn each side into something computable.

### 8.1.3 Vector Similarity: Dot Products

First, we'll handle the left side: how do we know if vectors are **similar**?

- We've come across this problem multiple times, and we'll solve it the same way as always: using the **dot product**.

**Concept 409**

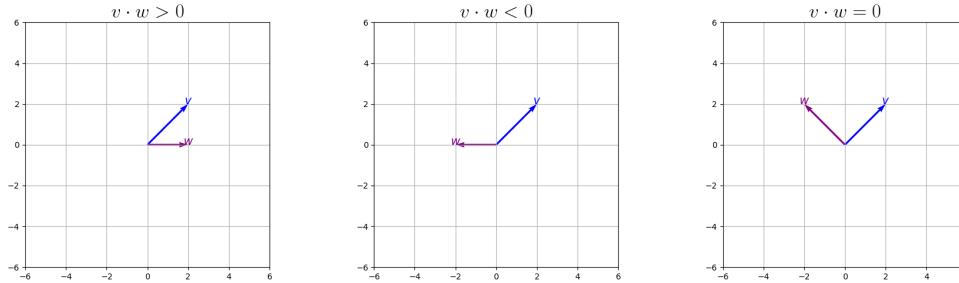
*Review from the Classification chapter*

You can use the **dot product** between vectors  $u$  and  $v$ , **normalized by their magnitudes**, to measure their "cosine similarity".

$$S_C(u, v) = \frac{u \cdot v}{|u| \cdot |v|}$$

If two vectors are more **similar**, they have a **larger** normalized dot product.

- This function ranges from -1 (opposite vectors) to +1 (identical vectors). Perpendicular vectors receive a 0.



We call it "cosine similarity", because this is equal to the cosine of the angle  $\alpha$  between  $u$  and  $v$ .

We can see here what we mean by "similar" or "dissimilar".

**Clarification 410**

You can use  $S_C(u, v)$  to measure the **similarity** between two vectors, ignoring magnitude.

But for simplicity, we'll skip the **normalizing** step, and just take the **dot product**:

$$S_D(u, v) = u \cdot v = u^T v$$

We're getting closer to a computable form:

$$\overbrace{(v_a \cdot v_b)}^{\text{Similar vectors}} \text{ is large} \iff a \text{ and } b \text{ are semantically similar words} \quad (8.5)$$

### 8.1.4 Word2vec

Next, we should get into the math of how to determine which words are likely to be **similar**.

- But this is a bit cumbersome, and isn't really necessary for understanding transformers.

So, we relegate this mathematical labor to Appendix D, where we'll get into the details of **skipgram** and **word2vec**.

The short version: we expect words which frequently appear in the same contexts, to be similar.

#### Definition 411

We can think of **word2vec** as a system for **word embeddings** where words which have **similar meanings**, have **similar vector embeddings**.

- Most commonly, we measure "vector similarity" with the **dot product**.

Instead, we'll skip a couple steps, and look at things from a high level.

### 8.1.5 Probability

Our goal is to be able to numerically talk about the "similarity" or "relatedness" of words.

Above, we represent this with a **dot product**: this gives us a **real number**  $u \cdot v \in \mathbb{R}$ .

- This number isn't very **meaningful**, though. For example, what does a "similarity of 37" even mean? Is that high? Is that low?

Generally, our best bet for understanding a number like this is to **compare** it to other numbers.

So, let's think about the **relative** similarity of words: if we have two words,  $w_1$  and  $w_2$ , which one is  $v$  **more related** to?

You know that someone who is 6'5" is really tall, because you know how tall other people tend to be.

We'll focus on one simple tool for comparison: **probability**.

- One way to think about it is, "how likely is  $w_i$  to be the **most relevant** word to  $v$ , in any given context?"
- Alternatively, "how **confident** are we that these words are actually closely related, compared to others?"

In skipgram, our probability comes from asking, "how likely is  $w_i$  to show up in the same context?"

The **higher** the probability of word  $w_1$ , the **lower** the probability of word  $w_2$ , and vice versa.

We'll represent this "relatedness of word  $w_i$  to word  $v$ " as probability  $P(w_i | v)$ .

### Concept 412

One way to describe the **relatedness** of different words  $w_i$  is with a **probability**  $P(w_i | v)$ .

This has a few advantages:

- A probability is easier to **interpret** than a real number.
- We can directly **compare** different words.
- We can systematically **convert** our dot product to a probability.

How do we turn a **real number**  $v_a \cdot v_b$  into a **probability**  $P(b | a)$ ?

This "probability" interpretation is a bit better justified if you read the skipgram section.

- For a probability, we need to compare  $b$  to every other word: this is a **multi-class problem**, using the **softmax function**.

$$\text{Softmax}(z_k) = \frac{e^{z_k}}{\sum_i e^{z_i}} \quad (8.6)$$

Let's review the concept behind "softmax":

### Definition 413

Suppose that we have **n possible words** ( $n$  "classes"), and we want to figure out which one is **correct**.

The  $k^{\text{th}}$  class has a score,  $z_k$ , used to compute probability.

- The bigger  $z_k$  is, the **more likely**  $k$  is to be the **correct class**.

To keep it **positive**,  $z_k$  is converted to  $e^{z_k}$ : each  $e^{z_i}$  competes to see which class is more likely.

- To create a probability, we **compare** the score of class  $k$  to all of our other classes, using **softmax**.

$$\overbrace{e^{z_k}}^{\text{Class } k} \quad \text{vs} \quad \overbrace{\sum_i e^{z_i}}^{\text{All classes}} \quad \Rightarrow \quad \text{Softmax}(z_k) = \frac{e^{z_k}}{\sum_i e^{z_i}}$$

We repeat this process for every possible word  $i$ , to get all of our predictions.

What is our "score"  $z_k$ ? We could use  $(v_a \cdot v_b)$ :

- The higher  $(v_a \cdot v_b)$  is, the more **similar/related** we expect a and b to be.
- The same is true for  $z_k$ : if  $z_k$  is larger, then our **probability** goes up.

So, we can use our dot product as a "score"  $z_k$ :

$$z_b = v_a \cdot v_b \quad (8.7)$$

We'll plug this into our probability equation:

#### Key Equation 414

The **more similar** (bigger dot product) a and b are, the **more likely** we predict to find them together.

- We use a **softmax** to compute this probability for each possible word b.

$$P\{b \mid a\} = \frac{e^{v_a \cdot v_b}}{\sum_i e^{v_a \cdot v_i}}$$

Or, in alternate notation:

$$P\{b \mid a\} = \frac{\exp(v_a \cdot v_b)}{\sum_i \exp(v_a \cdot v_i)}$$

This kind of interpretation makes our word embeddings a bit more **useful**.

- Later, we'll find that it's the most important part of making **transformers** work!

### 8.1.6 "Adding" words together

Our word2vec system works under the hope that these vector embeddings can accurately represent the **meanings** of words.

- In practice, this assumption works **surprisingly well**, for being so simple.

One example is the idea of "**adding**" words together. Normally, it's hard to say how to "add words" together, but we *do* know how to add **vectors**.

Consider the following example:

$$v_{\text{king}} - v_{\text{man}} + v_{\text{woman}} \approx v_{\text{queen}} \quad (8.8)$$

This sort of reasoning makes sense to most english speakers:

$$\underbrace{v_{\text{king}} - v_{\text{man}}}_{\text{ruler}} + v_{\text{woman}} \approx \underbrace{v_{\text{queen}}}_{\text{female ruler}} \quad (8.9)$$

We can repeat this process for other words:

$$v_{\text{paris}} - v_{\text{france}} + v_{\text{italy}} \approx v_{\text{rome}} \quad (8.10)$$

Paris is the capital of France, and Rome is the capital of Italy.

#### Concept 415

Transforming a word into a **vector** allows you to use vector operations, like **addition** and **subtraction**.

- The result can be surprisingly **meaningful**, for some word combinations.

This approach doesn't always work, but the fact that it works *sometimes* suggests that our vectors might capture real information about the "**meanings**" of words.

That said, this approach is often an over-simplification:

#### Concept 416

Reducing a word to a **single vector** can cause problems, because the same word might change its meaning, based on **context**.

- **Example:** For example, the word "bank" has a very different meaning when you compare "bank account" to "river bank".

This idea of "context" is what we hope to solve next.

### 8.1.7 Tokenization

One clarification, before we move on: so far, we've talked about predicting whole **words**, because it's easy to work with.

- But often, for language analysis, we break up words into parts, called **tokens**.
- These are the objects we study/predict, rather than whole words.

#### Definition 417

Rather than using/predicting entire **words**, we use **small parts of words**, called **tokens**.

- A "token" is the **smallest unit** in our language model.

- **Example:** You might break up the word "eating" into "eat" and "ing": both are meaningful by themselves.
- This process of turning words into tokens is called **tokenization**.

While "tokens" are used more often than "words", words often make for better examples, so we'll keep using them through the rest of this chapter.

#### Clarification 418

We'll continue using words (instead of tokens) for examples, when it's convenient.

## 8.2 Attention

Our **word embedding** technique has given us a basic way to talk about which words are "related".

- We can even use this to learn some about the "**meanings**" of words.

But there's some work to be done:

### Concept 419

Our **word embedding** technique has two major problems, for representing the **meanings** of words:

- There's a lot of information we're **missing**: **similarity** to other words isn't enough. We'll need a vector to represent that information.
- The meaning of a word is **contextual**: the sentence you put a word in, will affect its meaning.

It may not look like it, but our **word embedding** technique has already given us the basic tools we need to solve these problems.

Here's the basic idea, for how we handle each problem:

### Concept 420

We'll create a system that solves both of these problems, using **3 word embeddings**:  $v$ ,  $k$ , and  $q$ .

- We'll **embed information** about each word in a **value vector**  $v$ .
- When finding the **meaning** of a word, we'll calculate **context** from nearby words.
  - We'll use **word similarity** to figure out which parts of the context are **most important**.
  - For this purpose, word will need **two embeddings**: a **key vector**  $k$ , and a **query vector**  $q$ .

The result is a powerful model called the **attention mechanism**.

This description is over-simplified, which is why we'll need to go into detail below.

### 8.2.1 The Attention Mechanism: queries, keys

Let's consider an **example**, to get used to the idea of  $k$ ,  $v$ , and  $q$ .

Suppose we want a general idea of what "mexican" food is like. We'll need to consider lots of foods, and take an **average** of those we consider to be "mexican".

This problem comes in three parts: let's consider the first two, "query" and "key".

- **Query  $q$** : we're searching for "mexican" food. The word "mexican" is represented by a **query vector**  $q$ .
  - This is like our previous **word2vec** embedding: if two vectors are similar, then we expect them to have similar/related **meanings**.
  - So, we'll **compare**  $q$  to each food, to see which foods are 'close' to mexican.

Admittedly, we're turning "mexican-ness" into a number, which can be a bit strange.

It may help to think this way: "if someone is talking about mexican food, how often are they talking about this food?"

#### Definition 421

The **query vector**  $q$  represents a word, that we're **comparing** to **several other words** ("keys").

- It answers the question, "what kinds of words are we **searching** for?"

Using word embeddings, we design  $q$  to be "meaningful": similar words, should have similar vectors.

- And we expect **similar words** to be **more relevant** to our query.

- **Key  $k$** : Each food (apple, burrito, sushi...) has a **key vector**  $k$ , representing it.
  - A **word2vec**-style embedding, just like the **query**.
  - Combining  $k$  and  $q$  will tell us which foods are '**more**' **mexican**.

#### Definition 422

The **key vector**  $k$  represents a word, that we want to **compare** to the **query**  $q$ .

- It answers the question, "what kinds of searches does this word **match**?"

Because it's a word embedding, which encodes meaning, we expect that, if  **$k$  and  $q$  are similar**, then our key word is **more relevant** to our query.

Each embedding has a role: a **query** is used to search for relevant words, and a **key** is responding to that search, on behalf of one word.

Reminder that when we say "word", we're simplifying: we could talk about any kind of token.

**Concept 423**

Another way we could view keys vs. queries:

- **Query vector q**: asks, "how relevant are these words/tokens **to me**?"
- **Key vector k**: asks, "how relevant is my word **to the query**?"

Notice that we've made a perspective shift, in how we view word embeddings:

**Concept 424**

When we were developing word2vec, we wanted **similar vectors** to represent **semantically similar** words.

- But, in this case, we're less focused on "similarity", than **relevance**.

We look for **keys** that are the **most relevant** to our **query**.

These two ideas don't necessarily conflict, but they have somewhat different goals.

### 8.2.2 The Attention Mechanism: attention weights

How do we compute how similar k and q are? The same way as we did for word2vec: we use a **dot product**.

**Key Equation 425**

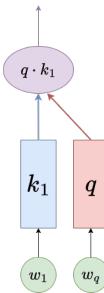
We can get a score for how **relevant** the word b is to word a, by taking the **dot product** between **b's key**, and **a's query**.

$$q_a \cdot k_b$$

We can also write this as matrix multiplication:

$$q \cdot k = q^T k$$

This gives us a "**score**": the higher  $k \cdot q$  is, the more similar they are.



We convert  $w_1$  and  $w_q$  into a key and query, respectively, before taking the dot product.

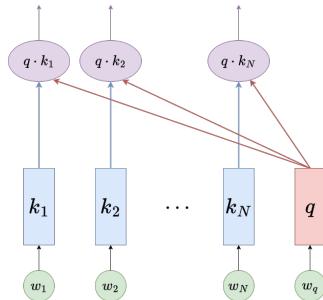
#### Notation 426

Note that  $k$  and  $q$  have to have the **same length**: they're both  $(d_k \times 1)$  column vectors.

But we're not just considering one key word: we're considering **all of them**.

If their lengths don't match, we can't take the dot product.

- In our "mexican food" example, we need to check every food, to see which ones best fit the category.



We re-use our query  $q$  for every single dot product.

#### Notation 427

We have  $N$  distinct keys.

How do we compare each of these keys?

In the official notes, we use  $n$  instead of  $N$ . This doesn't affect any of our math.

- Once again, we'll reuse a tool from word2vec: **softmax**.

### Key Equation 428

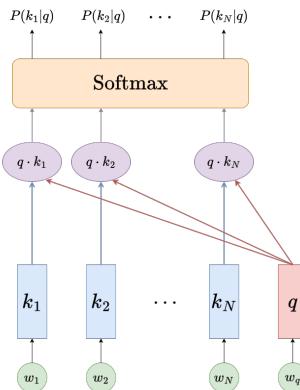
We can compute the relative **relevance** of a key  $k_j$ , by:

- **Comparing** each key  $k_i$  to  $q$  ( $q \cdot k_i$ )
- Use **softmax** to compute  $p(k_j|q)$ : given query  $q$ , how important is  $k_j$ ?

$$P\{k_j \mid q\} = \frac{e^{q \cdot k_j}}{\sum_i e^{q \cdot k_i}}$$

$P\{k_j \mid q\}$  tells you, "how much **attention** should  $q$  pay to  $k_j$ ?"

- Thus, we call  $P\{k_j \mid q\}$  an **attention weight**.



Finally, we've converted each word into their "probability" of being relevant.

One notational thing: we can write this a bit more densely.

- So far, we've been computing  $q^T k_i$  for each  $k_i$  term **separately**.
- But, one benefit of matrix multiplication, is that we can **combine multiple operations** into one.

First, we'll **combine** all of our key vectors into a matrix  $K$ :

$$K = \begin{bmatrix} | & | & & | \\ k_1 & k_2 & \dots & k_N \\ | & | & & | \end{bmatrix}^\top \quad (8.11)$$

This matrix has shape  $(N \times d_k)$ : the transpose of what you might expect.

With that, we can compute all of our dot products **at the same time**:

This product has shape  $(1 \times N)$ .

$$q^T K^T = \begin{bmatrix} q \cdot k_1 \\ q \cdot k_2 \\ \vdots \\ q \cdot k_N \end{bmatrix}^T \quad (8.12)$$

And we can combine all of these together into a softmax.

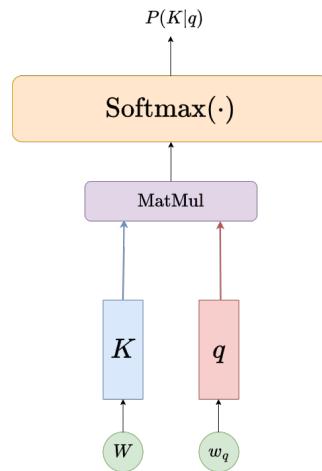
### Key Equation 429

By combining all of our keys into a matrix  $K$ , we can compute all of our **attention weights at the same time**.

$$P\{K \mid q\} = \begin{bmatrix} P(k_1 \mid q) \\ P(k_2 \mid q) \\ \vdots \\ P(k_N \mid q) \end{bmatrix}^T = \text{softmax}(q^T K^T)$$

It has shape  $(1 \times N)$ .

Note that here, softmax creates a row vector.



Now, our diagram is visually simpler, though it reflects the same information. "MatMul" means "Matrix Multiplication".

### 8.2.3 Scaling factor for softmax

One pragmatic detail. First, let's quickly define:

**Notation 430**

Reminder that keys and queries are both vectors of length ( $d_k \times 1$ ).

We have one problem: the larger  $d_k$  is, the more terms in our dot product: our dot product can grow unreasonably **large**.

- This can cause computational issues.

So, we **normalize** our dot product by a factor of  $\sqrt{d_k}$ .

**Key Equation 431**

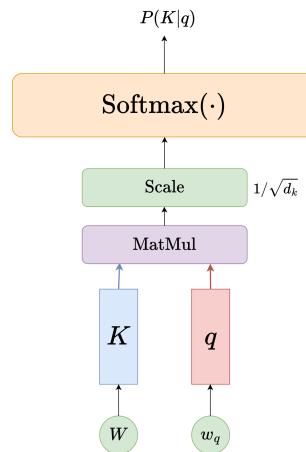
When computing **attention weights**, we **normalize** our dot product  $q^T k$  by a factor  $\sqrt{d_k}$ .

- This compensates for the fact that longer vectors will create larger dot products.

So, when computing our attention weights  $a$ , we use the formula:

$$a(q, K) = \text{softmax}\left(\frac{q^T K^T}{\sqrt{d_k}}\right)$$

It still has shape  $(1 \times N)$ .



We scale down our MatMul by the appropriate factor.

### 8.2.4 The Attention Mechanism: values, attention

Now, we have a collection of **attention weights**: each one tells us relevant each word is to  $q$ .

- Now, we want to make them useful. Our original goal was to get an **average** sense of what "mexican" food is like.

To make this concrete, we'll introduce our third embedding: the **value vector**.

- Value v:** Each food has a **value vector**, directly storing information about a word.
  - Unlike the key/query vectors, this embedding isn't based on **similarity** to other words.
  - Instead, it usually contains more direct **information** about our word: in this example, maybe it contains the price, calories, ingredients, etc.

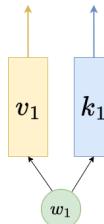
#### Definition 432

The **value vector**  $v$  represents a word, and stores useful **information** that it can contribute to the **query**.

Note that, in a real model, value vectors are often "learned" during training. So, they won't always contain such simple, easily explained data.

- It answers the question, "what useful data could this word contribute to the query?"

By **adding together** the value vectors from each word **relevant** to the query, we can get an overall "**averaged value**" for  $q$ .



Each word has both a value and a key attached to it.

For our example, let's suppose that the value vector contains price, calories, and salt.

$$v_i = \begin{bmatrix} \text{price}_i \\ \text{cal}_i \\ \text{salt}_i \end{bmatrix} \quad (8.13)$$

We want to get an "average" calorie count for mexican food.

- Some foods are **common** for mexican food, and some are more rare.
- So, to get an average, we'll need to **emphasize** more "common" mexican food.

How do we do that? Using our **attention weights**: the larger the attention weight, the more "**relevant**" a food is to our mexican food calculation.

If we use  $q$  to represent **mexican** food, and  $k_i$  is the **key** for the  $i^{\text{th}}$  food, we get:

$$\text{cal}_q = \overbrace{\sum_i P(k_i|q) \text{cal}_i}^{\text{Weighted average}} \quad (8.14)$$

Rather than repeating this process for each row of  $v$ , we can just do a **weighted average** of the whole vector, at the same time:

$$v_q = \sum_i P(k_i|q) v_i \quad (8.15)$$

### Key Equation 433

Each word  $i$  has a **value vector**  $v_i$ , which represents all of the useful **information** it can provide to the **query**.

- We can use a **weighted average** to combine all of these value vectors together: this provides the "**overall context**" for the query.
- Each value is weighted based on its **attention weight**  $P(k_i|q)$ : how likely it is to be relevant.

$$v_q = \sum_i P(k_i|q) v_i$$

This is the calculation for **attention**.

Just like we did for the  $k_i \cdot q$  operation, we can re-write this in terms of matrix multiplication.

- We'll change from  $P(k_i|q)$  to  $P(K|q)$ .

$$P\{K \mid q\} = \begin{bmatrix} P(k_1 \mid q) \\ P(k_2 \mid q) \\ \vdots \\ P(k_N \mid q) \end{bmatrix}^T = \text{softmax}(q^T K^T)$$

- We'll stack all of our value functions  $v_i$  into a matrix  $V$ .

$$V = \begin{bmatrix} | & | & & | \\ v_1 & v_2 & \dots & v_N \\ | & | & & | \end{bmatrix}^T \quad (8.16)$$

**Notation 434**

We'll assume that we have  $N$  value vectors of length  $d_k$ .

- $v_i$  has shape  $(d_k \times 1)$ ,  $V$  has shape  $(N \times d_k)$ .

Now, we can compute with every value vector at once: \_\_\_\_\_

**Key Equation 435**

We can compute **attention** using matrix multiplication:

$$\text{Attention}(q, K, V) = \text{softmax}\left(\frac{q^T K^T}{\sqrt{d_k}}\right)V$$

Where  $\text{softmax}\left(\frac{q^T K^T}{\sqrt{d_k}}\right)$  computes our **attention weights**.

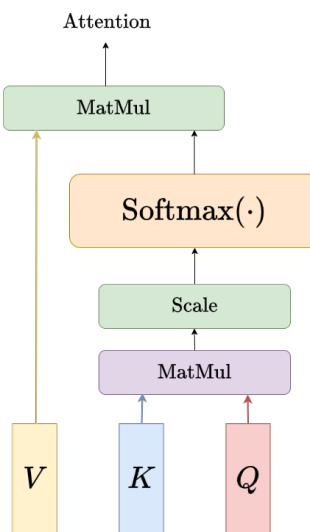
- Under this definition, attention has shape  $(1 \times d_k)$ .

If you study the classic "Attention is all you need" paper, you'll find that their version of  $k$  and  $q$  are transposed compared to ours.

**Definition 436**

$\text{Attention}(q, K, V)$  is the **weighted average** of all of our value vectors (transposed).

- Attention is the **result** of **aggregating** information from  $N$  different words: each word is represented by a key  $k_i$ , and a value vector  $v_i$ .



We now have a completed representation of attention.

With this, we can summarize the basic idea of attention:

In the "Attention is all you need" paper, this diagram is analogous to Figure 2 (left).

Here, we omit the "Mask" layer (discussed later).

**Concept 437**

**Attention** is a mechanism that allows you to **combine** information from multiple tokens, weighting each token by how **relevant** it is.

This mechanism is broken into three parts:

- **Value vector v**: **what information** are we trying to combine?
- **Query vector q**: what kinds of words are **relevant** to this search?
- **Key vector k**: what kinds of searches is this word **relevant** for?

Each token has a **value vector** (information from that token), and a **key vector** (used to compare this token to the query).

Note that this isn't the **only** way to do attention:

**Clarification 438**

There are multiple ways we can implement attention.

- For example, we use  $q \cdot k$  to measure similarity, but we could **replace** it with a different metric.

Reminder: a "metric" is just "a way of measuring something. The dot product is a **similarity metric** for vectors.

So far, we've mostly focused on the mathy details of **how** attention works: an abstract idea of "relevance" between words, "combining" the value ("meaning") of different words, etc.

- Here, we'll try something different: we'll focus more on **why** we use attention, and how it applies to a real, concrete situation.

### 8.2.5 Why we need context

Attention is designed to integrate information from other, **nearby** words. But why do we need to do this?

- Because language is heavily dependent on **context**.

Consider the task of **language translation**: we have a sentence in one language, and we want to convert it into another language, while **preserving the meaning**.

Let's translate the sentence:

I miss her **warm smile**.

We'll focus on the word "warm".

- Most commonly, "**warm**" means "higher-than-average temperature". For example, being under a blanket is warm.
- But most humans would say that, in this situation, the word "**warm**" means 'friendly' or 'kind'.

We know this because of the *context*: a "warm smile" usually means a "kind smile". The word '**smile**' has **changed the meaning** of the word '**warm**'.

#### Concept 439

The meaning of a word can change based on the other words which are **nearby**.

- This is why we need to integrate **context** for language processing.

If our machine blindly translated "warm", without context, we could've ended up with the wrong meaning in another language.

### 8.2.6 Why we need *attentive* context

So, we need to use context. But what makes attention special?

- It allows us to figure out **which words** are most important to us!

In the above sentence, the word "smile" changed the meaning of "warm".

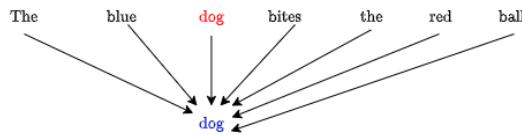
- How do we know that "smile" is the important context word? "her" is equally far from "warm".
- Attention handles this for us: we "**pay more attention**" to the word 'smile' than the word 'her', when we're trying to understand "warm".

**Concept 440**

**Attention** allows us to determine which parts of the **context** are **most important** to a particular word.

### 8.2.7 Self-attention

Attention has given us a tool for comparing one word to every other word in a sentence.

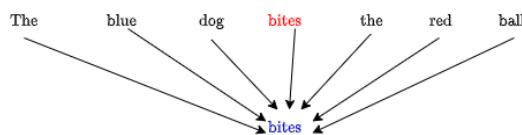


"dog" is represented by a query  $q_{\text{dog}}$ , compared to the key for every other word in the sentence. This gives us our attention weights.

Next, we combine these weights with the value vector for each word. This gives us our **attention**: the "contextual meaning" of the word **dog**.

This has a limitation: we're only focusing on a single word, "dog".

- But we need to get the meaning of **every word** in the sentence, based on the context from other words.



This time, we use the query  $q_{\text{bites}}$  for the word "bites". However, the key and value vectors are still the same for each word.

We need to repeat this attention process once for each word.

This is interesting: we're seeing how much each word affects each other word in the sentence. We're seeing how the sentence provides context for **itself**.

- This is why we call this **self-attention**.

#### Definition 441

**Self-attention** is the process of using attention on every word in a passage.

- For the  $i^{\text{th}}$  word, we compare it to **every other word** in the passage.

This allows us to interpret each word, based on the **context** provided by the rest of the sentence.

Technically, we also compare each word to itself.

### 8.2.8 Self-attention in matrix form

How do we handle this, mathematically?

- When we are getting the attention for word  $w_i$ , we use its query  $q_i$  to compare it to other words in the sentence.

We've gone from having a single query  $q$  to having many  $q_i$ : one for each word in the sentence.

$$Q = \begin{bmatrix} | & | & & | \\ q_1 & q_2 & \dots & q_N \\ | & | & & | \end{bmatrix}^T \quad (8.17)$$

Let's note some conventions:

#### Notation 442

A few useful **dimensions**: in an attention problem, we have...

- $n_k$  keys of length  $d_k$ .
- $n_q$  queries of length  $d_q$ .
- $n_v$  values of length  $d_v$ .

In **practice**, we usually take  $d_k = d_q = d_v$ , and simply refer to all three as  $d_k$ .

- Each column vector  $(k_i, v_i, q_i)$  has shape  $(d_k \times 1)$ .

In **self-attention**, we take  $n_k = n_q = n_v$ , and simply refer to all three as  $N$ .

- Matrices  $K$ ,  $V$ , and  $Q$  all have shape  $(N \times d_k)$ .

Each of these queries will create a separate set of attention weights,  $\text{softmax}(q_i^T K)$ .

#### Key Equation 443

We define the **self-attention weight** matrix  $A$ , to represent all attention weights:

$$A = \begin{bmatrix} \text{softmax}\left(\frac{q_1^T K^T}{\sqrt{d_k}}\right) \\ \text{softmax}\left(\frac{q_2^T K^T}{\sqrt{d_k}}\right) \\ \vdots \\ \text{softmax}\left(\frac{q_N^T K^T}{\sqrt{d_k}}\right) \end{bmatrix} = \text{softmax}\left(\frac{Q K^T}{\sqrt{d_k}}\right)$$

This is an  $(N \times N)$  matrix.

- Row  $i$  tells us all of the attention weights applied to query  $q_i$ .
- Col  $j$  tells us the attention weights for key  $k_j$ .
- Element  $\alpha_{ij}$  (row  $i$ , col  $j$ ) tells us, "how **important** is word  $j$  (key) as context for word  $i$  (query)"?

We can use this to get the total attention:

### Key Equation 444

The **self-attention** equation is given as

$$\text{Attention}(Q, K, V) = A V = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

It is a  $(N \times d_k)$  matrix.

Note that the elements in row  $i$  must add up to 1: we have softmax.

This is not true for column  $j$ : they're probabilities for different queries.

Row  $i$  gives the **averaged value vector**  $y^{(i)}$  for the  $i^{\text{th}}$  word, based on all of the surrounding **context**.

- We can write this in element-wise form:

$$y^{(i)} = \sum_{j=1}^N \alpha_{ij} v_j$$

We could view this as the "output" for the  $i^{\text{th}}$  word.

One theme we'll run into, many times in this chapter, is that attention-based models benefit from being able to **parallelize**:

### Concept 445

#### Transformer Parallelization I

Computing self-attention can be strongly parallelized:

- Each  $q_j^T k_i$  term is **independent** of the others: we can compute all of the key-query dot products **at the same time**, rather than waiting for one to finish before starting the others.
- We can compute each softmax term at the same time, as well.

This remains true if we're using **cross-attention**, where the keys and queries come from different words.

### 8.2.9 Positional Encoding

First, a problem we need to address:

- Currently, our key  $k_i$  is determined by asking the **identity** of the word at index  $i$ .
- This key doesn't encode information about the **position** of this word in the sentence.

But clearly, the position of a word will determine its meaning.

- **Example:** "The cat lies on the green table" and "the green cat lies on the table" are **not the same**: moving the word "green" to a different index changes its meaning.

We fix this by adding information to keep track of this position.

#### Definition 446

We apply **positional encoding** to each word embedding: each embedding includes information about the **position** of a word in the text.

- This allows our attention mechanism to use this information when deciding the **relevance** of different words.

### 8.2.10 Masking

One common use for transformer models is **text prediction**: learning what word should come next, based on what it has seen so far.

Typically, we would give our model the text, and give it a chance to try to **predict** each index, **before** it can see it.

We need to prevent our model from being able to **cheat**:

- We don't want our model to be able to see the words it's supposed to be predicting.



So, we'll hide those words, so our model can't see them. In this case, we want our model to predict the next word: "dog".

This is called **masking**.

**Definition 447**

**Masking** is a technique where we **hide** some information from our model, so it can't use that information.

- For example, if our model is being used to **predict text**, we hide the text that it's trying to predict.

However, the word "masking" can apply to **any** situation where we want to hide tokens from the model.

### 8.2.11 Attention Heads

We have a system for "attention": deciding which words provide the **most important** context/information, and paying more attention to those words.

But there's something we haven't considered: the "importance" of different words, depends on **what you're interested in**. Let's consider a couple examples:

- **Syntax:** which words are **subjects, objects, verbs, adjectives?**

**Example:** "The boy kicks the red ball": our focus is on the word "ball".

- "red" is important for color.
- "kicks" is important for knowing what's happening to the ball.
- "boy" is important for knowing who is acting on the ball.

- **Semantics:** which words change the **meaning** of our target word?

**Example:** "I miss her warm smile": our focus is on the word "warm".

- The word "smile" changes the meaning of warm from 'high temperature' to 'kind'.

- **Coreference:** which words are referring to the **same object?**

**Example:** "John said that he isn't hungry": our focus is on the word "John".

- "he" refers to the same object as "John": if we apply something to the word "he", it also applies to "John".

#### Concept 448

What is "**important**" in a sentence can **change**, based on what you're trying to study.

- And generally, these ideas of "important" won't agree with each other.

Above, we suggested several different perspectives on "what is important".

- Rather than having our attention mechanism try to handle all of these kinds of importance, we could create a **separate mechanism** for each one of them.

We'll do just that: each "perspective" will be represented by a different mechanism. We call each of these, **attention heads**.

**Definition 449**

A transformer model may use **multiple** attention mechanisms **at the same time**:

- Each attention mechanism is a different "**perspective**" on our data: it focuses on different aspects of the text (grammar, meaning, tone, etc.)
- To accomplish this, each one represents a word  $w$  with a different  $k$ ,  $q$ , and  $v$ .

We call each mechanism one **attention head**.

If we have 3 different attention heads, each one may **encode** the word "silly" differently. We could have three different keys for this one word:  $k^1$ ,  $k^2$ , and  $k^3$ .

- Each head will require a distinct word encoding:  $K^{(h)}$ ,  $Q^{(h)}$ , and  $V^{(h)}$ .

**Concept 450****Transformer Parallelization II**

Each attention head uses calculations which are **independent** from the others: we can compute each attention head at the same time!

## 8.3 Transformers

Now that we've built up attention, we'll use it to build a **transformer**. We'll assume our transformer uses self-attention, though the math works out similarly even if it doesn't.

### Definition 451

A **transformer block** is a collection of attention heads running in **parallel**, applied to the same text.

A **transformer** is composed of several transformer blocks in **series**: the output of one block is the input of another.

### 8.3.1 How to create embeddings

Something we've ignored for a while is, "how do we construct our **embeddings** K, Q, and V"?

- We aren't actually given them: we're given a sequence of **tokens**: each token is a vector  $x$  representing a word. So, our whole body of text is a matrix  $X$ .

We'll compute each embeddings by using a **linear transformation**:

Each vector is length  $d$ : this is different from the length of the embedding,  $d_k$ .

### Key Equation 452

We use **projection matrices**  $W_k$ ,  $W_q$ , and  $W_v$  to transform each **token**  $x^{(i)}$  into embeddings  $k, q$ , and  $v$ .

$$\begin{aligned} k_i &= W_k^\top x^{(i)} \\ q_i &= W_q^\top x^{(i)} \\ v_i &= W_v^\top x^{(i)} \end{aligned}$$

All three projection matrices have shape  $(d \times d_k)$ .

All of our tokens are stored in matrix  $X$ :

$$X = \begin{bmatrix} | & | & & | \\ x^{(1)} & x^{(2)} & \dots & x^{(N)} \\ | & | & & | \end{bmatrix}^T \quad (8.18)$$

We can get the keys, queries, and values for all of our vectors in matrix form:

Reminder that:

$d$  is the original length of  $x^{(i)}$

$d_k$  is the length after embedding.

Unlike our usual  $X$ , this is transposed: shape  $(N \times d)$ .

**Key Equation 453**

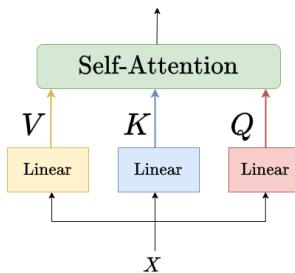
We can compute K, Q, and V:

$$K = XW_k$$

$$Q = XW_q$$

$$V = XW_v$$

Based on this linear transform, we modify our diagram:



We have to generate V, K, and Q before we can use them.

**Concept 454**

Once benefit of computing keys, values, and queries based on **weight matrices** is that we can **train** these matrices:

- Rather than manually designing the **embeddings**, we can allow our model to learn whichever embedding is most useful.

### 8.3.2 Attention Heads

What if we have **multiple** attention heads?

**Notation 455**

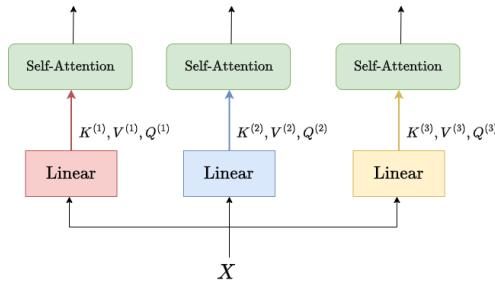
If we have H attention heads in a transformer block we'll indicate the h<sup>th</sup> head with:

$$K^{(h)} = XW_{h,k}$$

$$Q^{(h)} = XW_{h,q}$$

$$V^{(h)} = XW_{h,v}$$

Each attention head is applied in parallel:



Here's an example with  $H = 3$  attention heads. Each uses a distinct set of keys, values, and queries.

To finish off our multi-headed attention unit, we do two more things:

- Transform each token back into the original dimensions: going from length- $d_k$  to length- $d$ .
- Combine the results from each attention head: we'll do a **weighted average**.

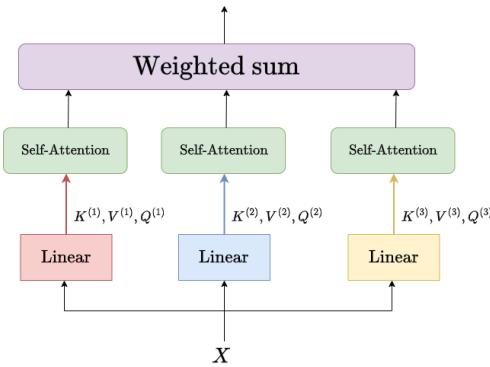
### Key Equation 456

After computing attention for each head, we take a **weighted average** of our heads, combining them together:

- For each head, we use matrix  $W_{h,c}$  to **scale** the weight of each head, and convert them back to their original **shape**.
  - $W_{h,c}$  has shape  $(d_k, d)$ .
- We **add** together the results, gathering information from each head.

$$u = \sum_{h=1}^H \text{Attention}(Q^{(h)}, K^{(h)}, V^{(h)}) W_{h,c}$$

$u$ , the final output of our **multi-headed attention**, has shape  $(N \times d)$ , where the  $j^{\text{th}}$  column represents the  $j^{\text{th}}$  token.



We have our completed multi-headed attention unit!

In the "Attention is all you need" paper, this diagram is analogous to Figure 2 (right).

Instead of directly doing a weighted sum, they concatenate each attention head, and then apply a linear weight  $W^o$ .

These are equivalent.

Note that this is the same shape as our original input,  $X$ :

$$u = \begin{bmatrix} u^{(1)} & u^{(2)} & \dots & u^{(N)} \end{bmatrix}^\top \quad (8.19)$$

In fact, we can compute this multi-headed attention, one  $u^{(i)}$  at a time.

Reminder that  $\alpha_{ij}$  is an attention weight from  $A$ , and  $v^{(j)}$  is a value vector of  $V$ .

### Key Equation 457

We can combine our multi-attention heads as

$$u^{(i)} = \underbrace{\sum_{h=1}^H W_{h,c}^\top}_{\text{Heads}} \underbrace{\left( \sum_{j=1}^N \alpha_{ij}^{(h)} v_j^{(h)} \right)}_{\text{Attention}}$$

This is a nested sum:  
 $\sum_{h,j}(\cdot)$ ,

not a product of two sums,  
 $(\sum_h(\cdot)) \cdot (\sum_j(\cdot))$

### 8.3.3 Residual Connections

Our next component will handle a problem with **deep** neural nets that we've addressed before: vanishing/exploding gradient.

#### Definition 458

(Review from Neural Networks 2)

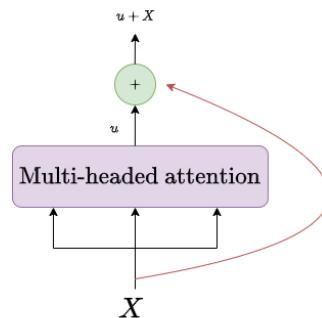
**Vanishing gradient** occurs when a deep neural network ends up with **very small gradients** in the **earlier** layers.

This happens because a deeper neural network has a **longer chain rule**: if all of the terms are **less than one**, they'll multiply into a very small value, "**vanishing**".

This means that our gradient descent will have **almost no effect** on these earlier weights, **slowing down** our algorithm considerably.

In short: the "further away" from our input layer, the messier our gradients get.

One simple solution is to include our original, **unmodified** input, deeper in the neural network: we just **add** it, so that our second layer gets to see the input data, too.



Our output contains direct information about the input. This hopefully improves training.

#### Definition 459

In a **residual block**, the **input**  $x$  is added to the **output**  $F(x)$  of the block (in our case, multi-headed attention).

$$\text{output} = F(x) + x$$

- This is designed to reduce the risk of **vanishing gradient**, by directly exposing deeper layers to the input
  - The long chain rule is what causes vanishing gradient: we've created a shorter chain rule.

If you ever hear someone refer to a "ResNet" or "Residual Network", this is a CNN that uses the same technique!

### 8.3.4 Layer Normalization

Another topic from the NN chapter: **batch normalization**.

#### Definition 460

(Review from Neural Networks 2)

**Batch Normalization** is a process where we

- Standardize the pre-activation for each layer **across data points in the batch** using mean  $\mu_i$  and standard deviation  $\sigma_i$  (for the  $i^{\text{th}}$  dimension).

$$\bar{Z}_{ij} = \frac{Z_{ij} - \mu_i}{\sigma_i}$$

- Choose the new mean and standard deviation for the pre-activation using  $(n \times 1)$  vectors  $G$  and  $B$

$$\hat{Z}_{ik} = G_i * \bar{Z}_{ij} + B_i$$

We would get the same kinds of benefits from **normalization** in transformers as we did before in NNs.

In short: we set the (mean, sd) to (0,1) and then scale it back up to  $(G_i, B_i)$ .

But rather than normalizing across multiple **data points** (batch), we'll normalize across the **features** (layer) of a single token.

Stabilizing our training process, mostly.

#### Key Equation 461

Suppose we have a  $(d \times 1)$  data point  $z = [z_1 \ z_2 \ \dots \ z_d]^T$ .

**Layer normalization** computes the mean  $\mu_z$  and standard deviation  $\sigma_z$  across our **features**  $z_i$

$$\mu_z = \frac{1}{d} \sum_i z_i \quad \sigma_z = \sqrt{\frac{1}{d} \sum_{i=1}^d (z_i - \mu_z)^2}$$

And then **normalizes** them.

$$z_{\text{norm}} = \frac{z - \mu_z}{\sigma_z}$$

Finally, we scale them back up, to have mean  $\beta$  and s.d.  $\gamma$ .

$$\text{LayerNorm}(z; \gamma, \beta) = \gamma \left( \frac{z - \mu_z}{\sigma_z} \right) + \beta$$

Now that we understand this process, we can apply this to our transformer model:

Layer normalization can be used on a single data point, while batch normalization requires many.

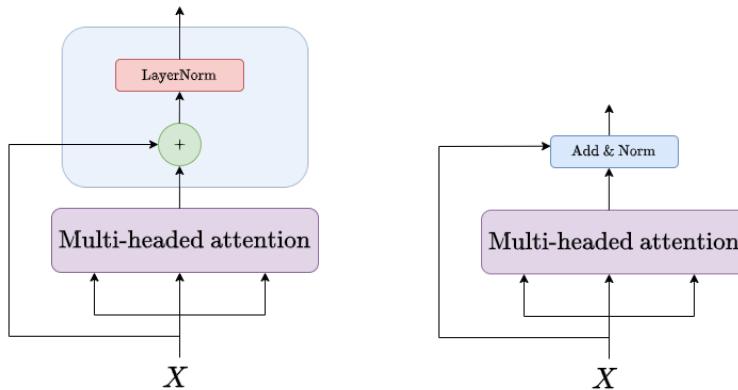
- After we get  $u + x$  (creating the residual block), we use layernorm on each token separately:

### Concept 462

At the end of our **residual block**, we apply LayerNorm to each of our tokens separately

- We take our  $(N \times d)$  object  $u + X$  and normalize the features of each of our  $N$  tokens (shape  $(d \times 1)$ ) **separately**.

$$u_{\text{norm}}^{(i)} = \text{LayerNorm}(u^{(i)} + X^{(i)}, \gamma_1, \beta_1)$$



We append a LayerNorm layer. We'll follow the convention from the "Attention is all you need" paper and combine these into a single unit: "Add+Norm".

With this, our Residual Connection is complete.

### 8.3.5 Feed Forward

In our CNNs, after convolution, we would use a **fully-connected feed-forward network** to analyze the processed data.

- We'll follow the same sort of pattern here: the main difference being that we apply feed-forward after only **one layer** of **multi-headed attention**.

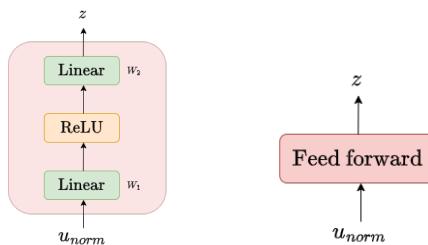
#### Key Equation 463

After we apply **Add & Norm** to our Multi-headed attention, we run the output through a **feed-forward** layer, processing the data it receives.

- We use a linear layer  $W_1$ , a ReLU layer, and another linear layer  $W_2$ .

$$z = W_2^T \text{ReLU}(W_1^T u_{\text{norm}})$$

We can think of this as applying a hidden FC layer to our network, followed by another linear transform.



Linear, ReLU, linear. Once again, following "**Attention is all you need**", we simply call this the "Feed forward" Layer.

We'll follow this up with another LayerNorm:

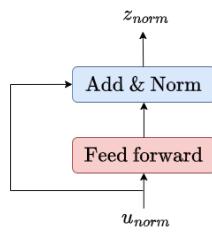
Meaning, we use another residual block.

#### Concept 464

After our **feed-forward layer**, we apply **Add & Norm** again.

$$z_{\text{norm}}^{(i)} = \text{LayerNorm}(z^{(i)} + u_{\text{norm}}^{(i)}, \gamma_2, \beta_2)$$

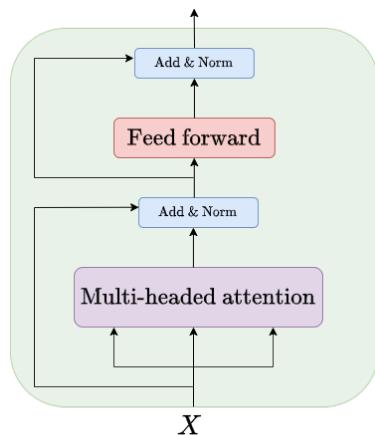
This is the final output of our **transformer block**.



$z_{norm}$  is the final result of our transformer block.

### 8.3.6 Transformer Block

With this, we can assemble our transformer block, top to bottom:



We have a transformer block!

#### Definition 465

A **transformer block** is made up of several functions composed together:

- **Multi-headed attention**
  - Each head encodes the input text  $X$  as keys  $K^{(h)}$ , a value  $Q_h$ , and vectors  $V^{(h)}$ : one for each token.
  - Based on these, we compute **attention**.
  - Finally, we linearly combine information from across all  $H$  heads.

- **Add & Norm**

- **Feed-forward**

- We apply a fully-connected layer (linear+ ReLU), then another linear unit.

- **Add & Norm**

Both "Add & Norm" layers accomplish the same thing: they create a **residual connection**.

- We add the input to the output, and then layer normalize.

**Concept 466**

Each layer of our transformer block serves an important function:

- The **multi-headed attention** layer explores connections between tokens, and provides information about the internal structure of our data.
- The **feed-forward** layer processes our information nonlinearly (via ReLU).
- The **add & norm** layers create residual connections between the input/output of the preceding layer, improving our gradient-training process.

From here, we can design a transformer model by combining many of these transformer units in series.

### 8.3.7 Translation Task: training

We just have one more layer of complexity, before we finish. Let's consider a training example, for the task of translating from english to spanish.

I'm not hungry yet  $\Rightarrow$  Todavía no tengo hambre

Our transformer will start by predicting the **first word** in the sentence: presumably "todavía".

- But not necessarily: if our model isn't **well-trained** yet, it might predict some random word, like "espacio".

Now, we want to predict the **second word** in our output. But we just brought up an important problem:

It's also possible for us to have multiple valid translations, but we'll ignore that for now.

- The best "second word" in our translation is **dependent** on the first word. We should factor that into our model, when predicting the second word.
- If our first word was wrong, then we're **more likely** to use an incorrect second word!

The solution? Instead of using the first word we predicted, we use the **correct** first word.

- Only one condition we need to remember: we need to **mask** the rest of the "correct" output sentence, so our model can't use it to cheat.

#### Concept 467

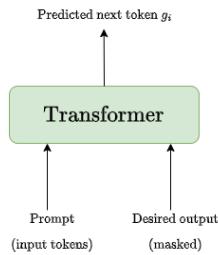
When **training** our model to complete a language task, our model predicts each word (token) one-by-one, based on two pieces of data:

- The entire **input prompt**
- The **desired output** sequence for every token **before** the one we want to predict.

**Example:** Suppose we're predicting the third word in our above sentence. We'll use the first two "correct" words as part of our model:

In this case, "tengo" is the word we want to predict.

$$\begin{bmatrix} \text{I'm not hungry yet} \\ \text{Todavía no} \end{bmatrix} \Rightarrow \text{tengo}$$



Our model is actually trained with *two* inputs.

Something to take note of: we predict the  $i^{\text{th}}$  token based on the input, and the first  $i - 1$  **desired inputs**.

- That means that, when predicting token  $g_i$ , we don't care what we predicted for the previous tokens!
- We don't need to finish predicting token  $i$  to predict token  $i + 1$ : we can do them at the same time!

#### Concept 468

##### Transformer Parallelization III

Predicting token  $i$  is an independent calculation from predicting a second token  $j$ .

- That means we can predict every token in our sentence at the same time!

This is a *huge* advantage in training transformers: it can essentially think about the entire sentence at the same time, massively speeding up training.

#### Clarification 469

We can't parallelize **token generation** when we're using our model **after** training:

- We can parallelize during training because we're using the **desired output** for the previous  $i - 1$  tokens.
- When using our model for unseen data, we don't have "desired output": we have to use our **actual output** for the previous tokens.

We have to **wait** for our model to predict the first  $i - 1$  tokens, before it predicts the  $i^{\text{th}}$  token.

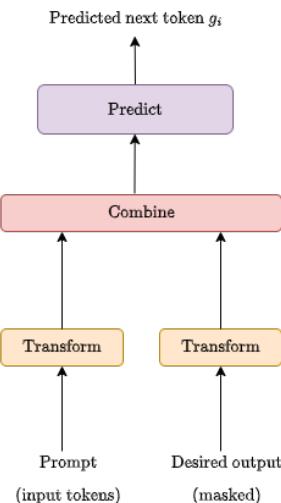
### 8.3.8 Encoder + Decoder Structure

Now, we have to structure our transformer model to be able to handle both the input and desired output.

#### Concept 470

There are three tasks we want our model to complete:

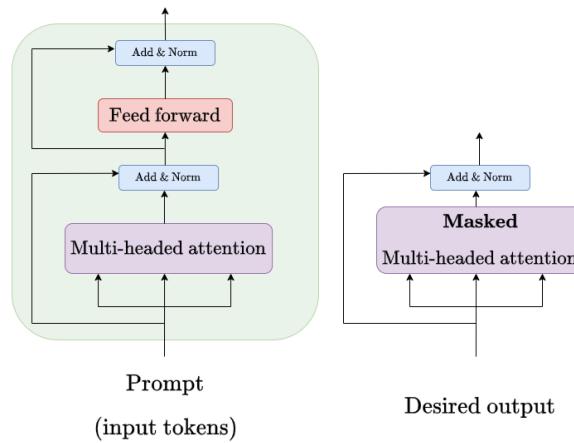
- **Process** our input sequence,
- **Process** our target sequence (desired output),
- **Combine** the two sequences of information
- **Predict** the next character based on this data.



We'll choose functions to handle each of these tasks.

- We'll process our prompt using a complete transformer block. This unit is our **encoder**: we encode our prompt in a form that is more **meaningful** to our computer.
- However, for our desired output, we'll only use **attention**, learning about the internal structure of the output.
  - We'll also use this unit to **mask** our output (so our transformer can't "look ahead" at future tokens).

Why not add the feed-forward layer? We'll add it later: there's another component we want to add first (see below).



We'll add the second feed-forward unit later. **First**, we want to **combine** information from our input, with the earlier tokens of our output.

We accomplish this with another attention unit: this time, we'll use **cross-attention**.

#### Definition 471

In **cross-attention**, our **queries** come from one sequence of text, while our **keys/values** come from a **different** sequence of text.

- As opposed to **self-attention**, where our keys/values/queries all come from the same sequence.

Our goal is to use the **earlier** part of our **output sentence** to determine which parts of our **input** we should **pay attention to** in our input sentence, when choosing the next token.

- Our **keys/values** represent the words we might want to **pay attention to**.
- Our **queries** help us decide **what** to pay attention to.

For example: if our output sentence already includes a word, it might be less likely we'll need to use that word again.

We can use "attend" as a verb meaning "pay attention to": this is common when talking about transformers.

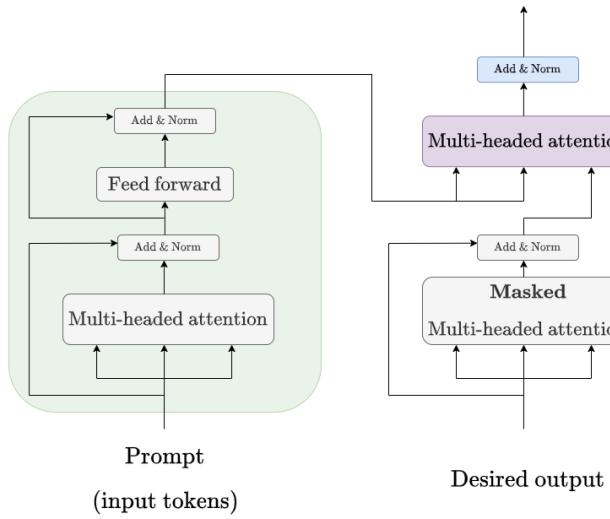
For example, in this case, we're deciding "which input tokens to attend to".

#### Concept 472

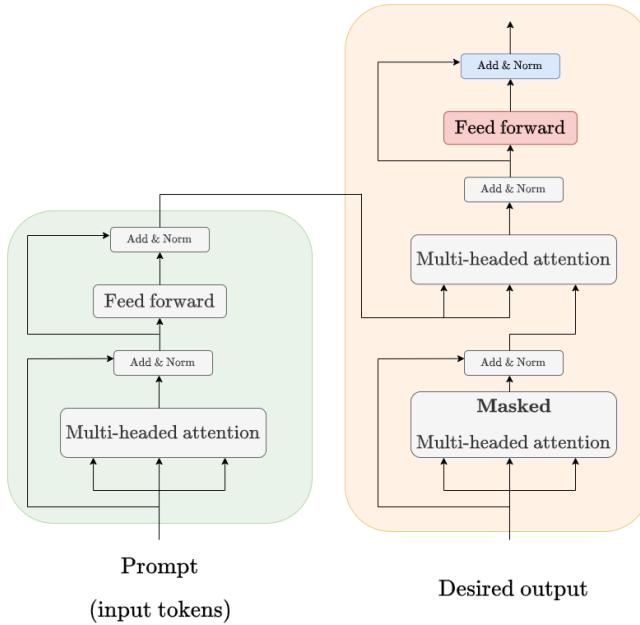
We integrate our **input tokens** and our previous **output tokens** using **cross-attention**: we apply attention, using

- Our encoded input as our **keys and values**.
- Our attended output as our **queries**.

This is our **encoder-decoder layer**.



Now that we've integrated information from both our input and output, we finally include our **feed-forward** unit: we'll process our integrated information.



This unit on the right is called our **decoder**.

We've got a complete encoder/decoder setup:

**Concept 473**

We break our transformer into an "**encoder**" and "**decoder**" unit:

- The encoder transforms our **input** into a representation that contains more useful information: connections between tokens in the prompt, etc.
- The decoder transforms that encoding into a **output**/response: this decoder takes the information we've gathered, and applies it to our problem.

Consider the translation example:

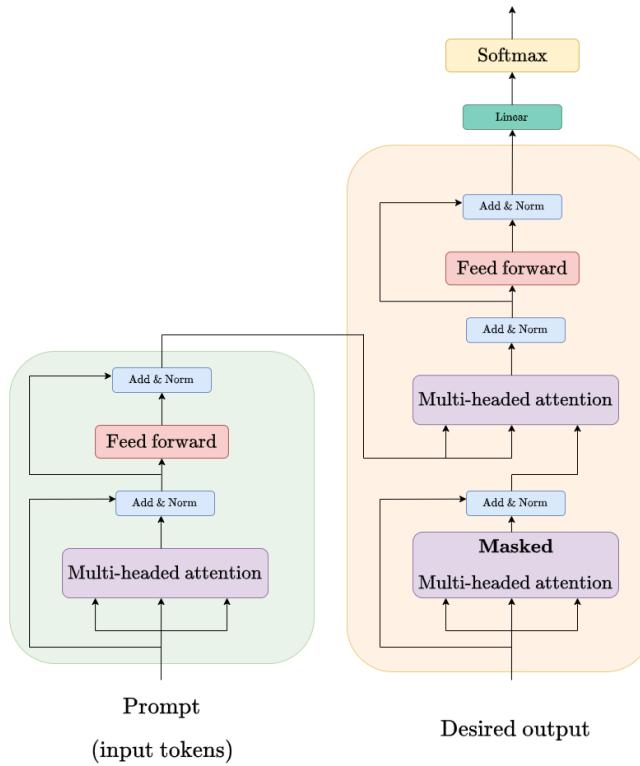
- The encoder stores our English text in a form that hopefully represents the **meaning**.
- The decoder "decodes" that representation into a form we can read, but in a **different language**: Spanish, in our example.

In this analogy, we've created a special "code" that we write in English, and read in Spanish.

### 8.3.9 Predicting a token

Only one step left: using this decoded information, we need to **choose** our token. This is the multi-class classification problem: we use the same protocol as we always do.

- We linearly transform our data: each token gets a "**score**", based on how likely we think it is to be the correct one.
- We apply **softmax**, to turn these scores into probabilities. We get a probability for every possible token.



This is essentially a completed transformer.

This is the (now-famous) diagram from the "Attention is all you need" paper! [\\_\\_\\_\\_\\_](#)

Only one detail still missing:

- Our decoder/encoder typically has **several copies** of the same unit in a row: for example, we might have 3 transformer blocks in a row for our encoder.

Notably, this is only one kind of transformer model: which architecture we use depends on the problem, cost constraints, etc.

We've excluded the initial embedding (turning words into vectors) and positional embedding (adding information about the position of each word in the sentence).

We could include those for completeness, but that would just take up more space.

### 8.3.10 Training Process

We typically train transformer models in two stages: pre-training and fine-tuning.

#### Definition 474

In **pre-training**, we expose our model to a very large dataset of human language, so it can learn **patterns** in that language.

- We can use **unlabelled** data in this stage: thus, we have an unsupervised/self-supervised problem.

This stage of training is typically expensive.

#### Definition 475

In **fine-tuning**, we take our pre-trained model, and train it for a **specific task**.

- We use **labelled** data in this stage.

It tends to be much faster and less expensive than pre-training.

### 8.3.11 Variations

We could make variations on this network:

- Use more/fewer decoder/encoder units.
- Use a different style of attention (rather than the dot product, we use some other similarity metric).
- Move LayerNorm to different parts of the network.

## 8.4 Terms

- Natural Language Processing (NLP)
- (*Review*) Convolutional Neural Networks (CNNs)
- Locality
- Recurrent Neural Networks
- (*Review*) Word Embedding
- Co-occurrence
- Context window
- Skipgram
- Word2vec
- Token
- Key Vector
- Query Vector
- Value Vector
- Attention Weights
- $d_k$
- Attention
- Self-attention
- Positional Encoding
- Masking
- Attention Head
- Projection Matrix
- Multi-headed attention
- Residual Block
- Residual Connection
- Layer Normalization
- Add & Norm
- Feed-forward layer (transformers)

- Transformer Block
- Cross-attention
- Encoder (Transformers)
- Decoder (Transformers)
- Encoder-Decoder Layer
- Pre-training
- Fine-tuning

# CHAPTER 9

---

## Non-parametric Methods

---

### 9.0.1 Parametric Methods

We've spent a large part of the course on models that rely on **parameters**:

- Linear regression/classification models, with parameters  $\Theta = (\theta, \theta_0)$ :

$$h(x; \Theta) = \theta^T x + \theta_0 \quad (9.1)$$

- Neural networks, with weights  $W^\ell$  and  $W_0^\ell$ :

$$A^\ell = f(Z^\ell) \quad Z^\ell = (W^\ell)^T A^{\ell-1} + W_0^\ell \quad (9.2)$$

These can be thought of as "parametric" methods:

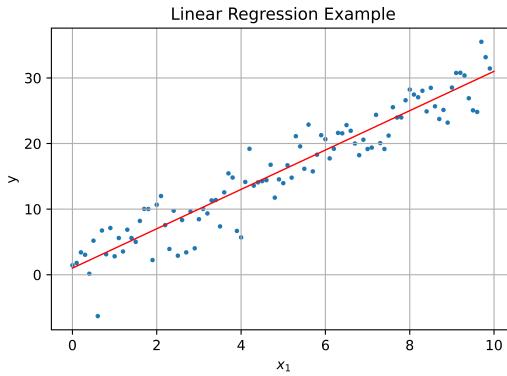
**Definition 476**

**Parametric methods** are those with a **fixed number of parameters**.

- Typically, we optimize these models by **changing** those parameters.

We can also phrase these as "models with a particular, known mathematical form".

**Example:** The equation  $y = mx + b$  has the **fixed** parameters  $m$  and  $b$ . It assumes our data will follow the "mathematical form" of a line.



This kind of model works best if the "true distribution" follows a similar shape.

Neural networks have the advantage of being very **expressive**: they can express **many different** kinds and distributions of data.

- The same NN architecture can be re-used to make tools for multiple different tasks.

These methods have been our primary focus, because they're very configurable, and have a wide range of applications.

But they aren't our only option.

And if our problem is too complex, we can systematically increase expressiveness: increase layer size and depth.

We should be careful adding layers blindly, though.

## 9.0.2 Non-parametric methods

We can consider models that are more *flexible*. Rather than assuming a particular structure, we can discover patterns, directly from our data.

### Definition 477

**Non-parametric methods** exclude models with a **fixed number** of parameters.

Instead, it includes:

- Models with **no parameters**
- Models with a **variable number of parameters**, depending on the data.

The name can be misleading: non-parametric models *can* have parameters!

As suggested in our definition, these methods often base their structure on the data they receive.

Here are some **examples**, each with a unique, **data-based** model structure:

- **k-means clustering**: We already discussed this in the **Clustering** chapter!

Note that k isn't a parameter: it's a *hyperparameter*.

- We want to cluster our training data as tightly (low-variance) as possible.
- **Nearest neighbor:** We predict the output  $g^{(i)}$  of a data point  $x^{(i)}$ , based on the **nearest** points of training data.
  - In this case, we **don't have** a model at all: we just directly use our data.
- **Tree models:** We **split** up our space into smaller pieces. Each region of inputs is assigned an output  $y$ .
  - We divide up space to get good accuracy on training data.

Reminder:  $y^{(i)}$  is the true value,  $g^{(i)}$  is our prediction.

And here are some examples that **combine** many simpler models, in a non-parametric way:

- **Ensembles:** We train multiple models on the whole data set, and we **average** them.
  - By combining multiple models, they'll (hopefully) average out to being more accurate, reducing estimation error.
- **Boosting:** in boosting, we use multiple models **consecutively**: we train one model, and then we use our second model to try to improve on the mistakes of the first.
  - If an earlier model struggled with a data point, that data point is **weighted more heavily**: future models will focus more on that mistake.
  - We will not discuss boosting further.

We'll specifically refer to "bagging" in this chapter.

### 9.0.3 Why learn about non-parametric methods?

Of course, neural networks are incredibly popular for a reason: they're **effective** at what they do.

So, why do we need non-parametric methods? Well, they come with several major benefits:

#### Concept 478

**Non-parametric methods** can have genuine benefits over parametric, **neural network** models:

- Fast to **implement**, few hyperparameters to tune.
- Often **human-interpretable**, easier to understand than a neural network computation.
- In some circumstances, **just as well or better** than neural networks, despite their relative simplicity.

## 9.1 Nearest Neighbor

Suppose that you're trying to figure out how to approach a problem. Maybe a medical problem, or just a personal situation with a friend.

- One question you can ask yourself is, "what's the most **similar** situation I can think of?"
- If that's not enough, you could think of 2, or 3, similar examples, and try to **guess** from that, what the best course of action is.

This is the basic idea of the **nearest neighbor** approach: we have a new data point. We want to use the most similar examples from our past **training data** to make a judgment.

- We judge the most relevant/"similar" data based on lowest **distance**: we call this the "nearest neighbor".

What's interesting is that we're **not** developing a model: we're directly using our training data to make predictions.

In the simplest variation, we only choose the **single closest data point**.

### Definition 479

**Nearest neighbor method** is a **non-parametric** method where we predict output  $g^{(i)}$  based on the **nearest** data point (**nearest neighbor**) in our training set.

- Our assumption is that nearby data points are **similar** to the situation we're currently dealing with.
- So, they should have similar **outcomes**.

This method works **exactly the same** for regression and classification:

- Whatever output  $g^{(i)}$  we find for the nearest neighbor, is our **prediction** for our data point.

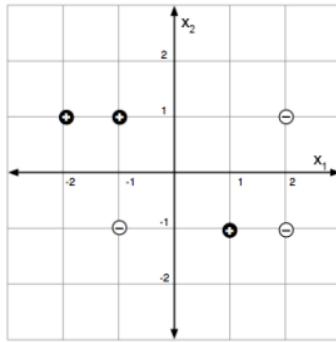
One nice benefit of nearest neighbor is that we require no training: we just check the training data, to make a judgment.

### Concept 480

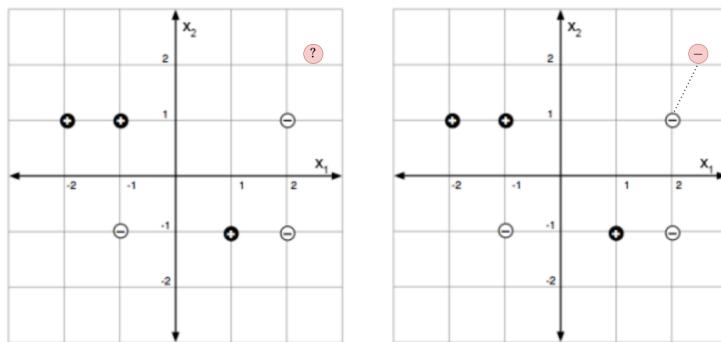
**Nearest-neighbor models** don't require training: you directly reference the training data to come to answers.

### 9.1.1 Nearest Neighbors: An example

Consider this classification problem.

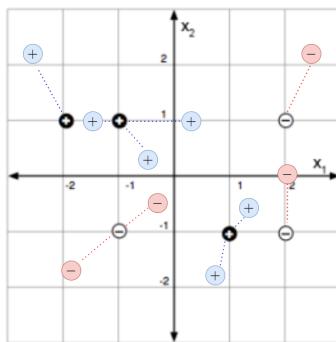


Rather than use a linear model, we'll assign any data point to the **same class** as whichever data point it's closest to.



This red data point is closest to a negative data point. So, we'll assign it negative.

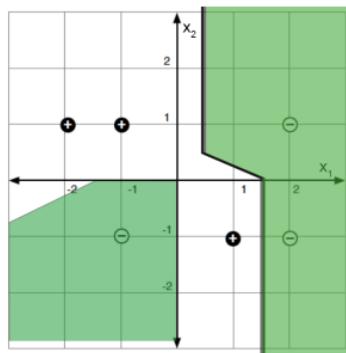
We could apply this to as many data points as we want:



We can see all of the data points that we assign our nearest neighbor to.

We're starting to fill the space: we see that some whole "regions" are positive or negative.

- We can even depict that: we'll highlight the regions which are labelled positive vs. negative.



We can see all of the data points that we assign our nearest neighbor to. Negative regions are green.

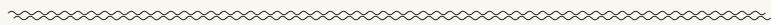
### 9.1.2 Simplified Voronoi Diagram (Optional)

This kind of diagram lets you classify data very quickly. It would be useful to be able to draw:

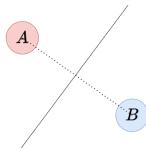
#### Concept 481

To help with finding **nearest-neighbor classification boundaries**, it can be helpful to draw a line **halfway** between opposite-label data points, A and B.

- This is where the distance to either data point is **equal**.
- We decide output based on closeness to A or B, so this is a **decision boundary**.

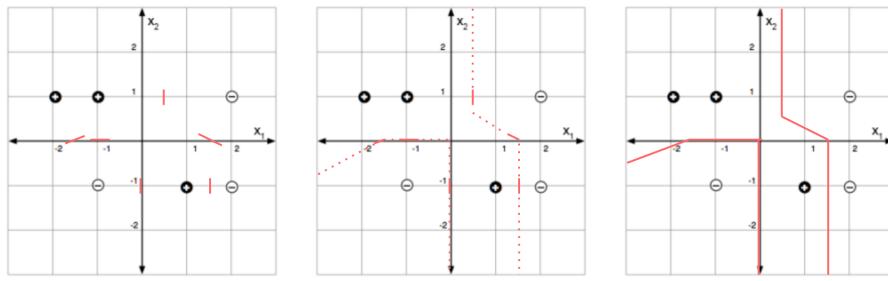


In order to split the space between "closer to point A" and "closer to point B", we'll draw a boundary line **perpendicular** to the line from A to B.



We drew a solid line, where the entire line is equally far from points A and B. Notice that it is 1. halfway between them and 2. perpendicular to the line from A to B.

If we apply this to our previous problem:



We just have to draw our perpendiculars, and extend them.

The resulting diagram is a simplification of a [Voronoi diagram](#).

#### Clarification 482

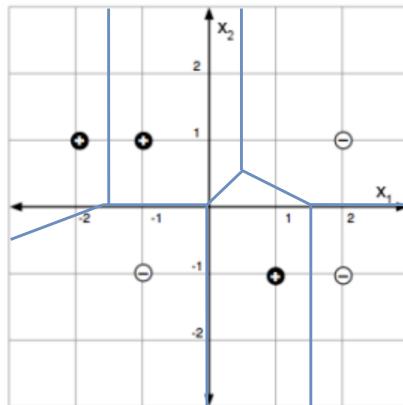
This system is for **nearest neighbors**.

You **CANNOT** use it for  **$k$  nearest neighbors** (discussed below).

If you try to use it this way, I will be sad.

Why is this a "simplified" voronoi diagram?

- In a real voronoi diagram, every single data point gets its own tile of "closer to this point than every other".
- In this one, we've simplified, so all of the + and - classifications join into one region.



Here's a "complete" voronoi diagram: each region represents all data which are closest to that particular data point.

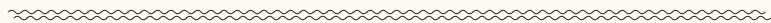
### 9.1.3 k-Nearest Neighbors

If one neighbor is too simplified, we can also use **multiple** neighbors: our number of neighbors is  $k$ .

#### Definition 483

In the **k-nearest neighbors** (kNN), we use  $k$  ( $k \in \mathbb{N}$ ) to give us the "size" of our neighborhood:

- When making a prediction, we only focus the  $k$  nearest data points.
- We ignore all data points beyond the  $k^{\text{th}}$  one.



This same approach can be applied to both **regression** and **classification**:

- In **regression**, you would **average** the output of those  $k$  nearest neighbors.
- In **classification**, you would pick the **majority**: the most common label of your  $k$  nearest neighbors.

**Example:** Suppose your  $k = 4$  nearest neighbors had output values,  $y = (3, 4, 5, 6)$ . You would predict your output by simply averaging them:

$$\frac{3 + 4 + 5 + 6}{4} = 4.5 \quad (9.3)$$

One reason to use k-nearest neighbors is that "nearest neighbor" ( $k = 1$ ) might **overfit**.

#### Concept 484

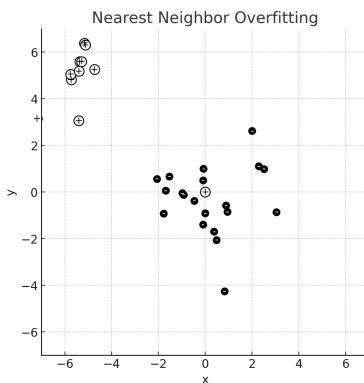
Having a low  $k$ -value could be seen as **overfitting**:

- If you're closest to an **outlier** data point, you'll **choose** that value, rather than the general trend you see with more data.

So, you're sensitive to **noise** in the data, if you happen to be too close to that noise.

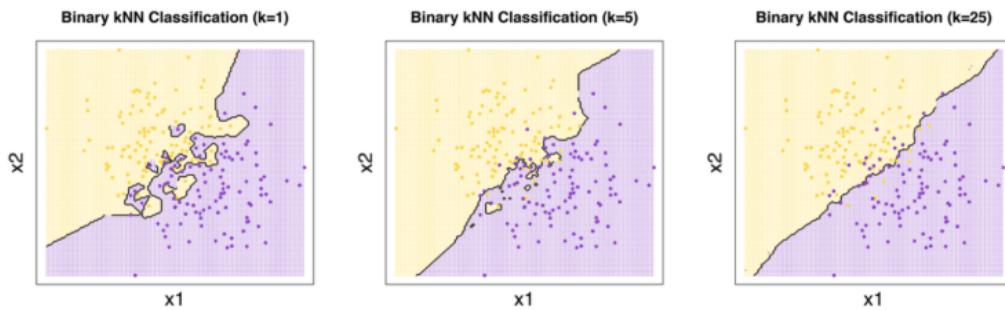
- With a larger  $k$ -value, you're **averaging** over more data points: this can reduce the effect of noise.

In other words: you're not creating a general pattern based on the data structure: you're closely matching the training data.



Suppose we happened to have a data point right next to (0,0). Based on the whole region, the result is probably a negative (-). But with  $k = 1$ , we would assume it was positive.

What happens with different  $k$ -values? Here's a different dataset to test this out on:



We can see that we get a more "general" pattern as  $k$  increases. (Credit to [blog.eduonix.com](http://blog.eduonix.com))

#### Concept 485

On the opposite end, having a **high  $k$ -value** makes it more difficult to see more detailed trends, or fit well to **complex** data.

- If the data really does have a complex shape, but our  $k$  is **too low**, we won't be able to match it.

In other words, we can "underfit" as well.

**Example:** In the extreme case: imagine that  $k$  equals the size of our training data. That means, you always include **all of your training data** to make a judgement.

- No matter what input you give, you'll **always** get the same output: the average of all data.
- That's not going to be very predictive.

### 9.1.4 Locally weighted regression

Another technique is *locally weighted regression*, where we use the local region to create a small regression model:

#### Definition 486

In **locally weighted regression**, we take the  $k$  nearest data points, and **fit** a regression model to only those data points.

We might **weigh** closer data points more heavily than those which are further away:

- Meaning, they're more important to the fitting process.

### 9.1.5 Tradeoffs

One major concern for  $k$ -nearest neighbors is that it's a pain to compare a new data point to the **entire training set**, to find the closest data points.

- So, it's important to use data structures (e.g., ball trees) that make it **easier** to quickly find possible "nearest neighbors".

On the other hand, it's relatively easy to **interpret** nearest neighbors:

- If you want to know why you got the answer that you did, you can directly view your "**nearest neighbors**": these exact data points gave you your answer.
- This makes it easier to check for strange **outliers**, or interesting patterns.

### 9.1.6 Distance metrics (Optional)

What do we mean when we say that two data points are "close" or "far away"?

- This requires us to have some kind of way to measure **distance**: a **distance metric**.
- Above, we were using the **euclidean** distance metric.

#### Definition 487

A **distance metric**  $d$  is one way of measuring the total distance between two data points.

It must follow three basic rules:

- The distance from  $x$  to **itself** is 0.

$$d(x, x) = 0$$

- Distance is always **positive**.

$$d(a, b) > 0$$

- The distance from  $a$  to  $b$  is the **same** as the distance from  $b$  to  $a$ .

$$d(a, b) = d(b, a)$$

- **Triangle Inequality:** A **direct path** from  $a$  to  $c$  is always the **shortest** – taking a detour from  $a$  to  $b$  cannot be faster.

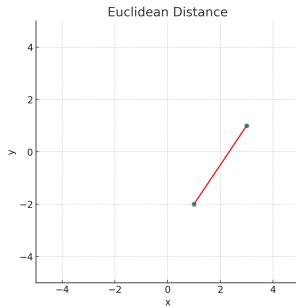
$$d(a, c) \leq d(a, b) + d(b, c)$$

Different metrics may be useful for different situations, or different data types.

Here's a few common distance metrics:

- **Euclidean** distance: the "shortest path" distance in space.

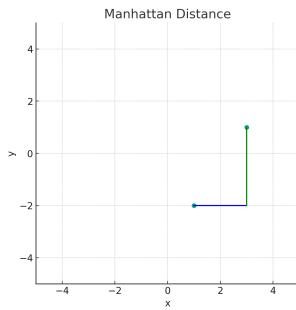
$$d(a, b) = \sqrt{\sum_i (a_i - b_i)^2} \quad (9.4)$$



Here, we have  $\sqrt{(3-1)^2 + (1-(-2))^2} = \sqrt{13}$

- **Manhattan** distance: the "shortest path", while only moving along one axis at a time.

$$d(a, b) = \sum_i (|a_i - b_i|) \quad (9.5)$$



Here, we have  $|3-1| + |1-(-2)| = 5$

- **Hamming distance**: if we have two binary codes, how many digits are different between them?

$$d(a, b) = \sum_i (\mathbb{1}(a_i \neq b_i)) \quad (9.6)$$

$$a = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} \quad b = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad \Rightarrow \quad d(a, b) = 2 \quad (9.7)$$

And as a bonus:

- **Minkowsky** distance: a generalized version of the euclidean/manhattan distance:

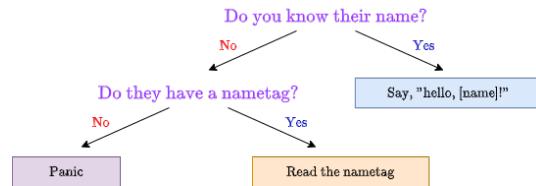
Notice that, for  $p = 1$ , it's equal to manhattan. For  $p = 2$ , it's equal to euclidean.

$$d(a, b) = \left( \sum_i (a_i - b_i)^p \right)^{1/p} \quad (9.8)$$

## 9.2 Tree Models

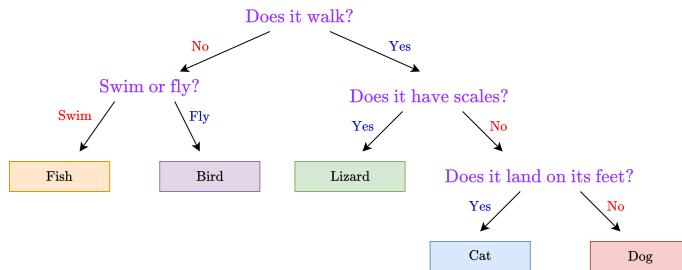
Here, we'll create another algorithm, based on a common way that humans solve problems.

- You might ask a series of **questions**, to narrow your situation down to a more specific example, that you know how to solve.



This tree answers the question, "how to start a conversation?". Panicking may not be the optimal strategy, but it's what this tree advises.

Each question is simple: a **binary**, "yes or no" question. But with enough questions, you can get pretty specific classifications:



A tree for classifying household pets. A bit too simple to be fully accurate.

### Definition 488

A **binary tree** is a structure where:

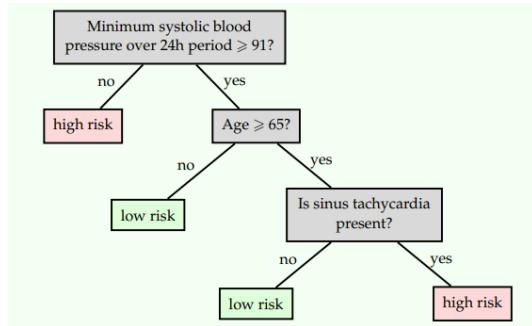
- You start at a **root node**: the **top** of the tree.
- At each node, you **split** your data into two "branches", based on a **binary** (yes-or-no) question.
- You **terminate** at a **leaf node**: each leaf node stops branching, and returns an output  $y$ .

The above diagrams show a strength of tree diagrams: they are **very interpretable**.

### Concept 489

**Tree models** tend to be more interpretable than most other types of models.

They're so interpretable, that they can even be used for day-to-day problems, like medical analysis:



Reproduced from Breiman, Friedman, Olshen, Stone (1984).

However, they tend to work best on data with a small number of input dimensions, or at least a small number that matter.

#### Concept 490

**Tree models** tend to work best on datasets with:

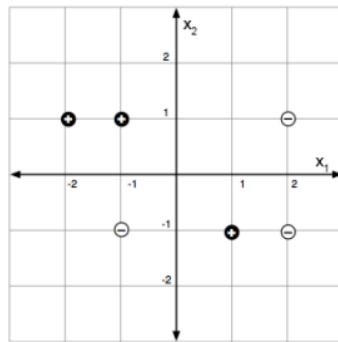
- Low dimensionality
- A small number of individual dimensions contain a lot of information

Otherwise, it can take a huge number of splits to get a productive result.

### 9.2.1 An example in 2D space

The examples we've shown so far have all been abstract, categorical questions. Now, let's consider an example where we have a **continuous** input space, 2D space.

We'll re-use the plot from the nearest-neighbor section:



Our goal is to split the space up, so that we can classify more easily.

This is the type of problem we'll deal with for the rest of the chapter: **n-dimensional, real-valued** space.

How do we want to split up our space? For simplicity, we'll only use **one axis** for each split.

**Definition 491**

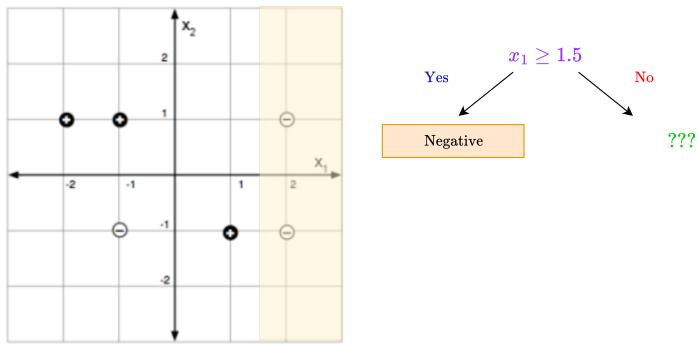
Our **tree-generating algorithm** will split the data along **one axis at a time**:

$$x_i \geq C \quad \text{or} \quad x_i < C$$

Let's give an example: what split would help you classify data?

- The two rightmost data points are both **negative**. So, let's separate them from the rest of the data:

$$x_1 \geq 1.5 \tag{9.9}$$

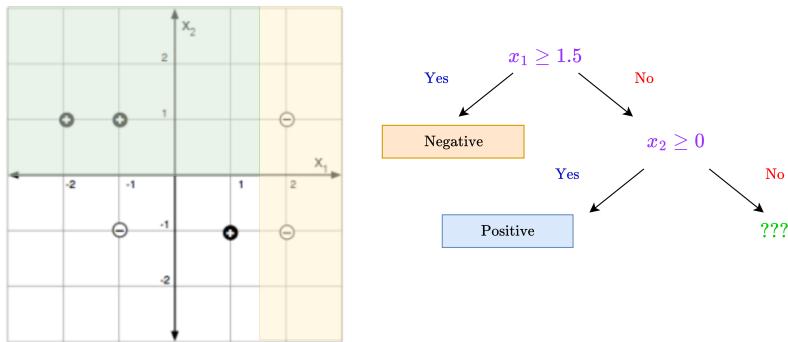


We've split our dataset once.

The right side is taken care of: everything is **negative**. Let's take our first attempt at splitting the left side.

- The top two data points are both **positive**.

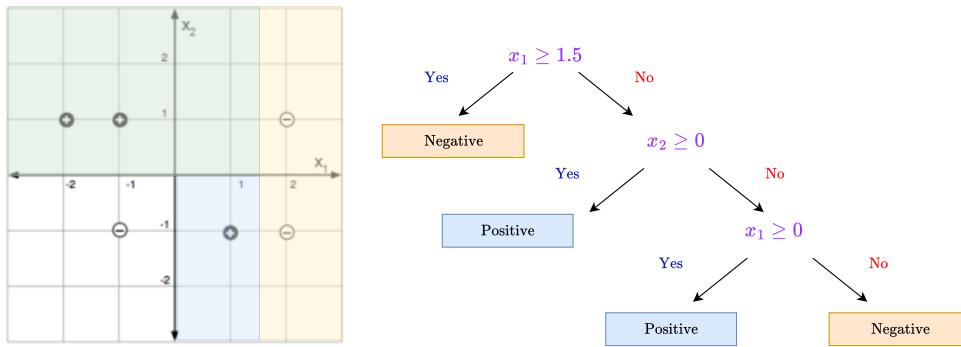
$$x_2 \geq 0 \quad (9.10)$$



Another split.

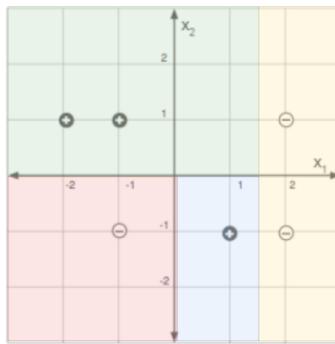
We only have to make one more split: between the positive and negative data point.

$$x_1 \geq 0 \quad (9.11)$$



And we're finished!

- Let's color in our last region.



### 9.2.2 Partitioning: Formalizing our Tree

In the above example, we've broken up our input space into chunks, or **partitions**.

#### Definition 492

Our tree is used to "partition" our data into multiple different chunks.

- Thus, we call one of these chunks, a single **partition**. Partitions are the "**leaf nodes**" of our tree.
- If we gather all of our partitions together, they should cover the **whole space**.

Each of these partitions is assigned a single **constant**  $O_m$ : every data point in that partition is given this constant as an output.

$$g^{(i)} = O_m$$

**Example:** Each of the differently colored regions above is one partition. The red and yellow partitions are assigned "negative", while the green and blue ones are assigned "positive".

Now that we understand how trees work, we can **formalize** our process, mathematically.

Our tree does two things:

- Put each data point in **one partition**  $R_m$ .
- Give each partition an **output value**  $O_m$ .

This is one of our regions, after splitting up the space.

- This is the output we'll use for any data point in this partition.

These two parts make up our **predictor**.

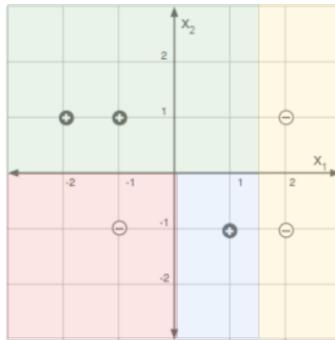
### Definition 493

Suppose our tree creates  $M$  distinct partitions. The **predictor** generated by our **tree** has two main parts:

- A **partition function**  $\pi$ : this function assigns each point  $x$  in the **input space** to a **partition**  $R_m$ .
- A **collection of outputs**  $O$ : the  $m^{\text{th}}$  **output**,  $O_m$ , is assigned to **all points**  $x$  in region  $R_m$ .

Each of our training data is assigned to one partition, by the partition function.

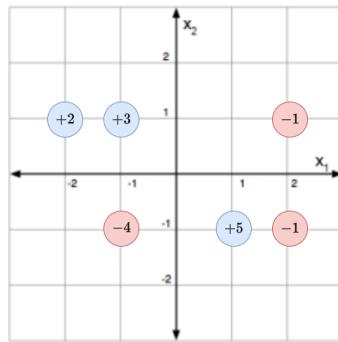
$$\pi(x^{(i)}) = R_m \implies x^{(i)} \in R_m \quad (9.12)$$



The data point at  $(2, 1)$  is assigned to the yellow partition. Since we created it first, we could call the yellow partition  $R_1$ . If we classify either  $-1$  or  $+1$ , we should choose  $O_1 = -1$ .

### 9.2.3 Regression

Now, we want to show how to measure, and create this tree. We'll start with the problem of **regression**.



For regression, our values aren't categorical: they're real numbers. We used integers here, but we could've used 2.5 or  $-\sqrt{2}$ .

What output value do we give for  $O_m$ ? Typically, the most accurate option would be the **average** of all outputs  $y^{(i)}$  for data points in region  $R_m$ .

- But we only want to include the data points  $(x^{(i)}, y^{(i)})$  where  $x^{(i)} \in R_m$ .
- To simplify things, we'll refer to each data point by its index.

#### Notation 494

Each training data point is referenced by its **index**  $i$ . So, rather than **partitioning**  $x^{(i)}$ , we'll partition indices  $i \in I$ .

- The training data which are **included** in  $R_m$ , will have their indices included in  $I_m$ .

$$x^{(i)} \in R_m \iff i \in I_m$$

We can use this to take our average.

**Definition 495**

In **regression**,  $O_m$  is the **average** of all outputs  $y^{(i)}$  for training data in **partition**  $R_m$ .

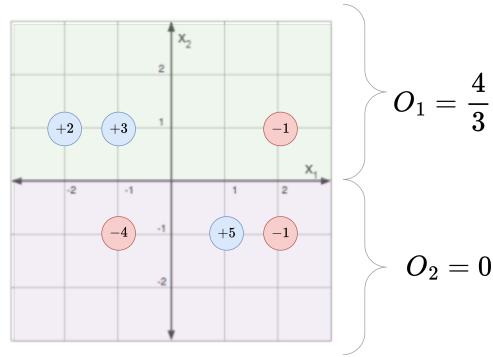
- So, we only include the data points  $x^{(i)}$  relevant to  $R_m$ .

$$O_m = \text{Average}_{i \in I_m} (y^{(i)})$$

or,

$$O_m = \text{Average} \left( \begin{cases} y^{(i)} & \text{if } x^{(i)} \in R_m \\ \end{cases} \right)$$

**Example:** Consider the following split.



We create a split at  $x_2 = 0$ . For each region, we average the value of all elements.

### 9.2.4 Regression Loss

How do we measure our loss? Same as usual for regression: we use **squared error**.

**Definition 496**

The **regression loss** for our region  $R_m$  is given by the **squared error** between our guess and the answer.

- We guess  $O_m$  for every data point in region  $R_m$ .

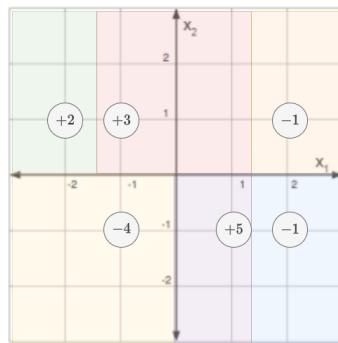
$$E_m = \sum_{i \in I_m} (O_m - y^{(i)})^2$$

To get our loss, we could add up this loss over all of our regions.

If you want to practice with the above example, the error is  $152/3$ .

$$\mathcal{L}_{\text{err}} = \sum_{m=1}^M E_m \quad (9.13)$$

But we have a concern: **overfitting**. What if we create too many regions? That wouldn't be very helpful.



If we wanted to, we could create a partition for every **single data point**. 100% accuracy.

This especially becomes a problem as we get a **larger dataset**: having a partition for every piece of training data will make you incredibly **sensitive to noise**.

So, we want to *discourage* this kind of behavior.

#### Concept 497

To reduce **overfitting**, we'll include a **regularization term** that discourages having too many regions.

- Our number of regions is  $M$ , so we want to **penalize** this: we'll add it to our loss function.

$$\mathcal{L}_{\text{reg}} = \lambda M$$

$\lambda$ , similar to ridge regression, is used to indicate how strongly we want to **regularize**:

- Too high  $\lambda$  can result in **underfitting**: we get structural error, not splitting enough to accurately represent our data.
- Too low  $\lambda$  can result in **overfitting**: we split more than we need, focusing on noise in the data.

Combining these, we find our loss function for tree-based regression.

**Key Equation 498**

The **objective function** for tree-based regression has two parts:

- Loss  $\mathcal{L}_{\text{err}}$ , telling us how **inaccurate** our predictions are.
- Regularization  $\mathcal{L}_{\text{reg}}$ , penalizing us for having **too many splits**, and overfitting.

$$J = \mathcal{L}_{\text{err}} + \mathcal{L}_{\text{reg}} = \sum_{m=1}^M E_m + \lambda M$$

Where

$$E_m = \sum_{i \in I_m} (O_m - y^{(i)})^2 \quad O_m = \text{Average}_{i \in I_m} (y^{(i)})$$

It's possible to search all partitions of our training data, and find the best one directly.

- But this is NP-hard. All you need to know is that it's incredibly time-consuming.

**Concept 499**

For large training sets, it's often **too expensive** to try all possible partitions of our training data.

Since partitions also aren't **smooth**, it'll be difficult to find the **optimal** solution via gradient descent.

So, instead, we'll come up with a (greedy) algorithm to find a pretty good solution.

### 9.2.5 Greedy algorithms

We need to design a tree-building algorithm. Our easiest bet is to be **greedy**:

**Definition 500**

A **greedy algorithm** is one that takes the choice that looks best, **immediately**.

- It doesn't factor in the **future** effects of our decision.

**Example:** In an MDP problem, this would be like looking for the best immediate reward  $R(s, a)$ , without at all consider what our future rewards look like.

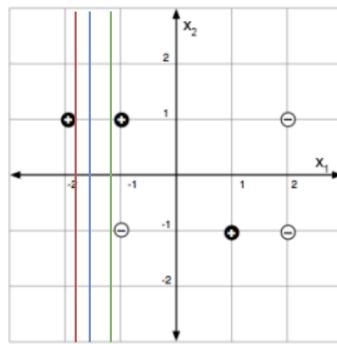
This is a very rough approach, but it's much faster than trying every possibility.

### 9.2.6 How to be greedy

So, does our "greedy" choice look like?

- Our first thought might be to try every possible split. But our input space is made up of **real numbers**: there are an infinite number of possible splits?

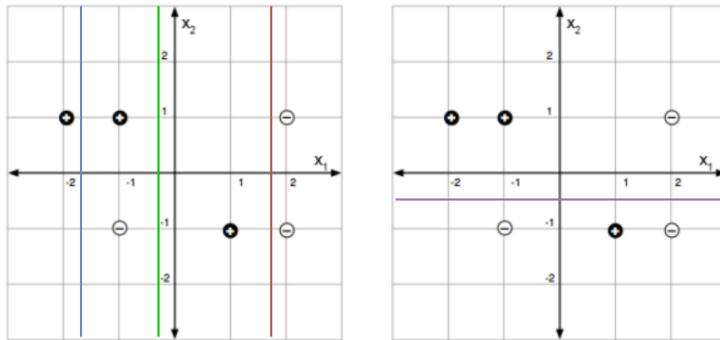
But are there an infinite number of splits that *matter*?



(Using classification ex., for visual clarity) All three of these splits are the same, as far as our training data is concerned.

This is useful: we don't have to try every possible split, because some splits are equivalent.

- We only create new splits each time we **move past** a new data point.
- So, we can iterate through our splits by going **one data point** at a time.



Here are all of the distinct splits between data points, on our two axes. We only create one split each time we cross data points on an axis.

This is the complete set of all possible "first splits".

- In order to be greedy, we just try all of these splits, and see which one works best.

Notice that we don't need to include splits that have no data on one side of our dataset: these splits do nothing.

### Definition 501

In our **greedy algorithm** for tree generation, we:

- List every **distinct** split on each axis.
- Try every one of these splits, and measure their **error**.
- Choose the split with the **lowest error**.

After splitting once, we repeat this algorithm in **both halves** of the tree.

- And, once we've split a second time, we split all quarters.
- We repeat this process **recursively**.

If many splits are equivalent to the ones shown above, how do we know which ones to use?

Our answer: for this class, we don't really care. If you need a mental image, you can place each split halfway between the closest data points on either side.

One thing left: **termination**.

- We need a stopping point: if we don't terminate, then we'll continue until **every data point** has its own region.

**Definition 502**

We **terminate** one region of our tree-building algorithm if that region has **fewer than**  $k$  data points in it.

- Suppose that  $I_m$  gives us all of the **indices** in a particular region. We terminate when:

$$|I_m| \leq k$$

$k$  is a **hyperparameter**.

Now, our algorithm is complete.

### 9.2.7 Tree regression pseudocode

We can also write this as pseudocode. But, let's establish some notation:

#### Notation 503

For this section, each split occurs on **dimension  $j \in J$** , at **position  $s \in S_j$** .

$$x_j \geq s$$

We're ready to go.

BUILDTREE( $I, k$ )

```

1  if  $|I| \leq k$           # If fewer than  $k$  data points: no splitting
2
3       $\hat{y} = \text{Average}_{i \in I}(g^{(i)})$       # Final output
4      return LEAF(output =  $\hat{y}$ )           # Leaf node: no more splits
5
6  else                      # Try every possible split
7      for dim  $j$  in  $J$                   # Check each dimension
8          for value  $s$  in  $S_j$             # Check each split on dim  $j$ 
9
10          $I^+[j, s] = \left\{ i \in I \mid x_j^{(i)} \geq s \right\}$       # Data points "above" the split ( $j, s$ )
11
12          $I^-[j, s] = \left\{ i \in I \mid x_j^{(i)} < s \right\}$       # Data points "below" the split ( $j, s$ )
13
14          $\hat{y}^+ = \text{Average}_{i \in I^+[j, s]} (g^{(i)})$       # Output "above" the split
15          $\hat{y}^- = \text{Average}_{i \in I^-[j, s]} (g^{(i)})$       # Output "below" the split
16
17          $E^+ = \sum_{i \in I^+[j, s]} (\hat{y}^+ - y^{(i)})^2$ 
18          $E^- = \sum_{i \in I^-[j, s]} (\hat{y}^- - y^{(i)})^2$ 
19          $E[j, s] = E^+ + E^-$           # Error for this split
20
21          $(j^*, s^*) = \arg \min_{j, s} (E[j, s])$       # Pick split ( $j, s$ ) with lowest error
22
23     # Recursion step
24     left_branch = BUILDTREE( $I^-[j^*, s^*], k$ )      # Split the left/lower half of data
25     right_branch = BUILDTREE( $I^+[j^*, s^*], k$ )      # Split the right/upper half of data
26
27     return NODE( $j^*, s^*$ , left_branch, right_branch)    # Our node contains the split, and the two halves after the split

```

Below, we use  $i \in I^+[j, s]$  to filter for the data points **above** the split, and  $i \in I^-[j, s]$  to filter for the data points **below** the split.

### 9.2.8 Pruning

It's possible that our tree has more branches than it needs. There are a couple ways we might try to avoid this:

- Set  $k$  relatively **high**,
- Stopping when splitting doesn't improve the **error** very much.

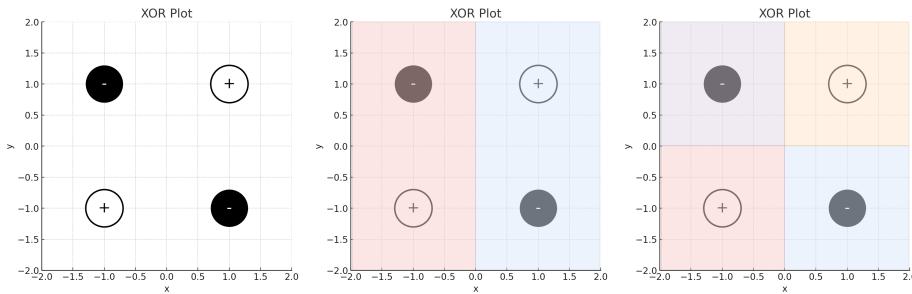
But stopping too early can be a problem:

#### Concept 504

One problem with **early stopping** in tree-building, is that some splits aren't obviously, immediately beneficial.

- But with one or two more splits after, they become very useful.

Consider the XOR problem.



With only one split, the accuracy isn't any better. But with two splits, we've fixed our problem!

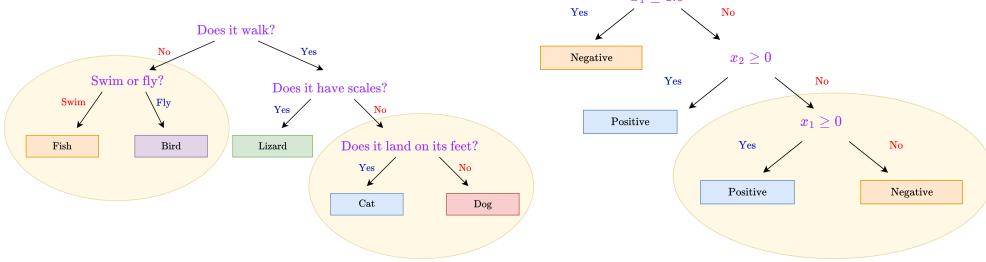
So, rather than stopping early, it's better to make too many splits, and then **prune**.

#### Definition 505

**Pruning** is to remove **branches** from your **tree**.

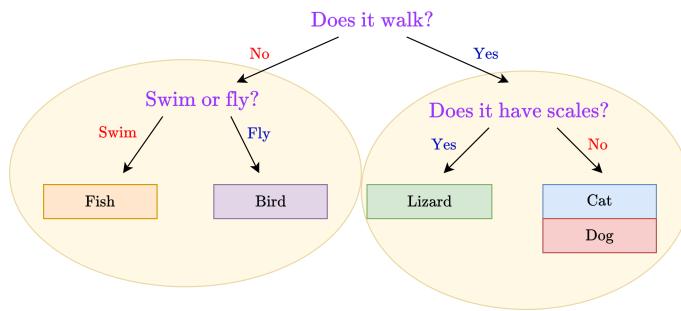
- We remove the "**lowest level**" branches first: splits that create two leaves.

**Example:** Consider our examples from earlier in the chapter.



The circled branches on each tree are the only ones we can prune.

Let's prune one from the pet tree:



Now that we've pruned the lowest branch, we can prune a new, higher branch.

How do we decide which branches to prune? With our **objective**:

$$J = \sum_{m=1}^M E_m + \lambda M \quad (9.14)$$

Here, we use a slightly different notation, so that we can express this as a **function**: \_\_\_\_\_

ML notation is a fickle thing.

### Notation 506

Rather than our objective/loss, we have a **cost complexity function**, for our tree  $T$ , with some key notational changes:

$$\lambda \rightarrow \alpha \quad M \rightarrow |T|$$

Giving us

$$C_\alpha(T) = \sum_{m=1}^{|T|} E_m + \alpha |T|$$

From here, we can **greedily** prune, until we have nothing left to prune.

$|T|$  implies that our tree  $T$ 's size is the number of partitions it has,  $M$ .

- We'll return the tree that has the lowest cost complexity.

### Concept 507

To **prune** our tree, we:

- Try to prune each of our **bottom-level** branches.
  - Actually prune the one with the **lowest cost complexity** (greedy algorithm).
- Repeat until we reach the **root note** (we've removed all of our splits).
- **Return** the pruned tree that has the lowest cost complexity.

Note that, just like how we keep **building** our tree past when it seems beneficial, we also **prune** past when it seems beneficial.

- For the same reason: it's possible for 1 prune to be worse, and 2 prunes to actually be much better.

How do we decide our "regularization term",  $\alpha$ ?

### Concept 508

$\alpha$ , much like  $\lambda$  in **ridge regression**, can be selected via **cross-validation**.

~~~~~

Reviewing cross-validation:

- Break our training data into disjoint **chunks**.
- For each α value, train the model with a **different chunk**.
- Whichever model that performs best on the **held-out** data (the data outside the chunk), used the best α (we hope).

This is the α we use for future training.

9.2.9 Classification

We can re-apply this process for **classification**. Thankfully, most of the steps are the same.

- Our first major difference is that you can't **average** different categories.
- Instead, we'll pick the **most common** category: the *majority*.

How would you average cat, toaster, and forklift? It's, at best, ambiguous.

Definition 509

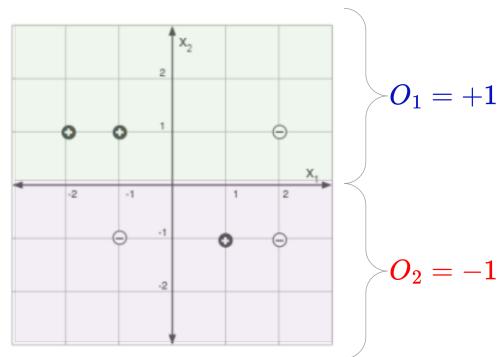
In **classification**, every point in a region is assigned O_m . O_m is the **most common** output for training data in **partition** R_m .

- In other words we want the **majority**.

$$O_m = \text{Majority}_{i \in I_m} (y^{(i)})$$

or,

$$O_m = \text{Majority} \left(\left\{ y^{(i)} \quad \text{if} \quad x^{(i)} \in R_m \right\} \right)$$



Technically, we're willing to take the *plurality*, if there is no majority. But we can just say we want the "most common".

We create a split at $x_2 = 0$. For each region, we choose the most common class.

9.2.10 Classification Loss: Misclassification Error

How do we measure our performance?

- We'll need this kind of tool for choosing the **best split**, **pruning**, and **evaluating** our tree.

The simplest way would be, to simply count how many data points are misclassified.

Key Equation 510

One way to measure loss of our classification tree in **region m** is **misclassification error**: the fraction of data points misclassified.

$$Q_m(T) = \frac{\#(\text{Incorrectly Labelled, region } m)}{\#(\text{All data points, region } m)}$$

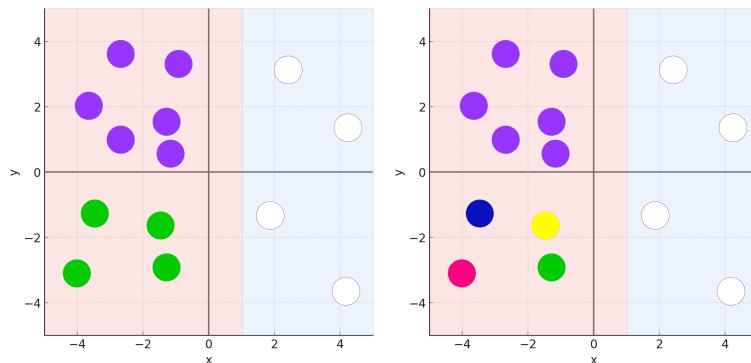
$Q_m(T)$ can also be seen as a **probability**: if you pull a random data point, what's the chance that it was predicted incorrectly?

This is a pretty simple metric, but there's something else we can try.

9.2.11 "Purity" of child nodes: Empirical Probability

One possible problem with "misclassification error" is that it could be **too simple**.

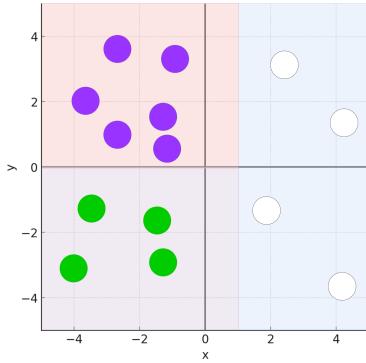
- **Example:** Suppose that we have 6 different classes, and one of our regions has a 60% accuracy rate.
- Our "incorrect" data points could be from **5 different classes**.



Let's focus on the left region. These are two very different situations!

For our more "pure" split on the left, we get **perfect accuracy** with only one more split:

This is an exaggerated, "lucky" example: having more of the same category just makes it **more likely** that we can find a good split.



While, this isn't true for the more "impure" split, with four different categories in the bottom-left.

- We can see that, despite having the same accuracy, the left split is generally "better".

Concept 511

Our **tree classification** problem is generally *easier* if our data is more **concentrated**, in fewer categories.

- If there are **fewer** categories, it'll be easier to find a large group of data points in the **same category**, to split from the rest.

So, we don't just prefer splits that have higher accuracy: we also prefer ones that create **more pure** regions ("child nodes").

So, we want to measure how "pure" a region is.

- A region that is more pure will have a **larger percentage** of data points in the same category.
- So, we want to measure the *fraction* of our data from a given category.

Definition 512

The **empirical probability** $\hat{P}_{m,k}$ tells us what fraction of data points in **region m**, are in **category k**.

$$\hat{P}_{m,k} = \frac{\#(\text{Category } k, \text{region } m)}{\#(\text{Region } m)}$$

- We call it "empirical probability", because we have the **probability** of getting a data point in category k, in region m.

High $\hat{P}_{m,k}$ means that **category k** is very **common** in **region m**.

9.2.12 Classification Loss 2: The Gini Index

Generally, we want a **high** empirical probability in a few categories: that'll mean that **most** of our data is in those few categories.

- We have **two** different metrics for measuring this property, of having our data "**concentrated**" in a few categories.
- These are loss functions, so if they're large, then we have very "**diluted**" data, with lots of categories.

Naturally, this means low empirical probability in all the other categories.

Concept 513

For our data to have high **purity**, we want our **empirical probabilities** $\hat{P}_{m,k}$ to all be either high, or low.

- In other words: a **few** classes have a lot of data, **the rest** have very little.
- The fewer the number of "popular" classes, the **more pure**.

Our first measure asks the question, "if we randomly select a data point, and then select **another** (with replacement), what's the chance that we get a **different class** the second time?"

- If we're likely to get a different category, each time we select one, that suggests that our data is "spread out" across several categories.

"With replacement" means that it's possible to select the same data point twice: after we select the data point the first time, we don't remove it.

There are two (equivalent) expressions that compute this.

- First version: we focus on the chance of them being the **same** class.

$$P\{2 \text{ different classes}\} = 1 - P\{\text{Both same class}\} \quad (9.15)$$

Conversely, if our data was only in one class, then you'd always get the same category, every time.

or,

$$\overbrace{Q_m(t)}^{\text{Region } m} = 1 - \sum_k \underbrace{\left(\hat{P}_{m,k} \right)^2}_{\text{Same class twice}} \quad (9.16)$$

- Second version: We select a first class k , and then, make sure our second class is **different**.

$$P\{2 \text{ different classes}\} = \sum_k P\{\text{Class } k\} \cdot P\{\text{Not class } k\} \quad (9.17)$$

or,

$$\widehat{Q_m(t)} = \sum_k \underbrace{\left(\widehat{P}_{m,k} \right) \cdot \left(1 - \widehat{P}_{m,k} \right)}_{\text{Class } k, \text{ then different class}} \quad (9.18)$$

Key Equation 514

Another way to measure the loss of our split is the **Gini index**:

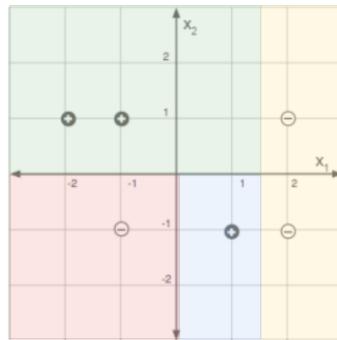
- "If we randomly sample **two** data points (with replacement), what's the probability they have **different classes**?"

$$Q_m(t) = \sum_k \left(\widehat{P}_{m,k} \right) \cdot \left(1 - \widehat{P}_{m,k} \right) = 1 - \sum_k \left(\widehat{P}_{m,k} \right)^2$$

9.2.13 Information 1: Uncertainty (Optional)

Entropy is another measure we can use. Entropy comes from **information theory**: we want each of our splits to give us the **most** information possible.

- But what do we mean by "information"?
- Let's consider our very "**informative**" tree example from the beginning of this section:



Based on which region you're in, you know the exact class of your data.

The tree provides perfect information, for **classifying** the training data: if we know the region of our data point, we know its classification.

- In other words, we have **no uncertainty** left.

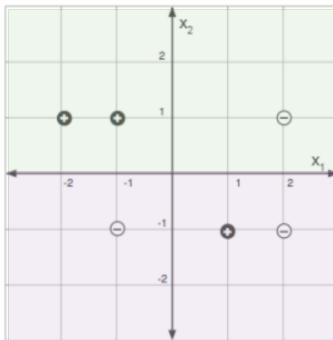
Definition 515

Uncertainty describes "how much" randomness we have in our outcome.

The more information we gain, the **less uncertainty** remains.

This is the kind of information we're discussing.

- Let's investigate further: what does "incomplete" information look like?



Even if we know the region we're in, we're still uncertain about the classification of our data.

This tree is less "informative", because we **don't know** exactly what class we're in.

- Still, it's better than no splits: in the top half, we have more +1 data than -1: we have **less uncertainty** than before.

Concept 516

Our tree is designed to provide **information** about our data:

If you had to guess the classification, and you guessed +1, you'd be right 2/3 of the time, rather than half the time.

- If you learn that a data point is in region R_m , you are **more certain** in guessing which class it's in.

The **less variation** we have in a region, the more **informative** it is:

- If a region R_m only contained a **single class**, you would know **exactly** the class of any data point you find there.
- But if there are many classes that are all likely, then you're pretty **uncertain** about which class to choose.

Example: You want to figure out if someone is sick. If I say, "she yawned earlier", that doesn't help: plenty of people (sick and non-sick) yawn.

- But, if I say, "she's sneezing and coughing", that narrows things down: a lot more

people who are sneezing and coughing, are sick.

9.2.14 Information 2: Entropy (Optional)

Now, we have an idea of what we want. Next, we need to figure out how to **quantify** information.

- Let's use **complete certainty** as a baseline: the situation where we know exactly what class we're in.
- How much more work do we need to do, before we know our class with **100% confidence**?

Concept 517

Our goal is to measure our "**distance**" from being **completely certain** in our classification.

Suppose we want to identify a random data point. We use our **tree structure** to narrow it down to R_m . We've gained some information.

But how far are we from **complete certainty**?

- Our binary tree recursively split our data into two parts, so that we could "**narrow down**" our classification.
- We'll re-use that idea here: we'll ask a **binary** yes-or-no question, to narrow it down further.

We'll assume the **best case**:

Concept 518

We'll **measure uncertainty** based on how many **binary questions** we have to ask about our data point, before we're completely certain of its class:

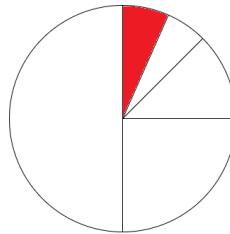
- Each binary question eliminates **half** of the data.
- We're careful with our split: we only remove data that is from a **different class** from our true data.

In other words: "how much work (how many questions) do you need to do, to get complete certainty?"

- If it takes more work, you're **more uncertain** of your answer.

Example: We select a data point, which happens to be in class A. Class A makes up $1/8$ of the data.

- First question: we've remove half, Class A is $1/4$ of the remaining data.
- Second question: we remove half again, Class A is $1/2$.
- After 3 questions, **all** of our data is class A: our data point must be in class A.



We have to split our data in half 3 times, to get down to our $1/8$.

So, each time, we **double** the proportion of class A.

- If $p = 1/8$, then we double 3 times. If $p = 1/16$, we double 4 times, and so on.
- This matches the behavior of the **logarithm** with **base 2**.

$$\log_2 \left(\frac{1}{p_i} \right) = -\log_2(p_i) \quad (9.19)$$

This can be a bit weird when our probability isn't a power of 2: for example, $1/3$. In which case, we could round up: we have to ask 2 questions.

But when we're computing uncertainty, we still consider $1/4$ "more uncertain" than $1/3$. So, we'll allow non-integer values.

Key Equation 519

Each class c_i contributes to the **uncertainty** within our region R_m :

$$\log_2 \left(\frac{1}{p_i} \right) = -\log_2(p_i)$$

We also sometimes call this...

- **Information:** if you need to ask n questions to narrow down your data, then you need n binary answers: n **bits of information**.
- **Surprisal:** if an outcome is less common/likely (lower p), then we're **more surprised** when it happens.

But we're not quite done: this is just our uncertainty associated with one of our outcomes.

- How do we aggregate this, over all of our possible classes?

We'll get the **expected value**: if we pull a random data point, what's the **average number of binary questions** until we've identified it?

$$\mathbb{E}[X] = \sum_i p(x_i) \cdot x_i \implies -\sum_i p_i \log_2(p_i) \quad (9.20)$$

This is called **Entropy**.

Definition 520

Entropy $H(X)$ tells us how **spread-out** our data is, across different outcomes.

- More entropy means that our data is more spread-out, and **uncertain**.

$$H(X) = -\sum_i p_i \log_2(p_i)$$



In the simplified case, we could think of this as answering the following question:

- If we pull a random data point, what's the **average number** of **binary questions** until we've identified it?

What do we do if we have a non-integer entropy? This still measures how "spread-out" or uncertain our data is:

- We're closer/further from having to add one more binary question.

9.2.15 Classification Loss 3: Entropy

Now, we move back to our tree. We need to make one notational change: *empirical probability*.

Key Equation 521

The entropy of a **region** is computed with our **empirical probabilities**:

$$H(I_m) = -\sum_i \hat{P}_{m,k} \cdot \log_2(\hat{P}_{m,k})$$

This is one way to measure the **purity** of our data.

One important caveat:

Clarification 522

$0 \log_2(0)$ is **not defined**. However, for calculations, we usually assume:

$$0 \log_2(0) = 0$$

- The **limit** supports this choice:

$$\lim_{n \rightarrow 0} n \log_2(n) = 0$$

- Additionally, we use entropy to indicate "**uncertainty**". With no data, there's no uncertainty.

Now that we have entropy, we can use it to determine the best splits.

- The lower the entropy, the less **spread-out** our data is.
- So, we want splits that **reduce entropy** the most.

Concept 523

In our greedy algorithm, we choose the splits with the **lowest entropy**:

- These are the splits which give us the most "**information**" about our data.

~~~~~  
Because we have fewer classes in each region, our next split may give better accuracy, as well.

How do we compute the "**change**" in entropy after we've split? We could just add, or **average**, the entropy of both regions.

- But, this could be **misleading**: if we split our data into a region of 2 data points, and a region of 20 data points, we probably care more about the region with 20.
- So, we'll do a **weighted average**: the region with more data points, contributes more to entropy.

We discussed the possible benefits of having more "pure" data in 12.2.11.

A region with only 2 classes of data, might be more easily split, than a region with 5 classes.

$$\frac{\#(\text{Data in region } m)}{\#(\text{Total data})} \cdot \overbrace{H(I_m)}^{\text{Entropy in region}} \quad (9.21)$$

Or, using more dense notation: \_\_\_\_\_

$$\left( \frac{\#I_m}{\#I} \right) \cdot \overbrace{H(I_m)}^{\text{Entropy in region}} \quad (9.22)$$

Remember that  $I$  represents all of your data points, while  $I_m$  represents data in a region.

**Key Equation 524**

The **entropy**  $\hat{H}$  after a split is the **weighted average** of the two splits:

$$\hat{H} = \left( \frac{\#I^+}{\#I} \right) \cdot H(I^+) + \left( \frac{\#I^-}{\#I} \right) \cdot H(I^-)$$

Our data  $I$  is broken into  $I^+$  (data points above split), and  $I^-$  (data points below the split).

If we want to be pedantic, we could create separate notation for splitting at value  $s$ , on axis  $j$ : we use  $I_{j,s}$  notation.

$$\hat{H} = \left( \frac{\#I_{j,s}^+}{\#I} \right) \cdot H(I_{j,s}^+) + \left( \frac{\#I_{j,s}^-}{\#I} \right) \cdot H(I_{j,s}^-) \quad (9.23)$$

We can also say that we "gain information" when we reduce entropy: it takes **less** additional information to completely know our classification.

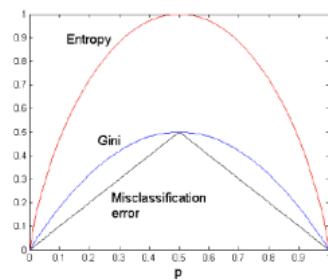
**Key Equation 525**

The **information gained** from a split comes from comparing the entropy, before and after the split:

$$\text{InfoGain} = \overbrace{H(I_m)}^{\text{Pre-split}} - \overbrace{\hat{H}}^{\text{Post-split}}$$

### 9.2.16 Which Loss function to use?

In the past, there's been a lot of debate over which loss function works best.



All three capture the basic idea of accuracy/purity:

- If a class matches **none** of our data ( $\hat{P}_{m,k} = 0$ ), it doesn't contribute to the loss.

- If a class matches **all** of our data ( $\hat{P}_{m,k} = 1$ ), it *also* doesn't contribute to the loss.

For our purposes, we'll consider the following.

**Concept 526**

Traditionally, we use:

- **Entropy** for tree-building
- **Misclassification error** for pruning

### 9.2.17 Bagging: General Concept

One major weakness of our tree model is their **sensitivity** to the data they receive.

- Suppose some noise in our data makes our first split **different**. That will affect **every split** that comes after.
- The second split **depends** on the regions created by the first split. The same is true for the third split.

So, a small change early in our tree can create a dramatically different overall structure.

#### Concept 527

Trees are **sensitive to noise**.

- If you have **slightly different** training data, you can end up with a **very different** tree structure.

This means that our trees are very vulnerable to **estimation error**:

- Even if it's *possible* to get a good tree (low structural error), random chance can often give you a **much worse** tree.

Our solution? Create **several different trees**, with modified training data.

- We **combine** the "opinion" from each tree, and give an answer based on that. This is a type of **ensemble method**.

#### Definition 528

When using an **ensemble method**, we use **multiple models** together, to solve a problem.

- There are multiple different kinds of ensembles: **boosting** and **bagging** are popular examples.

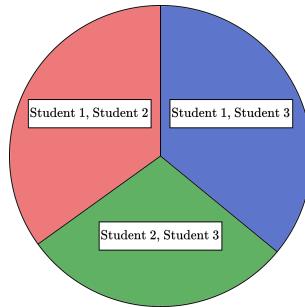
Here, we'll discuss **bagging**.

We hope that having multiple trees allows us to "**average out**" the randomness from each one.

- Each tree might have some estimation error problems, but we hope that most trees don't make the **same mistakes**.
- If they make different mistakes, then each tree can **cover** for the weaknesses of other trees!

**Example:** Suppose that you have 3 students, trying to do 3 homework problems together. For each question, they take the majority vote.

- Each student gets 1 in 3 questions wrong, but it's a different question.



Every student gets one question wrong, but by majority, they get all three right!

Despite each student having a 66% accuracy, they have a 100% accuracy together!

We call this "bootstrap aggregation", or "bagging".

This is a very optimistic version of why bagging could be beneficial. But the idea holds true in general.

#### Concept 529

**Bagging** is a particular kind of **ensemble method**, combining several models to compute an answer.

- In bagging, we train several models **separately**, and then combine all of their answers, to hopefully find a more **accurate** result.

### 9.2.18 Bagging: Bootstrapping (Optional)

So, now, we need to hammer out the details of this process:

- Create **multiple** datasets
- Train a **tree** on each of those datasets
- **Combine** the results of those trees

First: how do we create multiple datasets from our training data?

- We could try breaking our data into **chunks**, like we do in cross-validation.

But that's not what we want here:

**Concept 530**

In **bagging**, we don't want to break our data into **chunks** (partitioning).

This is because these chunks of data aren't independent: they're **correlated** with each other.

- If you include data point  $i$  in chunk  $k$ , you **know** that data point  $i$  **isn't** in chunk  $k + 1$ .
- Knowing about one chunk, provides information about the other chunks: that's how you know they're correlated.

We want each of our models to make its decisions, **independent** of our other models.

**Remark (Optional)**

If our chunks of data are correlated, then why do we use **partitioning** for our **cross-validation**?

- In cross-validation, we want to see how our model performs on data it **hasn't seen before**.
- So, it's important to make sure that the chunk we **train** on, is different from the chunk we **test** on.

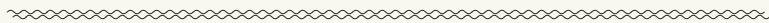
How do we create our datasets, then?

- We **sample with replacement**: after you sample a data point, you **put it back** in the dataset.
- So, you can sample the same data point **multiple times**, or not at all.
- Each dataset we make is called a **bootstrap sample**.

**Concept 531**

When creating data for **bagging**, we use **bootstrapping**:

- Each dataset is created by **sampling with replacement**. This dataset is a **bootstrap sample**.



This creates two major benefits:

- Each bootstrap sample is **uncorrelated**: knowing the data in one dataset, tells you nothing about the others.
  - That means our trees will be **fully separate** from one another.
- When bootstrapping our dataset, we randomly **modify** it, which helps us account for **estimation error**:
  - Each dataset can end up with multiple of a single data point, or missing several data points.
  - So, each tree uses a different "**variation**" of our dataset.

This bootstrapping process, along with "aggregating" opinions across multiple models, is why we call this **bootstrap aggregating** (shortened to bagging).

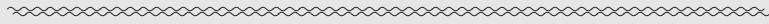
But why do we call it "bootstrapping"?

**Remark (Optional)**

Our training data comes from our true distribution. We could say that it was "**sampled**" from it.

- Meanwhile, our "**bootstrap sample**" comes from sampling our **training data**.
- In other words, we're **sampling a sample**.

This is kind of **circular**: we're getting "more data", without actually getting **new** data.



Thus, we call it **bootstrapping**, referencing the phrase "pull yourself up by your bootstraps".

- Because, it's circular to try to "pull yourself up".

Bootstrap sampling is also used for statistical analysis, when our data is limited.

- If we bootstrap from our sample, multiple times, we can compute the *mean* of those bootstraps.

- So, rather than just having the single mean of our whole sample, we find multiple possible means we could get from our data.

This can be useful: for example, suppose that your bootstrap means vary wildly. We might not be able to trust our sample mean, if it changes so easily.

### 9.2.19 Bagging: Completed

We have a procedure for bagging now.

#### Definition 532

**Bagging** (bootstrap aggregation) is an **ensemble method** for reducing **estimation error**, by combining the answers from  $B$  independent models.

- First, we create a **bootstrap sample** for each tree, sampling with replacement from our dataset  $\mathcal{D}$ .
- We train the  $b^{\text{th}}$  tree using the  $b^{\text{th}}$  bootstrap. The **predictor** we get from this is written as

$$\hat{f}_b(x)$$

When we're evaluating a particular data point  $x$ , we use each of our models, and then **aggregate** the results.

- Regression and classification use different methods of aggregation.

Our "aggregation" method is the same as it was for determining the output of one tree region  $R_m$ .

#### Key Equation 533

In a **regression** problem, we aggregate our  $B$  trees by **averaging** their results.

$$\hat{y}_{\text{bag}}(x) = \frac{1}{B} \sum_b \hat{f}_b(x) = \text{Average}_b(\hat{f}_b(x))$$

In a **classification** problem, we aggregate our  $B$  trees by taking the **majority** (most common) vote.

$$\hat{y}_{\text{bag}}(x) = \text{Majority}_b(\hat{f}_b(x))$$

We *hope* that bagging reduces estimation, but does it really? It turns out it does!

- But as a tradeoff, it's not easily interpretable, like a single tree is.

Imagine having to reference 20 different trees every time you need to make a decision... it's unwieldy.

**Concept 534**

Bagging can be shown to **reduce** estimation error.

- So, **bagged trees** will generally perform **better** than an individual tree.

*But*, we it's much more difficult to interpret a bagged tree, than a single one.

- You have to look at **several** different trees to understand a decision.

This isn't just saying that, in practice, bagging seems to reduce estimation error. We actually have theoretical results that suggest it should!

### 9.2.20 Random Forests

In bagging, we created several different "opinions" on our data, by training each tree on a modified version of our dataset.

Here, we'll consider a different approach:

- Instead of modifying our dataset, we'll modify which **dimensions** we split along.
- Each time we make a split, we'll **randomly select** a few dimensions, and only choose the best of those to split along.

This time, all of our trees have the **same dataset**, but they split in **randomly different** ways.

- Many trees, modified randomly: we call this a **random forest**.

So, if the original "best" splitting dimension is omitted, then the tree will choose the best one it has access to.

#### Definition 535

The **random forest** approach is an **ensemble method** where, instead of modifying the data for each tree, we modify the **splitting algorithm** for each tree.

- Normally, when training a tree, it selects the best split, on the best dimension.
- But in random forests, we **randomly restrict** our tree to only split along **some dimensions**.

So, our tree splits the "best dimension", among the ones those that are randomly selected.

Restricting our tree seems counter-intuitive: why would we deliberately prevent it from choosing the best split?

- This forces some of our trees to "**explore**" other splits, which might be worse short-term, but better long-term.
- Hopefully, this avoids the **estimation error** problem: by trying lots of trees, we can avoid getting "unlucky" with one tree.

Random forests often perform remarkably well, compared to many, much fancier methods.

### 9.2.21 Other types of tree models

There are tons of tree variations:

- We could split along **any hyperplane**, rather than restricting ourselves to only one axis.
  - In which case, we have to figure out how to limit our options: there are many more hyperplanes, than ways to split on only one axis.
- We could expand beyond hyperplanes: we could use **polynomial curves**, like paraboloids.
- We could use **linear regression** for each region  $R_m$ , rather than just averaging all of our data points.
- We could use a **probabilistic** split: rather than a data point belonging in exactly one region,  $R_m$ , it could *partly* belong to all of them.
  - **Example:** Our data point could 90% belong to  $R_1$ , 10% belong to  $R_2$ .
  - Because this is more continuous than our previous method, we can often use **gradient descent** to train.

For example, we could try  $2x_1 + 3x_2 \geq 0$ , rather than  $x_1 \geq 10$ .

Similar to our "polynomial features" method.

This is called a "hierarchical mixture of experts".

### 9.2.22 Benefits of Trees

We've shown lots of reasons you might want to use a tree:

- Easy to interpret, fast to train.
- Very flexible with different loss functions, problem types.
- Often surprisingly effective, despite their simplicity.

#### Concept 536

It's often good practice to use **trees** as a **baseline**, to compare more **complex** models against:

- If your complex model doesn't perform much better than a tree, it may not be worth using.

### 9.3 Terms

- Parametric Methods
- Expressive (Review)
- Non-parametric methods
  - k-means clustering (Review)
  - Nearest neighbor
  - Tree models
  - Ensembles
  - Boosting (Optional)
  - Voronoi Diagram (Optional)
  - k-nearest neighbors (kNN)
  - Locally weighted regression (Optional)
- Distance metric
  - Euclidean distance (Review)
  - Manhattan distance (Optional)
  - Hamming distance (Optional)
  - Minkowsky distance (Optional)
- Binary tree
- Root node
- Leaf node
- Branch
- tree model
- Partition
- Partition function  $\pi$
- Collection of outputs  $O$
- Index
- Tree model objective function
- Greedy algorithm

- Early stopping (Review)
- Pruning
- Low-level/high-level branch
- Cost complexity function
- Majority function
- Misclassification Error
- Empirical Probability
- Purity
- Gini Index
- Information (Optional)
- Uncertainty (Optional)
- Surprisal (Optional)
- Entropy
- Information Gain
- Ensemble Method
- Bagging
- Correlated
- Bootstrapping
- Sampling with Replacement
- Random forest
- Hierarchical Mixture of Experts

# CHAPTER 10

---

## Markov Decision Processes 0 - State Machines

---

### 10.1 State Machines

#### 10.1.1 How to Model Time

How do we want to model time?

The simplest way is one we've used before: keeping track of the current **timestep**  $t$ .

- But this is too little information to be useful: it doesn't tell us much.
- **Example:** If I told you "the current time is  $t = 1563$ ", that doesn't help you much with decision-making.

So, what would be a **useful** representation of time? We've already shown that we don't really care much about the exact **index** of time  $t$ .

Instead, we care about **what happened** in the past.

#### Concept 537

One simple way to record the past is to ask about **events**, and **when** they happened.

**Example:** You might keep track of a medical history, or the purchases made by a company over the last year.

### 10.1.2 States

Keeping a "history" of events is an **improvement**. In some contexts, though, it can become **expensive**: the **longer** our time frame, the more events will pile up.

We could ignore very **old** events, but whether an old event matters depends on the context.

- **Example:** If we omit all company profits/expenses from more than 3 years ago, we don't know our balance. What if we forgot a debt?

This particular example has a pretty simple solution: just keep track of the **total** amount of money you have.

And herein lies our *general* solution: rather than keeping track of every single event, we can keep track of the **state** that result from those past events.

#### Definition 538

A **state** represents information we use to keep track of the **current situation** you're in.

It allows us to store "**memory**" about the past:

- If an event changes our current situation, we'll **update** the state.
- Then, in future timesteps, we'll use that updated state.

A state can be almost **any information** that we want to keep.

- In practice, we want to exclude unhelpful, irrelevant, or outdated data.

**Example:** Suppose that you're an investor. Your state could include: 1. how much money you have, 2. the stocks you currently own, and 3. whether the market seems to be going up or down.

- Notice that, while these variables don't give you exact time, they do **remember** past events: if you have \$30, you at some point must have gotten those \$30.

There are many other kinds of states: position and velocity of an object, or the progress on a project, etc.

### 10.1.3 How states are stored

Now that we've introduced the idea, we'll start formally notating it.

**Notation 539**

Typically, a **state**  $s$  stores our information as a **vector**.

We represent the **set** of all possible **states** as  $\mathcal{S}$ .

- We can have a **finite** or **infinite** set of states, depending on the situation.
- If  $s$  is one of our states, we can express that as  $s \in \mathcal{S}$ .



Our state at time  $t$  is  $s_t$ .

Our **initial state** ( $t = 0$ ) is represented as  $s_0$ .

- Since it's a state,  $s_0 \in \mathcal{S}$ .

We now have two of the pieces of our state machine:

- $\mathcal{S}$  is a finite or infinite **set** of possible **states**  $s$ .
- $s_0 \in \mathcal{S}$  is the **initial state** of the machine.

#### 10.1.4 State examples

Let's show a couple examples of what states different systems might have.

- The game of chess.
  - The **finite** set  $\mathcal{S}$  is the set containing **every chess board**.
  - The initial state  $s_0$  is the **board** when you first **start playing**.
- A ball moving in space, with coordinates.
  - The **infinite** set  $\mathcal{S}$  contains **every pair [position, velocity]** for the ball.
    - \* For example, the ball might be in state  $[(1, 2), (5, 0)]$ :
    - \* at position  $(1, 2)$ ,
    - \* with velocity  $(5, 0)$ .
  - The initial state  $s_0$  is the **position and velocity** when you first **release** the ball.
- A combination lock with 3 digits.
  - The **finite** set  $\mathcal{S}$  contains every **sequence of 3 digits**, where only one sequence unlocks the lock.

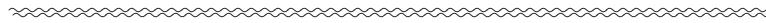
There are multiple different ways to represent the same set of states with a vector, so we won't specify the representation.

- \* For example: [0,0,0], [4,6,9], [9,0,2], etc.
- The initial state  $s_0$  is the sequence when you leave your lock; maybe [1,2,3].

### 10.1.5 Input

We now have a way to **store** our information in time. However, we need to know how to **update** our state: what happens if we learn new information?

We'll include some new variables to address this.



At each timestep, we get some kind of **input**  $x$ , which is our update: this is the newest information about our system. We'll *also* store this in a vector.

#### Definition 540

The **input**  $x$  represents **new information** we get from our system.

We represent the **set** of all possible **inputs** as  $\mathcal{X}$ .

- This set can be **finite** or **infinite**.
- We can say  $x \in \mathcal{X}$ .

Our input at time  $t$  is  $x_t$ .

### 10.1.6 Transition

Based on this new information, we need update the current **state** of the world.

- But often, this update depends both the new information, **and** the **old state**.

**Example:** If your timestep update tells you "got 50 dollars", you need to know how much money you had before, to get your new total.

$$\text{New balance} = \text{Old balance} + \text{Money added}$$

(10.1)



Here's a second example.

**Example:** Suppose you're taking care of a plant.

- If a plant is dry ( $s_t = \text{Dry}$ ), then watering it will make it healthier ( $s_{t+1} = \text{Healthy}$ ).
- If the plant is watered ( $s_t = \text{Healthy}$ ), then watering it more might make it sick ( $s_{t+1} = \text{Sick}$ ).

We're **transitioning** between states, so we use a **transition function**.

**Definition 541**

The **transition function**  $f_s$  tells us how to update our **state**, based on our new **input** information.

- Thus, our transition takes in two pieces of information:  $s$  and  $x$ .

$$f_s(s, x)$$

- We use this function at **every timestep**  $t$  to get our next state, at time  $t + 1$ .

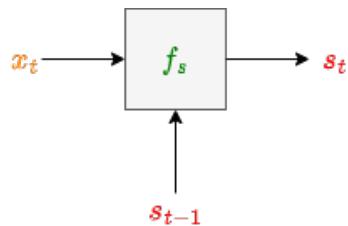
$$f_s(s_t, x_t) = s_{t+1}$$

~~~~~

We can treat each state-input **pair** as an object, (s, x) . Thus, the set of all of these pairs is $\mathcal{S} \times \mathcal{X}$.

$$f_s : \mathcal{S} \times \mathcal{X} \rightarrow \mathcal{S}$$

We can visualize this as:



Now, we have two more pieces of our state machine:

- \mathcal{X} is a finite or infinite set of possible **inputs** x .
- f_s is the **transition function**, which moves us from one state to the next, based on the input.

$$f_s : \mathcal{S} \times \mathcal{X} \rightarrow \mathcal{S} \quad (10.2)$$

10.1.7 Transition Examples

Now, we revisit our examples, and consider how they "transition":

- The game of chess.

- The input x is the **choice** our player makes, **moving one piece** on the chess board according to the **rules**.
- The transition function f_s applies this move to our current chess board, and produces a **new chess board**.
 - * If we moved our pawn, the transition function outputs the board **after** that pawn is **moved**.
- A ball moving in space, with coordinates.
 - The input x might represent a **push** changing the ball's velocity.
 - The transition function f_s uses the push to change our **velocity**, and the velocity to change the ball's **position**.
 - * If our ball wasn't moving before, and we **push** it, the new state is **moving** in that direction.
- A combination lock with 3 digits.
 - The input x is you **changing** one of the three digits on the lock: for example, **increasing** the first digit by 3.
 - The transition function f_s applies the **change** you make to the lock.
 - * If the first digit was 2, and you **increase** it by 3, the new first digit is 5.

10.1.8 Output

We now have a system for keeping **track** of our state, and **updating** that state: this is a really powerful tool for managing time!

We're still missing something, though: why do we **care** about our state? Typically, there's some **result** we actually want from storing our state.

Usually, the desired output is more simple than keeping track of everything we want to **remember**.

Just like how in CNNs, convolution wasn't the end goal: it was a transformation to help improve regression/classification.

Example: If we're storing a bunch of information about the stock market, and our own money, we might simply return "invest in X" or "do not invest in X".

This is what we call our **output**.

Definition 542

The **output** y represents the **result of our current state**.

What we use as "output" depends on what we are **trying to predict/compute**.

- Sometimes, the output is the **only** thing (aside from input) we can **see**. This happens when the state is **hidden**!

In other words, while the state **stores** relevant information to keep track of the situation, the **output** is the decision based on this information.

We represent the **set** of all possible **outputs** as \mathcal{Y} .

- This set can be **finite** or **infinite**.
- If y is a possible output, we say $y \in \mathcal{Y}$.

Our output at time t is y_t .

Note that we use y_t : we don't necessarily create a single output at the end of our runtime.

- Instead, we continuously create outputs at each timestep.

10.1.9 Output Function

So now, we need to actually **compute** our output. This will be based on all the data we have **stored** at the time we're asked for an output.

- We don't need to use the input, because the input data is already included in the state.

We can create an output for each timestep using the **output function**.

Definition 543

The **output function** f_o tells us what **output** we get based on our current **state**.

Thus, our **output function** only takes in the **state**.

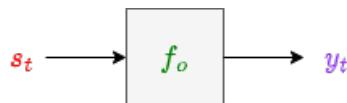
$$f_o(s_t) = y_t$$

It uses our current information (**state**) to produce a the result we're interested in (**output**).

Using sets, we can write this as:

$$f_o : \mathcal{S} \rightarrow \mathcal{Y}$$

We visualize this unit as:



This gives us the last two parts of our state machine:

- \mathcal{Y} is a finite or infinite set of possible **outputs** y .
- f_o is an **output function**, which gives us our output based on our state.

$$f_o : \mathcal{S} \rightarrow \mathcal{Y} \tag{10.3}$$

10.1.10 Output Examples

Again, we go to our examples, and give them outputs, completing our state machines:

- The game of chess.
 - The output y could be many things. But, what do we care about most: **winning!**
 - * So, \mathcal{Y} will have four options: "ongoing", "draw", "player 1 win", "player 2 win".
 - The output function f_o will give us our output. Thus, it represents the **chess rules** for whether there is a winner or a draw.
 - * So, f_o looks at a board, and tells us whether someone has won, or there's a draw.

- A ball moving in space, with coordinates.
 - We want output y .
 - * Sometimes, the **output** is the same as the **state**: all we want to know is what's **happening**!
 - * In this case, we'll say our **output is the state**: we return the **position** and **velocity** of the ball.
 - If our state and output are the same, then the output function f_o should just **copy** the state it receives!
 - * Our function is the **identity function**: $f_o(s) = s$.
- A combination lock with 3 digits.
 - We want our output y .
 - * Our goal is more clear: we want the combination lock to be **open** or **closed**. So, those are our outputs y .
 - Our function f_o will tell us the lock is open if the current digits exactly **match the correct sequence**.

We could have chosen a different output if we had a specific goal in mind!

10.1.11 A Completed State Machine

Finally, we can assemble our completed state machine.

Definition 544

A **State Machine** can be formally defined as a collection of several objects

$$(\mathcal{S}, \mathcal{X}, \mathcal{Y}, s_0, f_s, f_o)$$

We have three sets:

- \mathcal{S} is a finite or infinite **set** of possible **states** s .
- \mathcal{X} is a finite or infinite **set** of possible **inputs** x .
- \mathcal{Y} is a finite or infinite **set** of possible **outputs** y .

And components to allow us to transition through time:

- $s_0 \in \mathcal{S}$ is the **initial state** of the machine.
- f_s is the **transition function**, which moves us from one state to the next, based on the input.

$$f_s : \mathcal{S} \times \mathcal{X} \rightarrow \mathcal{S}$$

- f_o is an **output function**, which gives us our output based on our state.

$$f_o : \mathcal{S} \rightarrow \mathcal{Y}$$

We have:

- Our **state** to store information,
- Our **input** to update information,
- Our **output** gives us the result of our information.

And to combine these, we need:

- Our **initial** state,
- How to **change** states,
- How to **get** an **output**.

10.1.12 Using a State Machine

How do we work with a state machine? Well, we have all of the tools we need.

We start with our initial state, s_0 . For our **first** timestep, we get a new input: new **information**. We use this to get a new state.

$$s_1 = f_s(s_0, x_1) \quad (10.4)$$

With this state, we can now get our **output**.

$$y_1 = f_o(s_1) \quad (10.5)$$

We've calculated everything in our **first** timestep! Now, we can move on to our **second** timestep, and do the same thing.

In general, we'll repeatedly follow the process:

$$s_t = f_s(s_{t-1}, x_t) \quad (10.6)$$

$$y_t = f_o(s_t) \quad (10.7)$$

Concept 545

To move through time in a state machine, we follow these steps from $t = 1$:

- Use the **input** and **state** to get our **new state**.

$$s_t = f_s(s_{t-1}, x_t)$$

- Use the **new state** to get our **output**.

$$y_t = f_o(s_t)$$

- Increment the time from t to $t + 1$.

$$t_{\text{new}} = t_{\text{old}} + 1$$

- Repeat.

10.1.13 Example Run-Through of a State Machine

To make this more concrete, we'll build our own simple state machine and run a couple iteration steps.

Suppose you're saving up money to buy something. At each timestep, you gain or lose some money.

You want to know when you have enough money to buy it.

What are each of the parts of our state machine?

- The state s : how much money do we have right now?
- The input x : the money we add to our savings.
- The output y : we want to know when we have enough money. Maybe our goal is 10 dollars.
- Initial s_0 : we start with 0 dollars.
- Transition f_s : we just add the new money to how much we have saved up.

This example is simple enough that you might feel like a state machine is unnecessary. However, this is just for demonstration!

$$f_s(s, x) = s + x \quad (10.8)$$

- Output f_o : do we have enough money?

$$f_o(s) = (s \geq 10) = \begin{cases} \text{True} & \text{If } s \geq 10 \\ \text{False} & \text{Otherwise} \end{cases} \quad (10.9)$$

We'll run through our state machine for the following input:

$$X = [x_1, x_2, x_3, x_4] = [4, 5, 6, -7] \quad (10.10)$$

Let's apply the steps above:

- Get new state from (old state, input).
- Get output from new state.
- Increment time counter.

For our first step, we get:

$$\begin{aligned} s_1 &= 4 + 0 = 4 \\ y_1 &= (4 \geq 10) = \text{False} \end{aligned} \quad (10.11)$$

For the others, we get:

$$\begin{array}{ccc} s_2 = 9 & \longrightarrow & s_3 = 15 \\ y_2 = \text{False} & & y_3 = \text{True} \end{array} \quad \begin{array}{ccc} s_3 = 15 & \longrightarrow & s_4 = 8 \\ y_3 = \text{True} & & y_4 = \text{False} \end{array} \quad (10.12)$$

Though our transition and output functions might become more complicated, this is the basic idea behind all state machines.

10.1.14 State Machine Diagram

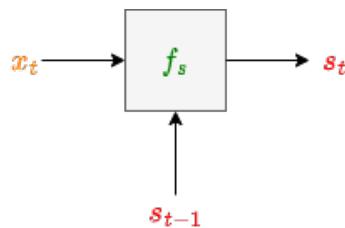
Finally, we'll create a visualization that represents our state diagram.

10.1.14.1 Transition Function

Our **transition function** follows this format:

$$s_t = f_s(s_{t-1}, x_t) \quad (10.13)$$

We can diagram this component as:



Note that the state appears **twice**: once as an input, once as an output.

In the *next* timestep, s_t will be the **input** to f_s , even though it's currently the **output**.

- We'll create a way to represent this later.

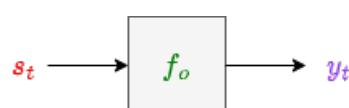
If $t = 10$, then s_{10} is the output. If $t = 11$, then s_{10} is the input!

10.1.14.2 Output Function

Our **output function** takes in the state we just got from the transition function:

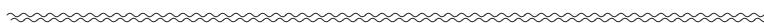
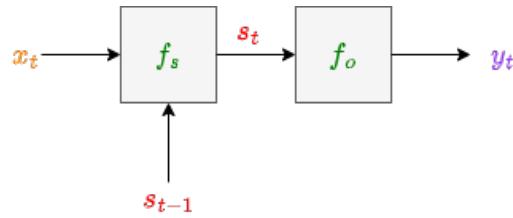
$$y_t = f_o(s_t) \quad (10.14)$$

So, we diagram it accordingly:



As we mentioned, the **output** function takes in the state as its input.

- That means that the **output** of f_s , is the **input** of f_o .



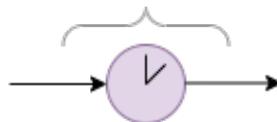
10.1.14.3 Time Delay

Only one thing is missing: we know that our current state s_t needs to be reused **later**: we'll need it to compute our *new state* s_{t+1} .

We don't want it to *immediately* send the state information back to f_s : we only use the function once per timestep. So, we'll *delay* by waiting one time step.

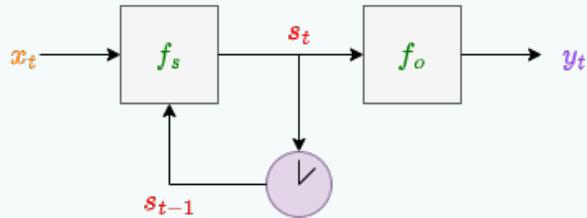
We'll use a little clock symbol to represent this fact.

Waiting one timestep to send information...



Notation 546

We can depict a **state machine** using the following diagram:



At every timestep, we use x_t and s_{t-1} to calculate our new state, and our new output.

The circular "clock" element represents our **delay**: s_t becomes the input to f_s on the **next** timestep.

10.1.15 Finite State Machines

To get used to state machines, we'll start with a simpler, special case, the **finite state machine**.

Definition 547

A **finite state machine** is a state machine where

- The set of states S
- The set of inputs X
- The set of outputs Y

Are all **finite**. Meaning, the total space of our state machine is **limited**.

Each aspect of our state machine can be put into a finite list of elements: this often makes it easier to *fully* describe our state machine.

This seemingly limited tool is more powerful than it seems: **all computers** can be described as finite state machines!

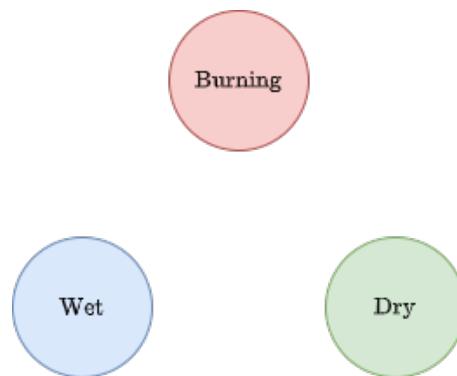
Even when a computer seems to be describing "infinite" collections of things, it only has a finite amount of space to represent them.

10.1.16 State Transition Diagrams

One nice thing about the simplicity of a finite state machine is that we can represent it **visually**.

Let's build one up: we'll pick a simple, though not entirely realistic example.

Example: We have a blanket. It can be in three states: either **wet**, **dry**, or **burning**. We can represent each state as a "node".

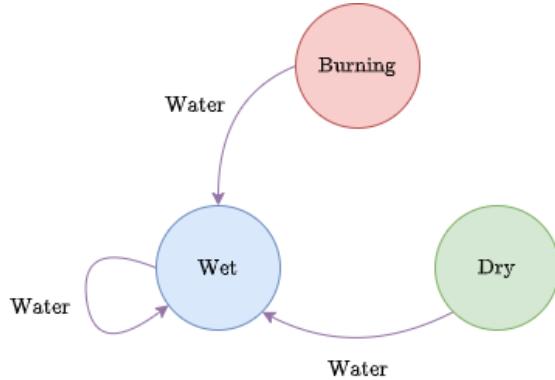


Concept 548

In a **state transition diagram**, states are represented as **nodes**, or points on the graph.

We have our states down. The other important thing is our **transitions**. How do we go between states?

Well, one input could be **water**: it would stop the blanket from burning. In any case, the blanket will be wet.



Now, we can see: each arrow represents a **transition** between two states. Each **input** gets its own transition.

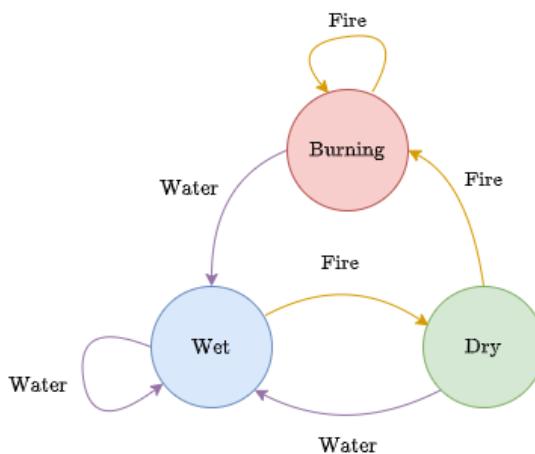
Concept 549

In a **state transition diagram**, transitions are represented as **arrows** between states.

We usually label these with whichever **input** will cause that transition.

Also notice that a state can transition to itself: a wet blanket **stays wet** when you add water.

What if we add **fire**? That would make a dry blanket **burn**. But, we could also use it to **dry off** the wet blanket!



And now, we have a simple **state transition diagram**!

Note that our diagram doesn't show the output. In this case, that's not a problem: the output is the state.

Each transition, as usual, is based on two things: the **current state** (where the arrow starts) and the **input** (which arrow you follow).

Definition 550

A **state transition diagram** is a **graph** of

- Nodes (**points**) representing **states**
- Directed edges (**arrows**) representing **transitions**

Where each input-state pair has one arrow associated with it.

These arrows show one **transition**, with the properties:

- The start and end **states** represented by the start and the end of the arrow
- The **input** that causes this transition is labelled.

This diagram does not have to show the input.

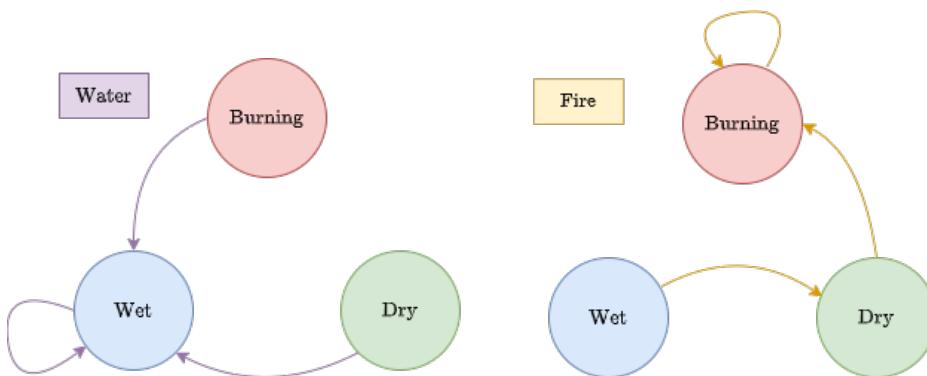
If you're not familiar with "nodes" or "edges", don't worry about it! For our purposes, "point" and "arrow" are good enough.

10.1.17 Simplified state transition diagrams: One-input graphs

One more consideration: the graph above is helpful, but it's a bit **complicated**.

In fact, if we added more **states**, or more **inputs**, it could get too complicated to read!

Our solution: if a system is too complicated, we create a separate state-transition diagram for **each input**.



The left diagram only uses **water** as an input, while the right diagram only uses **fire** as an input.

Each of our diagrams is much more readable now! Not only do we have less arrows, but we don't have to label each arrow.

As a tradeoff, we have two graphs to keep track of, instead of one. However, this is usually

necessary.

In the next chapter,
MDPs, we'll need this!

Concept 551

We can simplify our **state transition diagrams** by creating a **separate diagram** for each input.

This makes it easier to visualize what's going on.

10.1.18 Linear Time-Invariant Systems (LTI)

A wide range of problems can be modelled by a simplified, **linear** version of this system.

That means we'll work entirely with vectors and matrices: no non-linear functions.

- Our **states** are all vectors of length m .
- Our **inputs** are all vectors of length ℓ .
- Our **outputs** are all vectors of length n .

$$\mathcal{S} = \mathbb{R}^m \quad \mathcal{X} = \mathbb{R}^\ell \quad \mathcal{Y} = \mathbb{R}^n \quad (10.15)$$

To transition between states, we'll **linearly** combine state s_{t-1} , with our input x_t : A and B are **matrices**.

$$s_t = As_{t-1} + Bx_t \quad (10.16)$$

Notice that we **exclude the offset terms**: this is due to our definition of **linear**.

Clarification 552

There are two related, but **distinct** definitions for what it means to be **linear**:

- The kind of linear we're more used to: "an equation that draws a line". This is allowed to have an **offset**.

$$f(x) = W^T x + W_0$$

- The kind of linearity used in **linear combinations**, where you're only allowed to **scale** and **add** the inputs together: **no offset**.

$$f(x) = W^T x$$

The latter allows us to use the **linearity** property:

$$f(a + b) = f(a) + f(b) \quad f(c a) = c f(a)$$

While the former definition, with the offset, does not.

Our output will simply be a **linear** scaling of our state: C is a matrix.

$$y_t = C s_t \quad (10.17)$$

We'll also find a second, interesting property:

Definition 553

Time-invariance is the property of our input having the **same** effect on our system, no matter what **time** we apply it.

Thus, we call this restricted model a **Linear Time-Invariant System (LTI)**.

Definition 554

A **Linear Time-Invariant System (LTI)** is a variant of a **state machine**, where

- Our input x , state s , and output y are all vectors

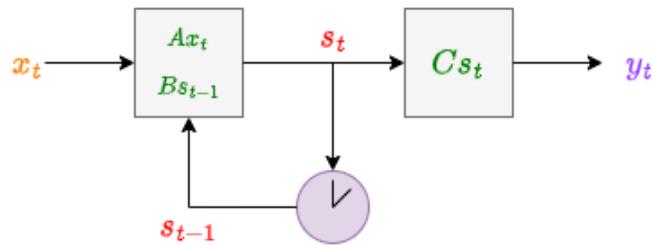
$$\mathcal{S} = \mathbb{R}^m \quad \mathcal{X} = \mathbb{R}^\ell \quad \mathcal{Y} = \mathbb{R}^n$$

- Our transition and output functions are linear (with A , B , and C being matrices)

$$s_t = f_s(s_{t-1}, x_t) = As_{t-1} + Bx_t$$

$$y_t = f_o(s_t) = Cs_t$$

We can depict this with a modified version of our state machine diagram from above:



This kind of model is often an excellent approximation of simple systems in physics, signals, and other sciences.

CHAPTER 10

Markov Decision Processes 1 - Value Functions, Policies

We've developed a notion of a **state machine (SM)**: a system for keeping track of our *current* situation, using a "state".

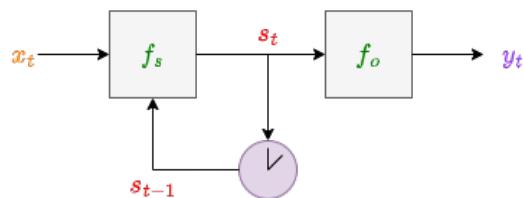
- This state also contains information about the **past**: our **present** state is influenced by past data.

In particular, our focus is on **finite state machines**.

Clarification 555

For the rest of this chapter, we'll assume that our state machines are finite:

- Our set of **states** and set of **inputs** are finite.



A reminder of what our state machine looks like.

10.0.1 A new perspective: the "outside world"

Now, we'll build on our state machine, to create something more specialized for what we need: this will require a new perspective.

- Our "state" has been referred to as our "current situation". This could suggest that it's **representing** something about the **world**.
- In this perspective, our state machine is representing how the world **changes** over time.
- In this case, the state of the world is what we're **interested** in: that's going to be our new **output**.

Concept 556

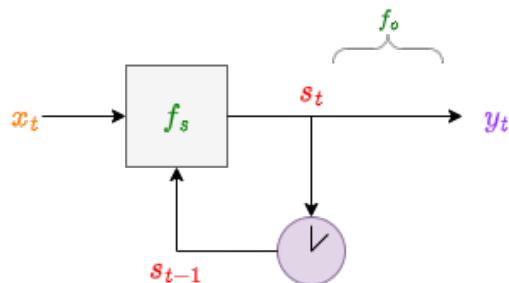
We can view the **state** of our state machine as the current state of the **outside world** that we're modeling.

If we're interested in this "state" of our world, that is the **output** we want:

$$y_t = s_t$$

- f_s is the identity function $f_s(z) = z$: our state is returned, the same way it entered.

Example: In a game, you might want to know **where** you are: so, we keep track of "position" as a state, and return it as an output (on screen).



We can basically remove f_o : it has no effect.

10.0.2 Making "decisions"

This is already interesting, but now, we'll build on this perspective:

- Often, we don't just want to simply **observe** the world, we want to **interact** with it.

- We might want to experiment with different ways to interact with, and change, our **model world**.
- Our state machine modifies its state ("world") through the **input**. We'll use this input to interact with our world: we'll call it an **action**.

Concept 557

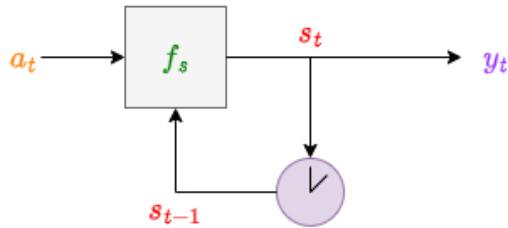
We want to be able to deliberately **modify** the state of the **outside world** through our interactions.

- With an SM, **modify** the state of the world through our **inputs**.

Thus, we'll replace our input x_t with an **action** a_t .

- Our set of actions \mathcal{A} will replace \mathcal{X} .

- **Example:** If you were playing a game, your actions might be "move up", "move down", "move left", or "move right".
- These would affect your **position**, which we could encode in our state.



Structurally, nothing has really changed. x_t has become a_t .

10.0.3 Transitioning between States

One limitation of our state machines so far is that they're **deterministic**: the same inputs will always lead to the same outputs.

Definition 558

A **deterministic state machine** is one where the transitions between states are **deterministic**: given the same inputs, we always get the same output.

- In a realistic setting, the same actions won't always have the same effect: we might end up with **different** states, even if we take the same action.

Thus, we'll use **probabilistic** state transitions. Instead of outputting the same result every time, there will be a certain **probability** of a given outcome.

Example: You have a plant you want to keep healthy. It's currently dry, so you choose the action "**water**".

- 95% of the time, the plant becomes "healthy": it's been watered, and has what it needs to grow.
- 5% of the time, the plant becomes "sick": you just got unlucky, and the plant is sick now.

This model doesn't require giving up on state machines: our function f_s has just changed, returning a **random variable**.

Maybe you watered more than it needed, or maybe something about the environment changed... it doesn't really matter.

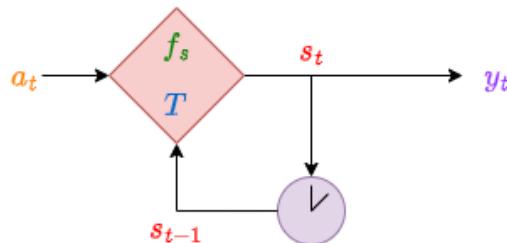
Concept 559

We want to be able to represent the **stochastic** (randomized) nature of our world.

So, we include some randomness in our **state transitions**:

- Given a particular state s_{t-1} and action a_t , **you don't know** exactly what state you'll get for s_t .
- Instead, we have a **distribution**, which assigns a **probability** to each possible next state.

We have a **probabilistic state machine**.



A probabilistic state machine is one type of "non-deterministic" state machine.

T represents the probability distribution of possible output states.

This kind of state machine is very similar to something called a **markov chain**.

Remark (Optional) 560

Our **probabilistic finite state machine (PFSM)** is roughly equivalent to an important mathematical model: a **markov chain**.

In order to be a markov chain, however, it must fulfill the **Markov Property**:

- The state transition is **memoryless**: it only depends on our most recent state s_{t-1} , not any earlier states.

This requirement is already met by our PFSM, but some more complex models may not.

The main difference from a markov chain is that, instead of having inputs, we have actions: a mechanism for making **decisions**.

This remark doesn't fully, rigorously define markov chains. However, we've already built a model that behaves very similarly.

10.0.4 Introducing Rewards

Fundamentally, all we've done so far is choose a particular type of state machine. But now, there's something we'd like to add:

- We have introduced the idea of an "action", but currently, we have reason to choose one action over another.
- To resolve this, we'll introduce an idea of which actions are "good": we'll give a **reward** based on your state, and action.

Concept 561

In addition to our markov chain/PFSM, we'll include a **reward function**, which tells us which situations are more or less desirable.

This depends on both your **action** and current **state**.

- These **state-action pairs** can be compared to each other using the reward function.

Example: A "reward" in a game might be represented by a change in your score.

This reward creates two kinds of decision-making:

Concept 562

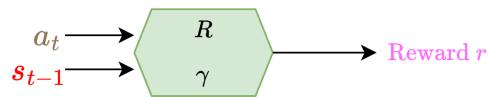
Our reward is determined by the **state-action pair** it receives. This creates two different aspects that weigh into our decision:

- **State:** how do we transition into the state(s) that give us the highest reward?
- **Action:** which actions give us the highest reward?

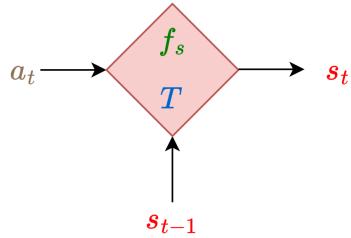
Often, our model must choose between an action that moves you into a "more rewarding" **state**, or an action that gives you higher **immediate reward**.

Example: Do you spend your time getting the easy reward in one area, or try to go to a different, possibly more profitable area?

Later, we'll introduce a **discount factor** γ , which influences this balance between immediate and long-term rewards.



R is a function computing rewards, γ is our discount factor we'll discuss later.



We can compare to the state machine, if we ignore the circular component: they take the same inputs, with different outputs.

Concept 563

Our **reward function** and **state machine** take the same inputs:

- The current state s_{t-1}
- The next action a_t

Based on this information, they tell us two different things:

- The **state machine** tells us how this action affects the world, in this situation.
- The **reward function** tells us how "good" this action is, in this situation.

The former tells us how the world has **changed**, while the latter tells us how **immediately** desirable this action was.

Together, they give us a more complete understanding of this state-action pair.

10.0.5 Markov Decisions Processes

Taking all of these modifications together, we create a new model: the **Markov Decision Process**.

Definition 564

A **Markov Decision Process (MDP)** is a model building upon **state machines**.

First, we make one labelling change:

- Our **inputs** $x_t \in \mathcal{X}$ are replaced with **actions** $a_t \in \mathcal{A}$.

Then, we select a particular variation of state machine:

- Our **state** is returned as the **output**: $f_o(z) = z$
- Our **transition** between states is now **stochastic**: we have a certain probability of ending up in each new state.

Finally, we add one new structure outside the state machine:

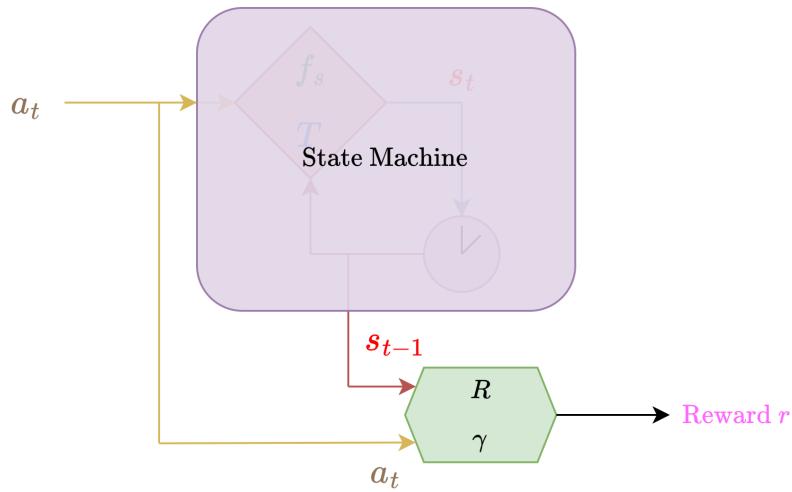
- We include a **reward** function to evaluate the quality of these decisions, based on the **state-action** pair.

Remark (Optional) 565

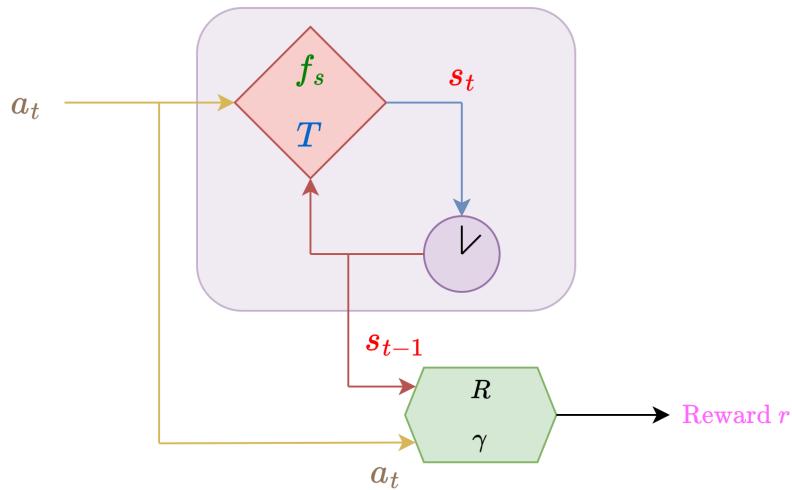
Alternatively, we can view an MDP as a **markov chain**, with two modifications:

- **Actions** replacing inputs, allowing for **decisions**
- A **reward function**, that allows us to evaluate those decisions.

Above, we depicted our state machine and our reward function separately. Here, we'll combine them:



This represents our complete MDP, albeit simplified. We see our real goal: to figure out the relationship between action a_t , and our reward r .



This view shows all the working parts: it's more complex, but more complete.

Our eventual goal is to find out how to **maximize** our reward: we find out which actions provide the most reward.

10.1 Definition and Value Functions

10.1.1 States and Actions in our MDP

We've laid out the general structure of our MDP, but now, we formalize each object.

First, the familiar parts:

Definition 566

For our **MDP**, we have a **finite** action space \mathcal{A} and state space \mathcal{S} .

Thus, every action $a \in \mathcal{A}$, and every state $s \in \mathcal{S}$.

- Reminder that a "space" is just a set, with some extra structure.
- So, our action space is our set of actions, and our state space is our set of states.

Remember: $a \in \mathcal{A}$ means "object a is in the set \mathcal{A} ".

The "structure" depends on what set we choose.

10.1.2 Transition Model

Now, we need to represent the **transition** between states.

- Each *possible* state has a probability p of being our *next* state.
- We'll compute the probability with our **transition model**.

Our transition model will give us a probability. But in order to know the probability, we need three pieces of information:

- s : What is our current state? (Previously s_{t-1})
- a : What action did we take? (Previously a_t)
- s' : What is the **possible next state** we want to get the **probability** of? (Possible s_t)

Our transition function T takes these three pieces of information, and gives us the probability:

$$T(s, a, s') = \text{Probability that, in state } s, \text{action } a \text{ results in new state } s' \quad (10.1)$$

In more mathematical terms:

$$T(s, a, s') = P\{S_t = s' \mid S_{t-1} = s, A_t = a\} \quad (10.2)$$

Because our state S_t is now a random variable, we'll represent it with a capital letter.

Definition 567

The **transition function** T gives the probability of

- Entering state s' ,
- Given that we chose action a in state s

$$T(s, a, s') = P\{S_t = s' \mid S_{t-1} = s, A_t = a\}$$

After a transition, we will be in **exactly one** new state s' .



We can represent it using function notation by considering the following:

- T has input (state, action, state): $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$
- T returns a probability: a real number between 0 and 1: $[0, 1]$

$$T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$$

It would also be valid to write $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$, because $[0, 1]$ is part of the real numbers \mathbb{R} .

Example: We'll return to the example of our plant.

- Its current state is $s = \text{Dry}$.
- We choose action $a = \text{Water}$.

We have two outcomes:

- 95% chance of becoming healthy.

$$T(\text{Dry}, \text{Water}, \text{Healthy}) = 0.95 \quad (10.3)$$

- 5% chance of becoming sick.

$$T(\text{Dry}, \text{Water}, \text{Sick}) = 0.05 \quad (10.4)$$

10.1.3 Comments on our Transition Function

Note that we said that we transition to **exactly one** new state. This means two things:

- Each new state s' is **disjoint**.
- We will definitely end up in **one** of those sets.

Combined, we can say that the probability of all of our states s' adds to 1.

Concept 568

Given a particular **state** s and **action** a , the probabilities for all new sets s' adds to 1:

$$\sum_{s' \in \mathcal{S}} T(s, a, s') = 1$$

One more comment: we use our transition to determine the probability of state transitions.

However, T is **not** our state transition function f_s .

Clarification 569

While T and f_s are both involved in **state transitions**, they serve different functions:

- T gives the **probability** of entering a new state, based on our old states.
- f_s actually **gives us** the new state, according to those probabilities.

In other words, T is a function which **describes** how f_s behaves.

They even have different inputs/output sets:

$$\begin{aligned} T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} &\rightarrow [0, 1] & f_s : \mathcal{S} \times \mathcal{A} &\rightarrow \mathcal{S} \\ T(s, a, s') = p && f_s(s, a) = s' \end{aligned} \tag{10.5}$$

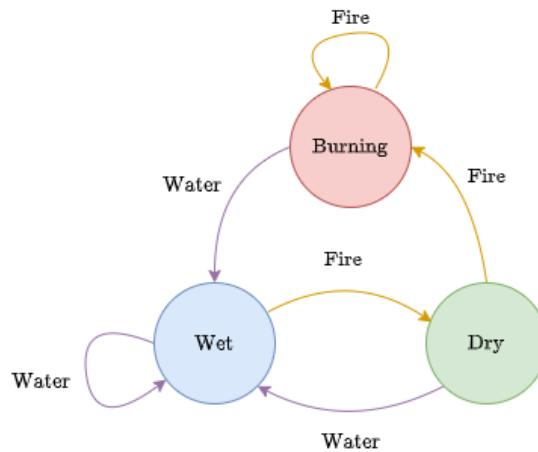
10.1.4 State-Transition Diagram: Review

Our state-transition diagram needs an upgrade, now that our transitions can be probabilistic.

First, we'll review our example from the [RNN chapter](#).

Example: We have a blanket. It can be in three states: either **wet**, **dry**, or **burning**. We can represent each state as a "node".

- To change its state, we can either add "water" or "fire".



We want to update this to include transition probabilities.

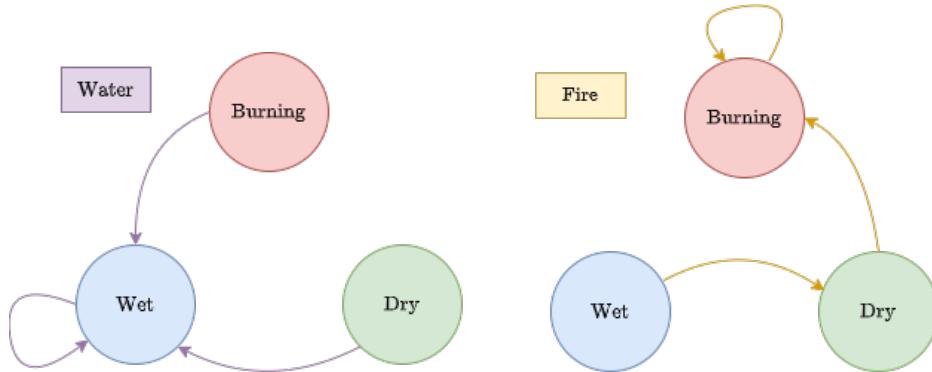
- But this would get pretty dense and **complex**: each arrow would require a **state**, and a **probability**.

- Even worse: right now, we only have one possible outcome for each state-action pair.

- But our probabilistic version allows for multiple outcomes: more arrows, more complexity.

So, we'll split up our diagram based on the **action**, like we did in the RNN chapter:

There could be 2 or more outcomes in the same situation, based on probability.



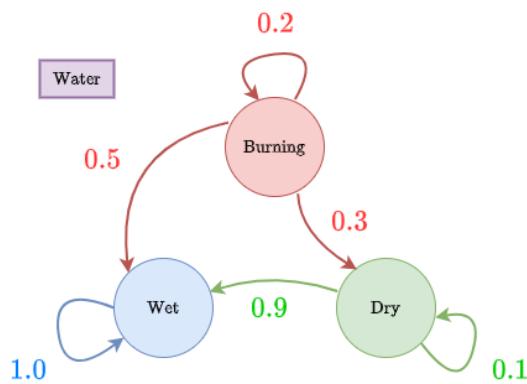
The left diagram uses **water** as an input, while the right diagram uses **fire** as an input.

10.1.5 State-Transition Diagram: Probabilistic

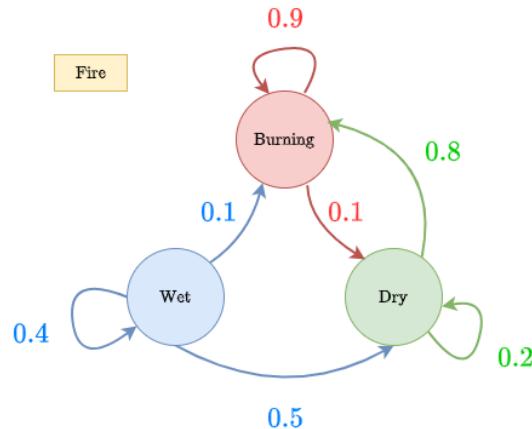
Now, we can extend these diagrams with probabilities.

We'll use the following:

- Add water
 - Burning** blanket: 30% dry, 50% wet, 20% burn.
 - Wet** blanket: 100% wet.
 - Dry** blanket: 10% dry, 90% wet.



- Add fire
 - **Burning** blanket: 10% dry, 90% burn.
 - **Wet** blanket: 50% dry, 40% wet, 10% burn.
 - **Dry** blanket: 20% dry, 80% burn.



These would be almost impossible to represent in a readable way if we include every action on the same graph. So, we create a separate graph for each action.

- If we add more actions, our graph becomes no more complex: we just create more graphs.

Concept 570

For MDPs, we usually have a separate **state-transition diagram** for each **action**.

A second comment: Notice that, when adding water to a wet blanket, it has a 100% chance to stay wet.

- This example is **equivalent** to the deterministic state machine from the RNN chapter: based on our state and action, we know exactly what state we end up with next.

Concept 571

Our **MDP** can reproduce a **deterministic** state machine by setting the probability for every outcome to 0 or 1.

Of course, we still need 1 valid action for each state-action pair.

10.1.6 Transition Matrix

Representing our transitions is made complicated by the fact that we have three parameters: $T(s, a, s')$.

- If we wanted to represent the outputs, with each parameter on one axis, we'd need a 3-tensor to depict the whole thing.

But above, for graphing purposes, we found a solution: separating our transitions based on our **action** a .

- If we only consider one action a , we only have two parameters: s and s' .
- We can represent this with a **matrix** T .

One axis will indicate the previous state s , and the other axis will represent the new state s' .

- We'll use rows for s (input state), and columns for s' (output state).

$$T(a) = \underbrace{\begin{array}{c} \text{Input state} \\ s \end{array}}_{\left\{ \begin{array}{c} \text{Output state } s' \\ \overbrace{\begin{bmatrix} ? & ? & ? \\ ? & ? & ? \\ ? & ? & ? \end{bmatrix}}^{\text{?}} \end{array} \right\}} \quad (10.6)$$

- The element in our matrix will represent the **probability** of this transition.

$$T(a)_{ij} = T(s_i, a, s_j) \quad (10.7)$$

Notation 572

One way to represent our **transition** T is to create a separate **matrix** T for each action a where

- Row i starts in state s_i
- Column j moves us to state s_j

In this cell, we have:

$$T(a)_{ij} = T(s_i, a, s_j)$$

- The probability that, in state s_i , action a takes us to state s_j .

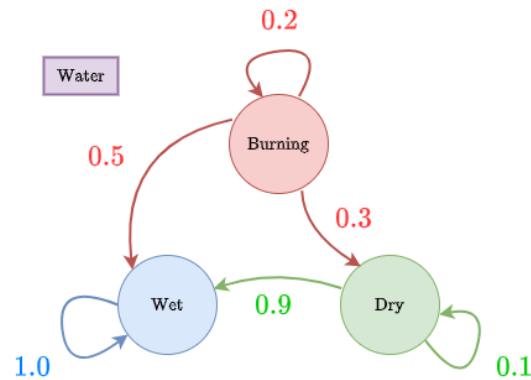
Example: Suppose we have 3 states: s_1, s_2, s_3 . Our matrix for action a looks like:

$$\mathcal{T}(a) = \begin{bmatrix} T(s_1, a, s_1) & T(s_1, a, s_2) & T(s_1, a, s_3) \\ T(s_2, a, s_1) & T(s_2, a, s_2) & T(s_2, a, s_3) \\ T(s_3, a, s_1) & T(s_3, a, s_2) & T(s_3, a, s_3) \end{bmatrix} \quad (10.8)$$

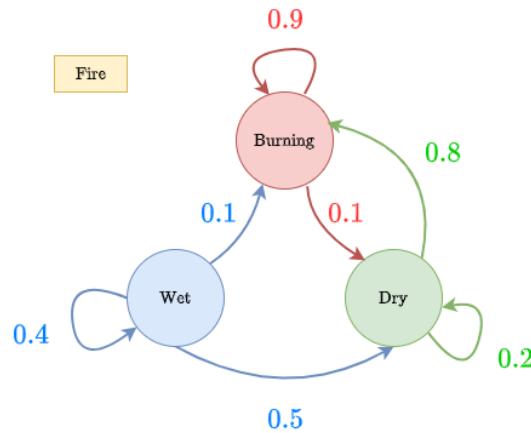
We'll practice on our usual blanket example. We label each of our states with an index:

$$\begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix} = \begin{bmatrix} \text{Burning} \\ \text{Dry} \\ \text{Wet} \end{bmatrix} \quad (10.9)$$

With this, we can create a matrix for each action.



$$\mathcal{T}(\text{Water}) = \begin{bmatrix} 0.2 & 0.3 & 0.5 \\ 0 & 0.1 & 0.9 \\ 0 & 0 & 1.0 \end{bmatrix} \quad (10.10)$$



$$\mathcal{T}(\text{Fire}) = \begin{bmatrix} 0.9 & 0.1 & 0 \\ 0.8 & 0.2 & 0 \\ 0.1 & 0.5 & 0.4 \end{bmatrix} \quad (10.11)$$

10.1.7 Reward Function

We'll represent our **reward function**:

- We compute the **reward** of each state-action pair as with a number: $r \in \mathbb{R}$.

Definition 573

Our **reward function** R gives the **reward** of a particular **state-action** pair.

This indicates how desirable it is to

- Choose action a
- From state s

$$R(s, a) = r$$

We determine our function notation by analyzing the input/output pair.

- R has input (state, action): $S \times A$
- R returns a "reward" as a real number: $R(s, a) \in \mathbb{R}$.

$$R : S \times A \rightarrow \mathbb{R}$$

Example: In our blanket example, we may only care about the state, not the action.

$$R(s, a) = \begin{cases} 10 & s = \text{Dry} \\ 0 & s = \text{Wet} \\ -20 & s = \text{Burnning} \end{cases} \quad (10.12)$$

Maybe we're trying to use the blanket. A dry blanket can be used, a wet blanket cannot, and a burning blanket is an active problem.

Concept 574

Sometimes, our **reward function** may only depend on the **state** we are in.

For consistency, we still use the notation $R(s, a)$.

One thing to be careful of:

We'll procrastinate the discussion of our discount factor γ to our discussion of **infinite horizon**.

In the meantime, we'll include it in our formal definition, but we won't discuss it.

10.1.8 MDP Formalized

Finally, we've built all the pieces we need for a mathematical definition of our MDP.

Definition 575

We formally define **Markov Decision Process (MDP)** as a list of 5 objects: $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma)$

- \mathcal{S} is our **state space**, and \mathcal{A} is our **action space**.

$$s \in \mathcal{S} \quad a \in \mathcal{A}$$

- $\mathcal{T}(s, a, s')$ is our **transition function**, which gives us the **probability** of transitioning from state s to state s' , if we take action a .

$$\mathcal{T}: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$$

$$\mathcal{T}(s, a, s') = \mathbf{P}\{S_t = s' \mid S_{t-1} = s, A_t = a\}$$

- $\mathcal{R}(s, a)$ is our **reward function**, which tells how **desirable** a particular state-action pair is.

$$\mathcal{R}: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$$

Lastly, we have our discount factor: _____

Some definitions treat the discount factor as separate from the MDP. We do not.

Definition 576

γ is our **discount factor**, which tells us how much we value future rewards.

$$\gamma \in [0, 1]$$

The **higher** γ is, the **more** we value future rewards.

- A reward t timesteps in the future, is worth γ^t times as much.



Because γ is never larger than 1, our discount factor can only:

- Treat future rewards as **equal** to current rewards ($\gamma = 1$) or
- Treat future rewards as **lesser** than current rewards ($\gamma < 1$)

10.1.9 Policies

We've discussed taking **actions** a : these are the **decisions** we want to model with our MDP.

- But we haven't actually created a model for **choosing** these actions.

At best, we could select a **sequence** of actions, and then see what reward we get on average. That way, we can compare them.

$$\begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix} \quad (10.13)$$

This strategy is based on our "sequence of inputs" from the RNN chapter.

But this approach doesn't make use of the **current state**:

- In an MDP, our state transitions are **stochastic**: we don't know exactly what the next state is going to be.
- Based on what state we get, the "best" sequence would change.

Concept 577

One weakness of using a pre-planned sequence of actions is that our model is **stochastic**:

- Depending on how our state **randomly** changes, the "best next action" changes.

Example: Imagine we model poker as an MDP. If you used the **exact same strategy** every time, while ignoring your cards (your "state"), you'd be a pretty terrible player.

It would be better if we could **adapt** our choice of **action**, based on the **state** we ended up in.

We'll construct a new function that does just that, called a **policy** π .

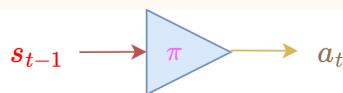
Definition 578

Our **policy** π is a function used to compute **our next action** based on our **current state**.

$$\pi(s) = a$$

One policy represents one particular **strategy** for interacting with the MDP.

$$\pi : S \rightarrow A$$



Example: Here's one possible policy for our blanket example.

$$\pi(s) = \begin{cases} \text{Add fire} & s = \text{Wet} \\ \text{Add water} & s = \text{Dry} \\ \text{Add water} & s = \text{Burnning} \end{cases} \quad (10.14)$$

To be fair, we should probably add a third action that is neither "add water", nor "add fire". But this is an example policy: it isn't necessarily the best policy.

This system allows our policy to "react" to changes in our state.

- If we run our MDP twice, and our state changes are **different**, our **policy** will choose different actions accordingly.
- **Example:** In our poker example, our policy chooses actions based on the cards in your hand, and on the table.

Note that a policy π is not necessary **optimal**:

Clarification 579

As long as you provide an **action** for every **state**, you have created a valid **policy**.

- This means that a policy doesn't have to make sense, or even be a **good** policy.

Which policy is "good" depends on entirely on our **reward function**.

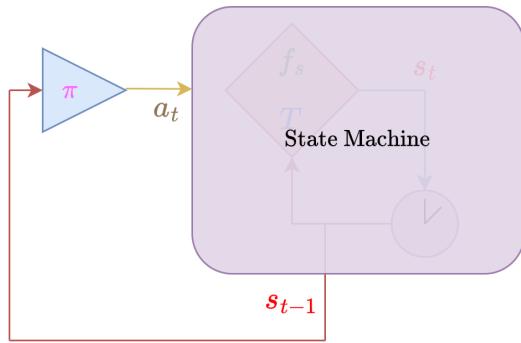
Example: Here's our "pyromaniac" policy. This is probably not optimal unless you really like setting things on fire.

But it could be, depending on the reward function.

$$\pi_p(s) = \begin{cases} \text{Add fire} & s = \text{Wet} \\ \text{Add fire} & s = \text{Dry} \\ \text{Add fire} & s = \text{Burnning} \end{cases} \quad (10.15)$$

Our policy is used by our MDP to process through time, generating actions.

- So, let's add it to our MDP diagram.



Our policy chooses an action based on the state, and the state machine answers with our new state. And so, the cycle continues.

In the coming sections, we'll consider two different kinds of problems for our "game" (MDP):

- Finite horizon – we have exactly T timesteps remaining before we finish our "game".
- Infinite horizon – we have no upper limit on when the "game" ends.

But it might end eventually.

10.1.10 Value Functions

We've built up our MDP, a model to help us making **decisions**. Now, we want the **best policy**.

Concept 580

Typically, the **optimal policy** is the one that gives us the **highest rewards**, on average.

- Averaged over our **stochastic**, random state transitions.

Now, we need a way to directly **compute** this average reward, for each policy: that way, we can compare them.

We'll create a function for this purpose. It'll depend on two things:

- Our chosen **policy**
- Our current **state**

With this information, we'll compute the average reward: this represents the worth of our policy, so we call it our **value function**.

Different states earn **different** rewards, even with the **same** policy:

A perfect robot (same policy) could do better in the stock market if they had **more money** (different state).

Definition 581

A **value function** $V_\pi(s)$ gives us the **average reward** of our policy π , starting from state s .

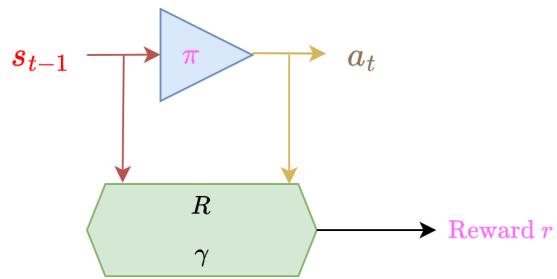
This allows us to compute which policies are "more valuable" in a particular **state**.

- Let's use **function** notation. Reward $r \in \mathbb{R}$ is a real number. So, the same is true for our average reward $V_\pi(s)$.

$$V_\pi : S \rightarrow \mathbb{R}$$

Now, we have a general idea of what a value function is. But there's plenty of details we left out.

- For example: if we're averaging over our rewards, we need our **reward function**.
- This function is specifically designed to **evaluate** our policy π :



Our reward is based on the input/output of our policy: it can be seen as directly "observing" our policy function.

Concept 582

While our value function $V_\pi(s)$ only directly depends on π and s , there's information we need from the MDP itself:

- The **reward function** $R(s, a)$: our average reward is made up of many individual rewards, for different situations.
- The **transition function** $T(s, a, s')$: we need to know how likely different outcomes are, for our average.

In other words: our value function is specific to our MDP: two different MDPs will have completely separate value functions.

Lastly, how we compute our value function depends on whether we're using **finite** or **infinite** horizon.

10.1.11 Finite Horizon

For our first setting, we'll assume that we are going to run our MDP for a precise length of time, h .

- This is our **horizon**: it tells us how "far away" the end of our MDP run is.
- After that time, we don't do anything with our MDP: no actions, no state changes, no rewards.

Definition 583

The **horizon** h of our MDP is the number of timesteps until it **terminates**.

Upon each **timestep** t , our policy π chooses an **action** a_t , and our state machine transitions to **state** s_t . Then, t increases to $t + 1$, and h **decreases** to $h - 1$.

- When our MDP terminates, we **stop** taking actions, or receiving any reward.
- If $h = 0$, our MDP has already terminated: we can't take any more actions.

We can run our MDP as pseudocode, with MDP functions already defined:

```
MDP_RUN(H, s0, π)
1  h = H          #H is the initial horizon
2  t = 0
3
4  while h > 0      #Not yet terminated
5      at = π(st)        #Choose action
6      st+1 = fs(st, at)    #Next state based on fs, described by T
7
8      yield R(st, at)        #New reward
9
10     h = h - 1        #Update horizon
11     t = t + 1
12
13 yield None        #h = 0, Program terminated
```

Meaning, we've defined f_s and R already.

Note that horizon goes down, and time goes up.

Clarification 584

t indicates the **number of timesteps** since **initialization** (the start of the MDP).

h indicates the **number of timesteps** until **termination** (the end of the MDP).

- t increases, while h decreases, as our MDP runs.

Because our horizon is a finite, natural number $h \in \mathbb{N}$, we call this the **finite horizon** case.

Definition 585

If $h \in \mathbb{N}$, then we are dealing with a **finite horizon problem**.

10.1.12 Finite Horizon Value Function

Now, we've set up the problem. In this **finite-horizon** setting, we need to compute the value of different policies.

The first thing to address: we now have a third variable in our value function: **horizon**.

- The amount of **time** remaining affects the total amount of rewards you can receive.
- **Example:** If a button gives you 10 dollars, it's more valuable to be able to press it 5 times ($h = 5$) than only 2 times ($h = 2$).

Definition 586

Your **finite-horizon value function** depends on three factors: **current state**, **policy**, and **horizon**.

We note this combination as $V_{\text{policy}}^{h_{\text{horizon}}}(\text{state})$. Typically, we write

$$V_{\pi}^h(s)$$

This is our function is V_{π}^h , taking s as an input.

- Each horizon-policy combination has its own function.

We'll use **function** notation again: it's the same as without horizon.

$$V_{\pi}^h : \mathcal{S} \rightarrow \mathbb{R}$$

How do we get started? This is actually pretty complicated:

- Suppose you have $h = 5$, and there are **three** possible options for your next state.
- Then, there are three possible options for the state after that. This **repeats** until $h = 0$.
- We have $3^5 = 243$ different possible outcomes, each with their own **rewards**, and **probability** of occurring.

Needless to say, it would be a massive headache to compute all of these possible outcomes at once, especially as h gets larger.

Concept 587

Because MDPs are **stochastic**, they can take many possible paths, even with the same starting state s , and same policy π .

Computing every possible outcome directly can become very **expensive**.

Instead, let's try a different tactic: we'll think of the simplest example possible. What would that be? $h = 0$.

- When $h = 0$, there are no further actions to be taken, and no more rewards.
- So, the expected reward is 0.

Key Equation 588

The total reward for a horizon-0 MDP is always 0, no matter what.

$$V_{\pi}^0(s) = 0$$

10.1.13 Finite Horizon, $H = 1$

We've gotten a start. Now let's try the *second* simplest example, $H = 1$. We have two stages:

- When $h = 1$, we take exactly **one** action, and get **one** reward.
- After that, $h = 0$, and we're in the same situation as before: **no more** rewards.

We'll break our problem into these two parts:

$$V_{\pi}^1(s) = (h = 1 \text{ reward}) + (h = 0 \text{ reward}) \quad (10.16)$$

Concept 589

After taking our first action at horizon h , we're in the **same situation** as if we started in horizon $(h - 1)$, with a new state s' .

Thus, we can **separate** the rewards for each timestep.

We'll find the $h = 1$ reward first.

- Our reward for each step is given by $R(\text{state}, \text{action})$.
- Our state is given by s . What about our action?

- Our **policy** determines our action, based on our state: $\pi(s)$.

$$(h=1 \text{ reward}) = R(s, \pi(s)) \quad (10.17)$$

This same reasoning works no matter which timestep you're starting on.

Concept 590

In our MDP, our **action** will always be determined by our **policy**:

$$a = \pi(s)$$

Thus, our reward $R(s, a)$ will always take the form

$$R(s, \pi(s))$$

Now, the $h = 0$ reward. We've moved into state s' , but that doesn't matter: our reward is always 0.

$$(h=0 \text{ reward}) = V_{\pi}^0(s') = 0 \quad (10.18)$$

At $h = 1$, we gather our reward $R(s, \pi(s))$, but we also change states from s to s' .

Taken together, we get:

$$V_{\pi}^1(s) = \overbrace{R(s, \pi(s))}^{h=1 \text{ reward}} + \overbrace{V_{\pi}^0(s')}^{h=0 \text{ reward}} \quad (10.19)$$

Or,

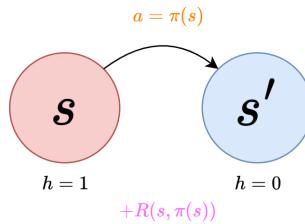
$$V_{\pi}^1(s) = R(s, \pi(s)) + 0 \quad (10.20)$$

Key Equation 591

The total reward for a horizon-1 MDP equals the reward for taking one **action**, according to **policy**.

$$V_{\pi}^1(s) = R(s, \pi(s))$$

We can represent this similar to a state-transition diagram:



We take one action $\pi(s)$, and get a reward $R(s, a)$.

Note that, while we show only one future state s' , this is just one *possible* outcome.

- You could think of this as one run of our MDP.

10.1.14 Finite Horizon, $H = 2$

Next, we'll tackle $H = 2$.

- When $h = 2$, we take action $\pi(s)$, moving from s to s' .
- When $h = 1$, we take action $\pi(s')$, moving from s' to s'' .

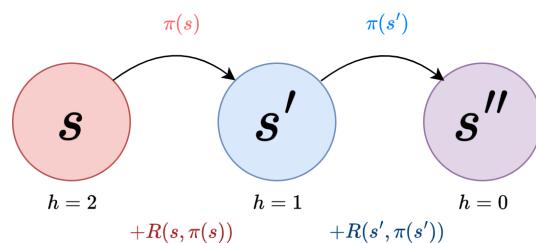
We already know $h = 0$ gives no reward.

We can separate out the reward for each timestep:

$$V_{\pi}^2(s) = (h=2 \text{ reward}) + (h=1 \text{ reward}) + (h=0 \text{ reward})$$

And remove $h = 0$:

$$V_{\pi}^2(s) = (h=2 \text{ reward}) + (h=1 \text{ reward})$$



Two actions, two rewards.

First, we'll compute the $h = 2$ reward.

- This is the same as the the $H = 1$ problem: we start in state s , and take action $\pi(s)$.

$$(h = 2 \text{ reward}) = R(s, \pi(s)) \quad (10.21)$$

Now, we want to compute the $h = 1$ reward.

- After $h = 2$, we've moved to state s' . We try to do what we did before:

$$(h = 1 \text{ reward}) = R(s', \pi(s')) \dots \text{Maybe?} \quad (10.22)$$

But we have a serious problem: which state is s' ?

Concept 592

Our **state transition** is **probabilistic**.

- That means that there are **several new states** s' we could transition to.

So, we can't just directly use $R(s', \pi(s'))$: we don't know which s' will be, and thus our reward.

In other words, the reward for $h = 1$ will depend on the random outcome of our **transition function** T .

How do we resolve this? The key is to remember that V is computing **average reward**: meaning, we'll average our reward over all possible transitions.

A quick review on how to do that:

Concept 593

Suppose we have a **random variable** X with multiple possible outcomes x_i . We want to get the average, or **expected value** $\mathbb{E}[X]$.

- The more **likely** an outcome, the more it **contributes** to the average.

Thus, $\mathbb{E}[X]$ is a **weighted average** of each outcome, **times** its probability.

$$\mathbb{E}[X] = \sum_i^{\text{All outcomes}} x_i \cdot P(x_i)$$

Example: Suppose you're betting on something: there's a 60% chance you make \$100, and a 40% chance you lose \$40. The average earnings (average **reward**) is:

$$\mathbb{E}[X] = \sum_i x_i \cdot P(x_i) = 0.6 * 100 - 0.4 * 40 = 44 \quad (10.23)$$

Now, we have the tools we need to proceed.

Let's apply this averaging to our current situation:

- We want to get the average reward, $R(s', \pi(s'))$.
- We'll average over possible states s' .

$$(h = 1 \text{ reward}) = \underbrace{\sum_{s_i \in S} R(s_i, \pi(s_i)) \cdot P(s' = s_i)}_{\text{Averaging over possible transitions to } s_i} \quad (10.24)$$

What's our probability $P(s' = s_i)$?

- Probabilities are determined by our transition function $T(s, a, s')$.
- $T(s, \pi(s), s')$ represents our transition from $h = 2$ to $h = 1$.

In this situation, $T(s, \pi(s), s_i)$ is "the probability of moving to state s_i ".

Because we already know our initial state s and our policy π .

$$P(s' = s_i) = T(s, \pi(s), s_i) \quad (10.25)$$

$$(h = 1 \text{ reward}) = \sum_{s_i \in S} \underbrace{T(s, \pi(s), s_i)}_{\text{Chance of state } s_i} \cdot \underbrace{R(s_i, \pi(s_i))}_{\text{Reward for state } s_i} \quad (10.26)$$

Key Equation 594

Suppose you know the **state** s of your MDP at horizon H .

- You can compute the reward for the next timestep, horizon $H - 1$, by **averaging over** all of the possible state transitions, to each state s' .

$$(h = H - 1 \text{ reward}) = \sum_{s'} T(s, \pi(s), s') \cdot R(s', \pi(s'))$$

We can assemble our full solution:

Key Equation 595

The total reward for a horizon-2 MDP adds two rewards:

- The reward for taking one **action**, according to **policy**.
- **Average** of all possible outcomes, based on the result of the **state transition**.

$$V_{\pi}^2(s) = \overbrace{R(s, \pi(s))}^{\text{h=2 reward}} + \sum_{s'} \overbrace{T(s, \pi(s), s') \cdot R(s', \pi(s'))}^{\text{h=1 reward}}$$

Without the annotations, we see:

$$V_{\pi}^2(s) = R(s, \pi(s)) + \sum_{s'} T(s, \pi(s), s') \cdot R(s', \pi(s'))$$

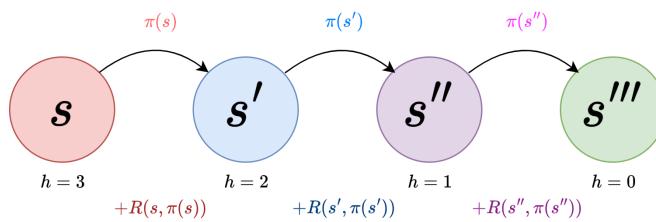
10.1.15 Finite Horizon, $H = 3$ and beyond

Finally, we tackle $H = 3$. And with this step, we come up with a strategy for all finite horizons.

- For $h = 3$, we start in state s , and use our policy to act.
- For $h = 2$, our action depends on the state s' we end up in, **stochastically**.
 - The same is true for $h = 1$: our action depends on the state s'' .

Remember: "stochastic" describes our random state transition, with known probabilities.

$h = 0$ gives **no reward**.



Three actions: s' and s'' can vary based on our random transitions.

The reward for $h = 3$ is the simplest: we know what **state** we're in, and our policy's action.

What about $h = 2$? Well, we could **separately** get the reward for $h = 2$, and $h = 1$. But *instead*, we'll consider the following:

- Suppose that we've already selected our new state, s' . If we know our new state, this is **identical** to the $H = 2$ problem.

Concept 596

If we know our **current state**, we actually **don't care** about what happened during earlier timesteps.

- That means, if our current horizon is $h = 2$, we can pretend as if we're starting a new MDP run with $H = 2$.

Let's forget $h = 3$ for a moment:

- We're starting with $H = 2$, from state s' .
- We want to compute the total reward from this state, using our policy π .
- We **already computed** this above: the result is $V_{\pi}^2(s')$.

Concept 597

$V_{\pi}^H(s')$ includes all the rewards for all timesteps between $h = H$ and $h = 0$.

$$V_{\pi}^H(s) = \sum_{h=0}^H (\text{Immediate reward for horizon } h)$$

Or,

$$V_{\pi}^H(s) = (\text{Rewards for } h \leq H)$$

That means, if we've already computed $V_{\pi}^H(s)$, it can handle the last H steps of our MDP:

$$V_{\pi}^{H+1}(s) = (\text{Reward for horizon } H+1) + (\text{Rewards for } h \leq H)$$

Remember to keep V and R separate in your mind:

Clarification 598

Our value function $V_{\pi}^h(s)$ gives us the **expected reward for all current and future steps**, given a **state** and **policy**.

Our reward function $R(s, \pi(s))$ gives us the **reward for one step**, given a **state** and **action**.

There's one important detail missing, though: this all assumes that we know s' . But we don't, because of our random transitions.

- So, once again, we have to **average** over all possible s' .

$$\left((h \leq H-1) \text{ rewards} \right) = \sum_{s_i \in S} P(s' = s_i) \cdot \overbrace{V_{\pi}^{H-1}(s_i)}^{\text{Future rewards if } s' = s_i} \quad (10.27)$$

Again, we use our **transition function**:

$$\left((h \leq H-1) \text{ rewards} \right) = \sum_{s_i \in S} T(s, \pi(s), s') \cdot V_{\pi}^{H-1}(s_i) \quad (10.28)$$

Concept 599

Often, we **don't know** our next state, but we know the **probability distribution** of states.

In this situation, we can:

- Use T to average all of our value functions, to get the **average future rewards**.

Now, we combine this with our first reward, for the initial H step. Because we **know** our state and policy, we don't have to use the transition function for our **first** step.

This is always true for the $h = H$ step.

$$\left((h \leq H) \text{ rewards} \right) = \left((h = H) \text{ reward} \right) + \left((h \leq H-1) \text{ rewards} \right) \quad (10.29)$$

$$V_{\pi}^H(s) = R(s, \pi(s)) + \sum_{s_i \in S} T(s, \pi(s), s_i) \cdot V_{\pi}^{H-1}(s_i) \quad (10.30)$$

If we apply this to our $H = 3$ example, we get:

$$V_{\pi}^3(s) = R(s, \pi(s)) + \sum_{s_i \in S} T(s, \pi(s), s_i) \cdot V_{\pi}^2(s_i) \quad (10.31)$$

Key Equation 600

The total reward for a horizon- H MDP $V_{\pi}^H(s_i)$ is broken into two parts:

- The reward for our **first action**.
- The reward for **all future steps**, based on the **first transition** $s \rightarrow s'$.
 - The "all future steps" reward has already been computed for the **horizon- $(H-1)$** MDP, $V_{\pi}^{H-1}(s')$.
 - We'll average over each possible "first transition".

$$V_{\pi}^H(s) = \overbrace{R(s, \pi(s))}^{\text{First reward}} + \sum_{s'} \underbrace{T(s, \pi(s), s')}_{\text{Chance of } s \rightarrow s'} \cdot \underbrace{V_{\pi}^{H-1}(s')}_{\text{Reward if } s \rightarrow s'}$$

If we remove the clutter, our equation is:

$$V_{\pi}^H(s) = R(s, \pi(s)) + \sum_{s'} T(s, \pi(s), s') \cdot V_{\pi}^{H-1}(s') \quad (10.32)$$

10.1.16 Finite Horizon MDP Solution

We've created a general solution:

- If $H = 0$, the reward is 0.
- If $H > 0$, then we **break** the reward into two parts:
 - The **immediate** reward
 - The **future** rewards, based on the $H - 1$ case.

Since horizon H depends on horizon $H - 1$, we typically "**build up**" our solution, starting from $H = 1$:

- We compute V_π^H , and then V_π^{H+1} , gradually increasing horizon H .
- That way, when we do horizon H , we already have the **value functions** we need (from $H - 1$).

Let's show this gradual progression:

$$V_\pi^0(s) = 0 \quad (10.33)$$

$$V_\pi^1(s) = R(s, \pi(s)) + 0 \quad (10.34)$$

$$V_\pi^2(s) = R(s, \pi(s)) + \sum_{s'} T(s, \pi(s), s') \cdot V_\pi^1(s') \quad (10.35)$$

And if we keep going...

$$V_\pi^H(s) = R(s, \pi(s)) + \sum_{s'} T(s, \pi(s), s') \cdot V_\pi^{H-1}(s') \quad (10.36)$$

Definition 601

Here, we present the equations for computing the **value functions** for a **finite-horizon MDP**:

$$V_{\pi}^0(s) = 0$$

$$V_{\pi}^H(s) = R(s, \pi(s)) + \sum_{s'} T(s, \pi(s), s') \cdot V_{\pi}^{H-1}(s')$$

We start by computing $V_{\pi}^1(s)$, then $V_{\pi}^2(s)$, and so on, incrementing our horizon each time.

- And at each step, we use the previous value function.

Once we know how to compute the value function for horizon h , we can compare different policies:

Concept 602

Suppose we have a given horizon h .

If policy π_1 is **better than** π_2 , then:

- For every state s , π_1 is at least as good as π_2 .

$$\forall s \in \mathcal{S} : V_{\pi_1}^h(s) \geq V_{\pi_2}^h(s)$$

- There is at least one state \hat{s} where π_1 is better than π_2 .

$$\exists \hat{s} \in \mathcal{S} : V_{\pi_1}^h(\hat{s}) > V_{\pi_2}^h(\hat{s})$$

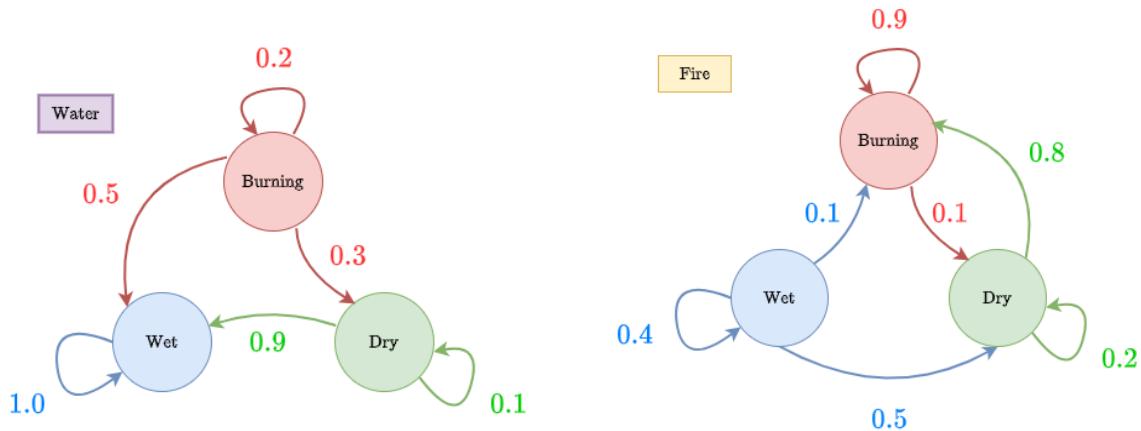
In other words: " π_1 is always at least as good as π_2 , and sometimes better".

\forall just means "for all / every", while \exists means "there is at least one".

10.1.17 Finite-Horizon, using our Blanket Example (Optional)

Example: We can apply this to our blanket problem.

First, our transitions:



Now, our rewards:

$$R(s, a) = \begin{cases} 10 & s = \text{Dry} \\ 0 & s = \text{Wet} \\ -20 & s = \text{Burning} \end{cases} \quad (10.37)$$

Finally, our policy:

$$\pi(s) = \begin{cases} \text{Add fire} & s = \text{Wet} \\ \text{Add water} & s = \text{Dry} \\ \text{Add water} & s = \text{Burning} \end{cases} \quad (10.38)$$

We have 3 states: 3 value functions for each horizon h .

Notation 603

Often, it's easier to write each **value function** in a **condensed** form, to make equations easier to read.

- This is best used for your personal calculations: you need to be clear about what system you're using to **abbreviate**.

In this case, we'll use:

$$V_{\pi}^h(\text{Dry}) = d_h \quad V_{\pi}^h(\text{Wet}) = w_h \quad V_{\pi}^h(\text{Burning}) = b_h$$

The reward for $h = 0$ is 0, no matter our state.

$$d_0 = 0 \quad w_0 = 0 \quad b_0 = 0 \quad (10.39)$$

The reward for $h = 1$ uses the formula

$$V_{\pi}^1(s) = R(s, \pi(s))$$

This is a simplified version of our general equation: because $V_{\pi}^0(s) = 0$, the future steps can be skipped.

Our immediate reward is only based on our current state, according to $R(s, a)$.

$$d_1 = 10 \quad w_1 = 0 \quad b_1 = -20 \quad (10.40)$$

For $h = 2$, we have to use our general formula

$$V_{\pi}^H(s) = R(s, \pi(s)) + \sum_{s'} T(s, \pi(s), s') \cdot V_{\pi}^{H-1}(s')$$

- If it's dry, add water.

$$d_2 = \overbrace{10}^{R_d} + \overbrace{\left(0.1d_1 + 0.9w_1\right)}^{\sum_{s'} T(s, \pi(s), s') V_{\pi}^1(s')} = 11 \quad (10.41)$$

- If it's wet, add fire.

$$w_2 = \overbrace{0}^{R_w} + \overbrace{\left(0.5d_1 + 0.4w_1 + 0.1b_1\right)}^{\sum_{s'} T(s, \pi(s), s') V_{\pi}^1(s')} = 3 \quad (10.42)$$

- If it's burning, add water.

$$b_2 = \overbrace{-20}^{R_b} + \overbrace{\left(0.3d_1 + 0.5w_1 + 0.2b_1\right)}^{\sum_{s'} T(s, \pi(s), s') V_{\pi}^1(s')} = -21 \quad (10.43)$$

We can repeat this process, with the same equations, for **larger and larger** horizons, up to any particular h .

10.1.18 Infinite Horizon

In finite horizon, we knew exactly how long our MDP would run for.

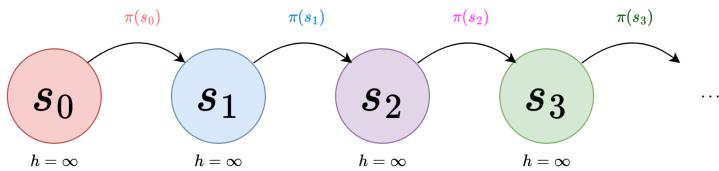
- But often, we don't know how long it'll run for: or at the very least, it'll be a **very long time**.

So, we'll consider an **infinite horizon**:

Definition 604

In the **infinite horizon problem**, we **don't know** how long our MDP will run: it can run for a long time.

In this situation, our horizon **never changes**: if you take n steps, you still don't know when the MDP will end.



We call this "infinite" because we've moved our 100% certain **horizon** infinitely far away.

No matter how many steps we take, the horizon is always the same.

So, we don't use it as a variable.

Definition 605

Our **infinite-horizon value function** V_π^∞ doesn't require a horizon, because it's always the same.

This value function behaves the same as before: input is state s , output is **average reward** r .

$$V_\pi^\infty : \mathcal{S} \rightarrow \mathbb{R}$$

Note:

Notation 606

We can notate our **infinite-horizon value function** two ways:

$$V_\pi^\infty(s) = V_\pi(s)$$

For readability, we'll use $V_\pi(s)$ in following sections.

10.1.19 Discounting

This model has a major bug we need to work out: infinite rewards.

Concept 607

If we run our MDP for an "infinite" amount of time, it can get an **infinite** amount of **rewards**.

This makes it difficult to **compare** two different policies that both have infinite value.

Example: Imagine you have two policies: one presses the "earn \$10" button, and the other presses the "earn \$20" button.

- Over an infinite timescale, both policies have a value of ∞ , even though one is obviously better.

Our solution is to consider the **discounted infinite horizon**: we treat future rewards as less valuable ("discounted") than present rewards.

This brings us back to our γ factor from the beginning of the chapter:

Definition 608

γ is our **discount factor**, which tells us how much we value future rewards.

$$\gamma \in [0, 1]$$

The **higher** γ is, the **more** we value future rewards.

- A reward t timesteps in the future, is worth γ^t times as much.

Because γ is never larger than 1, our discount factor can only either:

- Treat future rewards as **equal** to current rewards ($\gamma = 1$, **finite horizon**)
- Treat future rewards as **lesser** than current rewards ($\gamma < 1$, **infinite horizon**)

Example: We'll re-use the above example, and use $\gamma = 0.9$.

- So, for each timestep, we scale down the reward by 0.9^t .

If we have the same reward r for every time step, we get:

For simplicity, our first timestep is $t = 0$, not $t = 1$.

This is a geometric sum!

$$V_{\pi}(s) = r + r\gamma + r\gamma^2 + \dots = r \sum_{t=0}^{\infty} \gamma^t = \frac{r}{1-\gamma} \quad (10.44)$$

Let's compare our \$20 button to our \$10 button.

$$V_{\$10}(s) = 100 \quad V_{\$20}(s) = 200 \quad (10.45)$$

Now, it's clear which policy is better.

Definition 609

For finite horizon, our value function is given by the **average total rewards**.

$$V_{\pi}^h(s) = \mathbb{E} \left[\sum_{t=1}^h R_t \right]$$

For infinite horizon, our value function is given by the **average discounted rewards**.

$$V_{\pi}(s) = \mathbb{E} \left[\sum_{t=1}^{\infty} \gamma^t R_t \right]$$

In both cases, R_t is based on our **state** and **policy**.

There are a few justifications for why we could discount future value:

Remember that $\mathbb{E}[\cdot]$ gives our average, or **expected value**.

- Economics: **money** is worth more now than in the future.
- Predictability: in the real world, it's harder to accurately predict the distant future than near future: distant rewards are thus less reliable/valuable.

10.1.20 Discount factor: Termination

One of the most useful interpretations is **termination**:

We call our horizon "infinite", meaning that we don't have a **definite stopping point** $h = 0$ for our MDP.

- But, that *doesn't* mean it never terminates: we just don't know **when**.

Instead, consider this: at every timestep, our MDP has a $(1 - \gamma)$ chance of terminating. We can only get our **reward** if it **doesn't terminate**: probability γ .

- The chance of our model surviving for t timesteps, and thus receiving reward R_t , is γ^t .

Concept 610

We can interpret our infinite-horizon MDP having an "opportunity" to fail("terminate"), after every timestep.

- Our **discount factor** γ is the chance of our MDP **continuing**.
- $1 - \gamma$ is the chance of our MDP **terminating**.

So, while our MDP is unlikely to run "forever", there's no fixed horizon h that it will *definitely* terminate at.

This is why we scale down our reward for R_t by a factor of γ^t : because we're computing **expected reward**.

$$\mathbb{E}[r_t] = \underbrace{\gamma^t \cdot R_t}_{\text{Model continues}} + \underbrace{(1 - \gamma^t) \cdot 0}_{\text{Model terminates}} \quad (10.46)$$

- Note that we can "miss" the reward for R_t by failing at **any step** before t .

If the model terminates, we get 0 reward. So, we can ignore that part of the reward.

Make sure not to confuse $(1 - \gamma^t)$ with $(1 - \gamma)^t$: compare **exponents**.

$$\mathbb{E}[r_t] = \gamma^t \cdot R_t \quad (10.47)$$

And this is how we get our $V_\pi(s)$ expression above.

10.1.21 Lifespan of our MDP

As we mentioned, the horizon of our infinite-horizon model never changes.

But we also find something more surprising: **average future lifespan** (average number of timesteps until termination) is the **exact same**, before and after one step.

Let's find out why.

We could think of this as the "average horizon": on average, how long does our MDP have left?

Example: Suppose you're repeatedly flipping a fair coin. You count the number of tails you get. If you ever get heads, you **stop playing**.

A "fair coin" is 50% heads, 50% tails.

Suppose you flip your coin once, and get tails. What are the odds of your next coin being heads?

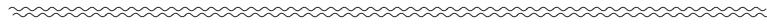
- They're **the same** as they were before that first coin toss: 50%.

What if you flip your coin, and get 10 tails in a row?

- The odds of your next coin being heads is **still** 50%.

Concept 611

In an MDP, the **odds of termination**, $(1 - \gamma)$, are **constant**.



We can apply this to the **average amount of time left** until the model terminates.

- **Example:** Suppose you just got 10 tails in a row. Technically speaking, the **average remaining time** until the game ends is unchanged.
- The coin doesn't "**remember**" that it just won 10 times: you have exactly the same coin as if you just started playing.

We can describe this argument for all MDPs:

- This "average future lifespan" is based on the odds of termination at each turn.
- Since the odds of termination γ are the same, the **average future lifespan** is unchanged.

Consider the opposite case: if "winning 10 times" decreased the remaining time, that means that our coin is suddenly more likely to lose. Why would the coin change?

Clarification 612

The fact that the **average future lifespan** of our MDP is **unchanged**, is often counter-intuitive.

We'll address two points of confusion:

- As a real machine gets older, it's **more likely** to break.
 - But that's because the odds of it breaking **each day** are going up, because it wears out.
 - Nothing in our MDP is "**wearing out**": it just might spontaneously stop.
- It feels like the model should've "**lost**" some of its life, after time passes.
 - This "lost" life is accounted for by the situation where the model immediately **failed**.
 - What's changed about the situation is we know the model model **succeeded**: we're biasing towards a longer total lifetime.

So, your "average future lifespan" is the same, while your "average total lifespan" has increased by 1.

The "average future lifespan", by the way, is always $\frac{1}{1-\gamma}$. Why?

- $(1 - \gamma)$ gives our failure odds.

- If the odds of something happening are 1/4, you would **expect** to wait 4 times for it to happen. _____
- So, you just take the **reciprocal**.

It happens 1 in 4 times, after all.

Key Equation 613

The **average time until termination** ("average future lifespan") of our model is always

$$\frac{1}{1 - \gamma}$$

We always have the same horizon, and the same properties. This is a result of the **markov property**:

Definition 614

MDPs are **memoryless**.

- **Future states** are **only** affected by the **current state and action**.
- Information in the **past** has no effect: thus, our MDP doesn't have "memory" of these past events.

This is also called the **markov property**.

~~~~~  
One effect of this property is that the **average future lifespan** is **always the same**.

One more note:

### Remark (Optional) 615

Often, in probability, we're referring to a "**weaker**" (having less requirements) version of **memorylessness**:

- It **doesn't matter** how long we've been waiting for an event: the time until that event (waiting time) **does not change**.

This is equivalent to the idea that "the **average future lifespan** never changes".

### 10.1.22 Infinite Horizon Value Function

Let's figure out our value function, starting from the finite horizon one:

$$V_{\pi}^H(s) = R(s, \pi(s)) + \sum_{s'} T(s, \pi(s), s') \cdot V_{\pi}^{H-1}(s') \quad (10.48)$$

First of all: our **horizon never changes**: if our model didn't terminate, we're in the same situation after each timestep.

#### Concept 616

**Past events** do not matter to our **infinite-horizon MDP**.

- If we already survived  $t$  timesteps in the **past**, that has no effect on the odds of surviving in the **future**.
- We're in the same situation as when we started.

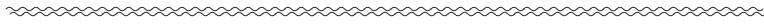
**Example:** Suppose that every round of a game, you **flip a coin** to decide whether you'll continue.

- Knowing that you've already survived 5 rounds has **no effect** on how likely you are to survive the next round.
- We can **ignore** those earlier rounds, and pretend as if we've just started playing.

This simplifies our work: we just remove the horizon.

$$V_{\pi}(s) = R(s, \pi(s)) + \sum_{s'} T(s, \pi(s), s') \cdot V_{\pi}(s') \quad (\text{if } \gamma = 1) \quad (10.49)$$

$V_{\pi}(s')$  doesn't care how many rounds it's been since we started.



Now, we need to account for  $\gamma$ .

Originally, we added  $T$  to represent the **odds** of ending up in state  $s'$ , and thus getting the reward from  $V_{\pi}(s')$ .

$$P\{s' \text{ reward}\} = T(s, \pi(s), s') \quad (10.50)$$

But that was when we knew that our model would continue. Now, we have to include two **independent** events:

- We end up in state  $s'$ :  $T(s, \pi(s), s')$
- Our model doesn't **terminate** immediately:  $\gamma$ .

We just have to multiply these odds together.

$$P\{s' \text{ reward}\} = \overbrace{\gamma}^{\text{MDP doesn't terminate}} \cdot \overbrace{T(s, \pi(s), s')}^{\text{End up in state } s'} \quad (10.51)$$

### Concept 617

In our **infinite MDP**, we have to include our  $\gamma$  factor, **diminishing** the value of future rewards by that factor.

We get:

$$V_\pi(s) = R(s, \pi(s)) + \sum_{s'} \gamma \cdot T(s, \pi(s), s') \cdot V_\pi(s') \quad (10.52)$$

Or, we can even pull  $\gamma$  out of the sum: all future rewards assumes the model doesn't terminate.

### Key Equation 618

We can compute the **infinite-horizon value function** using the following, **recursive** equation.

$$V_\pi(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') \cdot V_\pi(s')$$

We call it **recursive** because  $V_\pi$  references itself in this equation.

#### 10.1.23 Solving the Infinite-Horizon Value Function

We can't solve this version like how we'd solve the **finite-horizon** case: it doesn't make as much sense to do  $h = 1, h = 2$ , etc.

But this version is, in some ways, **simpler**:

- The equation for  $V_\pi(s)$  contains other  $V_\pi$  functions.
- And we have one equation for each state  $s$ .

So, if we gather all of the equations for every  $s \in \mathcal{S}$ , we can **solve** for each value function.

We're also in luck: our value function equations are **linear**!

**Concept 619**

In order to compute the **infinite-horizon value function** for every state, we can follow these steps:

- Write out the equation for value function  $V_{\pi}(s)$  for **every state**, using the other value functions  $V_{\pi}(s')$ .
- If we have  $n$  states, we now have  $n$  equations.
- Solve our  $n$  **equations** for our  $n$  **variables**, as a **linear system**.

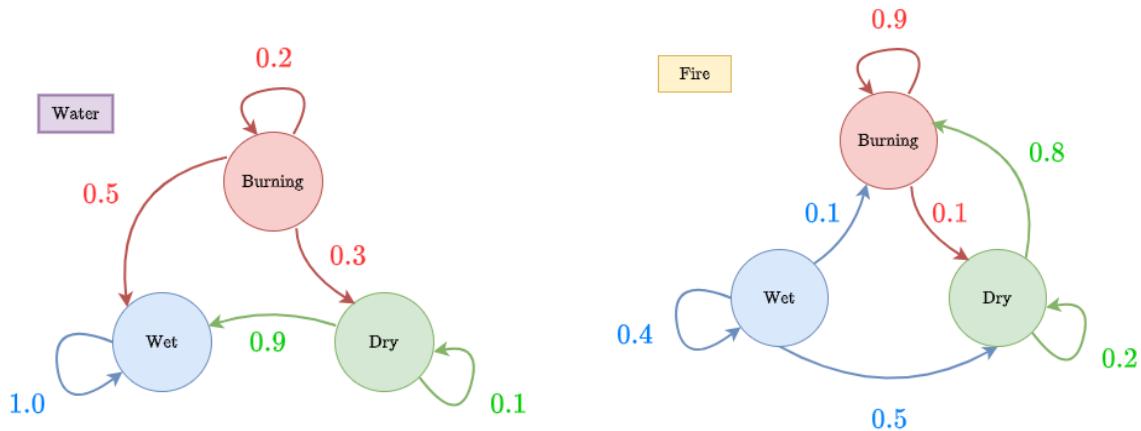
You can use any of our known tactics for solving linear systems.

The number of states is the same as the size of our set  $S$ . So, we can also write it as  $n = |S|$ .

### 10.1.24 Infinite-Horizon, using our Blanket Example (Optional)

**Example:** We can apply this to our blanket problem. Let's set our discount factor at  $\gamma = 0.8$ .

First, our transitions:



Now, our rewards:

$$R(s, a) = \begin{cases} 10 & s = \text{Dry} \\ 0 & s = \text{Wet} \\ -20 & s = \text{Burning} \end{cases} \quad (10.53)$$

Finally, our policy:

$$\pi(s) = \begin{cases} \text{Add fire} & s = \text{Wet} \\ \text{Add water} & s = \text{Dry} \\ \text{Add water} & s = \text{Burning} \end{cases} \quad (10.54)$$

We want three value functions.

#### Notation 620

Often, it's easier to write each **value function** in a **condensed** form, to make equations easier to read.

- This is best used for your personal calculations: you need to be clear about what system you're using to **abbreviate**.

In this case, we'll use:

$$V_\pi(\text{Dry}) = d \quad V_\pi(\text{Wet}) = w \quad V_\pi(\text{Burning}) = b$$

Let's get the equations for each value function.

- If it's dry, add water.

$$d = \overbrace{10}^{R_d} + \underbrace{0.8}_{\gamma} \cdot \overbrace{\frac{\sum_{s'} T(\dots) V_\pi(s')}{(0.1d + 0.9w)}}^{\text{Expected Value}}$$
(10.55)

- If it's wet, add fire.

$$w = \overbrace{0}^{R_w} + \underbrace{0.8}_{\gamma} \cdot \overbrace{\frac{\sum_{s'} T(\dots) V_\pi(s')}{(0.5d + 0.4w + 0.1b)}}^{\text{Expected Value}}$$
(10.56)

- If it's burning, add water.

$$b = \overbrace{-20}^{R_b} + \underbrace{0.8}_{\gamma} \cdot \overbrace{\frac{\sum_{s'} T(\dots) V_\pi(s')}{(0.3d + 0.5w + 0.2b)}}^{\text{Expected Value}}$$
(10.57)

If we solve this **linear system**, we get:

$$\begin{aligned} d &= V_\pi(\text{Dry}) \approx 17.6 \\ w &= V_\pi(\text{Wet}) \approx 8.7 \\ b &= V_\pi(\text{Burning}) \approx -14.6 \end{aligned}$$
(10.58)

# CHAPTER 10

---

## Markov Decision Processes 2 - Optimal Policies, Q-Values

---

### 10.2 Finding policies for MDPs

In the previous section, we computed the total value of different **policies**: strategies for how to act, to get the **most rewards**.

- We designed **value functions** in order to **evaluate** policies, and find the best one.
- So, we'll do that: we search for the **optimal** policies, in the **finite** and **infinite** cases.

**Definition 621**

The **optimal policy**  $\pi^*$  is better than (or as good as) every other policy  $\pi$ , for **every state**.

$$\forall s \in S : V_\pi(s) \leq V_{\pi^*}(s)$$

There can be **multiple** optimal policies.

#### 10.2.1 Optimal Policies – Finite Horizon, $H = 0, 1$

We could try every possible policy, and compare their **values** directly.

- But that's way too expensive: the number of policies is typically **huge**.

Instead, let's do what we did before: we start with the optimal policy for  $h = 0$ , and build up a larger **horizon**.

### Notation 622

Note that our **policy** can depend on our **horizon**. We'll add notation to accommodate this:

- The **optimal policy** for horizon  $h$  is  $\pi_h^*$ .

**Example:** If it takes 10 steps to reach a very valuable state, you should have a different policy if

- You have  $h = 3$  (not enough time to reach it)
- You have  $h = 100$  (more than enough time to reach it)

First:  $H = 0$ . There's no reward, no matter what we do: all policies are the same.

### Concept 623

All **policies** for  $h = 0$  are **optimal**.

Next,  $H = 1$ : we only have to take one action. Thankfully, this is simple: we just take the **action** that **maximizes** our reward.

- This looks like a job for **argmax**.

In the regression chapter, we discussed argmin, but the principle is exactly the same.

### Notation 624

*Review from the Regression chapter:*

The **argmax function** tells you the value of the {input **variable** that gives the **maximum output**.

$$\Theta^* = \arg \max_{\Theta} J(\Theta)$$

The **function we want to maximize** is written to the right, while the **variable we adjust** is written below.

So, we want to know which **action** maximizes the **reward** function.

- We just compute this by comparing all the actions for a single **state**.

$$a^* = \arg \max_a (R(s, a)) \quad (10.1)$$

If we're in state  $s$ , we want our **optimal policy** to give this action.

### Key Equation 625

The **optimal policy** for horizon  $H = 1$  is:

$$\pi_1^*(s) = \arg \max_a (R(s, a))$$

Remember that we can have multiple optimal policies.

### 10.2.2 Finite Horizon: $h = 2$

Now, we want to find the policy if  $H = 2$ .

This has a few complications:

- We have to choose **two** separate actions.
- Our  **$h = 2$  action** will affect our state (and reward) at  $h = 1$ .
  - So, we can't just choose the **action** that gives us the best **immediate reward**: we need to account for *future* reward.

~~~~~  
 This is what we designed our **value function** V^h for! It allows us to compare policies, while accounting for **future** actions/states/rewards.

- But we mentioned that there are **too many** possible policies to compare all of them.
 Is there a way we can **narrow it down**?

Here's the trick: we use the same concept from before –

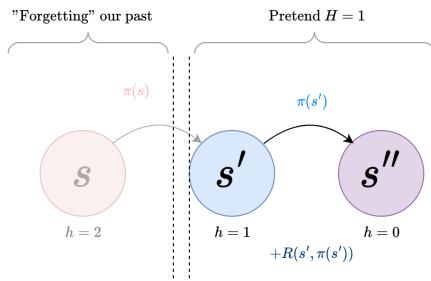
Concept 626

Review from chapter "Markov Decision Processes 1":

If we know our **current state**, we actually **don't care** about what happened during earlier timesteps.

- That means, if our current horizon is $h = 1$, we can pretend as if we're starting a new MDP run with $H = 1$, **ignoring** our original horizon.

Let's ignore whatever our first choice ($h = 2$) was, or our state transition. All we know is that we ended up in state s' .



Our problem is simpler if we simply "forget" our first step: we just pretend we started at s' .

- Now we're in $h = 1$. Thankfully, we already computed the optimal policy, π_1^* .
- Our reward will be _____

$$R(s', \pi_1^*(s')) = \max_{a'} (R(s', a')) \quad (10.2)$$

Since we want the reward, not the action, we'll use "max", not "argmax".

- Now, we know the **second half** of our optimal policy.

Concept 627

The **optimal** choice of action is independent of previous actions: it's only based on our **state** and **horizon**.

- That means that "**the last h steps** of a horizon- H MDP" is **the same** as "**every step** of a horizon- h MDP"
- If we've already computed the optimal policy for a horizon- h MDP, we can **reuse** them at the end of a **longer horizon**.

This **simplifies** the search for our **optimal policy**: we can only focus on policies where those last n steps **match** policy π_n^* .

Example: Suppose you know the best way to finish a game of chess in h turns, starting from any position.

- You can use that knowledge earlier in the game: those will be the end of a longer, winning strategy, starting earlier. _____

Since we already know the best action, for each state, at $h = 1$, we don't have to explore as many possible policies!

One weakness of this analogy: in chess, we don't have a well-defined "horizon" for the end of the game. But the same general idea applies.

Now, we can return to our $h = 2$ step, with the knowledge of how we'll act in the future.

We'll re-introduce our value function, so we can figure out which policy is best:

$$V_{\pi}^2(s) = \overbrace{R(s, \pi(s))}^{h=2} + \sum_{s'} T(s, \pi(s), s') \cdot \overbrace{R(s', \pi(s'))}^{h=1} \quad (10.3)$$

We'll adjust this:

- Our current action is chosen by our policy π_2 .
- Our next action is chosen by the **maximum** reward for the $h = 1$ step.

$$\overbrace{R(s, \pi_2(s))}^{h=2} + \sum_{s'} T(s, \pi_2(s), s') \cdot \max_{a'} \left(R(s', a') \right) \quad (10.4)$$

10.2.3 Q-Values

This is a bit different from V , though. Instead of using the **same policy** for every step, we do something different:

- First, we take one **chosen action** $\pi_2(s)$
- Then for our **second action**, we choose the optimal policy automatically.

We'll come up with a new name for this: a **Q-value function**.

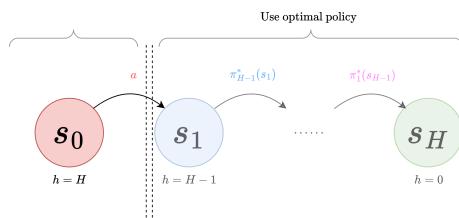
Definition 628

A **Q-value function** $Q^h(s, a)$ (a.k.a "**state-action value function**") is **similar** to a value function $V_{\pi}^h(s)$, but instead of using the **same policy** for every step, we:

- Choose one action a
- Every choice afterwards is **optimal**.

This is why the policy π has been replaced by a single action a : you only make **one choice**, and all following steps are optimal.

- $Q(s, a)$ tells us the **expected reward**, under these conditions.



We choose our first action, and the rest are optimal.

Despite appearing different, the Q-value function serves all of the roles we previously needed our value function V for.

Concept 629

We can think of our Q-value function $Q^H(s, a)$ as a value function V^H with a **special kind of policy**:

$$Q^H(s, a) \longrightarrow V_\pi^H(s)$$

This "special policy" now depends on **horizon**.

$$\pi_h(s) = \begin{cases} a & h = H \\ \pi_h^*(s) & \text{otherwise} \end{cases}$$

This perfectly matches our current strategy: we try out one action a , and then rely on the **optimal policies** we developed previously.

$$\overbrace{R(s, \pi_2(s))}^{\text{One action}} + \sum_{s'} T(s, \pi_2(s), s') \cdot \overbrace{\max_{a'}(R(s', a'))}^{\text{Optimal policies}} \quad (10.5)$$

Q can be seen as a way to "try out" one single action, to add onto your, so far, optimal policy.

If we replace $\pi_2(s)$ with a , we have our Q-value function:

$$Q^2(s, a) = \overbrace{R(s, a)}^{\text{One action}} + \sum_{s'} T(s, a, s') \cdot \overbrace{\max_{a'}(R(s', a'))}^{\text{Optimal policies}} \quad (10.6)$$

10.2.4 $H = 2$ completed

Our new Q-value function already takes care of optimizing $h = 1$, so now, we just need to optimize $h = 2$: we need to find the **best action** at $h = 2$, a^* .

$$a^* = \arg \max_a (Q^2(s, a)) \quad (10.7)$$

Key Equation 630

We can use $Q^2(s, a)$ to determine the **optimal policy** for $H = 2$.

- $Q^2(s, a)$ encodes information about **immediate** and **future** rewards, while **optimizing** those future steps.

This allows us to directly compare different actions a , searching for an optimal policy:

$$\pi_2^*(s) = a^* = \arg \max_a (Q^2(s, a))$$

This action will be chosen for our optimal policy π_2^* .

10.2.5 $H = 2$ Extended Solution (Optional)

Here's the un-compressed version: it's a lot messier, but it shows more of what's going on.

Remark (Optional) 631

The **optimal policy** for horizon $H = 2$ comes in two stages:

- We compute the **optimal policy** π_1^* and **reward** for our second step, $h = 1$.
 - This tells us how **valuable** each $h = 1$ state s' is.
- We compute the **optimal action** a^* for our first step, $h = 1$, by factoring in
 - The **immediate** reward, $R(s, \pi(s))$
 - The **average rewards in the next step**, based on all possible outcomes.

$$\pi_2^*(s) = a^* = \arg \max_a \left\{ \overbrace{R(s, a)}^{\text{Immediate Reward}} + \sum_{s'} T(s, a, s') \cdot \overbrace{\max_{a'} (R(s', a'))}^{(\text{Optimized}) \text{ Future Reward}} \right\}$$

We can see that we **separately** optimize our two steps:

- π_1 first, to get π_1^* : this gives us our action a' .

$$\max_{a'} (R(s', a')) \tag{10.8}$$

- Then, we use π_1^* to find π_2^* : this gives us our action a .

$$\arg \max_a (\dots) \tag{10.9}$$

10.2.6 $H = 3$ and beyond

Introducing Q-values has given us the last tool we need to complete our finite-horizon MDP solution.

Let's use everything we've built so far:

- After our first step, we're in state s' , with $H = 2$: we already determined what our policy should be, and what the **value** of that policy is: _____

The value includes both $h = 1$ and $h = 2$.

$$\pi_2^*(s') = \arg \max_{a'} (Q^2(s', a')) \quad (10.10)$$

$$\text{Optimal Value} = \max_{a'} (Q^2(s', a')) \quad (10.11)$$

- Our **first step** is the only one we choose, giving us an immediate reward:

$$R(s, a) \quad (10.12)$$

This matches our description for the Q-value function: all we have to do is account for different possible s' values, with $T(s, a, s')$.

$$Q^3(s, a) = R(s, a) + \sum_{s'} T(s, a, s') \cdot \max_{a'} (Q^2(s', a')) \quad (10.13)$$

This strategy, conceptually, works exactly the same, for any possible value of H :

Key Equation 632

The Q-value function for horizon H can be written as:

$$Q^H(s, a) = \underbrace{R(s, a)}_{\text{First reward}} + \sum_{s'} \underbrace{T(s, a, s')}_{\substack{\text{Chance of } s \rightarrow s' \\ \text{Future rewards}}} \cdot \underbrace{\max_{a'} (Q^{H-1}(s', a'))}_{\substack{\text{Optimize next step}}}$$

Or, without extra annotation:

$$Q^H(s, a) = R(s, a) + \sum_{s'} T(s, a, s') \cdot \max_{a'} (Q^{H-1}(s', a'))$$

Note that horizon H relies on optimizing horizon $H - 1$, similar to value iteration.

And based on this Q-value, we can determine the optimal action for our current state (and thus, our **optimal policy**):

Key Equation 633

We can use $Q^H(s, a)$ to determine the **optimal policy** for horizon H.

$$\pi_H^*(s) = \arg \max_a (Q^H(s, a))$$

We can use this form of equation to get **every optimal policy**.

Note some important reminders. First, make sure to distinguish between Q and V.

Clarification 634

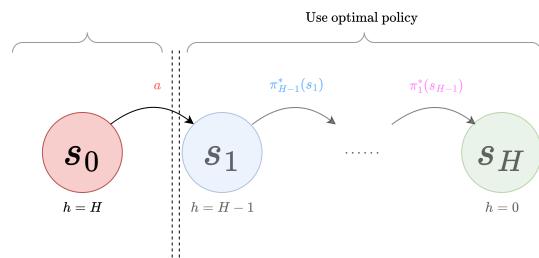
Let's compare our **value function** equation V, to our **Q-value function** equation:

- For our Q-value, we choose **one action**, and the remainder are **optimal**.

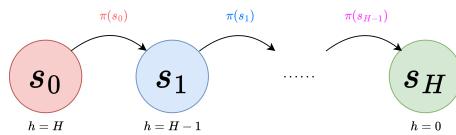
$$Q^H(s, a) = R(s, a) + \sum_{s'} T(s, a, s') \cdot \max_{a'} (Q^{H-1}(s', a'))$$

- Meanwhile, V relies on the **same policy** π for all actions.

$$V_\pi^H(s) = R(s, \pi(s)) + \sum_{s'} T(s, \pi(s), s') \cdot V_\pi^{H-1}(s')$$



Our first action is chosen manually: after that, we use an optimal policy.



All of our actions are chosen by our (potentially non-optimal) policy π .

10.2.7 Finite-Horizon Q-Value MDP solution

We have a method for creating the optimal policy:

- Choose any action for $\pi_0^*(s)$.
- Maximize $R(s, a)$ for $a = \pi_1^*(s)$
- Maximize $Q(s, a)$ for $a = \pi_h^*(s)$, if $h \geq 2$.

For horizon H , we can **re-use** the optimal policies for shorter horizons.

- This concept is encoded by the way Q-values work.
- So, we can use these Q-values to find the optimal policy:

So, our focus is on those Q-values. We'll compute them as we progressively increase the horizon:

$$Q^0(s, a) = 0 \quad (10.14)$$

$$Q^1(s, a) = R(s, \pi(s)) + 0 \quad (10.15)$$

$$Q^2(s, a) = R(s, a) + \sum_{s'} T(s, a, s') \cdot \max_{a'} (Q^1(s', a')) \quad (10.16)$$

And if we keep going...

$$Q^H(s, a) = R(s, a) + \sum_{s'} T(s, a, s') \cdot \max_{a'} (Q^{H-1}(s', a')) \quad (10.17)$$

10.2.8 Dynamic Programming

For value function V and Q-values, we've been using a very particular strategy for computing our solutions.

- We solve **simpler subproblems** (like V^1) and **save the results**.
- Later, we **re-use** those solutions for more complicated problems (like V^2).

This saves us a lot of work:

- **Example:** We could re-compute all the V^1 values, every time we need a V^2 value.
- But if, instead, we just store those values and use them later, we save them.
- The benefits are more obvious for computing V^{200} : it would be a nightmare to have to consider every possible sub-problem.

Last Updated: 09/03/24 03:53:41

We don't have to pay attention to V^2 when doing V^{200} : that information is stored in V^{199} .

We call this strategy **dynamic programming**, despite it being neither "dynamic", nor "programming".

Definition 635

Dynamic Programming is a strategy for solving complex problems that can be broken up into simpler, similar problems ("subproblems").

- First, we solve the **subproblems**, and **save** the result.
- These "mini-solutions" are re-used as a part of larger, more **complicated** problems.

Because a single sub-solution can be used **many times**, you save time by storing the answer, rather than re-computing it every time you need it.

We use it for every step of our value/Q-value calculation:

$$Q^H(s, a) = R(s, a) + \sum_{s'} T(s, a, s') \cdot \max_{a'} (Q^{H-1}(s', a')) \quad (10.18)$$

Q^H always requires us to use Q^{H-1} .

- $Q^H(s, a)$ has an equation for each state-action pair (s, a) .
 - Each of these equations will use Q^{H-1} : that involves using Q^{H-1} **many times**.
- It would be much slower if, for every $Q^H(s, a)$, we had to **re-compute** all of Q^{H-1} .

10.2.9 Dynamic Programming Performance (Optional)

Let's compare the performance of working with, and without dynamic programming. Our goal is to compute $Q^h(s, a)$.

$$|\mathcal{S}| = m, \quad |\mathcal{A}| = n \quad (10.19)$$

We'll need this concept:

Concept 636

If we have a pair of elements, (a, b) , the total number of **possibilities** is:

- The number of possible a values, **times** the number of possible b values.

This idea works for larger sequences: you just multiply together the **possible choices** at each step.

We'll count the amount of work we need with, and without dynamic programming.

First: **without dynamic programming**.

One way to compute the value of $Q^H(s, a)$ is to consider **every possible outcome** of that action, and find the optimal one.

- How many** outcomes are there? One outcome has one **state-action pair** (mn pairs), for each **timestep** (h timesteps).

The total number of possible outcomes is $(mn)^h$. So, our total work is proportional to that:

$$O((mn)^h) \quad (10.20)$$

m states, n actions: mn pairs of states and actions.

At each step, we have mn possible pairs. So, for two timesteps, we have to choose from mn options, twice: $(mn)^2$.

We won't teach O-notation here.

Let's **use dynamic programming** this time.

- For our **final timestep**, we have mn possible state-action pairs. We pick the optimal action for each state, and we **save** its value as $Q^h(s, a)$.
- For our previous timestep, we **don't care** about the possibilities in the final timestep: we just trust $Q^h(s, a)$, and use it to select one of our mn actions.

So, at each timestep, we compare mn elements: we don't have to consider combinations across different timesteps.

However, we do still have to do this calculation once per timestep: h times. We get mnh .

$$O(mnh) \quad (10.21)$$

Concept 637

Using **dynamic programming** for Q-value computation dramatically increases **efficiency**.

Instead of taking $O((mn)^h)$ time, we need $O(mnh)$ time.

- That's **exponentially** faster!

In short: the difference is that dynamic programming allows us to "ignore" other timesteps, and just rely on Q-values.

- Without dynamic programming, we have to "remember" other timesteps, and consider **combinations** of state-action pairs.

10.2.10 Optimal Policies – Infinite Horizon

Now, we need to figure out how to get the optimal **infinite-horizon** policy.

We'll start off with our finite Q-value equation:

$$Q^H(s, a) = R(s, a) + \sum_{s'} T(s, a, s') \cdot \max_{a'} (Q^{H-1}(s', a')) \quad (10.22)$$

All we do is **remove the horizon**, and **add the discount factor**.

- In this situation, we take one action, and then take **optimal** actions for every step after, forever (or until our model terminates).
- We denote this "optimal value" as Q .

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \cdot \max_{a'} (Q^*(s', a')) \quad (10.23)$$

Definition 638

$Q^*(s, a)$ is the **total discounted reward**, assuming you:

- Take action a in state s .
- Choose the **optimal action** for all future states.

Now, our Q-value function is an equation of **itself**.

Key Equation 639

The **Q-value function** for **infinite horizon** can be written as:

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \cdot \max_{a'} (Q^*(s', a'))$$

Something similar happened before, when we were doing value functions. Why is that?

- We're in the **same situation** before and after we take one (successful) step: we **don't know** how long it will be until the model terminates.
- In fact, nothing has changed: the chance of failing after one more step is still γ .

Because "maximizing Q " is how we determine our policy, we see that:

Concept 640

The optimal policy for an **infinite-horizon MDP** is **the same**, at any timestep, regardless of past events.

We call this a **stationary** optimal policy, because it doesn't change over time.

Our goal is to look for one of these "stationary" optimal policies.

There can be several optimal policies. We're only looking for one.

10.2.11 Finding an Optimal Policy: Value Iteration

We have a problem, though. When computing the **infinite value function**, we were able to solve a system of **linear equations**.

- But our Q-value function is **non-linear** this time, because of the **max** operation. We can't solve that!

Is there even a solution, in this complex system? It turns out there **definitely is**:

Theorem 641

Our system of Q-values has a **unique solution**.

This allows us to compute an **optimal policy**.

~~~~~  
Instead of linear solving, we'll try something different:

There may be more than one optimal policy, but they'll all have the same, **unique** Q-values.

In other words, each policy is worth the same average reward.

- Our Q-value function could be said to have an "infinitely far" horizon, with a  $1 - \gamma$  chance of terminating every timestep.
- We could **approximate** this by computing horizon  $h = 1, 2, 3, \dots$ : as our horizon gets really big, we hope this is **similar** to an infinite horizon.

**Concept 642**

We want to approximate an **infinite horizon** with a **really large, finite horizon**.

To match the infinite case, we'll include a chance of **termination**,  $(1 - \gamma)$ .

This is identical to our equations for finite horizon, but with a  $\gamma$  term included.

$$Q^H(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \cdot \max_{a'} (Q^{H-1}(s', a')) \quad (10.24)$$

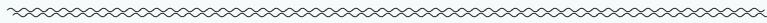
This is called **value iteration**.

However, we don't really care about the  $H$  of our fake, **finite** horizon: our goal is to approximate the infinite-horizon  $Q$ .

**Notation 643**

For each step of our **value iteration** process, we'll distinguish between two parts of our  $Q^*$  approximation:

- $Q_{\text{old}}$ : the  $H - 1$  horizon: it's our **previous**, less-accurate  $Q^*$  approximation.
  - We use it to compute  $Q_{\text{new}}$ .
- $Q_{\text{new}}$ : the  $H$  horizon: it's our **newest**, more accurate approximation.



After each timestep,  $Q_{\text{new}}$  **replaces**  $Q_{\text{old}}$ :

- Every approximation is used to create a new, better approximation.

With this, we can properly represent  $Q^*$ .

$$Q_{\text{new}}(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \cdot \max_{a'} (Q_{\text{old}}(s', a')) \quad (10.25)$$

**Definition 644**

**Infinite-horizon value iteration** is a process where we approximate the **infinite horizon Q-value**  $Q^*$  with a **large, finite horizon Q-value**  $Q^H$ .

To accomplish this, we compute  $Q^H$ , combined with a **termination** chance ( $p = 1 - \gamma$ ), from the **infinite horizon** problem.

- Each  $Q^h$  term is used to aim for a **better approximation** of  $Q^*$  than the one before.



It's more accurate to call these approximations of  $Q^*$ , rather than finite-horizon values  $Q^h$ . So, we'll change up our notation:

- We use  $Q_{\text{old}}$  (our old approximation) to compute  $Q_{\text{new}}$  (a new approximation), using the **finite-horizon equation** (discounted with  $\gamma$ ).

Throughout the process, we use the following equation:

**Key Equation 645**

The equation for **infinite-horizon value iteration** is:

$$Q_{\text{new}}(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \cdot \max_{a'} (Q_{\text{old}}(s', a'))$$

Once our Q value is **close to  $Q^*$** , we can find an **optimal policy**:

$$Q_{\text{new}} \approx Q^* \implies \pi^*(s) = \arg \max_a (Q_{\text{new}}(s, a))$$

Next, we'll figure out when to **stop** value iteration: when are we ready to use our  $Q^*$  approximation?

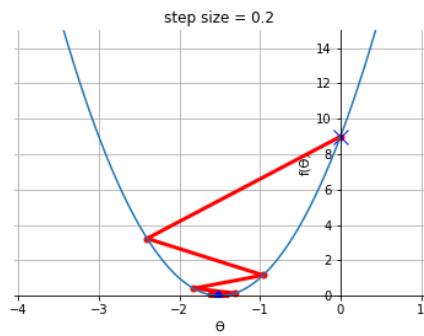
How do we know that value iteration converges at all? We have a **theorem** for that, in 12.2.14.

### 10.2.12 Convergence of Value Iteration

How do we know when to stop running our value iteration? We want  $Q_{\text{new}}$ , to **converge** to  $Q^*$ : it gets **close** to the answer we want!

How do we know *when* we're converging?

- We can use gradient descent as an example:



As we get **close** to the solution, our steps have to become **very small**.

We can see a pattern as we converge:

#### Concept 646

As we get close to **convergence**, our sequence will typically **stabilize**: the values become more and more **similar**.

We can use this "slowing down" progress to **detect** convergence.

We start at the blue x on the right. Each red line shows us "moving" to a new, better position.

If we're close to the solution and take a **big** step, we'd get **further away** again! So, our steps **must** get smaller.

Let's translate this to **value iteration**:

- As we converge, each  $Q_{\text{new}}$  value will be **very close** to the previous one.
- So, we'll detect convergence by seeing when our approximation is **changing by a very small amount**.

#### Concept 647

We **terminate** our value-iteration process when our **newest approximations**  $Q_{\text{new}}$  become **very similar** to each other.

If our approximations are similar (they change very little between timesteps), that means we're **converging** to  $Q^*$ .

Even if  $Q_{\text{new}}$  converges, how do we know it converges to  $Q^*$ ? We have **another theorem** for that, in 12.2.14.

### 10.2.13 Value Iteration: Termination Condition

Let's figure out how to represent this mathematically.

- We'll compare the two most recent approximations:  $Q_{\text{new}}$  and  $Q_{\text{old}}$ .
- We want these two **value functions** to be **similar**. How do we measure this?

#### Concept 648

To compare two functions, we see how different their **outputs** are, for each **input**.

$$\|f(x) - g(x)\|$$

So, we'll compare the values that we get for each  $(s, a)$  pair.

$$\|Q_{\text{new}}(s, a) - Q_{\text{old}}(s, a)\| \quad (10.26)$$

We want our value functions to be **similar**: that means they need to be similar for **all inputs**.

- We'll set a **maximum** for how different each output can be, called  $\varepsilon$ .

#### Definition 649

We want to see when  $Q_{\text{new}}$  and  $Q_{\text{old}}$  are **similar**.

We'll say that they're similar if **every output** is closer than a distance of  $\varepsilon$ :

Every  $(s, a)$  pair  $\forall s : \forall a :$  ...Must be at least this similar  $\|Q_{\text{new}}(s, a) - Q_{\text{old}}(s, a)\| < \varepsilon$

If we want to check if **all** of them are similar, we can focus on the **worst case**:

- Which  $(s, a)$  makes them the **most different**?

$$\max_{s,a} \left( \|Q_{\text{new}}(s, a) - Q_{\text{old}}(s, a)\| \right) \quad (10.27)$$

**Notation 650**

When we need to take the maximum over **multiple inputs**, we include **both of them** underneath the **max** notation.

$$\max_{x,y} (f(x,y))$$

The above asks, "if we can choose  $x$  and  $y$  to be anything, what's the **largest value** of  $f$  we can get?"

This equation will tell us whether we're finished.

**Key Equation 651**

The following equation tells us, "if we compare  $Q_{\text{new}}$  to  $Q_{\text{old}}$ , what's the **biggest difference** between their outputs?"

$$\max_{s,a} \left( \|Q_{\text{new}}(s,a) - Q_{\text{old}}(s,a)\| \right)$$

~~~~~

If their "biggest difference" is small ($< \varepsilon$), then you could say the two functions are **similar** for every input.

$$\max_{s,a} \left(\|Q_{\text{new}}(s,a) - Q_{\text{old}}(s,a)\| \right) < \varepsilon$$

This is our **termination condition**: once we reach this condition, our Q^* approximation is "good enough".

Finally, we have a complete value-iteration process.

Definition 652

To do **Q-value iteration**, we follow these steps:

1. Start with $Q_{\text{new}} = 0$.
2. Set $Q_{\text{old}} = Q_{\text{new}}$. This moves us forward 1 timestep.
3. Update Q_{new} .

$$Q_{\text{new}}(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \cdot \max_{a'} (Q_{\text{old}}(s', a'))$$

4. Repeat step 2-3 until we **converge**:

$$\max_{s,a} \left(\left| \left| Q_{\text{new}}(s, a) - Q_{\text{old}}(s, a) \right| \right| \right) < \epsilon$$

5. Return the optimal policy.

$$\pi^*(s) = \arg \max_a (Q^*(s, a))$$

If we write this as pseudocode:

```

INFINITE-HORIZON-VALUE-ITERATION( $\mathcal{S}, \mathcal{A}, T, R, \gamma, \epsilon$ )
1  for  $s \in \mathcal{S}, a \in \mathcal{A}$  :           # Every input pair
2     $Q_{\text{old}}(s, a) = 0$           # Start from  $Q^0 = 0$ 
3
4  while True: # Continue until converged
5
6    for  $s \in \mathcal{S}, a \in \mathcal{A}$  :       # Update  $Q_{\text{new}}$ 
7
8       $Q_{\text{new}}(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \cdot \max_{a'} (Q_{\text{old}}(s', a'))$ 
9
10   if  $\max_{s,a} \left( \left| \left| Q_{\text{new}}(s, a) - Q_{\text{old}}(s, a) \right| \right| \right) < \epsilon$            #If converged, we're finished
11    return  $Q_{\text{new}}$ 
12
13   $Q_{\text{old}} = Q_{\text{new}}$           #Re-use for next iteration

```

10.2.14 Convergence Theorems

We've made a couple pretty large **assumptions**: thankfully, each one has a **theoretical** justification.

In order to discuss these theorems, we need to clarify some notation:

Notation 653

Suppose we have run our value iteration for n steps (not necessarily enough for convergence within ε). For simplicity, let's say $Q = Q_{\text{new}}$: it's our current best approximation of Q^* .

Our policies:

- π^* is the optimal policy, based on Q^* .
- π_Q guesses the optimal policy, based on Q .

Our value functions:

- V_{π_Q} is the reward of using π_Q forever.
- V_{π^*} is the reward of using π^* forever.

If we succeed, then $Q \approx Q^*$, and $\pi_Q = \pi^*$.

~~~~~  
Our first assumption: "value-iteration makes our approximation better".

#### Theorem 654

$$\max_s \left( \left| V_{\pi_Q}(s) - V_{\pi^*}(s) \right| \right) \quad \text{never increases}$$

or,

$$\overbrace{\max_s \left( \left| V_{\pi_Q}(s) - V_{\pi^*}(s) \right| \right)}^{\substack{\text{Our worst-performing state } \pi(s) \\ \text{Never gets worse than this}}} \quad \underbrace{\text{never increases}}_{\text{never increases}}$$

By taking the max, we're getting the "worst" performing state for  $\pi_Q$ : we'll call this  $s_{\text{bad}}$ .

- This sets a **limit** on how bad our model can be.
- Iteration can't make any state worse than  $s_{\text{bad}}$  currently is.
- $s_{\text{bad}}$  can only stay the same, or get better!

This seems pretty weak: it's nice to know that value-iteration can't make  $\pi_Q$  much worse, but how do we know it gets better? How do we know it converges?

Another way to say "never increases" is "decreases monotonically".

### Theorem 655

When we run **value iteration** with  $\epsilon$ , we will eventually **improve**  $Q$  enough to meet our requirement:

$$\text{After value iteration, } \max_s \left( \overbrace{\left\| V_{\pi_Q}(s) - V_{\pi^*}(s) \right\|}^{Q \text{ will approach } Q^* \text{ (within distance } \epsilon)} \right) < \epsilon$$

This means that if we run value iteration long enough, it **will converge**, and get as close to  $Q^*$  as we want.

That's great!  $Q$  approaches  $Q^*$ , when we run value iteration.

We just have one more problem: just because  $Q$  is **close** to  $Q^*$ , doesn't mean they're close enough to get the optimal policy we want,  $\pi^*$ .

- Can get that policy?

### Theorem 656

If we choose the right  $\epsilon$ , our policy  $\pi_Q$  **will be** the optimal policy  $\pi^*$ .

$$\text{There is an } \epsilon \text{ where } \exists \epsilon > 0 : \quad \text{We find the optimal policy } \pi_Q = \pi^*$$

We don't know what value of  $\epsilon$  is required, but it's good to know that it always exists: we can always find the optimal policy.

One last useful computational trick:

### Concept 657

We can run the different parts of value-iteration **in parallel**, out-of-sync with each other.

As long as we update each  $(s, a)$  "infinitely many times" (as many times as we need), we will still **converge**.

## 10.3 Terms

### Section 12.0

- Timestep  $t$  (Review)
- State  $s_t$
- Input  $x_t$
- Transition function  $f_s$
- Output  $y_t$
- Output function  $f_o$
- State Machine
- Finite State Machine
- State Transition Diagram
- Linearity
- Time-Invariance
- Linear Time-Invariant System (LTI)
- Actions  $a_t$
- Deterministic State Machine
- Probabilistic State Machine
- Reward function  $R(s, a)$
- State-action pair
- Stochastic (Review)

### Section 12.1

- Action space
- State space
- Transition Model  $T(s, a, s')$
- (Probabilistic) State-Transition Diagram
- Transition Matrix  $\mathcal{T}(a)$
- Markov Decision Process  $(\mathcal{S}, \mathcal{A}, T, R, \gamma)$

- Policy  $\pi(s)$
- Value Function
- Finite Horizon
- Horizon  $h$
- Timestep  $t$  (Review)
- MDP Termination
- Finite Horizon Value Function  $V_\pi^h(s)$
- Expected Value
- Weighted Average (Review)
- Probability Distribution (Review)
- Infinite Horizon
- Infinite Horizon Value Function  $V_\pi(s)$
- Discounting
- Discount Factor
- Discounted Average
- MDP Lifespan
- Memorylessness
- Markov Property (Optional)

## Section 12.2

- Optimal Policy
- Argmax (Review)
- Q-Value Function (State-action value function)
- Dynamic Programming
- Value Iteration
- Convergence (Review)

# CHAPTER 11

## Reinforcement Learning

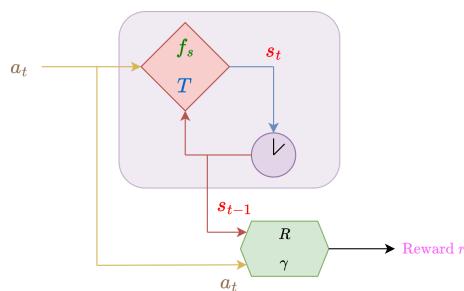
### 11.0.1 MDP Review

Last chapter, we explored MDPs, a tool for simulating a "game". We, the "player", choose which **actions** we take.

- Different **actions** can
  - Change the **state** of the world: what our system looks like.
  - Provide us with **rewards**, based on our actions.

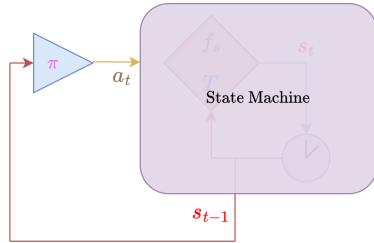
However, our system isn't perfectly consistent:

- The **transitions** between states are **probabilistic**: we don't know our exact next state, but we know the **odds** of each possible next state.



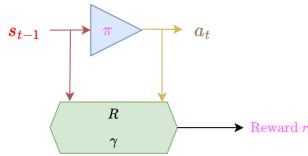
Here's our MDP: our model of a system that we can change over time.

Given complete knowledge of our system, we wanted to come up with the best possible strategy (**policy**) for getting the most reward, over time.



Our policy chooses different actions, based on what state our state machine gives us.

We evaluate our policies based on the **average expected reward**, for each state.



Combining these three parts (state machine, reward function, policy), we would find the best policy, using **value functions**, and **Q-value functions**.

### 11.0.2 What if we don't know as much?

There's a major limitation of this approach:

- It assumes we have know everything about our system.

#### Concept 658

**Value functions** can only be computed if you have **complete knowledge** of your MDP:

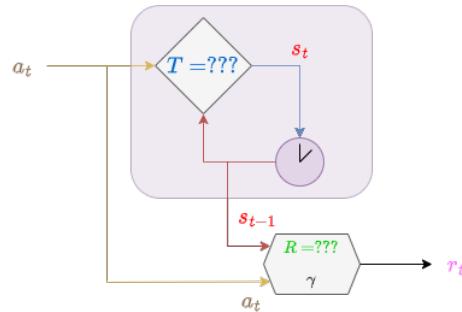
- What are the odds of your **state transitions**?
- What **rewards** will you get in different situations?

Without this information, it's not possible to compute the "value" of a policy, using our previous techniques.

$$V_{\pi}^H(s) = R(s, \pi(s)) + \sum_{s'} T(s, \pi(s), s') \cdot V_{\pi}^{H-1}(s') \quad (11.1)$$

- This equation is impossible to compute without those crucial variables T and R.

- But in plenty of real situations, you won't know exactly what effects your actions might have.



Often, we don't know T and R.

### 11.0.3 Learning about our MDP

If we don't know our transitions, or our rewards, our model is reduced to a simple box: based on an action, you see the next state  $s_t$ , and your reward  $r_t$ .



This simplified object is called the "environment" for our player.

Since we don't know what's inside, we reduce our MDP to a simple input-output machine.

The only way to learn our MDP is to **exploring** and gathering data.

The only way we can interact with our MDP is by taking **actions**. So, we do that:

- We are given the initial state  $s_0$ .
- We experiment, and take an action  $a_1$ .
- We learn some information:
  - We get reward  $r_1$ .
  - We transition to new state  $s_1$ .

Repeat.

We continue until we're satisfied, choosing actions and getting feedback.

**Concept 659**

In **reinforcement learning** (RL), we want to learn more about our MDP, so we **experiment**, by taking different **actions**.

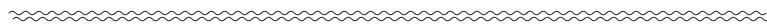
- We take an **action**, and see what it **does** (state transition, reward).
- We do it again. And again.

By **experimenting**, and continuously getting feedback, we slowly learn about our MDP. This gathers up our data:

$$\begin{bmatrix} s_0 & s_1 & s_2 & \cdots s_n \end{bmatrix} \quad \begin{bmatrix} r_1 & r_2 & \cdots r_n \end{bmatrix}$$

**Example:** You have a panel of buttons. You ask yourself, "what does this one do?", and press one of them.

- Then, you might ask: what if I press them in a different order? In different situations?
- As you learn more, you gradually figure out a "better" way to play.



This is very similar to how you might play a video game when you first pick it up.

#### 11.0.4 Reinforcement Learning

Now that we know what to do, we need to **formalize** it.

We can divide up this process into two:

Represent things with math, give each part a name, etc. Things that will make it easier to talk about.

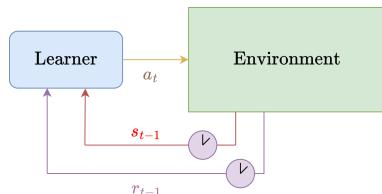
**Definition 660**

Our **reinforcement learning** (RL) problem can be divided into two main parts:

- The **learner**: this is the "player" of the game.
  - The learner chooses which **actions** to take: they decide the policy.
  - Based on what they observe from the environment, they learn to make **different choices**.
  - Eventually, the learner's goal is to make better choices, to get the most **rewards**.
- The **environment**: this is the "game" that the player is interacting with.
  - The environment reacts to the learner's actions, responding with a **reward** and **state change**.

The learner is trying to **learn** about the environment, and discover the best **policy**.

The learner chooses the action, and the environment teaches the learner. So, they work in a feedback cycle:



"Learning a better policy" is what we wanted in the MDP chapter: this time, it just takes more work.

Why does our diagram use  $s_{t-1}$  and  $r_{t-1}$ ? Because **past** data is used to make **future** decisions, like  $a_t$ .

Our learner makes decisions, while the environment gives feedback on those decisions. This feedback is used in the future to make better decisions.

**Notation 661**

In this chapter, we use capital R to represent the reward **function**, and lower-case  $r_t$  to represent a **single** reward at time t.

$$R(s_{t-1}, a_t) = r_t$$

If we expect our environment to behave like an MDP, that's what we'd put inside the "environment" block. But, RL isn't necessarily limited to that framework:

**Clarification 662**

So far, we've used MDPs as a concrete example for RL, but RL can be used for some other related systems, as well.

**Example:** One alternative environment is the "partially observable (PO) MDP".

This requires more inference than we'll cover in this class.

### 11.0.5 Supervised vs. Unsupervised vs. RL

RL is a bit different from our previous training frameworks: "supervised" and "unsupervised".

Let's review:

- **Supervised learning:** you're explicitly given an input  $x^{(i)}$ , and a desired output  $y^{(i)}$ 
  - **Example:** This is similar to being given a test, with the answer key.
- **Unsupervised learning:** you're given inputs, but you're not given an output: you have to look for patterns or structure without outside help.
  - **Example:** You're given a set of photos, and asked to sort them, based on what object is in the image. You aren't given labels.
- **Semi-supervised learning:** you're given *some* answers, but not most of them.

Reinforcement learning is a bit different:

**Concept 663**

**Reinforcement learning** (RL) provides data to the model differently from supervised / unsupervised frameworks:

- The model has some **choice** in which data it sees: it chooses **action**  $a_t$ , which affects the feedback  $s_t$  and  $r_t$ .
- This means the model doesn't just learn by observing the data: it **interacts** with it.

Over the course of training, our model can make **different choices** about what to learn, based on what it's already seen.

This approach forces our model to not only learn the structure of the data, but how to ask questions.

Just like how a student learns when to ask questions, and what they need to practice.

## 11.1 Reinforcement Learning Algorithms Overview

Our "learner" is more complex than our previous system for choosing actions:

- It gradually **learns** the rewards and state transitions.
- It has to not only choose the best rewards, but also choose what parts of the environment to **explore**.

But even so, it only really makes one decision: choosing actions. It's still a type of **policy**.

### Concept 664

A **reinforcement-learning (RL) algorithm** is a type of **policy**.

It chooses our **next action** based on all of its past data:

- States, actions, rewards

Similar to  $\pi(s)$ , the goal is to **maximize rewards**.

- However, our RL algorithm first has to **explore** different parts of the environment, to know what the rewards and transitions are.
- This is why it needs to use all of our past data: to **learn**.

When we're finished training, it may be possible to just keep the policy  $\pi : \mathcal{S} \rightarrow \mathcal{A}$ , and discard all of our past data.

This is similar to how, when we finish training our NN, we just give people the model, not the training data.

### 11.1.1 Evaluating RL algorithms

How do we evaluate our RL algorithm? There are a couple different ways:

#### Concept 665

We can **evaluate** our learning algorithm based on **how long** it takes for it to learn a mostly-**optimal policy**.

- In other words: "how long does it take to train?"

Our trained model contains the information we need to make decisions: we don't need all the training data.

In this scenario, we ignore the rewards we get while learning: our model is allowed to make **mistakes**.

In other situations, we want our model to do well while training:

**Concept 666**

We can also **evaluate** our learning algorithm based on **expected rewards** while training.

- In other words, "can it perform well, while still training?"

In this scenario we're focused on rewards **while learning**: we don't want our model to make as many mistakes.

Which do we usually use?

- We use the first one more **often**, because it's often easier to design and measure: we just train the model first, and keep track of the total time.
- The second one is more **challenging**: it's difficult to create a model that can perform reasonably well, while still learning.

But, sometimes the latter is necessary: you may need to train in real situations, where the rewards really matter.

**Concept 667**

If we have a **safe**, cheap environment to train in, it's easier to train first, and then figure out performance later.

But if you're training in a **costly** environment, you need to make sure your model performs well, even while still training.

**Example:** Suppose that you want to train a car in real traffic environments: simulations aren't good enough.

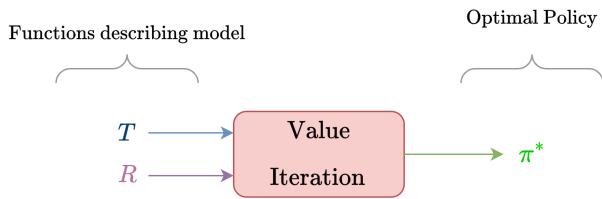
- You really don't want your car to make major mistakes in real traffic: even if you're "training", the accidents are very real.

### 11.1.2 Different types of RL models

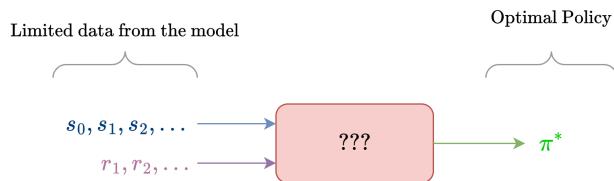
In a reinforcement learning situation, what we're missing is **information** about our model.

- We don't know our transitions  $T$ , or our reward function  $R$ .
- In our value-iteration setting, we used these to compute what's **optimal**:  $Q$  and  $\pi$ .

Value iteration can use **full information**:



Reinforcement learning is more restricted. We have **limited data**: some data points  $s_t$  and  $r_t$ .

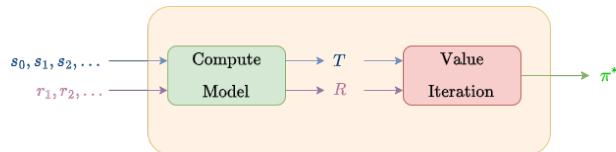


### 11.1.3 Types of Reinforcement Learning

There are multiple ways we can solve this problem. We'll focus on two types of approaches: **model-based** and **model-free** methods.

In a **model-based** RL algorithm, we use our data to try to **guess** the MDP model: we compute an approximation of  $T$  and  $R$ .

- Once we've computed  $T$  and  $R$ , we can use **value iteration**.



In this approach, we can re-use our previous logic.

#### Definition 668

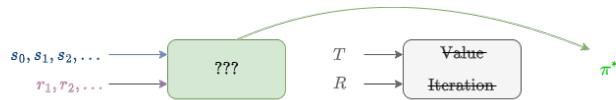
A **model-based** RL algorithm uses our previous MDP techniques to find the **optimal policy**. This approach requires **knowledge** of our model.

- First, we **approximate** our **model** ( $T$  and  $R$ ) based on data ( $s_t$  and  $r_t$ ).
- Then, we use that to do **value iteration**.

Our other approach is to use a **model-free** RL algorithm, where **skip** trying to compute  $T$  and  $R$ .

- Instead, we **directly** compute either  $Q$  or  $\pi^*$ .

We don't even bother learning our MDP.



We don't need  $T$ ,  $R$ , or value iteration.

### Definition 669

A **model-free** RL algorithm gives up our previous MDP techniques. Meaning, we **don't try** to directly compute our model ( $T$  and  $R$ ).

Instead, we find other ways to use our data ( $s_t$  and  $r_t$ ):

- In **Q-learning**, we approximate the **state-action value function**  $Q(s, a)$ , using **Q-learning**.
  - We find the best policy by maximizing  $Q(s, a)$ .
- In **policy search**, we represent our policy  $\pi$  with a **computable function**  $f(\theta)$ , and try to **optimize** that function.
  - We might use gradient descent, for example.

In this chapter, we will go in the following order:

- Model-free methods
  - Q-learning
  - Policy Search
- Model-based methods
- Bandit problems

## 11.2 Model-free methods

As we've already discussed, model-free methods are those where we don't learn  $T$  and  $R$  (our model). Instead, we skip over that, more directly learning our solution.

We generally boil these down into two kinds of approaches:

### Definition 670

We can sort **model-free methods** into two basic types:

- **Value-based** methods: we compute the value function  $V$  or  $Q$ .
- **Policy-based** methods: we compare policies  $\pi$  directly.

These two approaches aren't necessarily completely separate from one another:

### Clarification 671

Often, in more detailed models, the line between **value-based** and **policy-based** methods is blurry.

- Some techniques are somewhere in between.

We can even **combine** these into a single, more detailed algorithm.

- Some complex algorithms incorporate all of these elements: value functions, policies, transition/reward models.

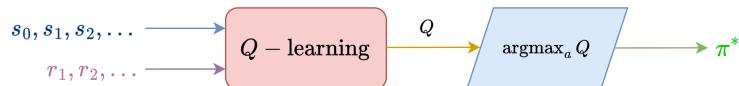
### 11.2.1 Q-learning: Computing Q from new data

We'll start with a popular **value-based** approach: Q-learning. Our goal is to compute Q directly.

- Then, we can find the optimal policy by maximizing Q.

$$\pi^*(\mathbf{s}) = \arg \max_a (Q(\mathbf{s}, a))$$

This is our process:



Rather than use T and R to compute Q, we compute Q directly.

Note the major difference:

#### Clarification 672

**Value iteration** and **Q-learning** can seem similar, because they both use our **model** to compute Q.

The main difference is that:

- Value iteration is used when you **fully understand your model** (T and R).
- Q-learning is used when we have **data points** (s<sub>t</sub> and r<sub>t</sub>).

In Q-learning, we determine Q based on our **experiences**.



How do we do Q-learning? Well first, let's remind ourselves of how we traditionally compute Q.

$$Q(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \cdot \max_{a'} (Q(s', a')) \quad (11.2)$$

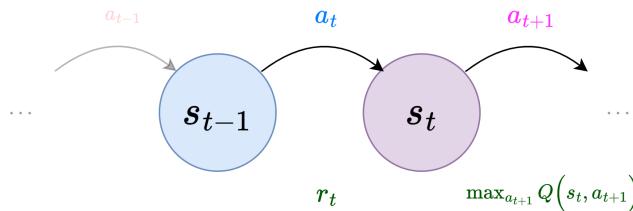
We don't have T or R. But, we can still use the basic idea:

- We take **one action** a, and end up going from **state** s to s'.
- Then, we use the **optimal policy** starting from state s': that's why we take the **max** of Q.

Let's try to apply this to one timestep of our "exploration data":

- We started in state  $s_{t-1}$ , and took action  $a_t$ .
- We moved to state  $s_t$ , and got reward  $r_t$ .

Our future rewards come from picking the **best** next action  $a_{t+1}$ .



We'll apply this to our Q equation, to get an **approximation** of Q.

$$Q_{\text{data}}(s_{t-1}, a_t) = r_t + \gamma \cdot \max_{a_{t+1}} Q_{\text{old}}(s_t, a_{t+1}) \quad (11.3)$$

Notice that we don't average over possible states ( $\sum_s TQ$ ):

- This "expected value" was used because we **didn't know** what our next state,  $s'$ , looked like.
- But in this case, we know which state we moved to:  $s_t$ .

### Key Equation 673

When deriving our Q-value, we broke our reward into two parts:

$$Q(s, a) = (\text{immediate reward}) + (\text{future reward})$$

If we apply this to one timestep of our simulation, we get:

$$Q_{\text{data}}(s_{t-1}, a_t) = r_t + \gamma \cdot \max_{a_{t+1}} Q_{\text{old}}(s_t, a_{t+1})$$

### 11.2.2 Q-learning: Making an update rule

Now, we have an approximation. But this approximation is only based on one data point. How do we incorporate **multiple**?

- We could **update** our Q-value every time we get new data for that state/action pair.

- That way, we can update it repeatedly, to incorporate multiple data points.

Our current equation doesn't allow us to "update" our Q-value, though: it **replaces** it. We'll modify the above equation to get our **update rule**:

- We want to **average** our new Q-value with our old one.

How much do we emphasize our new Q-value, versus our old one? We'll represent this with a **learning rate**  $\alpha$ .

#### Definition 674

When we **update** our Q-value, our **learning rate**  $\alpha$  ( $\alpha \in (0, 1]$ ) tells us how much we emphasize our new Q-value, based on **one data point**.

- $(1 - \alpha)$  tells us how much we emphasize our old Q-value, based on all past data points.

$$Q_{\text{new}}(s_{t-1}, a_t) = \alpha \cdot Q_{\text{data}}(s_{t-1}, a_t) + (1 - \alpha) \cdot Q_{\text{old}}(s_{t-1}, a_t)$$

We can also describe  $\alpha$  a little more conceptually:

#### Concept 675

If  $\alpha$  is **small** ( $\alpha \approx 0$ ), we care **very little** about new data.

If  $\alpha$  is **large** ( $\alpha \approx 1$ ), we are almost entirely **focused on** new data.

- 
- We call this a "**learning rate**", because it tells us how much we **learn** from new data.
  - But we could also think of it as a "**forgetting rate**": in order to learn from new data, we **pay less attention** to older data.

Our equation for Q is going to be messy, so let's change notation:

#### Notation 676

We update all of our variable names:

- $s = s_{t-1}$ ,  $s' = s_t$ ,  $a = a_t$ ,  $a' = a_{t+1}$ ,  $r_t = r$

As well as our value function:  $Q = Q_{\text{old}}$

If we plug in our previously calculated  $Q_{\text{data}}$ , we have our Q-learning equation:

**Key Equation 677**

In **Q-learning**, all of our Q-values start as 0 (similar to value iteration):

$$Q_{\text{new}}(s, a) = 0$$

With each new data point, we **update** our Q value:

$$Q_{\text{new}}(s, a) = (1 - \alpha) \cdot Q(s, a) + \alpha \cdot Q_{\text{data}}(s, a)$$

$$Q_{\text{new}}(s, a) = (1 - \alpha) \cdot Q(s, a) + \alpha \cdot \left( r + \gamma \cdot \max_{a'} (Q(s', a')) \right)$$

If we're being specific, we sometimes call this approach **tabular Q-learning**.

### 11.2.3 Selecting our action: $\epsilon$ -greedy

Now, we have a way to **update** our Q-value, based on a data.

- But we need a way to actually **get** our data: we need to start **exploring** the space.
- Which means our model **decides** what **data** it wants to **see**.

We could try always exploring whichever action seems most optimal. But this is a bad strategy when you're starting out:

**Concept 678**

It's not usually a good idea to **always** use the most "**apparently optimal**" strategy during training.

- There may be plenty of strategies that don't seem good *at first*, but will look more rewarding after some **exploration**.
- Your Q values are often very inaccurate, early in training.

**Example:** Suppose that there's some treasure at the end of a path.

- You might take 3 steps, and give up: walking around takes work, and you're not immediately rewarded.
- You'll miss out on that treasure, because you don't know it's there yet.

But exploring blindly isn't entirely helpful: it's *too slow*.

- It's often more useful to search near **high-reward** areas: these are more likely to be searched by (and be useful to) a **good policy**.

This is the **exploration vs. exploitation problem**.

### Definition 679

When we're trying to find the **best policy** for exploring a space, we run into a problem called **Exploration versus Exploitation**.

- **Exploration**: you're trying to **learn** more about the space, and you're not as focused on maximizing reward. You *explore* your options.
- **Exploitation**: based on what you've learned, you want to get the **maximum reward** from it. You *exploit* your knowledge.

If you explore more, you might learn how to get better rewards. But if you explore for too long, you'll waste time you could've spent taking advantage of that knowledge.

How much of each should we use? It depends on the context:

- If you only have **10 seconds** left in a game, it might not be worth it to explore anymore: you might as well cash in what you know how to do.
- But if you have **5 hours**, you're more likely to find something useful before the game ends: maybe you should explore more.

For Q-learning, our simplest option is to **randomly** alternate between the two modes: *explore* with probability  $\epsilon$ , and *exploit* with probability  $(1 - \epsilon)$ .

### Definition 680

The  **$\epsilon$ -greedy strategy** for Q-learning chooses our **actions** for interacting with the environment, **randomly**:

- With probability  $\epsilon$ , we choose an action  $a \in \mathcal{A}$  **uniformly, at random**.
  - We are equally likely to choose any action: we're **exploring**.
- With probability  $(1 - \epsilon)$ , we choose the action that gives us the **most reward**, based on what we know:

$$\arg \max_{a \in \mathcal{A}} Q(s, a)$$

- We're getting the most reward we can: we're **exploiting**.

How long do we want to run our Q-learning algorithm? It depends on the situation:

**Concept 681**

We can choose our **termination condition** for Q-learning based on our needs.

We could, for example:

- Terminate after a **fixed** number of timesteps T
- Terminate when our Q-values **aren't changing** much on successive iterations
- Terminate if we get **stuck** in the same state, or a loop

### 11.2.4 Q-learning

Based on this, we now have a completed Q-learning algorithm:

**Definition 682**

**Q-learning** is a strategy for learning the Q-values of our MDP, so we can find the **optimal policy** for our model.

We use the following steps:

- We set all Q-values to 0. Start from some initial state,  $s_0$ .

$$Q(s, a) = 0$$

- We repeat the following, until we reach our **termination condition**:

- Select an **action** based on Q and current state (possibly with  $\epsilon$ -greedy)

$$a_t = \text{select\_action}(Q, s)$$

- Record the **result**

$$\text{MDP}(s_{t-1}, a_t) = s_t, r_t$$

- Update our **Q-values** accordingly.

$$Q(s, a) \iff (1 - \alpha) \cdot Q(s, a) + \alpha \cdot \left( r + \gamma \cdot \max_{a'} (Q(s', a')) \right)$$

- Compute our **policy** based on Q.

Q-learning is guaranteed to **converge** under surprisingly simple conditions:

**Theorem 683**

Q-learning **converges** if

- Over an **infinitely-long run**, we visit every state an **infinite number of times**.

With this requirement, we ensure that our model doesn't decide on a sub-optimal strategy, without checking out other possibilities.

Guaranteed convergence does require our learning rate  $\alpha$  to **decay**, or gradually shrink over time.

- But typically, we set  $\alpha$  to a constant, for convenience.

select\_action isn't a specific function: in our case, it could just be  $\epsilon$ -greedy. But we could choose other options.

Okay, so an infinite amount of time isn't exactly promised to us... but it's better than a lot of other stricter convergence requirements!

We can set it to decay, but this also slows down the learning process.

Now that we have our completed Q-learning strategy, let's go through some details that we skipped over.

### 11.2.5 Initialization

When we're starting our Q-learning process, we have to choose some initial state,  $s_0$ .

- For some problems (like a chess game), there's a **natural choice** of initial state.
- For other problems (like a robot moving across terrain), there may be **multiple** possible "initial states".

In the latter case, we often **randomly** select our initialization.

#### Concept 684

When we're uncertain what **initial state  $s_0$**  to use, we often **randomly sample** from our state space.

This choice of initialization often **biases** what we learn about the state space: which sections we **visit**, what we **learn**, etc.

So, it's often helpful to run Q-learning through **several initializations**: we have one "run" of our MDP for each initialization.

- So we don't lose all of our progress, we usually modify it so that our Q-table (computed Q values) is **carried over** between different "runs".

#### Concept 685

To explore our **state-action space** (possible options) more thoroughly, we may take **several different paths** through our MDP.

- Each path starting with a **different initialization**,  $s_0$ .

We **share** our Q-values between these "runs" of our MDP, so that we can build up a more complete representation of the environment.

### 11.2.6 Action and state space

Our previous approach to Q-learning assumes that our **action space** and **state space** are both **discrete**.

- But this might not always be a realistic assumption. We might need a **continuous** space.

**Concept 686**

Our above approach to Q-learning is called **tabular Q-learning**.

It assumes that we have a **discrete** (typically finite) **state space** and **action space**.

- Other versions of Q-learning, on the other hand, allow this space to be **continuous**.

**Example:** A discrete state space might be  $\{1, 2, 3, 4, 5, 6\}$ . A continuous one might be  $[1, 6]$ .

We call it "tabular Q-learning" because our values could be stored in a **table**.

- $1 + \sqrt{2}$  is allowed in the latter, but not the former.

This causes us problems, though: if we have a continuous state/action space, we have an **infinite** number of possible states/actions.

- It's impossible to get the Q-values for all of these state-action pairs.

Many Q-learning variations enable continuous action/state spaces. Later, we'll focus on one example: **Deep Q-learning**.

### 11.2.7 An alternate view of Q-learning (Optional)

Consider our basic, conceptual Q-learning equation:

$$Q_{\text{new}}(s, a) = (1 - \alpha) \cdot Q(s, a) + \alpha \cdot Q_{\text{data}}(s, a) \quad (11.4)$$

We're averaging our immediate data, with our past experience with this Q-value.

We get something interesting if we rearrange it:

$$Q_{\text{new}}(s, a) = Q(s, a) + \alpha(Q_{\text{data}}(s, a) - Q(s, a)) \quad (11.5)$$

The right term could be seen as the **disagreement** between our new data, and past experience.

- And thus,  $\alpha$  tells us how much we **care** about that disagreement, and want to account for it.

$$Q_{\text{new}}(s, a) = Q(s, a) + \overbrace{\alpha(Q_{\text{data}}(s, a) - Q(s, a))}^{\text{"Error" of our old answer}} \quad (11.6)$$

This is an **update rule**: the difference between our new and old answer decides how we want to update.

**Concept 687**

We can view **Q-learning** as a direct **update rule**:

- We "update" our current Q value based on the **difference** from what the **newest data point** predicts.

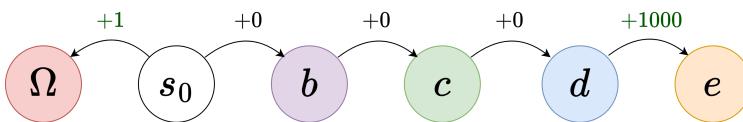
$$Q_{\text{new}}(s, a) = Q(s, a) + \alpha \overbrace{(Q_{\text{data}}(s, a) - Q(s, a))}^{\text{"Error" of our old answer}}$$

### 11.2.8 Problems with Q-learning: Slow Convergence

Because Q-learning only updates one state-action pair at a time, it often **converges slowly**.

Let's see an example:

- **Example:** We'll re-use our example of going down a long hallway, with treasure at the end.
- Each "state" is one tile of the hallway. We'll arrange them **left-to-right**: we can move left or right down the hallway. We start on  $s_0$ .



Above the arrows, we can see the reward we get for going left/right in each state.

If we go left, we get a small reward. If we go right, we'll *eventually* get a huge reward.

Being able to see from above, it's obvious to us that going **right** is better. But what does the robot see?

- Go right once. **No reward.**
- Go left once. **Reward!**
- We should go left!

Assume that every state transition we don't show (left/right) is +0.

So long as  $\gamma$  isn't really small: if our model is really likely to fail after 1 or 2 steps, then the right reward isn't worth it.

#### Concept 688

At first, our Q-learning algorithm will prioritize **short-term** rewards over long-term rewards.

- It hasn't had time to **find** rewards further from  $s_0$ .

Well, as the robot explores, it'll learn to get the reward, right? Let's see what happens as we move right, to our Q values.

$$Q(s, a) \quad \Leftarrow \quad (1 - \alpha) \cdot Q(s, a) + \alpha \cdot \left( r + 0.9 \cdot \max_{a'} (Q(s', a')) \right) \quad (11.7)$$

For simplicity, we'll use  $\alpha = 1, \gamma = 0.9$ .

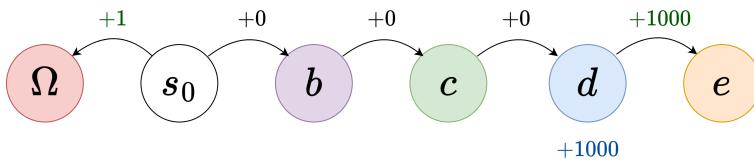
$$Q(\textcolor{red}{s}, \textcolor{brown}{a}) \quad \Leftarrow \quad \textcolor{violet}{r} + 0.9 \cdot \max_{\textcolor{brown}{a}'} (Q(\textcolor{blue}{s}', \textcolor{brown}{a}')) \quad (11.8)$$

Based on this model, our reward for going left ( $\leftarrow$ ) is simple:

$$Q(s_0, \leftarrow) = +1 \quad (11.9)$$

Let's go **right** ( $\rightarrow$ ) instead.

- Move from  $s_0$  to **b**: no reward.  $Q(s_0, \rightarrow) = 0$
- Move from **b** to **c**: no reward.  $Q(b, \rightarrow) = 0$
- Move from **c** to **d**: no reward.  $Q(c, \rightarrow) = 0$
- Move from **d** to **e**: reward!  $Q(d, \rightarrow) = +1000$



We learned that **d** is able to produce a +1000 reward.

We did it! We learned something, at the very end. Will our robot go the way we want?

- We start over from  $s_0$ . Let's **compare** the left and right rewards.

$$Q(s_0, \leftarrow) = +1 \quad Q(s_0, \rightarrow) = 0 \quad (11.10)$$

- Let's go left again!

No luck – it still prefers the **short-term** reward.

### Concept 689

Even once we find a reward, Q-learning will only update that **single state-action pair**.

- That means that nearby states, **don't know** about that reward!

We have to run Q-learning through a nearby state *again* to find the reward.

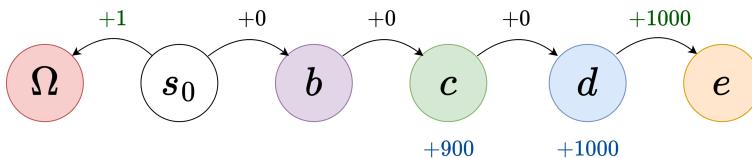
So, let's take another journey:

- Move from  $s_0$  to **b**: no reward.  $Q(s_0, \rightarrow) = 0$
- Move from **b** to **c**: no reward.  $Q(b, \rightarrow) = 0$
- Move from **c** to **d**: no reward. **However**, Q **remembers** that **d** can provide a reward!

$$Q(c, \rightarrow) \iff r + 0.9 \cdot \max_{a'} (Q(d, a'))$$

$$Q(c, \rightarrow) \iff 0 + 0.9 \cdot \overbrace{Q(d, \rightarrow)}^{+1000} = +900$$

We know that d is valuable. Thus, we've learned that c is valuable, because it's attached to d.



It's worth visiting c, because it allows you to visit d.

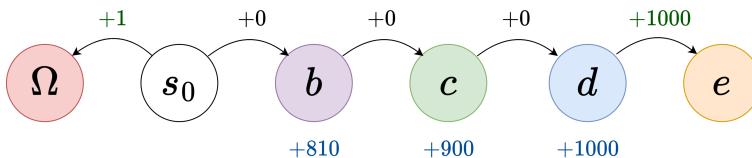
### Concept 690

Each time that we run though a path to a reward, **one more state** learns about the reward.

- If state d has an **action** with a **reward**, then **d is valuable**.
- If state c can move to d, **c is valuable**, because it gives a **path to reach d**.
- If state b can move to c, **b is valuable**, because it gives a **path to reach c**.
- This repeats until we reach  $s_0$ .

Each Q-learning run will update one more state.

If we continue, we get the result we're looking for:



We can now see that b has a lot of value. (The bottom number is the "expected value" we can get after reaching state  $s$ , if we make the best choice.)

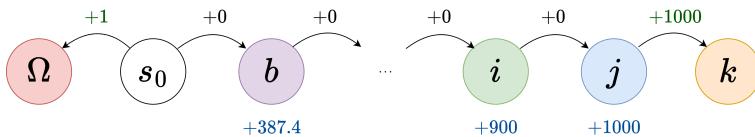
Let's compute  $Q(s_0, \rightarrow)$ , now that we know b is valuable:

$$Q(s_0, \leftarrow) = +1 \quad Q(s_0, \rightarrow) = +729 \quad (11.11)$$

Finally, we go right!

- But it took 4 trips right before we knew that.

This is already annoying, but it can get even worse: suppose we only reached the reward after moving right **10 times**.



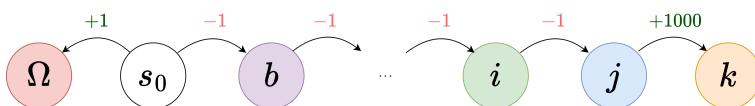
It's still worth it to go right, but it takes a painfully long time to figure that out.

Instead of making 4 trips right, we'll have to make 10 trips right.

- And imagine if the "reward" for going right was **-1** instead of **+0**.
- Our model would *always* prefer to go left, until the very end. Which means, it only has an  $\epsilon/2$  chance of moving right.
- Going right  $n$  times in a row has a chance of  $(\epsilon/2)^n$ .

Each trip is, thankfully, shorter than the last. But that's still really slow.

$\epsilon$  chance to move in a random direction, and  $1/2$  chance to randomly move right.



If moving left from ( $b, c, d, \dots$ ) is still **+0**, our model will try to avoid going right. It's even harder to make progress, now.

### Concept 691

The **longer** it takes to reach a distant reward, the more **difficult** it is to **propagate** that information back to  $s_0$ .

- This shows how *inefficient* Q-learning can be: only updating one **state-action** pair at a time, means that information travels **slowly** between states.

### 11.2.9 Deep Q-learning

Earlier, we mentioned that our state space  $\mathcal{S}$  and action space  $\mathcal{A}$  are **discrete** and relatively **small**.

- But what do we do if we need them to be **continuous**? \_\_\_\_\_

One solution is to treat Q like any other continuous variable we want to **predict**.

Or, just very large? A large finite space is still a pain.

- What do we do with complicated, continuous variables we want to predict? We use **neural networks**.

#### Definition 692

In **Deep Q-Learning**, we use **deep neural networks** to predict Q-values: a **regression** problem.

- This approach allows us to handle **continuous** state and action spaces.

To teach this network, we train it the way that we train any neural network, using data we receive while exploring:

- Input: **states** and/or **actions**
- Output: expected **reward**,  $Q_{NN}$ .

Our goal is to make the most accurate predictions of the Q-value. We determine Q based on each data point,

$$Q_{\text{data}}(s_{t-1}, a_t) = r_t + \gamma \cdot \max_{a_{t+1}} (Q(s_t, a_{t+1})) \quad (11.12)$$

So, we want our predicted Q value ( $Q_{NN}$ ) to be **as close** to  $Q_{\text{data}}$  as possible.

#### Definition 693

Our **deep Q-learning** neural network will use **squared error**:

$$(Q_{NN}(s, a) - Q_{\text{data}}(s, a))^2$$

In other words, our goal is for our NN ( $Q_{NN}$ ) to match the Q-values of our data points ( $Q_{\text{data}}$ ), as close as possible.

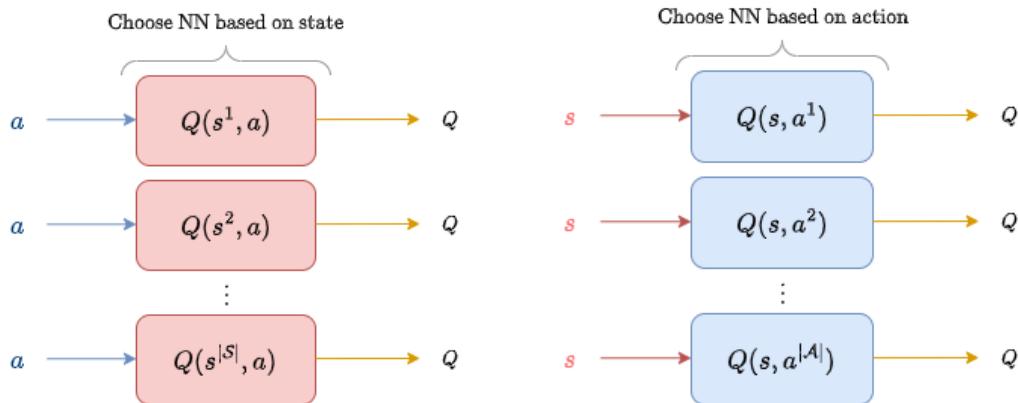
$$\left( Q_{NN}(s, a) - \left( r + \gamma \cdot \max_{a_{t+1}} Q_{NN}(s', a') \right) \right)^2$$

Note that, in our definition, we said states **and/or** actions: we might not have both as the input to our neural network. How?

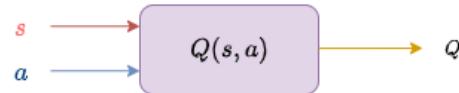
### Concept 694

There are three main ways we can design our **neural network**: in all cases,  $Q(s, a)$  is the **output**.

- Each **action**  $a$  has a separate neural network. **State**  $s$  is the input.
  - This only works with a **small, discrete action space**.
- Each **state**  $s$  has a separate neural network. **Action**  $a$  is the input.
  - This only works with a **small, discrete state space**.
- We have one neural network shared by **all inputs**. **State**  $s$  and **action**  $a$  are *concatenated* into the input.
  - This one is the most flexible, but it's **very hard** to find  $\arg \max_a Q(s, a)$ .



In one system, each **state**  $s^i$  has its own neural net. In the other system, each **action**  $a^j$  has its own neural net.



This version works for continuous state/action spaces, but comes with its own difficulties.

Unfortunately, deep Q-learning is often pretty unstable.

- But it's still useful enough to try, in a lot of contexts.

Improving, and then getting worse, for example.

### 11.2.10 Catastrophic Forgetting

Here, we'll address one of these forms of **instability**.

- When training a typical neural network, all of our data is **IID**: independent, and coming from the same distribution.
- But this is **not** the case for Q-learning.

#### Concept 695

In Q-learning, our data are **correlated in time** ("temporally correlated"). Meaning, **timing** affects our data.

- Why? Because two **states** which are "**near**" each other, typically behave **similarly**.
- If the **time** between two data points is **short**, they're probably **nearby** in state space. So, they're more likely to be similar.

**Example:** Consider a robot moving across the earth.

- **Example 1:** If, at time  $t$ , our robot is on a **mountain**, it's more likely to be on a mountain at  $(t - 1)$  and  $(t + 1)$ .
- **Example 2:** The 12 hours of daytime may seem very different from the 12 hours of nighttime.

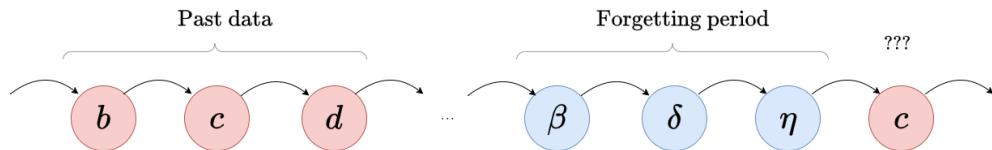
Why is this a problem? Because our neural network adjusts Q based on **new data**.

That means that our NN is capable of **forgetting**: if there's been a long time since we've used some information, it will be replaced by information from a **different context**.

#### Definition 696

**Catastrophic forgetting** occurs when our neural network hasn't seen a certain type of data in a long time, and **forgets** how to do a task.

- In deep Q-learning, it can occur when **recent** data doesn't reflect **past** data.
- So, our model *forgets* about the portion of the state space it visited in the past.



When we return to the red region, we've forgotten what we learned the first time!

This is still a problem, even if we don't return to the red region during this MDP run:

- It could be important for future runs.

### 11.2.11 Experience Replay

What do we do if a person is worried about forgetting something? You **remind** them of the past information.

- Perhaps they talk about that memory, or you periodically mention it to them.

This is our solution: we **keep track** of these past experiences, and re-use them later: we essentially "refresh" our NN, so that it doesn't forget.

This is called **experience replay**.

#### Definition 697

**Experience replay** is a technique for addressing **catastrophic forgetting**.

- In experience replay, we store our **past experiences** ( $s, a, s', r$ ) in a **replay buffer**.
- This buffer is used to "**remind**" our NN of past events.

After every timestep, we do two things:

- We store our newest experience in our replay buffer.
- We **randomly** pull  $n$  memories from our replay buffer, and "**re-experience**" them: we re-train our model, based on these past events.

This storage can get painfully large, though. This can be problematic:

- If the memories are too far back, they may just **not be relevant** anymore: they're in an undesirable part of the state space.
- It becomes **expensive** to store all of those memories.

So, we tend to only keep some of them.

**Definition 698**

Rather than storing *every* event in our **replay buffer**, we only keep the **k most recent memories**, in a **sliding window**.

- This prevents our memory from getting **too full**, or focusing on memories that are **too old**.

The best **size** for our sliding buffer **depends** on the problem, and what our state space is like.

Another reason that we like **experience replay** is for improving on a weakness we mentioned before:

**Concept 699**

Randomly reviewing **old memories** has a second benefit: it allows us to **propagate rewards** between states faster.

- Previously, we only updated **one state-action value**, for each experience.
- This means that, when we get our **reward**, we **only** update the state  $s_r$  we got the reward in.

With experience replay, we're more likely to **revisit** a state  $s_n$  "near" our reward:

- If  $s_n$  is near  $s_r$ , then  $s_n$  is more valuable, for being a **path** to the reward.

**Example:** This might, for example, help speed up the hallway problem.

By "nearby", we mean that there's an short series of actions  $a_t$  that moves us from  $s_n$  to  $s_r$ .

The one we used earlier in the chapter.

### 11.2.12 Fitted Q-learning

Here, we'll try a *different* approach for deep Q-learning, that avoids the "catastrophic forgetting" problem.

Our "forgetting" problem is caused by the fact that our data comes in a **particular order**: older data is learned earlier, and risks being forgotten, all together.

- Is there a way to "**shuffle**" the data we receive, before using it to train Q?

The problem is, we use Q to **choose** our data.

#### Concept 700

We want to gather data **before** training Q (so we can shuffle it).

- But we use Q to **decide** how to gather data.

We would need to have Q, in order to train Q – that seems paradoxical.

The solution? We have **two Q functions**: one we use to gather new data (but not train), another we train afterwards.

- $Q_{old}$ : trained on all **previous data**. We use this function to **decide** our actions, and gather more data.

- We **do not** re-train  $Q_{old}$  as we receive new data: we want to *avoid* training our data **in order**.

If we have no data yet,  $Q_{old}$  is just the "default" Q-value function:  $Q_{old}(s, a) = 0$ .

- $Q_{new}$ : once we've gathered enough data, we use **all of our data** (old and new) to train a new Q function **from scratch**.

- We **shuffle** our data, so that  $Q_{new}$  *also* avoids training our data in order.

Meaning, we start with  $Q_{new}(s, a) = 0$ , and then train.

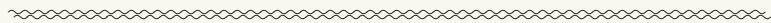
$Q_{new}$  can, then, be used to decide our future actions: it replaces  $Q_{old}$ .

$$Q_{old} \leftarrow Q_{new} \quad (11.13)$$

**Concept 701**

In typical Q-learning, we take an **action**, get data, and immediately **update** Q with that new data.

- This means our Q-value function is trained in the **order** we receive the data.



In **fitted Q-learning**, we separate the **data-gathering process** ( $Q_{old}$ ) from the **training process** ( $Q_{new}$ ).

- We use the same Q-value function,  $Q_{old}$ , to gather data for a while: **we don't re-train  $Q_{old}$  with our new data.**
- Then, using **all** of our data shuffled (new and old), we train a **new** Q-value function,  $Q_{new}$ .

We can compare the two processes. First, typical Q-learning:

- Use Q, get **one data point**.
- Update Q with **newest data point**.
- **Repeat**.

And now, fitted Q-learning:

- Use  $Q_{old}$ , get **many new data points**.
- Train  $Q_{new}$  with **all data**.
- Replace  $Q_{old}$  with  $Q_{new}$ .
- **Repeat**.

Using pseudocode:

**Definition 702**

**Fitted Q-learning** uses the following procedure:

```

FITTED-Q-LEARNING( $\mathcal{A}$ ,  $s_0$ ,  $\gamma$ ,  $\alpha$ ,  $\epsilon$ ,  $m$ )
1    $s = s_0$            # Initial state
2    $\mathcal{D} = \{\}$        # No data yet
3
4    $Q(s, a) = 0$         # Initial Q-values
5
6   while True:
7
8      $\mathcal{D}_{\text{new}} = \text{gather\_data}(Q, m)$       # Gather  $m$  points of data using  $Q_{\text{old}}$ 
9      $\mathcal{D} = \mathcal{D} \cup \mathcal{D}_{\text{new}}$           # Add new data to database
10
11     $\mathcal{D}_{\text{train}} = \text{convert\_data}(\mathcal{D})$       # Convert  $(s, a, s', r)$  to  $(x_i, y_i)$ 
12
13     $Q = \text{NN\_train}(\mathcal{D}_{\text{train}})$         # Train  $Q_{\text{new}}$ , ignore  $Q_{\text{old}}$ 

```

"convert\_data" turns each experience  $(s, a, s', r)$  into a data point  $(x_i, y_i)$ :

- Input  $x^{(i)}$ : **state** and **action**

$$x^{(i)} = (s, a) \quad (11.14)$$

- Output  $y^{(i)}$ : **expected reward** (based on reward  $r$ , and new state  $s'$ )

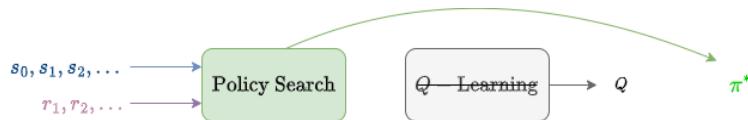
In other words, the Q-value, based on this data point.

$$y^{(i)} = r + \gamma \cdot \max_{a'} Q(s', a') \quad (11.15)$$

### 11.2.13 Policy Search

So far, we've been focused on methods for directly computing  $Q$ .

- But we could even go one step further: directly computing  $\pi$ .



Not only do we ignore  $T$  and  $R$ , but even  $Q$ .

Our strategy is to represent our policy as a **function** with **parameters** we can optimize.

#### Definition 703

In **policy search**, we represent our policy  $\pi$  as a *differentiable function*  $f$ , with **parameters**  $\theta$ .

- This approach treats our policy like a **hypothesis**.

$$\pi(s) = f(s; \theta) = a$$

Using this approach, we can **optimize** our parameters  $\theta$ , to get the greatest average reward.

- We need our function to be differentiable, for the same reasons as we needed for **gradient descent**.

One possible problem: often, we have a **discrete** action space. Our output will be a category: **not a continuous variable**.

- Discrete outputs aren't differentiable!

Our solution is the same as it was in classification: we use **probabilities**.

#### Concept 704

Rather than outputting the chosen action,  $a$ , we output the **probability** of that action.

- Because we chose our **action based on our state**, it's a **conditional probability**:

$$f(s, a; \theta) = P\{a | s\} = \text{Prob of choosing action } a, \text{ given state } s$$

This allows us to output a **continuous** variable.

Once we have our continuous function, we can use **gradient descent**.

### Key Equation 705

If  $\theta$  is **low-dimensional**, we can use **numerical gradient descent**/ascent to train our policy:

- Slightly **adjust**  $\theta_i$  by  $\varepsilon$ , see whether the **total reward**  $R$  is higher or lower: we approximate the derivative.

$$\frac{\partial R}{\partial \theta_i} \approx \frac{\Delta R}{\Delta \theta_i} = \frac{R(\theta_i + \varepsilon) - R(\theta_i)}{\varepsilon}$$

- Repeat for every  $\theta_i$  term, to get a **numerical gradient**.

$$\nabla_{\theta} R = \begin{bmatrix} \partial R / \partial \theta_1 \\ \partial R / \partial \theta_2 \\ \vdots \\ \partial R / \partial \theta_n \end{bmatrix}$$

- Apply **gradient ascent** (we want to maximize  $R$ , rather than minimize  $\mathcal{L}$ )

$$\theta \leftarrow \theta + \eta \cdot \nabla_{\theta} R$$

We could use **gradient descent** by choosing  $\mathcal{L} = -R$ .

For problems with higher-dimensional  $\theta$ , this is often too slow/inefficient.

- Instead, we use other, more complex algorithms, like REINFORCE.

But these algorithms are often tricky.

Policy search works best in those lower-dimensional cases.

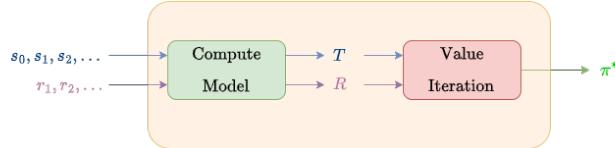
### Concept 706

**Policy search** works best when

- The policy's **functional** form is known and **simple**.
- Estimating the MDP would be **difficult**.

## 11.3 Model-based RL

Rather than try to directly compute  $\pi$  or  $Q$ , we could also *re-use* our previous techniques: we just need to compute  $T$  and  $R$ .



Once we compute  $T$  and  $R$ , we can do value iteration, like we did before.

### Notation 707

We want to **approximate**  $T$  and  $R$ .

We'll represent these approximations as  $\hat{T}$  and  $\hat{R}$ , respectively.

### 11.3.1 Computing $\hat{T}$

To compute  $\hat{T}$ , let's remind ourselves: what does  $T$  represent?

#### Definition 708

*Review from MDP Chapter, pt. 1:*

The **transition function**  $T$  gives the probability of

- Entering state  $s'$ ,
- Given that we chose action  $a$  in state  $s$

$$T(s, a, s') = P\{S_t = s' \mid S_{t-1} = s, A_t = a\}$$

After a transition, we will be in **exactly one** new state  $s'$ .

So, we want to compute this probability.

**Key Equation 709**

We can **approximate** the probability of event E happening, by **counting** the number of times it does/doesn't occur:

$$P\{E\} = \frac{\text{Number of times } E \text{ happens}}{\text{Total number of chances for } E \text{ to happen}}$$

or,

$$P\{E\} = \frac{\#E}{\#\text{Total events}}$$

Let's apply this to our situation: first, our "total events".

- We're computing the probability, **if** we chose action  $a$ , in state  $s$ .

$$\#\text{Total events} = \#(s, a) \quad (11.16)$$

If we were in a **different state**, or chose a **different action**, then that doesn't affect the probability.

- And we're looking for the chance that we **enter state  $s'$** .

$$\#E = \#(s, a, s') \quad (11.17)$$

So, we get:

$$\overbrace{\widehat{T}(s, a, s')}^{\text{Not our final equation}} \approx \frac{\#(s, a, s')}{\#(s, a)} \quad (11.18)$$

But there's a **problem** with this equation.

### 11.3.2 The Laplace Correction

In particular, we have *two* problems with this equation.

- If we have **no data**, what's our estimated probability?  $0/0$ .
  - This is **nonsense**: we're dividing by zero.
- If we have **one data point**, and didn't get E, what is our probability?  $0/1 = 0$ .
  - If we have only one data point, why should we be so sure that E **never** happens?

**Concept 710**

The equation

$$P\{E\} = \frac{\#E}{\#\text{Total events}}$$

Has two major flaws:

- It gives a **non-number** if we have no data.
- If we have **no events**  $E$ , it says that there's a **0% chance** that  $E$  will appear. Our model shouldn't be so confident.

**Example:** Imagine you flipped a coin 3 times, and happened to get heads 3 times. This will happen 1/8 of the time, on a fair coin.

- But our model has decided that there's a 0% chance of ever getting tails.

Let's solve each of these problems:

- We don't want to **divide by 0**. We need to add something to the **bottom**.
- We don't want to give a **0% chance of  $E$** , when we don't have enough data. We'll add something to the **top**.

$$\hat{T}(s, a, s') = \frac{\#(s, a, s') + b}{\#(s, a) + c} \quad (11.19)$$

How do we decide these constants? Well, let's return to the situation where we have **no data**.

$$\hat{T}(s, a, s') = \frac{b}{c} \quad (11.20)$$

We want  $b/c$  to be our "**default**" assumption: what do we think are the odds of transitioning to state  $s'$ , without any data?

- We have  $|S|$  different states, that we could **transition** to.
- Without any data, we have no reason to prefer one state over another. So, we assume all states to be **equally likely**.

If we split our probability evenly, we get:

$$\hat{T}(s, a, s') = \frac{1}{|S|} \quad (11.21)$$

We have our **correction terms**.

### Definition 711

The **laplace correction** is an adjustment to our **probability equation**, that solves the problems of

- Dividing by 0
- Computing probability to be 0, with very low data

The solution is to set a **default** probability for an event: we split probability **evenly** between all of our **N possible outcomes**.

$$P\{E \mid \text{No data}\} = \frac{1}{N}$$

Applying this to our general equation, we get

$$P\{E\} \approx \frac{1 + \#E}{N + \#\text{Total}}$$

As we gather more data, this correction term gradually **vanishes**.

**Example:** Let's say we have **5 possible outcomes**, and the odds of our event E are 40% (0.4). \_\_\_\_\_

- We'll compare the prediction for 5,50, and 500 data points.

And let's say our data exactly matches our probability, just to make things easier.

$$\frac{1+2}{5+5} = 0.3 \quad \frac{1+20}{5+50} \approx 0.381 \quad \frac{1+200}{5+500} \approx 0.398$$

- As we get more data, the laplace correction becomes less and less important.

Now, we can show our approximation for T.

### Key Equation 712

Our **approximation** for the transition function T is given by the equation

$$\hat{T}(s, a, s') = \frac{\#(s, a, s') + 1}{\#(s, a) + |S|}$$

### 11.3.3 Computing $\hat{R}$

Our reward function  $R$ , on the other hand, is much simpler to "approximate", because it's **deterministic**:

- The same state-action pair  $(s, a)$  will **always give the same reward**.
- So, we don't have to approximate our reward: if we get our reward once, we know **exactly** what it'll be.

#### Key Equation 713

Our "**approximation**" for the reward function  $R$  comes directly from our observations:

$$\hat{R}(s, a) = r_t \quad \text{if } s_{t-1} = s, a_t = a$$

This isn't really an approximation: it gives our **exact reward**.

$$\hat{R}(s, a) = r_t = R(s, a)$$

In some situations, our reward might not be deterministic.

- In which case, we can compute the reward probability function, or the expected reward for our state-action pair.

### 11.3.4 Solving our MDP

Once we've computed our approximations  $\hat{T}$  and  $\hat{R}$ , we can now construct the "approximated MDP":

$$\text{MDP}(\mathcal{S}, \mathcal{A}, \hat{T}, \hat{R}, \gamma) \tag{11.22}$$

And we can just solve it like any other MDP, using a technique like **value iteration**.

**Definition 714**

Our **model-based RL algorithm** has three basic parts:

- Computing our model  $(\mathcal{S}, \mathcal{A}, \hat{T}, \hat{R}, \gamma)$ .

$$\hat{T}(s, a, s') = \frac{\#(s, a, s') + 1}{\#(s, a) + |\mathcal{S}|}$$

$$\hat{R}(s, a) = r_t \quad \text{if } s_{t-1} = s, a_t = a$$

- Using that model to do **value iteration**.

$$Q(s, a) = \hat{R}(s, a) + \gamma \sum_{s'} \hat{T}(s, a, s') \cdot \max_{a'} (Q(s', a'))$$

- Using Q-values to find the optimal policy.

$$\pi^*(s) = \arg \max_a (Q(s, a))$$

The approach requires us to approximate T and R for every possible combinations of input variables.

- So we can't use it if our state space is too large.

**Concept 715**

**Model-based RL algorithms** work best when we have a **small, discrete** state space  $\mathcal{S}$ .

- They're difficult to generalize to **large, or continuous** state spaces.

## 11.4 Bandit Problems

Here, we'll move away from our MDP framework.

- We'll consider a different kind of problem: one without **states**.

### 11.4.1 Slot machines

Here's our general idea: we have  $k$  different **choices** we can make. We want to **explore** each option, and figure out which one is the best.

No states here: just actions you can take.

- This sounds simple: we just try each option **once**, and pick the best one.
- But it's not so easy: each choice has a **randomized** outcome.
  - And each lever has different odds of giving you a particular reward.

We can think of this problem like a "**slot machine**" with  $k$  levers: each one has different odds of giving you a reward.

#### Concept 716

In a **bandit problem**, you have a set of  $k$  independent **actions** you can choose from.

Slot machines have, in the past, been called "one-armed bandits", because they take your money. This is why we call these "bandit problems".

- Each action will give you a **randomized** reward.

Your goal is to maximize the **total rewards** you get (while training!)

Again, note that **states** have been completely removed from the problem.

### 11.4.2 Formalizing the Bandit Problem

We'll define each part of the bandit problem.

- First, we'll need our "options", or **actions**  $\mathcal{A}$ .
- What are the possible **rewards** we can get? That's our set  $\mathcal{R}$ .

$$a \in \mathcal{A} \quad r \in \mathcal{R}$$

Finally, we need the **probability** of getting a reward, if we take an action. This is similar to the **transition function**  $T$ :

- Rather than returning our next state  $s'$ , it instead gives us the **odds** of ending up in state  $s'$ .

If we take **action**  $a$  in **state**  $s$ .

In the same spirit, we'll have a function  $R_p$ , which gives us the **probability** of getting **reward  $r$** , from **action  $a$** .

$$R_p(a, r) = P\{\text{Getting reward } r \text{ from action } a\} \quad (11.23)$$

Using conditional notation,

$$R_p(a, r) = P\{r \mid a\} \quad (11.24)$$

Based on our inputs and outputs, we can write this with **function notation**:

$R_p : (\mathcal{A} \times \mathcal{R}) \rightarrow \mathbb{R}$ , where  $\mathbb{R}$  are real numbers, is also acceptable.

$$R_p : (\mathcal{A} \times \mathcal{R}) \rightarrow [0, 1] \quad (11.25)$$

These are the three parts of our bandit problem.

### Definition 717

A **bandit problem** has three parts:

- A set of actions  $\mathcal{A}$
- A set of rewards  $\mathcal{R}$
- A **reward-probability** function

$$R_p : (\mathcal{A} \times \mathcal{R}) \rightarrow [0, 1]$$

That tells us the **chance** of getting reward  $r$ , from action  $a$ .

Bandit problems are very, very important to reinforcement learning, and computer science.

- But we won't go through solutions/theorems here.

### 11.4.3 k-armed bandit problem

The simplest version of this problem is called the **k-armed bandit problem**:

- Every action ("arm") will either provide a reward ( $r = 1$ ) or not ( $r = 0$ ).

So, each action varies only by the **chance** that you get a reward.

In other words, we don't have "different types" of rewards.

**Definition 718**

The **k-armed bandit problem** is a simplified bandit problem, where

- You either get a simple **reward**, or you get nothing:  $\mathcal{R} = \{0, 1\}$
- You have **k actions** to choose from:  $|\mathcal{A}| = k$

#### 11.4.4 Exploration vs. Exploitation

In bandit problems, you have to balance **exploration vs. exploitation**:

- **Exploration:** Do we want to improve our **estimate**,  $\hat{R}_p$ ? The better our estimate is, the more likely we are to make optimal choices, moving forward.
- **Exploitation:** Do you want to maximize your rewards, **based on**  $\hat{R}_p$ ? You'll get more benefits short-term than if you keep exploring.

If you want to maximize your rewards, while still **learning**, you can't just explore, and you can't "exploit" blindly without data.

There's lots of interesting details we'll skip here, but the basic idea is:

**Concept 719**

The longer your horizon  $h$  (or the larger  $\gamma$  is), the longer you should continue to **explore**.

- The same amount of exploring takes up **smaller fraction** of your time: so, it takes away **less** of the exploitation reward.

**Example:** Consider a simplified version: you have **h turns** to play.

You spend  $n$  turns "exploring", and then  $h - n$  turns "exploiting". You get 0 reward for exploring.

- You get **\$10** for exploiting if you explored a **little** ( $n = 3$ )
- You get **\$15** for exploiting if you explored a **lot** ( $n = 10$ )

This example uses a lot of huge simplifications. Let's say these as "average" benefits for exploring/exploiting.

If you have only a little time ( $h = 15$ ), it's not really worth it to explore more.

$$\begin{array}{l} \text{More time to exploit} \\ \overbrace{(15 - 3) \cdot 10 = 120} \\ \text{More reward for exploiting} \\ \overbrace{(15 - 10) \cdot 15 = 75} \end{array} \quad (11.26)$$

If you have plenty of time ( $h = 100$ ), it's definitely worth it.

$$(100 - 3) \cdot 10 = 970$$

$$(100 - 10) \cdot 15 = 1350$$

(11.27)

Of course, this exploration/exploitation process is fairly sensitive to "luck": whether we get better or worse outcomes than the average.

**Concept 720**

"Bad luck" (getting low rewards for a valuable lever) is often more harmful than "good luck" (getting high rewards for a bad lever):

- If you get bad luck, you're **less likely** to keep trying that lever: you **won't find out** it's a good lever.
- If you get good luck, you'll probably **keep trying** that (seemingly profitable) lever: you'll get **lots of data** to learn that it isn't as good as you thought.

Which can happen very easily, with small sample sizes.

The longer we can train, the more likely we are to be near the true average.

### 11.4.5 Contextual Bandit Problems

Our typical bandit problems lack the concept of a "**state**". However, if we *do* need states, we can use a *contextual bandit problem*:

**Definition 721**

In a **contextual bandit problem**, we re-introduce **states  $s$** .

- Each state  $s$  has its own bandit problem.

## 11.5 Terms

- MDP (Review)
- Value function (Review)
- Q-value function (Review)
- Reinforcement Learning
- Learner
- Environment
- Supervised Learning (Review)
- Unsupervised Learning (Review)
- Model-based RL
- Model-free RL
- Q-learning
- Value iteration (Review)
- Learning rate  $\alpha$
- $\epsilon$ -greedy strategy
- Exploration vs. Exploitation
- Tabular Q-learning
- Deep Q-learning
- Temporally Correlated
- Catastrophic forgetting
- Experience Replay
- Replay Buffer
- Sliding Window
- Fitted Q-learning
- Policy search
- Conditional Probability (Review)
- Numerical Gradient Descent
- $\hat{T}$

- $\hat{R}$
- Laplace Correction
- Bandit Problem
- Reward-probability function  $R_p$
- k-armed bandit problem
- Contextual Bandit Problem

# CHAPTER 12

---

## Clustering

---

### 12.0.1 Why do clustering?

In chapter 4, we discussed **classification**: sorting data points into different groups, or **classes**.

**Example:** We might sort animals by **genetics**, or different sub-diseases that need different **treatments**.

Simplifying our data into categories can allow us to do better work, more easily.

This has lots of benefits:

- It could be used to make **decisions**. Sometimes, knowing the class of an object is enough to make a decision, by itself.
  - We could use this to understand the structure and **distribution** of our data.
  - We could **sort** different types of data to be processed **separately**.
- 

The problem is, this relied on us **knowing** what classes we plan to sort into.

This may seem obvious, but what if we're looking at something **new**? A disease we don't fully understand, or animals we've never seen before? How do we **classify** them?

In the past, we've done this ourselves, giving us lots of useful classifications. But, **computers** allow us to do this in new situations:

- **High-dimensional** datasets, with too much **complex** information for a human to make sense of.
  - **Example:** Looking for patterns in hundreds of genetic factors at the same time.
- Discovering **new classes faster** than ever using computers.
  - And thus saving human labor.
- Finding **patterns** in creative ways humans would never think to, especially for really **abstract** problems.

#### Concept 722

**Clustering** is like **classification**, where we want to assign things to **classes**: we call them **clusters**.

But, we use it when we **don't know** what groupings we want, so we have to **find** them.

We have some challenges ahead of us, though. How do we decide what things are "similar" or "different"? How do we create new classes, and know that they're meaningful?

## 12.1 Clustering Formalisms

### 12.1.1 Unsupervised Learning

The first thing we should note:

This problem is similar to classification, a **supervised** problem.

- It was **supervised** because we knew the **correct** labels for our data in advance.  
We just wanted to **teach** it to our computer.

The problem here: we **don't** know the correct labels! In fact, we're making them up as we go.

Because we aren't being "supervised" by a correct answer, we call this **unsupervised learning**.

#### Concept 723

**Clustering** is a type of **unsupervised learning**: meaning, we don't have a **correct** answer in advance.

The labels we create are not based on a **known** truth.

The **label** for data point  $x^{(i)}$  is written as  $y^{(i)}$ .

### 12.1.2 What is clustering?

So, if we don't know **what** our classes are, how do we figure out **which** classes to create?  
Well, we have to think of what we expect in a class.

Intuitively, we think of a class as a **collection** of things that are **similar** to each other, and more different from other classes.

So, two points in the **same class** are more similar: in RBF, we decided that "more similar" meant "low distance" in the input space.

- **Example:** Two people might look more similar if their heights are numerically closer: shorter distance.

Remember that input space is where we represent each data point using input variables.

Meanwhile, two points in **different classes** are more **distant**: they're further apart in input space.

- **Example:** Two people might look more different if their weights are numerically further: greater distance.

We'll call these "possible classes" that we discover, **clusters**.

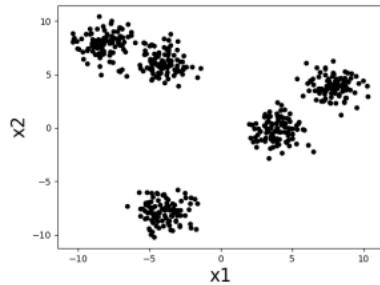
**Definition 724**

Informally, a **cluster** is a collection of **data points** that are all

- **Near** each other
- **Far** from the other clusters.

We use clusters as our way to discover **new classifications**.

**Example:** Below, we can visually mark out what looks like 5 distinct **clusters** in input space  $(x_1, x_2) \in \mathbb{R}^2$ :



This is an informal way to understand clusters, though. If we want to be more precise, we need to ask ourselves questions like:

- What does it mean for points to be "close" or "far"? How are we measuring distance?
- How many clusters do we want?
- How do we evaluate how "good" we are at clustering? Which clusters are closest to what we want?

## 12.2 The k-means formulations

In this section, we'll introduce a common way to do clustering, called the **k-means approach**.

### 12.2.1 Defining a cluster: The mean

We need to **define** what makes a "cluster" in order to move forward.

Suppose we have a collection of  $n$  points, which we informally think of as a "cluster".

We want the points within a cluster to be as **close together** as possible. So, you might ask, "how far is this point  $x^{(i)}$  from the rest of the cluster?"

- How do we measure this? Well, we could average the distance to every other point. That could get time-consuming.
- Instead, could we somehow use one point that "**represents**" the whole cluster?
  - With this system, if you want the **distance from the cluster**, we can just use that "representative" as the "position" of our cluster.
  - That way, you only need to compute 1 distance, not  $n - 1$  distances.
- This "representative" will be the **average** of all of our points: that way, it'll give us a rough idea of "where" all the points in the cluster are.

This is our **cluster mean**: when we want to know how far a data point is from the rest of the cluster, we'll compare it to the cluster mean.

#### Definition 725

We want to represent our **cluster** using its **mean**: the **average** of all of the data points in that **cluster**.

- This **cluster mean** will be treated as the "position" of our cluster.

Our goal is for the **cluster mean** to have the **minimum average distance** possible to all of our data points: it's as **close** to our points as we can get.

**Example:** We describe the "male lifespan" using **life expectancy**: the **average** time a male human lives for. Same for women as well.

### 12.2.2 k-means

Now, we've created **one** cluster. To extend this to **many** clusters, we just need each cluster to have its **own** mean.

We say that there are  $k$  of these clusters: this is why we call this the **k-means formulation**.

How do we decide which point goes in which **cluster**? Well, we want our points to be close to their cluster. So, we'll assign it to the **closest** one.

#### Concept 726

A **point** is assigned to the **closest cluster mean**.

For a point  $x^{(i)}$ , the **output** is which **cluster** ("new class") it has been assigned to:  $y^{(i)}$ .

Once we've successfully clustered using our **algorithm** below, we will find that both of these goals are met:

- Our points are **assigned** to the **closest** cluster mean.
  - This separates **different** clusters of points from each other:
    - If they're closer to different cluster means, they're in different groups.
- The cluster mean is the **average** of all of our points: the **minimum distance** to them.
  - This makes sure our cluster is made up of points that are **similar** to each other:
    - If our point is close to the **mean**, it's probably close to the **other** points in the cluster.

### 12.2.3 k-means loss

Now, we know what we **want** out of our clusters. But, the problem is, we don't know **how** to get those nice clusters.

So, first, we will have to **assign** our initial "cluster means": often, we **randomly** select some points from our dataset.

#### Concept 727

We **initialize** our clustering by **randomly** selecting one point to **represent** each cluster, which we call the **cluster mean**.

At first, each point is assigned to the **closest** cluster mean.

But as you'll notice, these points are **not** specially designed clusters! They're just a random **initialization**: we have to **optimize**.

**Clarification 728**

Notice that, when we **first** select our "cluster means", we don't get them by **averaging** any points: we choose them **randomly**.

That means, at first, is our cluster mean **isn't a true mean!**

Our k-means algorithm is designed to **fix** this problem.

In order to **improve** our clustering, it helps to have a way to measure the **quality** of a clustering: we need a **loss function**.

#### 12.2.4 One-cluster loss

Let's start with just one cluster: what do we want to **minimize**?

Well, we want the points within a cluster to be as **close together** as possible. So, we want to minimize the **distance** to the mean,  $\mu$ .

To make our function smooth, we'll use **squared distance** instead.

**Concept 729**

In **k-means loss**, we want to minimize the **square distance** from each point  $x^{(i)}$  to the **cluster mean**  $\mu$ .

$$D_i = \|x^{(i)} - \mu\|^2 \quad (12.1)$$

We'll add this up for each of the  $n$  data points in our cluster.

$$\mathcal{L} = \sum_{i=1}^n \|x^{(i)} - \mu\|^2 \quad (12.2)$$

#### 12.2.5 Building up to $k$ clusters

So, what do we do for each of our  $k$  clusters? Well, we can just **add** up the **loss** for them.

We'll use  $j \in \{1, 2, 3, \dots, k\}$  to represent our  $j^{\text{th}}$  cluster. Each cluster has a mean  $\mu^{(j)}$ .

$$\overbrace{\mathcal{L}_j}^{\text{Loss for only cluster } j} = \sum_{i=1}^n \|x^{(i)} - \mu^{(j)}\|^2 \quad (12.3)$$

Problem is, we're including **every** point  $x^{(i)}$  in **every** cluster! We want a way to filter by **cluster**: we only put one data point in each cluster.

Remember that we **label** clusters the same way we labeled **classes** before:

**Notation 730**

For a **data point**  $x^{(i)}$ , its **cluster** is given by

$$y^{(i)} \in \{1, 2, \dots, k\}$$

Where  $j$  represents the  $j^{\text{th}}$  cluster.

Cluster mean  $\mu^{(j)}$  **only** includes points in cluster  $j$ . So, when computing **loss**, we **only** want to include data point  $x^{(i)}$  when:

$$\underbrace{y^{(i)} = j}_{x^{(i)} \text{ is in cluster } j} \quad (12.4)$$

We'll do this using the following helpful **function**:

**Notation 731**

The **indicator function**  $\mathbb{1}$  tells you whether a statement  $p$  is true:

$$\mathbb{1}(p) = \begin{cases} 1 & \text{if } p \\ 0 & \text{otherwise} \end{cases}$$

Combined with our **condition** of matching clusters, this can be useful:

$$\mathbb{1}(y^{(i)} = j) = \begin{cases} 1 & x^{(i)} \text{ is in cluster } j \\ 0 & x^{(i)} \text{ is not in cluster } j \end{cases} \quad (12.5)$$

If we **multiply** this by our loss, it'll filter for situations where the clusters **match!** We can **eliminate** data points in a different cluster.

$$\mathbb{1}(y^{(i)} = j) \|x^{(i)} - \mu^{(j)}\|^2 = \begin{cases} \|x^{(i)} - \mu^{(j)}\|^2 & x^{(i)} \text{ is in cluster } j \\ 0 & x^{(i)} \text{ is not in cluster } j \end{cases} \quad (12.6)$$

### 12.2.6 k-mean loss: final form

So, we can **filter** by the data points in our cluster:

$$\mathcal{L}_j = \sum_{i=1}^n \underbrace{\mathbb{1}(y^{(i)} = j)}_{\text{Check cluster}} \cdot \underbrace{\|x^{(i)} - \mu^{(j)}\|^2}_{\text{Sq. dist from mean}} \quad (12.7)$$

And finally, we add up over many clusters:

$$\mathcal{L} = \sum_{j=1}^k \mathcal{L}_j \quad (12.8)$$

Using our equation, we get:

$$\mathcal{L} = \overbrace{\sum_{j=1}^k}^{\text{clusters}} \overbrace{\sum_{i=1}^n}^{\text{data points}} \underbrace{\mathbb{1}(y^{(i)} = j)}_{\text{Check cluster}} \cdot \underbrace{\|x^{(i)} - \mu^{(j)}\|^2}_{\text{Dist from mean}}$$

Let's clean that up:

### Key Equation 732

The **k-means loss** is given as:

$$\mathcal{L} = \sum_{j=1}^k \sum_{i=1}^n \mathbb{1}(y^{(i)} = j) \|x^{(i)} - \mu^{(j)}\|^2$$

Where:

- $\mu_j$  is the **cluster mean**: the **average** of the points in the  $j^{\text{th}}$  cluster.
- $\mathbb{1}(y^{(i)} = j)$  is the **indicator function**: meaning that we only **include** terms where the data point and mean are in the **same cluster**.

#### 12.2.7 Making further use of the indicator function (Optional)

We can actually use our **indicator function** to represent some of our **other** variables:

For example: the **cluster mean** is the average of data points, but **only** those belonging to that cluster.

So, we can use  $\mathbb{1}(\cdot)$  to **filter** those other data points out:

$$\mu^{(j)} = \frac{1}{N_j} \sum_{j=1}^k \underbrace{\mathbb{1}(y^{(i)} = j)}_{\text{check cluster}} \cdot \underbrace{x^{(i)}}_{\text{data points}} \quad (12.9)$$

And how large is  $N_j$ ? We can just **count** the number of data points in cluster  $j$ :

$$N_j = \sum_{j=1}^k \mathbb{1}(y^{(i)} = j) \quad (12.10)$$

One more loose end: we've been focusing on **square distance** as loss function. We want to

minimize this, but we're doing this over multiple data points.

So, really we want to minimize the **average** of that. This is a very common (and very useful!) property of a distribution called the **variance**.

### Definition 733

The **variance** of a dataset is the **average square distance** from the **mean**:

$$\text{Var}[X] = \frac{1}{N} \sum_{i=1}^N \|x^{(i)} - \mu\|^2$$

It tells us how **spread out** our data is.

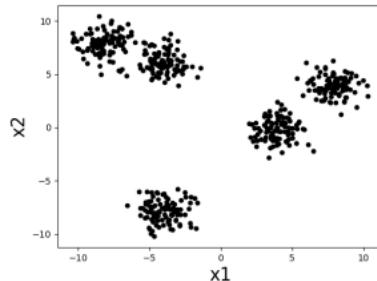
That means, our loss function is meant minimize the total **variance**.

Almost, the factor of  $1/N$  is missing: since this constant won't change much as we improve our clustering, we'll leave it alone.

### 12.2.8 Initializing the k-means algorithm

Now that we have our **clusters**, **means**, and a **loss** function for evaluating them, we can begin looking for a better **clustering**.

We'll start out with a **dataset** we want to cluster: we'll use the one from the **beginning** of the chapter:



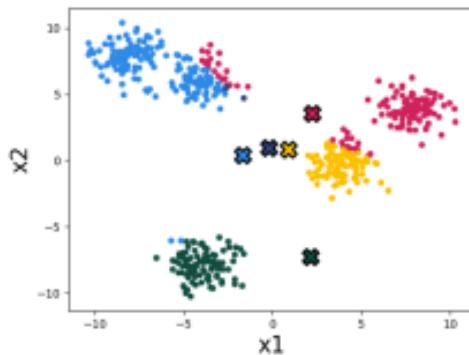
We could cluster this visually, but we want our machine to be able to do it for us.

First, we need to decide on our **number** of clusters. When you can't **visualize** it, this can be **difficult** - how many is too many or too few?

But, for now, we'll **ignore** that problem, and say  $k = 5$ .

Let's **randomly** assign our initial cluster means, and assign each point to the **closest** cluster:

Above, we suggested selecting a random data point as our cluster mean, but you can also just pick a random position, like we did here.



1) Initial assignments

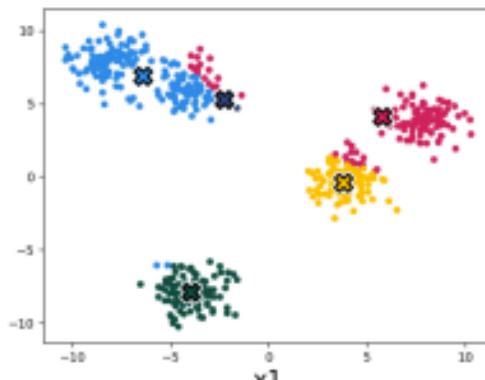
This is our starting point for the algorithm.

### 12.2.9 First step: moving our cluster means

As we mentioned before, these points aren't **actually** the average of their cluster: you can tell that by looking at it.

We want to **minimize** the variation in our cluster: that's why we're using the mean.

So, let's fix this: we'll take the **average** of all the points in each **cluster**, and **move** the cluster mean to that position.



2) Update means

And now, our cluster means are closer to all our data points!

#### Concept 734

One way **minimize** the **distance** between the **cluster mean** and its **data points** is:

- Take the **average** of all the points in the cluster, and **reassign** the cluster mean to that average.

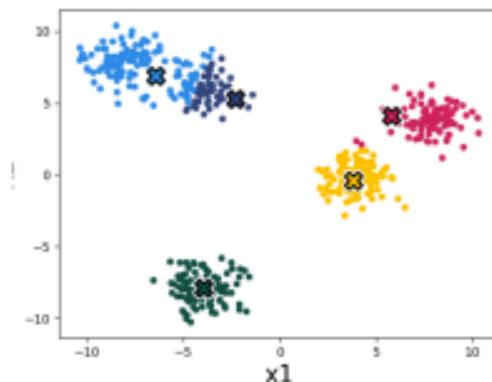
### 12.2.10 Second step: Reassign data points

We've **improved** our model by moving the cluster mean.

The problem is, we originally **assigned** every point to the **closest** cluster mean.

If the cluster means **move**, then some points might be closer to a **different** cluster now!

If so, we can **improve** our clustering further by reassigning points to the cluster they're **closest** to!



3) Update assignments

#### Concept 735

Another way **minimize** the **distance** between the **cluster mean** and its **data points** is:

- After the **means** have been **moved**, **reassign** the **data points** to whichever mean is **closest**.

### 12.2.11 The cycle continues

But wait - now that we've changed the points in each cluster, our cluster mean might not be the **true** average!

So, we can, again, improve our loss by taking the average of each cluster, and moving the cluster mean.

This creates a cycle that continues until we **converge** on our final answer.

#### Concept 736

Together, both of our steps for **improving** our clusters create a **cycle** of **optimization**:

- **Moving** our cluster mean **changes** which point should go in each cluster.
- **Reassigning** points to different clusters **changes** our cluster mean.

### 12.2.12 The k-means algorithm

These two steps make up the **bulk** of our algorithm:

#### Definition 737

The **k-means algorithm** uses the following steps:

- First, we **randomly** choose our **initial** cluster means.

Then, we **cycle** through the following two steps:

- **Reassign points** to the cluster mean they're closest to.
- **Move** each **cluster mean** to the average of all the points in that cluster.

Until our clusters means **stop** changing.

When we run our algorithm on the above dataset, we get:

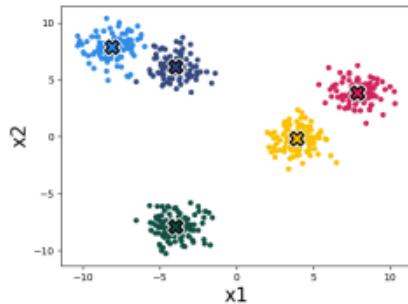


Figure 6.3: Converged result.

Note that, our cycle **works** because changing **either** cluster mean or point assignments allows you to further **improve** the **other** step.

So, if we're **not** changing one of them, the other one won't **change** either: the cycle is **broken**, and we can **stop**.

Another nice fact: it can be shown that this algorithm does **converge** to a local minimum!

This is our termination condition!

#### Concept 738

The **k-means algorithm** is guaranteed to **converge** to a **local minimum**.

### 12.2.13 Pseudocode

```

K-MEANS( $k, \tau, \{x^{(i)}\}_{i=1}^n$ )
1  $\mu, y = \text{randinit}$       #Random initialization
2 for  $t = 1$  to  $\tau$       #Begin cycling
3
4      $y_{\text{old}} = y$       #Keep track of last step
5
6     for  $i = 1$  to  $n$ 
7          $y^{(i)} = \arg \min_j \|x^{(i)} - \mu^{(j)}\|_2^2$       #Reassign data point to closest mean
8
9     for  $j = 1$  to  $k$ 
10         $\mu^{(j)} = \frac{1}{N_j} \sum_{i=1}^n \mathbb{1}(y^{(i)} = j) x^{(i)}$       #Move cluster mean to average of cluster
11
12    if  $\mathbb{1}(y = y_{\text{old}})$ 
13        break      #If nothing has changed, then the cycle is done. Terminate
14
15 return  $\mu, y$ 

```

### 12.2.14 Using gradient descent: minimizing distance to $\mu$

We can also use **gradient descent** to solve this problem!

We want to **minimize** our loss  $\mathcal{L}$ , and we do this by **adjusting** our cluster means  $\mu^{(j)}$  until they're in the **best** position.

**Concept 739**

We can solve the **k-means problem** using **gradient descent**!

So, we want to **optimize**  $\mathcal{L}$  using  $\mu$ :

$$\mathcal{L}(\mu) = \sum_{i=1}^n \mathbb{1}(y^{(i)} = j) \left\| \mathbf{x}^{(i)} - \mu^{(j)} \right\|^2 \quad (12.11)$$

Rather than dealing with the indicator function  $1(\cdot)$ , we could instead just consider whichever  $\mu$  is closest: **minimum** distance.

$$\underbrace{\min_j}_{\text{Minimizing}} \underbrace{\left\| \mathbf{x}^{(i)} - \mu^{(j)} \right\|^2}_{\text{distance}} \quad (12.12)$$

This **automatically** assigns every point to the closest **cluster** before we get our loss! So, all we need to worry about is  $\mu_j$ .

**Notation 740**

Instead of using an **indicator function**  $\mathbb{1}(p)$ , we can represent **cluster assignment** another way: using the **function**  $\min_j$ .

It can give **minimum distance** from  $x^{(i)}$  to one of the cluster means: it picks the **closest** mean.

This **automatically** assigns the point to the **closest** cluster, making our job easier.

$$\mathcal{L}(\mu) = \sum_{i=1}^n \widehat{\min}_j^{\text{Nearest cluster}} \|x^{(i)} - \mu^{(j)}\|^2 \quad (12.13)$$

Now, we can do gradient descent using  $\frac{\partial \mathcal{L}(\mu)}{\partial \mu}$ .

$\mathcal{L}(\mu)$  is **mostly** smooth, except when the cluster assignment of a **point** changes. So, it's usually smooth **enough** to do gradient descent.

We move our means until they're minima!

**12.2.15 Getting labels**

Once we've finished gradient descent, and we've **minimized** our loss, we can get our **labels**.

The "min" function gives the **output** value that we get by minimizing. In this case, average **squared distance** from the cluster mean: the **loss**.

Meanwhile,  $\text{argmin}$  gives us the **input** value that gives us the minimum output. In this case, the **choice of cluster means** that gives the minimum distance.

So,  $\text{arg min}$  gives us the cluster closest to each point: that's our **label**!

We can use this notation to get our **labels**.

**Notation 741**

After **optimizing**  $\mu$ , our **labels** are given by:

$$y^{(i)} = \arg \min_j \|x^{(i)} - \mu^{(j)}\|^2$$

Using gradient descent can give us a **local** minimum, but our surface is not fully **convex**: so, we don't necessarily get a **global** minimum.

Of course, this is also true for the k-means algorithm.

Even though individual terms of squared distance may be convex, adding min terms may not be convex.

## 12.3 How to evaluate clustering algorithms

The biggest problem with clustering algorithms is that they're **unsupervised**: this makes it much harder to know if we've gotten a **good** result.

This is partly because our **loss** function doesn't necessarily tell us if clustering is **useful**, or represents the data **accurately**.

It just tells us if our points are **close** to their cluster **mean**. That doesn't always mean the clustering is **good**.

**Example:** Imagine **every** single point in the dataset gets its **own** cluster mean. The **distance** to the cluster mean would be 0 (low loss), but this isn't very **useful**!

### Clarification 742

The **k-means loss function** does **not** tell us if we have a good and **useful** clustering or not.

This isn't useful because nothing has changed: we've gone from having  $n$  separate data points, to having...  $n$  separate clusters.

It only tells us if the points in our clusters are **close** to their **cluster mean**.

This can help us make **better** clusters, but that does not mean they are **good** or what we **want**.

Without having "true" labels, we have to find other ways to **verify** our approach.

We'll do two things to **approach** this problem:

- We'll look at some of the ways our **algorithm** can go wrong (or right).
- Then, we'll find **better** ways to evaluate our clusterings than just looking at the **loss**.

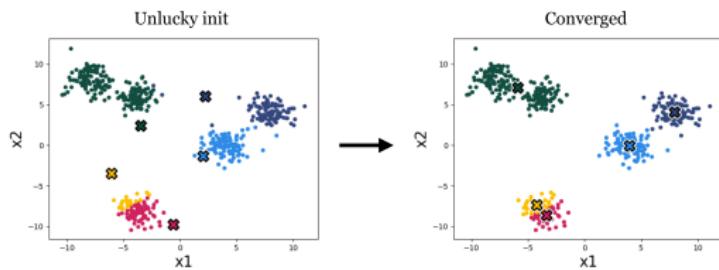
But, always remember that a "good" clustering is partly **subjective**, and depends on what you **want** to accomplish.

### 12.3.1 Initialization

The first problem we have is related to something we mentioned at the end of the last section: k-means is not **convex**.

That means we can find **local** minima that are not the **global** minimum: our **initialization** (our **starting** clusters) can affect whether we end up in a useful minimum.

The reason why is, mathematically, the same as when we first introduced the idea of a local minimum.



In this example, notice that we ended up with convergence on some very **bad** clusters: the bottom cluster is split in **half**!

The easiest way to resolve this is to run k-means multiple times with different initializations.

Other techniques exist, but this is the simplest one.

#### Concept 743

Getting an **unlucky initialization** can result in **clusters** that aren't **useful**.

We try to **solve** this by running our algorithm **multiple times**.

### 12.3.2 Choice of k

One important question we decided to **ignore** earlier was: **how many** clusters should we pick in advance?

Especially for **complex** data, we **don't know** how many natural clusters there will be.

But our number of clusters matter: because it's a parameter determines **how** our learning algorithm runs (rather than being chosen *by* the algorithm), it's a **hyperparameter**:

#### Concept 744

Our **number of clusters** k is a **hyperparameter**.

And, choosing too high *or* too low can both be **problematic**:

- If we set k too **high**, then we have more clusters than actually **exist**.
  - This can cause us to **split** real clusters in half, or find **patterns** that don't exist.
  - In a way, this resembles a kind of **overfitting**: we try to **closely** match the data, but end up fitting **too closely** and not **generalizing** well: **estimation error**.
  - **Example:** The **extreme** case looks like the example we mentioned **before**: when labeling animals, we could make... a different **species** for every single instance of **any** animal we find.
- If we set k too **low**, we don't have **enough** clusters to represent our data.

That doesn't sound very helpful.

- This means some clusters will be **lumped together** as a single thing: we **lose** some information.
- In this case, it's **impossible** to cluster everything in the way that would make the most **sense**: we have **structural error**.
- **Example:** Let's say we wanted to **sort** fish, birds, and mammals into **two** categories: we might just **divide** them into "flies" and "doesn't fly".

That's some information, but often not enough!

#### Concept 745

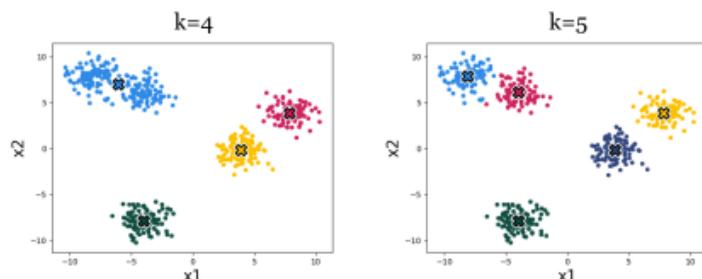
When choosing  $k$  (our **number of clusters**), we can cause **problems** by picking an inappropriate **value**:

- **Too many** clusters (large  $k$ ) can cause **overfitting** and **estimation error**: we find patterns we don't want.
- **Not enough** clusters (small  $k$ ) causes **structural error**: it prevents us from correctly **separating** data.

### 12.3.3 Subjectivity of $k$

Not only is it hard to choose a "good" value of  $k$ , what a good value of  $k$  is can really depend on your opinion, and what you know about reality.

For example, consider the following example:



Which of these two clusterings is more accurate?

Should the top left be **one** cluster, or **two**? It's hard to say!

Even if you're **sure**, you might **disagree** with others, or find that the best one depends on your **needs**.

So not only can  $k$  values be too high or too low, they can also be **debatably** better or worse!

**Concept 746**

The **best** choice of **clustering** is not entirely objective: it can depend on your **opinion**, or how you plan to **use** the clustering.

What do we mean by, what we're "**using**" the clustering for? We'll get into that later, but in short: we might use **clusters** to make sense of **information**, or to make better **decisions**.

Different clusterings might be good when you want a different kind of understanding.

**Example:** The understanding you get from high-level comparisons (plants vs animals vs bacteria) is different from low-level comparison (cats vs dogs).

#### 12.3.4 Hierarchical Clustering

That last example reveals something: not all types of groups are the same! Some are much broader than others, for example.

If two types of groups are different, then why do we only have to have one type in our clustering? We don't have to restrict ourselves to a single k.

Instead, we could treat some groups as inside of other groups: we call this a *hierarchy*, because some groups are "higher" on the scale.

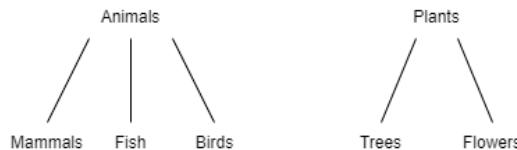
**Definition 747**

**Hierarchical Clustering** is when we cluster at multiple different **levels**.

Some groups are **high-level**, or **coarse**: they are groupings that contains **more elements**: items in the same group can be more **different**.

Some groups are **low-level**, or **fine**: they are groupings that contain **fewer elements**: items in the same group have to be very **similar**.

**Example:** Categorizing living things is done using hierarchical clustering: some groupings **contain** other groupings.



The top row of clusters are more **coarse**, while the bottom row is more **fine**.

We can **split** our groupings into **smaller** and smaller ones, to be as **fine-grained** as we need. Useful!

### 12.3.5 k-means in feature space

One **important** consideration: when working with **feature** representations, we found that sometimes, feature transformations made it **easier** to **classify** data.

Clustering is very **similar**, so, could we do the same here? It turns out, we **can!**

Often, it wasn't even possible without those transformations!

Rather than directly clustering in **input** space  $x$ , we can **process** our data using features, and then cluster in **feature** space  $\phi(x)$ .

#### Concept 748

**Feature transformations** can be used to make it easier to **accurately** cluster our data in a **meaningful** way.

There are some **other** reasons to do feature transformations, though: imagine that our data is **stretched** out along axis  $x_1$ , but not  $x_2$ .

$x_1$  distances would be **larger** in general: it would contribute more to our distance metric! We could correct for this by **standardizing** our data: **scaling down** the more stretched axis.

This would be a **feature** transformation, and would make it **easier** to do our **clustering**.

### 12.3.6 Solutions: Validation

Now, we start trying to answer the **question**: how do we **check** whether we have a **good** clustering?

Well, first, we can check for a **poor fit** (or overfitting) using new, **held-out** testing data: do we get **low loss** on that testing data?

If we **don't**, then our clusters definitely aren't **representative** of the overall **dataset**: they don't **generalize** to new data.

#### Concept 749

If our clusters give **large testing loss**, then they aren't **generalizing** well, and are probably **not representative** of the overall distribution.

So, we already know our clusters **don't fit the distribution**.

### 12.3.7 Solutions: Consistency

But, just like for classification/regression **validation**, we don't only run our algorithm **one time**: we'll run it **many** times, with different training and testing sets.

We can't **just** use the loss, though: having **more** clusters could make our error lower, without making a better clustering, for example.

Another thought: we're trying to find some patterns **inherent** in the data. The idea is: if the pattern we're finding is **real**, we should find a similar pattern **each time!**

So, we look to see if our clusters are **consistent** when we generate them using different training data: if they aren't, then it's possible we're not finding the "real" patterns in the data.

Different training data from the same distribution, of course.

### Concept 750

If our **clusters** accurately **reflect** the underlying classes of data, then we should expect some **consistency** of which clusters we **generate** by running k-means many times.

If our clusters aren't **consistent**, then we might doubt if any of them especially reflect the **distribution**, rather than **noise**.

If our clusters are **consistent**, then we're probably seeing something about the **real** dataset.

### 12.3.8 Solutions: Ground Truth

But, even if we're getting something **consistent**, that doesn't mean we're seeing the patterns that **matter**.

If it was based on random noise, then the odds of getting matching results would be really low!

One way to **check** this is, if we have some idea of what the "**true**" clustering looks like for just a few data points, we can compare those results to ours.

We call this "real" clustering the "ground truth".

### Definition 751

In machine learning, the **ground truth** is what we know about the "real world".

In general, we want our models to be able to **reproduce** this reality: it is the data that we tend to **trust** the most, if it is gathered correctly.

That way, we can use a very **small** amount of **supervision** to get an idea of whether our clustering is on the **right track**.

### 12.3.9 Applications: Visualization and Interpretability

We've discussed some ways to **abstractly** test whether our clustering might be **accurate** the data.

But, when it comes down to it, often, the **quality** of a clustering is based on how **useful** it is. So, what sorts of **uses** does clustering **have**?

Well, we're organizing our data into **groups**: this **simplifies** how we look at our data. And when we can **look** at our data, we can better **understand** what's going on.

In short: clustering allows humans to more easily make sense of data.

### Concept 752

One of the main **goals** of **clustering** is to make it easier for humans to **understand** the data.

This happens in two ways:

- We can **visualize** the data: we can **see** it, and more easily use our **intuition** to make sense of it.
- We can **interpret** our data: by seeing what sorts of **groupings** we create, we learn about the **structure** of the data.

So, machine learning experts judge partly based on how well a clustering **helps** them **achieve** these two goals.

Evaluating clusterings is **subjective** for exactly this reason: what is **good** "visually", or is the **best** "interpretation" of data, is often up to **debate**.

So, **human judgement** is important for this type of **problem**.

### 12.3.10 Applications: Downstream Tasks

Finally, there's one more way to think about clustering that is more **practical**, and closer to **objective**.

We use clustering to **sort** different data points that need different processing: this can make our model more **effective**, since different parts of the dataset may work better with different **treatment**.

**Example:** We could train a different regression model on each cluster: this can create a more accurate model.

We call this next problem a **downstream application**.

### Definition 753

A **downstream application** is a **problem** that relies on a **different** process to make its work better or easier.

In this case, **clustering** has **downstream applications** that can **take advantage** of the **structure** that clustering reveals.

- These "applications" rely on clustering for this improvement.

If our clustering is **good**, we would expect it to **improve** the performance of downstream tasks.

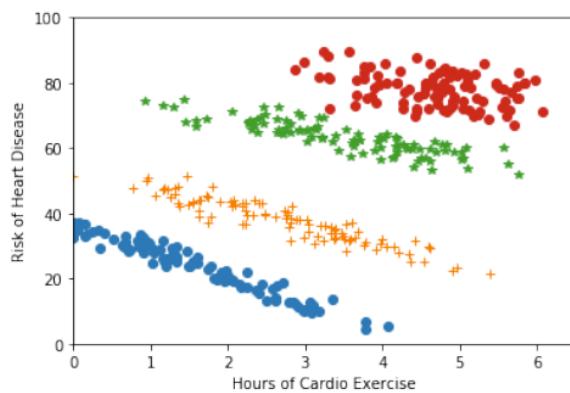
**Concept 754**

We can indirectly **evaluate** a **clustering algorithm** based on how **successful** the **downstream application** is.

If it **improves** the performance of a downstream application, we could say it works **well**.

### 12.3.11 A benefit of clustering

One advantage for downstream applications is, there might be patterns that are more obvious if you only look at a related segment of the data. For example:



If we take the data as a whole (**no clustering**), we would draw a **positive** regression: it seems that exercise and heart disease increase **together**. That doesn't make sense!

But, if we divide it into **clusters**, based on age, we see a **negative** relationship: each individual group experiences **benefits** from exercise.

This particular issue is called **Simpson's Paradox**.

**Definition 755**

**Simpson's Paradox** is when a **trend** that appears in groups of data either **vanishes** or **reverses** when we look at all the data **together**.

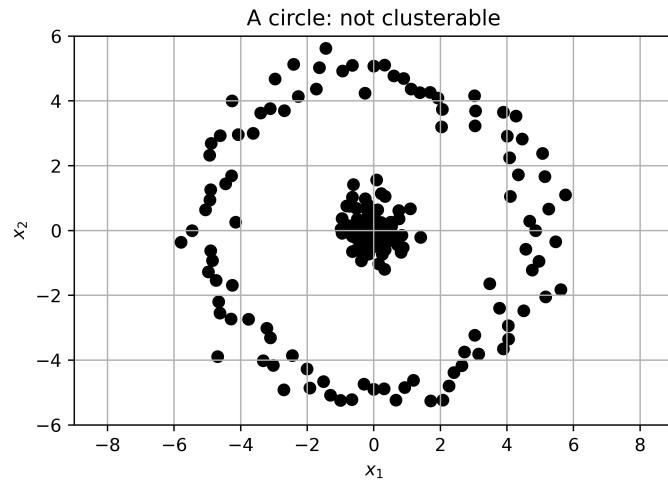
It shows that sometimes, **patterns** that we see may reflect how we're **looking** at the data.

Rest assured, you don't need to know this paradox by **name**. But it's important to **understand** possible problems like it: it'll help you make more responsible judgements in the future.

### 12.3.12 Weaknesses of k-means

There are some **weaknesses** to k-means clustering. Some patterns that a human eye can see, aren't easily **clustered** by our algorithms.

We can see this with a few **examples**:



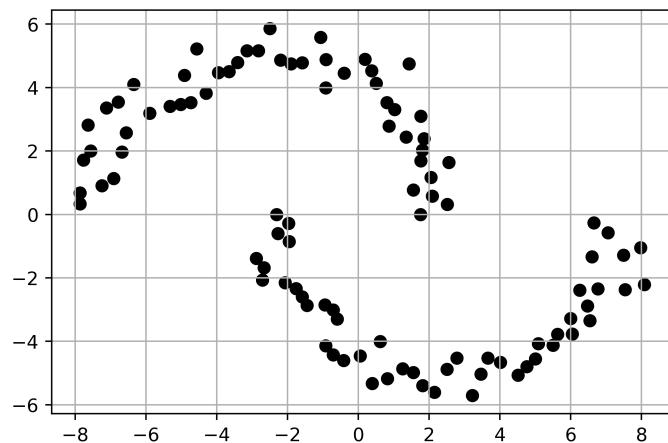
This data clearly seems to have a pattern. But does it have "clusters"?

This example can't be effectively **clustered**: and yet, most people would agree that the "outer ring" should be **one** cluster, while the "inner circle" should be **another**.

Assuming we (correctly) place one cluster mean in the **center**, there's **nowhere** we can put our other cluster mean to be **closest** to all of the **outer** points, but **not** the inner points.

We might be able to **resolve** this using a **feature** transformation. But, the problem remains.

Another example works for clusters that aren't very centralized:



For example, we could have a feature represent the radius! But then, we would still struggle with a ring not centered on the origin. Or, two rings with different centers.

This data can't be clustered either!

The edge of one cluster is too close to the other: we can't easily create a good pair of cluster means for each semi-circle.

These sorts of cases are often approached with either

- Attempts to directly visualize them
- Feature transformations
- Other algorithms not discussed here

## 12.4 Terms

- Clustering
- Unsupervised Learning
- Cluster
- Cluster mean
- k-means problem
- k-means loss
- Initialization
- Indicator Function
- Variance (Optional)
- k-means algorithm
- Hierarchical Clustering
- Consistency
- Ground Truth
- Visualization
- Interpretability
- Downstream Application
- Simpson's Paradox (Optional)

# CHAPTER 13

---

## Autoencoders

---

Through neural networks, we've **upgraded** the set of models we can use to do classification and regression tasks.

- Neural networks become more complex and **expressive** as we add more **layers**.

We'll spend the next few chapters exploring NNs: creating new variants (CNNs, RNNs), for example.

In this chapter, we'll investigate a different kind of application: **autoencoders**.

### 13.0.1 Unsupervised Learning

In chapter 6, we discussed an **unsupervised** learning problem: clustering.

- Classification tells us which classes to use. By contrast, clustering **doesn't** "know" the classes we're looking for.
  - Instead, we discover new classes ("clusters"), using the k-means algorithm.
- This lack of guidance is what makes clustering **unsupervised**.

Clustering was used as a form of **data analysis**: we were able to learn more about the **structure** of our data, by finding what sorts of "groups" existed.

We'll use autoencoders for a different task, that follows the same theme: learning more about the data, by attempting to look for an **unknown solution** to a simple problem.

### 13.0.2 Autoencoders: Compression

This time, our problem is not cluster-finding, but instead, **compression**. We want to take our input data, and find a more **space-efficient** way to represent it.

Typically, this means reducing the number of **dimensions**/variables.

- **Example:** turning a 10-dim data point into a 4-dim data point.

Why would we do this? Because of what we gain from this task:

- A good compression algorithm should be able to be **decompressed**, while keeping the result mostly similar.
  - That means that our compression must preserve **essential** information, so it's possible to retrieve later.
- By observing what's the algorithm decides to preserve and discard, can teach us what matters, and what doesn't.

#### Concept 756

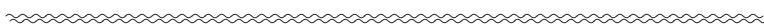
Learning a **compression algorithm** for your dataset creates a **simplified** representation, that still keeps the most important, distinct information.

Based on the information we find in that representation, we can figure out what components were "**important**".

Our compression/decompression system needs to do two things:

- **Reproduce** the original data well
- Do so in a way that lets us **distinguish** one data point from another

Based on this, a trained autoencoder can provide us some unexpected insights.



We can better see this with an example.

**Example:** Consider a database of human faces.

- It's more space-efficient if you can just re-use a "**template**" for a face, and just modify it.
  - So, your model might memorize what a face "generally" looks like.
  - That way, it doesn't have to waste space in the compression for data that appears frequently.

- Then, your compressed data only needs to store what's special, or different, about the face it represents.
  - So, you learn what separates different faces from each other, based on the info in the compressed model.

### 13.0.3 Training

This representation might even be easier for a new model to **train** with:

- We've omitted some unnecessary information that can **distract** our model.
- With a simpler input, it's faster to train, and compute answers.

The model can overfit to **noise** in these extra variables.

#### Concept 757

Just like how **clustering** is used to improve downstream tasks, **compression** can be, too.

Compressing your input can improve learning, so it takes less data to train.

#### Clarification 758

Not all compression is created equal!

Auto-encoding compresses in a way that contains **relevant information**.

However, in the Feature representation chapter, we discussed **binary code**: representing a number in **binary**, because it requires fewer features.

- This kind of compression doesn't emphasize what's "important" about the feature representation.

Binary compression is **worse**, not better! Instead of isolating useful information, it **obscures** it, forcing the model to learn binary.

## 13.1 Autoencoder Structure

### 13.1.1 Visualization

Our encoder is a **compression** algorithm, which takes an input  $x^{(i)}$ , and returns a new "transformed" input  $a^{(i)}$ , with a lower dimensionality.

$$\underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_d \end{bmatrix}}_{k < d} \longrightarrow \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix} \quad (13.1)$$

Note: you don't have to take the vectors in equation 8.1 literally.

You're not **required** to have  $\text{Dimension}(a) > 2$ , as in 8.1.

In other words, we're going from  $x \in \mathbb{R}^d$  to  $a \in \mathbb{R}^k$ .

**Example:** Take the classic problem of the MNIST dataset: identifying the identity of a digit based on a  $28 \times 28$  grid of pixels.



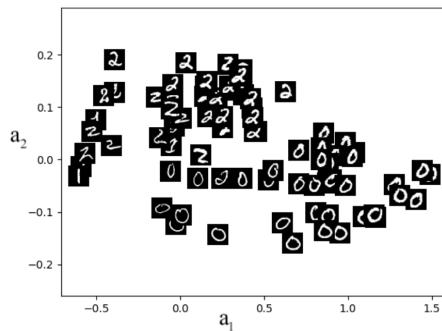
Here's an example of one data point: this represents the digit 9.

- That means our input data has 784 dimensions:  $x \in \mathbb{R}^{784}$ .
- Below, we've used an encoder to compress it down to 2 dimensions:  $a \in \mathbb{R}^2$ .

Each pixel is actually **restricted** to  $[0, 255]$ , but this statement is still technically true.

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{784} \end{bmatrix} \longrightarrow \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \quad (13.2)$$

- The x-axis and y-axis indicate the two dimensions of our output,  $a$ .



Notice that digits with the same identity are near each other: our compressed representation seems to be storing some useful data about digits!

Somehow, we've created a **simplified** representation that, despite having only 2 variables, has a lot of useful information!

- With only a couple labelled data points, we could guess the digits of most of these pictures, based on how close they are.

#### Clarification 759

This property, of "similar" data points, being **close** in the latent space, is **not** guaranteed, for any **compressed representation** you create.

However, it's a property we *hope* to find, in a useful one.

But whether or not our data "organizes" nicely like this, we've still demonstrated another benefit of compression:

- It allows us to transform our data into a lower-dimensional, more **digestible** format.
- So, we can create better visualizations.

#### Concept 760

One application of **compression** is using it for **visualization**.

A lower-dimensional version of a dataset is usually easier to diagram in a way humans can **interpret** directly.

- Because this representation tends to be more **dense** with information, it's often easier to draw useful conclusions.

It can also let us make simple predictions, or find patterns, since there are **fewer** variables to pay attention to.

This compressed version of data is called a "latent representation".

**Definition 761**

The **latent representation** of our data is the **compressed** version, containing as much useful information as possible, with fewer variables.

- We call it **latent** because original data  $x$  is in a "hidden" form, but we can retrieve it by **decompressing**.



The **latent space** represents all of the possible latent representations.

- This "space" follows the tradition of "input/feature spaces": sets with structure. In this case, the **distance** between our data gives us the **structure**.

Generally, a good latent space preserves **information**: the reconstructed input still **communicates** what we were interested in, from the original input.

### 13.1.2 Anatomy of an Autoencoder

Our autoencoder's purpose is **compression**, but we need to make sure that this compression preserves the information that we want.

**Definition 762**

An **autoencoder** comes in two parts:

- An **encoder**, which **compresses** our ( $d$ -dim) input into the ( $k$ -dim) latent representation.
- A **decoder**, which **de-compresses** our latent representation, to try to re-create the input.
  - This is used to make sure that our representation contains the information we need, to accurately re-construct the input.

The goals are:

- To create a **smaller** latent representation, with dimensions  $k < d$
  - To make sure that this latent representation can **accurately** re-construct our input
- 
- The encoder and decoder can be any kind of function, but we will use **neural networks**.

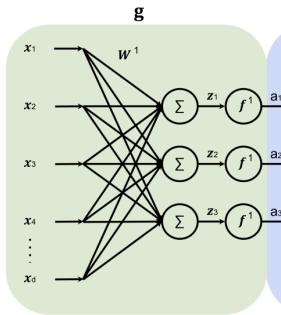
This is where neural networks shine: with more powerful model class, we can do more complex math to create our "encoding".

### 13.1.3 One layer encoder/one layer decoder

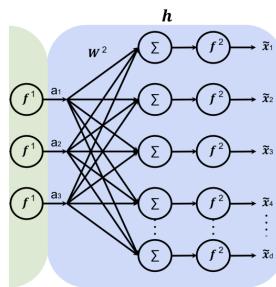
For a simple demonstration, we'll use a version with a one-layer encoder and a one-layer decoder.

We'll take our ( $\text{d-dim}$ ) input, and compress it into a ( $3\text{-dim}$ ) latent representation.

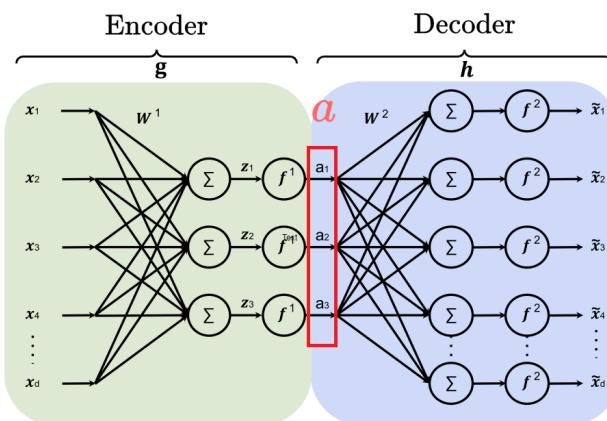
- Encoder: input  $x \in \mathbb{R}^d$ , output (compressed)  $a \in \mathbb{R}^3$ .



- Decoder: input (compressed)  $a \in \mathbb{R}^3$ , output (decompressed)  $\tilde{x} \in \mathbb{R}^d$ .



Taken together, we get our autoencoder:



Note that, in addition to  $W^1$  and  $W^2$ , we have a set of offsets that are not shown:  $W_0^1$  and  $W_0^2$ .

Let's run through our network:

- The **input**  $x$  goes through the encoder network. This compresses into the **latent representation**  $a$ .

- $W^1$  has shape  $(d \times k)$ , or equivalently,  $W^1 \in \mathbb{R}^{d \times k}$
- $W_0^1$  has shape  $(k \times 1)$ , or equivalently,  $W_0^1 \in \mathbb{R}^k$

$$\textcolor{brown}{z}^1 = (W^1)^T \textcolor{green}{x} + W_0^1 \quad \textcolor{red}{a} = f(\textcolor{brown}{z}^1) \quad (13.3)$$

- The **latent representation** goes through the decoder network. This de-compresses it into the **re-constructed input**.

- $W^2$  has shape  $(k \times d)$ , or equivalently,  $W^2 \in \mathbb{R}^{k \times d}$
- $W_0^2$  has shape  $(d \times 1)$ , or equivalently,  $W_0^2 \in \mathbb{R}^d$

$$\textcolor{brown}{z}^2 = (W^2)^T \textcolor{red}{a} + W_0^2 \quad \tilde{x} = f(\textcolor{brown}{z}^2) \quad (13.4)$$

The red layer in the center, is the "**latent representation**": the latent representation is **not** the output.

Our autoencoder can have more layers than we do here: this was just an example.

### 13.1.4 Autoencoders in general

Let's focus on that point about the "red layer" in the center:

#### Clarification 763

When we train an autoencoder, our goal is to create a **latent representation**.

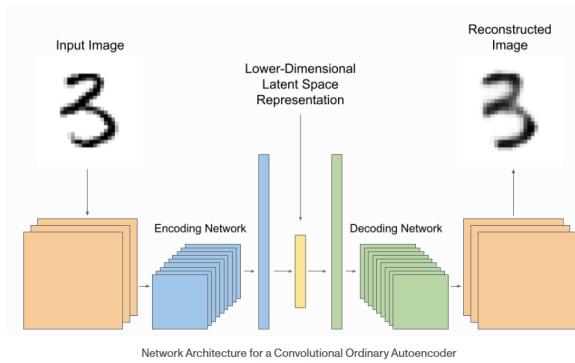
Deceptively, however, that representation is **not the output** of our autoencoder.

- Rather it's in the **middle** of the network: the output of the **encoder**, input to the decoder.

The actual autoencoder output is an **approximate** re-construction of our **input**.

Note that point: we're *approximately* re-constructing our input.

- The **quality** of the re-construction depends on our choice of encoder, and the size of our latent space.
- A bigger latent space (more dimensions) generally allows for a **better** re-construction, but takes up **more** space.



An example of the original vs reconstructed image from MNIST. Credit to [assemblyai.com](https://assemblyai.com)

The reconstructed 3 above is still very recognizable, but it's clearly been "degraded" somewhat: it's not the same.

#### Clarification 764

The reconstructed input is usually **not exactly the same** as the input.

$$x \neq \tilde{x}$$

Our re-construction is an **approximation**.

We're essentially running our input through a "**bottleneck**", in a lower dimension.

- We do so, hoping that the right compression size "squeezes out" only unnecessary information.

#### Concept 765

An autoencoder is, technically, just a normal neural network, where:

- We want the **input** to equal the **output** (re-constructed input)
- We monitor an **intermediate layer** (latent representation), where the layer output size ( $k$ ) **decreases** below the input size ( $d$ )
- If our input and output match well enough, then we use the **output** of that **intermediate layer**.

## 13.2 Autoencoder Learning

Now that we've defined our autoencoder, it's time to **train** it.

What's our objective? Well, we want to create a lower-dimensional representation that can be used to **re-construct** the original.

- We've handled the lower-dim aspect with the **structure** of the neural network: the last layer of the **encoder** will output  $k$  values, giving our **latent representation**. 
- So now, we need to show that this representation contains the information we want: it's able to **re-construct** the original.
  - This is the job of the **decoder**.

Where  $k < d$ , given  $d$  is the original input dimension.

That latter point is what we need to address: we need to check the quality of our re-construction,  $\tilde{x}$ .

### Concept 766

We measure the **quality** of our autoencoder by measuring the **similarity** between the original input  $x$ , and the re-constructed input  $\tilde{x}$ .

- If the re-construction is similar to the original input, that means our latent representation successfully **encodes** information about  $x$ .

So, we need some kind of **similarity** metric. We'll encode this into our loss function,  $\mathcal{L}$ .  $\mathcal{L}$  tells us how **different** our re-construction is, from the original.

- For **continuous** variables, you might use **squared distance** between  $x$  and  $\tilde{x}$ : loss  $\mathcal{L}_{SE}$ . 

$$\mathcal{L}_{SE} = \|x - \tilde{x}\|^2 = \sum_{i=1}^d (x_i - \tilde{x}_i)^2 \quad (13.5)$$

"SE" stands for "squared error". It's different from MSE, "mean squared error", because we're not dividing by  $d$ .

Note that often, an input does not only contain one data type: it may include different types of **discrete** data.

So, you may need to use different loss functions for different dimensions of the input.



Now that our problem is fully framed, we can simply **optimize** it, to minimize loss, using our parameters.

**Concept 767**

Once we've chosen our loss function, we can **optimize** our autoencoder as an ordinary neural network, in order to create our **encoder** for latent representations.

- $W_{en}$  and  $W_{de}$  are our encoder and decoder weights, respectively.

Our goal is to optimize these weights, with respect to the **loss**, over our  $n$  data points:

$$W_{en}^*, W_{de}^* = \operatorname{argmin}_{W_{en}, W_{de}} \left( \sum_{i=1}^n \mathcal{L}(\tilde{x}^{(i)}, x^{(i)}) \right)$$

Or, if we want to be more explicit,

$$W_{en}^*, W_{de}^* = \operatorname{argmin}_{W_{en}, W_{de}} \left( \sum_{i=1}^n \mathcal{L}(\text{NN}(x^{(i)}; W_{en}, W_{de}), x^{(i)}) \right)$$

~~~~~

- As usual for neural networks, we often optimize using gradient descent via **back-propagation**.

Example: For our one-layer encoder/decoder above, we would optimize over W^1, W_0^1, W^2, W_0^2 .

Remember that W^* notation is used to indicate "optimal" parameters.

13.3 Evaluating an Autoencoder

After **training** our autoencoder, we want to be able to **confirm** that it does what we want.

What do we **want**?

- A representation that contains **fewer** dimensions than the input: $k < d$.
 - Remember that k , in this case, is the dimension of our **latent** representation, and d is the dimension of our **input**.
- For this representation to contain useful **information** about our input.
 - This second aspect is (hopefully) addressed by our **loss** function.
 - If our **re-constructions** are really good, then we've managed to preserve our information.

Often, a latent representation can be **much** smaller than the input.

Remember our MNIST example at the start of the chapter, going from 784 dimensions, to 2.

13.3.1 Dimensionality of a

Our loss function handles the latter of these two problems, but the **dimensionality** is based on our **choice** of NN structure.

Well, if we're compressing our data, we want to reduce our number of dimensions, typically. But there's a tradeoff:

Concept 768

The **smaller** our latent representation is, the **less information** we can store in it.

- So, our re-constructions will be generally **worse**.

But a **larger** latent representation uses more **space**/computation, and doesn't **filter** out as much unnecessary, distracting information.

Clarification 769

Because we want to **compress** our data, we require $k < d$.

If we allow $k = d$, then our "latent representation" could be the **exact same** as our original representation.

- In fact, that would be the most **efficient** way to always be correct.
- If $k > d$, then we have extra dimensions, prone to overfitting.

If we **remove** the $k < d$ requirement, we're not forcing our network to do any real work, and our representation **doesn't** have to become more efficient.

In short: if we're not making the dimensions smaller, then we're not really compressing our data!

But what if we wanted to try making the dimensions **larger** on purpose? Does that have applications?

Concept 770

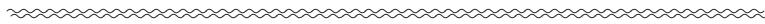
You might wonder if $k > d$ would let us "**unpack**" some of our input data into those extra dimensions.

An encoder with $k > d$ is called an **overcomplete autoencoder**.

Usually, this is **not** desirable:

- Our goal of "recreating the input" doesn't lend it itself well to "unpacking the data": it's usually easier to just **copy** the input.
- Moreover, **overfitting** makes these autoencoders hard to train.

That said, this sort of approach does have applications in de-noising, and learning sparse representations.



Just like in clustering, your exact choice of latent dimensionality k (usually $k < d$) is often subjective or task-based, and requires some trial and error.

We might have different considerations:

- How well does the plot seem to organize our data?
- Are we missing some crucial kind of information?
- Have we encoded information we don't care about on accident?

And more.

13.3.2 Data Analysis

One of the reasons we wanted to design autoencoders was to learn more about our data.

So, a useful encoder might be one that gives us some new or interesting insights.

Concept 771

We can learn more about our (already trained) encoding by **experimenting** with it.

How? By **modifying** the latent representation a , and seeing how that affects the reconstructed version \tilde{x} .

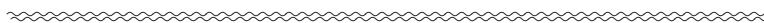
- We could start with a known data point, and modify one **dimension** of it, a_j .
- If we increase a_j and get a noticeable change, we can make guesses about what that dimension "**represents**".

Example: Suppose that we embedded the MNIST digits with k dimensions.

- We select one random data point, $x^{(i)}$. This data point happens to be a picture of the number 6.
- We scale up/down one dimension, a_j .
- We might see that the line thickness of the 6 increases. Maybe a_j represents line thickness.

There could be many possible features: how "angular" the number is, how it loops, etc.

Not to say that those are the particular features you *will* find. In fact, sometimes, it's totally unclear what your latent representation is "representing".



A few other examples of how to do our data analysis:

Concept 772

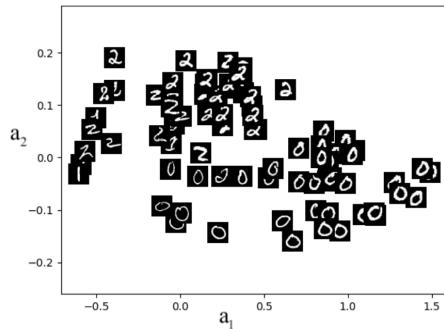
Rather than "experimenting" with individual axes, we could **plot** data points in the **latent space**.

There are two ways to do this:

- Take **real** data points, and plot where they appear in the latent space, compared to how the **input** looks.
- Directly **sample** points from the latent space, and see what their **re-construction** looks like.

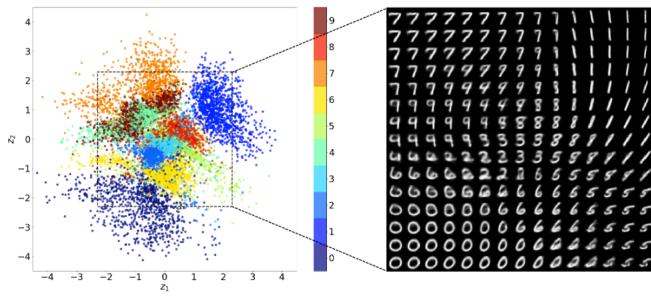
In both cases, we get an idea of how the autoencoder interprets the data it receives.

Example: We started the chapter with an example of the former technique:



Here, we take real data points, and plot them in the latent space.

Example: We could also use the latter: seeing how our re-construction appears, as we modify our latent variables.



The left shows the position of real digits in latent space, while the right shows our reconstructions, moving in a grid across the space. Credit to [argmax.ai](#).

Concept 773

Lastly, we could see what the model thinks the data **generally** looks like, based on the biases/offsets in the network.

- The offset values (W_0^ℓ) are the **same**, regardless of the data point.
- We could set all of the values of α to 0: in a **linear** autoencoder, this would give the **average** of our data points.

If our data clusters around the average, it would be reasonable to expect the average to be **somewhat similar** to all of our data.

- And if it looks similar to each data point, it could somewhat **represent** what it looks like in general.

In a **non-linear** autoencoder, our above approach doesn't give us the **average**, but could be useful regardless.

After analyzing all our dimensions, we can try to figure out what **information** the encoding decided was "important", or at least, necessary for re-construction.

This kind of analysis has a wide array of applications, including natural language processing, disease subtyping, and image processing.

13.3.3 Downstream Tasks

If we're using the encoder for downstream tasks, we can evaluate the autoencoder based on the performance of those downstream tasks.

- We mentioned the same kind of metric when we discussed clustering.

Concept 774

Performance on **downstream tasks** is one way to evaluate the quality of an encoding.

One simple approach, is to use **semi-supervised learning**.

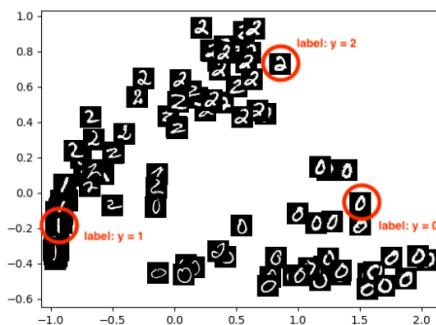
Definition 775

Semi-supervised learning provides labels for only **some** of the data we use to train. The rest of it is unlabelled.

So, the model has to **extrapolate** from that data, to figure out a pattern for the rest of the data.

If we have a **meaningful** latent representation, we could use it to help with this type of problem.

Example: We'll re-use our MNIST example. Suppose we were only given a **few** labelled digits:



Only the three labelled data points are "supervised".

- In this case, we could label many of these digits, just based on what we have.

3 data points is a really small amount of information! This makes it even more impressive.

The fact that this approach works so well, with only 2 dimensions, tells us that our encoding is impressively effective.

13.4 Linear Encoders and Decoders

Even simpler than our one-layer example above, is the **linear** autoencoder.

It turns out that, despite its simplicity, even a linear autoencoder can be useful! In some contexts, it's even **better**:

- A linear autoencoder often has a **closed-form** solution: we don't have to do gradient descent.
- The linear autoencoder tends to create a very simple kind of **interpretability**.

This closed form solution is equivalent to **principle component analysis** (PCA), which you might be familiar with from linear algebra.

- We obtain this solution using a technique called **singular value decomposition** (SVD).
-

We can draw some parallels to a paradigm we've seen before:

- **PCA** can be thought of as the simplified, linear version of the **autoencoder** problem.
- This is similar to how **linear regression** is the simple, linear version of a **neural network**.

13.4.1 Principle Component Analysis (Optional)

Remark (Optional)

The following section briefly covers the concepts of PCA.

These may provide some intuition for what autoencoders do, in a simple, linear environment.

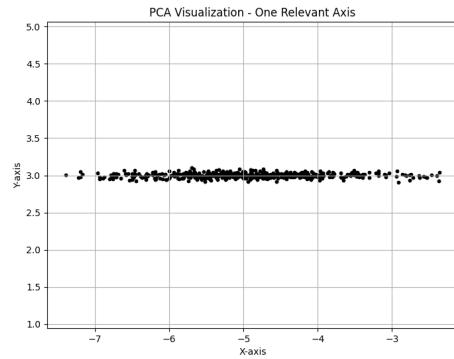
We mentioned that autoencoders need to make their data points clearly **distinct** from one another.

- In other words, it's best to focus on ways that they're **different** from each other.

The easiest way to do that is to focus on sources of **variance** in the data.

13.4.2 Low-variance: less important (Optional)

Let's give a motivating example:



Almost all data is encoded on the x-axis.

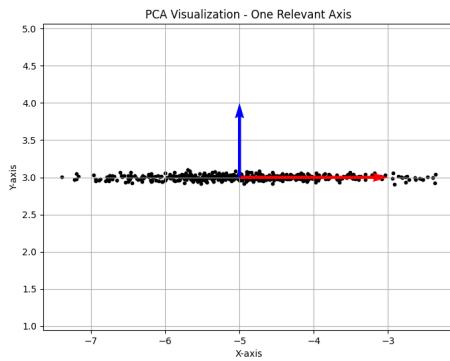
If we look at this dataset, there's a clear **difference** between our two axes: one is very high-variance (x), one is low (y).

If I told you "the y coordinate of this data point is roughly 3", that tells you essentially **nothing**: that's true of all of our data.

Meanwhile, the x coordinate is much more **informative**.

- It seems that **high-variance** dimensions tend to contain **more information** than low-variance dimensions.

So, if we break our data up based on coordinates:



We could remove the y-axis while preserving most of our information! This would be a good target for **omitting** from the latent space.

$$\begin{bmatrix} x \\ y \end{bmatrix} \longrightarrow \begin{bmatrix} x \end{bmatrix} \quad (13.6)$$

Concept 776

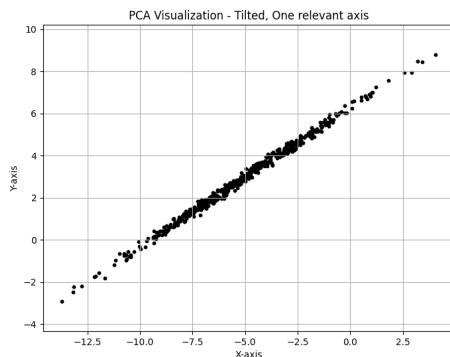
When doing PCA, we tend to focus on axes of **high variance**.

Axes with low variance carry **less information**.

So, if we want to **compress** our data, we can remove those low-variance dimensions.

13.4.3 Different axes (**Optional**)

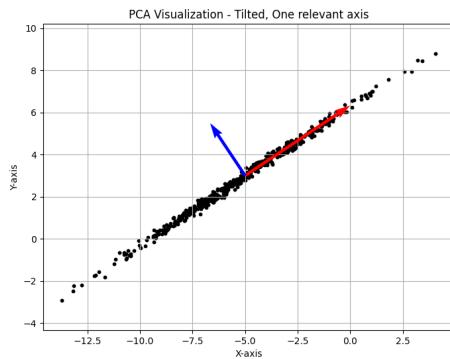
Our data doesn't always (or even usually) line up perfectly with our axes, though.



Almost all data is on a single line.

In this case, it looks very **similar** to our x-axis data. But the problem is, we can't just reduce it to one of our two axes.

The solution? **Change coordinate systems.**



Now, we have a high variance axis, and a low variance axis.

Now, we have the same situation as before! Almost all of our data is on **one axis**: we can omit the other.

Concept 777

Often, the most "**information-heavy**" directions in our data, aren't our default axes.

So, we look for a **new coordinate system**, where most of our variance is contained ("can be explained") by only a few dimensions.

- That way, we can **remove** the other remaining dimensions, which contribute very little to our understanding of the data.
- With fewer dimensions, our data is often more **interpretable**.

This idea, of finding high-variance components, and discarding low-variance components, is the core principle of **principle component analysis**.

- This is equivalent to how our linear autoencoders remove extra dimensions.

Sometimes, it can be a bit difficult to **interpret** these high-variance components: what does it mean to have high variance along a "different axis", conceptually?

But other times, we find that two "separate" variables, are both giving us the **same information**: they would correlate strongly, and thus, would give us the line we see above!

- This is why we can **remove** one dimension: we're using two axes to represent basically the same thing.

Example: If something is sold for a fixed price, then "number of units sold" and "profits from sales" are telling us **the same thing**.

So, we can compress into one dimension with low information loss.

Last Updated: 09/03/24 03:53:41

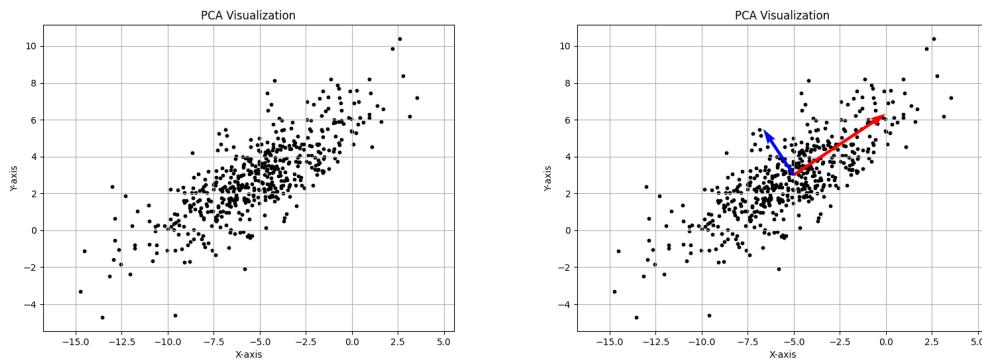
Many examples aren't perfectly correlated like this, but are still pretty similar: for example, sales and advertising spending at a company.

13.4.4 General example (Optional)

How many dimensions we need to capture most of our variance **depends** on the situation.

The goal is often, for visualization, to reduce it to **2 or 3** primary components.

But even if we can't remove axes, PCA can give us a useful way to focus in on the **relationship** between our variables:



Here, we have identified which axis matters "more" to our data, and how much.

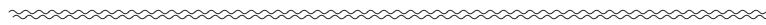
13.4.5 Non-linear encoders (Optional)

Of course, things aren't always so simple: we often won't have this nice blob, which we can measure across a few perpendicular axes.

Our real data might take on a different form, that has a coherent "shape", or "structure", but not one that fits our **linear** model.

This is where our **non-linear** autoencoders come in. They follow the same kind of principles as PCA:

- We want to distinguish patterns and curves that contain **variance**, and allow us to **distinguish** data points.
- Those which contain more information, we **keep**. The rest, we discard.
- We can then **visualize** those remaining dimensions, or use them for data analysis.



One more detail of PCA we've ignored so far: all of variance comes from some "baseline": an **average** position, which all of our data points are **shifted** from.

- We could view this as a "template", which our PCA components point away from, for any given data point.

In PCA, we didn't focus on this as much, because it was just the average of all of our data points: useful, but simple.

This concept becomes interesting in the more complex case of non-linear autoencoders: it isn't so easy to **guess** what's the "average"/typical example in the latent space.

- **Example:** Suppose we plot all of the data points labeled "7", based on their latent representations.
- We could see what happens if we average those, and then convert it back into a picture: it would teach us about how our model sees the number 7 in general.

The result depends on your exact choice of encoder, but it's not always what you expect: that's why it's so **informative!**

13.5 Advanced Encoders and Decoders (Optional)

We can build on this framework to create models for different, practical tasks.

13.5.1 Generative Networks (Optional)

One useful application of our latent space is to **artificially** create more data which is **similar** to the data we already have.

How?

We saw that, with **some** very effective autoencoders, we find some useful **structure** in our latent space:

- Data points which were **nearby** in latent space, appeared **similar** in the input space.

If we successfully train an autoencoder with this property, then creating new data is possible:

- We take our real data, and slightly modify it in the latent space. This should be similar to our real data, but not exactly the same.

How do we train our autoencoder to do this? We'll discuss one popular approach below: the "variational autoencoder" (VAE).

This is the basis of a **generative network**:

Definition 778

Generative networks are networks which are used to **generate** artificial data, which is similar to **real** data used to train it.

These often come in the form of an autoencoder:

- We start by using real data to **train** the autoencoder, and create a **latent space**.
- We use those same data, and **create** new data that is "close" to the real data, within our latent space.
- Finally, we use the decoder to **reconstruct** our **new data**.

~~~~~  
This technique relies on the assumption that data points which are **close** in the latent space, are **similar** in the input space.

These models have many applications:

- Augmenting (increasing the size of) **smaller** datasets
- Reducing **overfitting**, by perturbing data
- Art and Media

- **Example:** Superresolution: filling in "details" in images where they need to be believable, not necessarily real

And more, which we'll discuss below.

This, of course, can cause problems if you don't know you're working with fictional data!

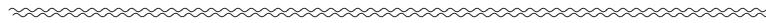
### 13.5.1.1 Variational Autoencoders (Optional)

We've introduced the general outline of a **generative network**: we create artificial data, using our latent space.

But we have **two problems**:

- First: how do we **generate new** data points, "close" to the real data? What's our procedure?
- Second: this relies on the assumption that our **encoding** is **smooth**: nearby points in latent space represent similar information. How do we guarantee that?

We'll find that one technique addresses both of these problems: representing our data as a **probability distribution**.



One simple way to generate new data, is to **randomly** perturb a data point in the latent space, by a small amount.

The outcomes of random processes, like this, have a certain **probability** of occurring.

- So, the output of our encoder isn't a point in the latent space, but a **probability distribution** of possible points.

#### Definition 779

A **variational autoencoder**, rather than creating a **single** encoding for a data point, encodes a **probability distribution**.

- This distribution represents different possible encodings, for the **same** input.

Then, we **sample** from this distribution, to randomly select an encoding.

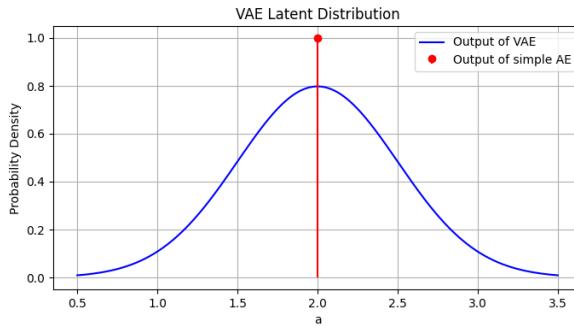
Finally, we **decode** this to get our "modified" data point.



This process actually **regularizes** our model: it needs to be able to re-produce the input, even if it's slightly modified in the latent space.

- We're also restricting our probability distribution to have a certain shape/structure (like a normal distribution).

**Example:** Suppose that we created a 1-D encoding, and for this particular data point, we encode it as  $a = 2$ . If we compare the simple AE to the VAE:

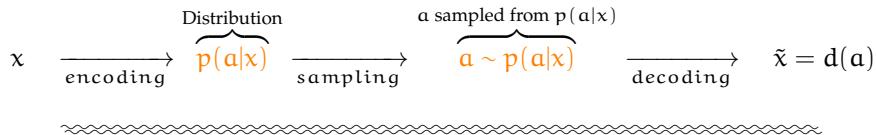


We've gone from "always output 2" to "output a normal distribution centered on 2".

We can directly compare the process in both cases. First, the **simple autoencoder**:

$$x \xrightarrow{\text{encoding}} a = e(x) \xrightarrow{\text{decoding}} \tilde{x} = d(a) \quad (13.7)$$

Next, the **variational autoencoder**:



Now, we can handle the second problem: how do we ensure that nearby points in the latent space are **similar**?

Well, if we want our model to do something, we **train** it with that goal in mind.

- As we've designed it, our VAE **already** generates points "nearby" to our encoding, using its probability distribution.
- Because they're nearby, we want the **original** data and the **sampled** data to be relatively similar.

Technical comment, that isn't important to this class:

The simple AE distribution depicted is the **dirac delta function**, where all the probability is 'stored' at  $a = 2$ .

### Concept 780

Unlike a simple AE, our VAE **doesn't** return exactly the same data that it started with:

- Instead, our VAE **samples** the latent space, **nearby** to the "pure" encoding.

We usually use something like the normal distribution above: most data is close to the mean.

Because the sampled and original data are nearby, we want them to be **similar**.

How do we compare the original and sampled data? The original data is given as the input. Meanwhile, we convert the sampled data by **decoding** (reconstructing) it.

We want our original (input) data  $x$  and sampled (+decoded) data  $\tilde{x}$  to be **similar**.

- This is similar to our goal for the simple autoencoder: there, too, we wanted our original and re-constructed data to match.

### Concept 781

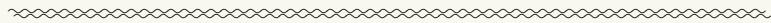
Just like with a simple autoencoder, we want the **input** and **output** of our VAE to be close to **equal**.

$$x \approx \tilde{x}$$

This serves a few functions:

- Same goal as we had before: being able to **reconstruct** the input shows that our encoding preserves meaningful **information**.
- The simple encoding of  $x$ , and the encoding that produces  $\tilde{x}$ , are **not** the same, but they are **similar**.
  - This helps us train our model to "smooth out" its surface: we teach it that nearby points should be similar.

So, we **train** our model with this goal in mind.



To help "smooth out" the VAE representation further, we often add a **regularizer** term to the loss.

- We won't go into detail on this feature here.

### 13.5.2 Adversarial Optimization (Optional)

We'll make a brief detour, to discuss a different way we can generate **artificial** data.

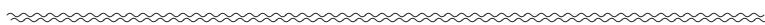
In order to "test" your network, and see how **robust** it is, you might want to deliberately find examples it **struggles** with, but *shouldn't*.

- We could do this by just running a large amount of data, and manually looking for examples that fail, despite seeming "obvious" to humans.
- But this is labor-intensive, and **expensive**.

Instead, we introduce a different way to create so-called "**adversaries**": examples specifically designed to "trick" our model.

We want a data point that:

- **Should** be classified correctly, and would be classified correctly by humans
- But the machine fails, despite looking **similar** to valid data points.



We'll start with a valid, **correctly** labelled data point: this handles the first condition: "should have been labelled correctly".

But we want to **modify** it so that it's labelled incorrectly.

- If our model isn't **robust** enough, we might be able to **confuse** it, without changing the data very much at all.

How do we modify it most efficiently? Using the **gradient**.

Previously, we used the gradient to compute, "what's the most efficient way to change my **model**, to **decrease** my loss?"

$$W - \eta \frac{\partial L}{\partial W} \quad (13.8)$$

This time, we want to ask, "what's the most efficient way to change my **data point**, to **increase** my loss?"

So, we want the gradient between the loss and the data point.

$$x + \eta \frac{\partial L}{\partial x} \quad (13.9)$$

We take **steps** in this process, repeatedly changing our gradient to match the model.

We continue this process until our data point is successfully classified **incorrectly**.

- Often, we can accomplish this without significantly modifying our data point.

This time, instead of traveling across the parameter space, we're traveling across the input space!

**Definition 782**

**Adversarial training** is a way to "trick" a model into making inaccurate predictions:

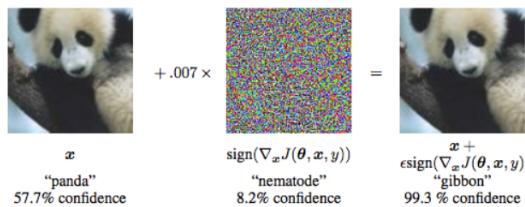
This is done by **training** data points that appear very similar to real data, but exploit weaknesses of the model.

To accomplish this, we take a real, correctly labelled data point, and slowly apply gradient descent **to the data point**, gradually increasing the loss:

$$x_{\text{new}} = x_{\text{old}} + \eta \frac{\partial L}{\partial x}$$

With some training, we can design a new data point that our model evaluates incorrectly, despite being very similar to the original data point.

**Example:** Here's a *real example* of applying this to image classification:



This panda, despite no visible change in appearance, is now a gibbon. (Source: Ian Goodfellow et al., 2014)

This kind of data is actually incredibly valuable:

- If we use it to train our model further (i.e., teaching it that it's wrong about these data points), it tends to become **more robust** against this kind of attack.

But, we need to be careful in this process:

**Clarification 783**

When creating adversarial data, we have to set a **termination** condition:

If we continue gradient descent too long, our data point will change **classification**, but it can also actually change enough that it **should** be identified differently.

**Example:** If you do gradient descent to turn a picture of a "9" into a "1", and it looks like a "1" to a human, it should be reasonable that the model makes the same decision.

We want to find the point where the data is different enough to be deceptive, but not different enough that it's obviously, genuinely a new data point.

Adversarial data is widely useful:

- Improving model robustness
- Defending against security threats
- Learning more about the nature of your model

### 13.5.3 Generative Adversarial Networks (**Optional**)

We've developed two useful ideas:

- Generative networks: used to generate artificial, plausible data
- Adversarial data: data designed to exploit weaknesses in a model

We can combine these ideas to create a powerful tool, called, appropriately, a **Generative Adversarial Network**.

Here's the general idea:

- **Generators**: we want to generate artificial data, that closely reflects the **original** distribution.
- We found that an **adversary** could teach us (and in turn, our models), their weaknesses, by actively seeking them out.
  - So, we'll create an **adversary** for the generator: the **discriminator**. This punishes our model for creating data that is "detectably different" from real data.

The generator will create "adversarial data": this data is designed to **trick** the discriminator into thinking it's real.

The discriminator will try to learn how to tell the two apart, and provide feedback to the generator, telling us how it made a mistake the previous time.

This feedback loop gradually improves how "realistic" our data generated is, through a form of **unsupervised** learning.

The two models are "supervising" each other!

**Definition 784**

A **Generative Adversarial Network** is a model that comes into two opposing, or "adversarial", parts:

- The **generator** is trained to create "fake" data, that looks as plausible and realistic as possible.
- The **discriminator** is trained to detect fake data

When one of these models fails, they teach the other model what they did wrong, to improve.

- This process repeats until the discriminator accuracy reaches 50%: if it's equally likely to mix up real/fake data, our generator has become indistinguishable from real data.

This is a sort of "arms race" between the two halves of our model.

**Clarification 785**

Why is best performance for the generator at **50%**?

If it was 100%, then the model always assumes the generator is real, and the real data is fake.

- But that's not true: the real data is *also* real. The discriminator isn't doing its job correctly anymore.

Another perspective: if you knew your discriminator was always **wrong**, then you could easily create a discriminator that was always right: just do the opposite of what you got before.

In short: 50% is the best you can do, because your discriminator is completely **unsure** of its answer: which is what it really means to be "**indistinguishable**".

GANs have been highly successful: this procedure not only improves robustness, but often creates a generally improved model.

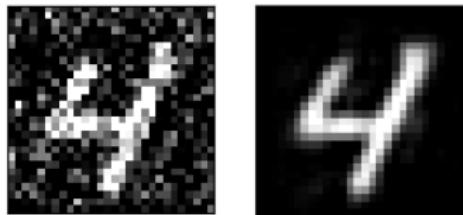
They're useful for creating very effective generators of new data, and have been particularly useful in the past for image generation.

GANs are still relevant, but if you want a more **modern** approach, you could look into **Diffusion** models, like Stable Diffusion.

### 13.5.4 De-noising (Optional)

One useful application of autoencoders is **de-noising**.

- We want to turn a noisy input into a less-noisy input.



The left is our input, right our desired, cleaner output.

The process for creating this kind of autoencoder is straightforward enough: we give noisy data and encourage the model to match the original, noiseless version.

#### Concept 786

Autoencoders can be used for **de-noising**:

By training with

- Noisy data as input
- Noiseless data as desired output

You can teach the model to create a latent representation that's resistant to this kind of noise.

### 13.5.5 Attention (Optional)

Transformer networks are a very modern, very powerful approach to many problems, most famously **language processing**.

In order to create detailed context within a sentence, these models use a technique called "**self-attention**", which has proven to be incredibly powerful for working with language.

As of writing, GPT-4 is the most famous example.

- Attention is an in-depth topic that deserves its own section; we'll **skip** the math.
- The main idea: attention helps determine how words create **context** for each other.

#### Concept 787

**Attention**, at a very high level, is based on the idea that:

- Different **words** in a sentence have different levels of **importance** to each other.

The words that are more **important** to a particular word X, are a bigger part of the **context** you use to understand that word.

With that in mind, you can apply the **context** from each other word Y, to better understand the meaning of word X.

Transformers aren't **limited** to language, but we'll focus on that use case, for ease of explanation.

**Example:** "The **blue dog** bites the **red ball**": the word '**red**' is describing the '**ball**', so it is less important for understanding '**dog**' in this sentence.

Based on that...

#### Definition 788

**Attention** is a mechanism for determining how, for a **pair of words**, one word X might be **important** to understanding the other word Y, contextually.

In other words, X might require your **attention**, if you want to understand Y.

- Attention is applied to **every** pair of words in a sentence: we measure every word's impact on every other word.
- Finally, for each word, we **integrate** information from the other words, based on how "important" they were.

~~~~~  
Attention has the benefit of allowing us to analyze our entire prompt simultaneously, speeding up the process.

Self-attention is used by a single sequence of text, by itself: it "learns about" itself.

Clarification 789

A few lingering comments:

- Word X and Y are typically asymmetric: X may not affect the meaning of Y the same way as Y affects X.
- This meaning is **contextual**, not the "importance" of each word in isolation.
- We "integrate information" by taking a linear combination of each word: a larger weight is given to words which are more "important" to word Y.

"Multi-headed attention" simply refers to having several of these attention mechanisms, applied to the same input in parallel.

So, each "attention head" receives the same data, like neurons in the same layer of a network.

13.5.6 Transformer Networks (Optional)

Now that we loosely understand attention, we can think about the bigger picture.

The goal of a transformer is to **predict text**.

At a very high level, transformers have a structure *similar* to **autoencoders**. They're broken into the same sort of two parts, though their internal structure is a bit **different**:

- **Encoder:** this is actually a **stack** of several encoders, one after another. This (presumably) creates an **internal representation** of the "meaning" of the input text, similar to a latent representation.

- Each encoder contains a fully connected network, as we've shown above, but also a "self-attention mechanism".

You could say that this converts the input "prompt" into something similar to the "latent representation", which retains our key information.

- **Decoder:** a stack of decoders, one after another. Based on the **internal representation**, it creates predicted text.

- Each decoder also contains an FC network+attention mechanisms.

This would create the "response" to your "prompt".

- Once our decoder predicts some text, that's part of the "**past text**": this gets included alongside the internal representation, for future predictions.

Concept 790

A transformer network follows a structure with some similarities to auto-encoders: using an **internal representation**.

- Converting its **input** (prior text) into an **internal representation**, similar to a latent representation.
- Converting that **representation** into an **output** (predicted text)

However, there are key differences:

- As the output creates new text, that is **included** with the internal representation.
- Our "predicted text" will **not** be the **same** as our past text.

Why multiple encoders?

Concept 791

The **multiple** encoders are used to gradually find "more **complex**" patterns (or concepts), as we move through more layers.

-
- The idea is that, the first encoder combines words to finds **basic**, simple language structures.
 - The second encoder has access to these simple structures, and can **combine** them together to create a more complex structure.
 - Each encoder is combining the results of the **previous** layer, to create something more complex.

This is similar to the structure and motivation behind **convolutional neural networks**, which we will cover later.

Definition 792

A **transformer network** is a model made of encoders and decoders that use **attention** to determine the **structure** and **context** of the input "prompt", and create an output "response".

The structure of a transformer network is:

- Several "layers" of **encoders**, that take the input and interpret it, creating the internal representation
- An **internal representation** of the input, and the previous output of the decoder
- Several "layers" of **decoders** that turn this representation+output, into more **predicted text**

Predicted text is given as a **probability distribution**, using softmax.

- We select an element from this probability distribution before moving on to the next word.

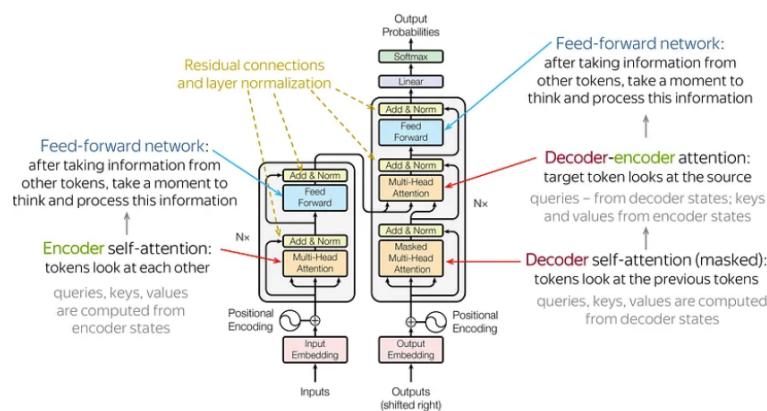
We've skipped over many important details of transformers, but this gives us the gist.

In particular, we've skipped some components, like normalization layers.

Clarification 793

A few key differences between an autoencoder and transformer networks:

- The input and output are the same in an autoencoder.
- Transformers use attention mechanisms.
- The internal representation in an autoencoder ("latent representation") is **smaller** than the input or output.
- They serve different purposes.



If you want a more in-depth explanation, go to [this very helpful resource](#), and the source of this lovely diagram!

13.6 Terms

- Unsupervised Learning (Review)
- Clustering (Review)
- Compression
- Decompression/Re-construction
- Encoder
- Decoder
- Autoencoder
- Latent Representation
- Latent Space
- Bottleneck
- Dimensionality (Review)
- Overcomplete Autoencoder
- Downstream Task (Review)
- Semi-supervised Learning
- Principle Component Analysis
- Singular Value Decomposition
- Generative Networks
- Variational Autoencoders
- Transformer Networks

Optional

- Adversarial Data
- Generative Adversarial Networks
- De-noising
- Attention
- Transformer Networks
- Internal Representation

APPENDIX A

Matrix Derivatives

A.1 Introduction and Review

Our goal here, is to combine the powers of matrices and calculus:

- **Matrices:** the ability to store lots of **data**, and do fast linear operations on all that data at the **same time**.

Example: Consider

$$\mathbf{w}^T \mathbf{x} = \begin{bmatrix} w_1 & w_2 & \dots & w_m \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} = \sum_{i=1}^m x_i w_i \quad (\text{A.1})$$

In this case, we're able to do m different **multiplications** at the same time! This is what we like about matrices.

In this case, we're thinking about vectors as $(m \times 1)$ matrices.

- **Calculus:** analyzing the way different variables are **related**: how does changing x affect y ?

Example: Suppose we have

$$\frac{\partial f}{\partial x_1} = 10 \quad \frac{\partial f}{\partial x_2} = -5 \quad (\text{A.2})$$

Now we know that, if we increase x_1 , we increase f . This **understanding** of variables

is what we like about derivatives.

Concept 794

Matrix derivatives allow us to find **relationships** between large volumes of **data**.

- These "relationships" are **derivatives**: consider dy/dx . How does y change if we modify x ? Currently, we only have **scalar derivatives**.
- These "data" are stored as **matrices**: blocks of data, that we can do linear operations (matrix multiplication) on.

Our goal is to work with many scalar derivatives at the **same time**.

In order to do that, we can apply some **derivative** rules, but we have to do it in a way that **agrees** with **matrix** math.

Our work is a careful balancing act between getting the **derivatives** we want, without violating the **rules** of matrices (and losing what makes them useful!)

Example: When we multiply two matrices, their inner shape has to match: in the below case, they need to share a dimension b.

$$\underbrace{X}_{(a \times b)} \quad \underbrace{Y}_{(b \times c)} \quad (A.3)$$

We can't do anything that would **violate** matrix rules like these: otherwise, we're not really doing "matrix **calculus**" anymore. This means we need to build our math carefully.

First, we'll look at the **properties** of derivatives. Then, we figure out how to usefully apply them to **vectors**, and finally, to **matrices**.

A.1.1 Partial Derivatives

One more comment, though - we may have many different variables floating around. This means we **have** to use the multivariable **partial derivative**.

Definition 795

The **partial derivative**

$$\frac{\partial B}{\partial A}$$

Is useful when there may be **multiple variables** in our functions.

The rule of the partial derivative is that we keep every variable **other** than A and B **fixed**.

- This is different from the "total" derivative $\frac{dB}{dA}$, where we consider how B affects A, without keeping other variables fixed.
- The total derivative is also used in multi-variable calculus, but serves a different role: we'll come back to this later.

Example: Consider $f(x, y) = 2x^2y$.

$$\frac{\partial f}{\partial x} = 2(2x)y \quad (\text{A.4})$$

Here, we kept y *fixed* - we treat it as if it were an unchanging **constant**.

Using the partial derivative lets us keep our work tidy:

- If **many** variables were allowed to **change** at the same time, it could get very confusing.
- Since this is too complicated, we'll instead change these variables *one at a time*. We get a partial derivative for each of them, holding the others **constant**.

Imagine keeping track of k different variables x_i with k different changes Δx_i at the same time! That's a headache.

Our **total** derivative is the result of all of those different variables, **added** together. This is how we get the **multi-variable chain rule**.

Definition 796

The **multi-variable chain rule** in 3-D ($\{x, y, z\}$) is given as

$$\frac{df}{ds} = \underbrace{\frac{\partial f}{\partial x} \frac{\partial x}{\partial s}}_{\text{only modify } x} + \underbrace{\frac{\partial f}{\partial y} \frac{\partial y}{\partial s}}_{\text{only modify } y} + \underbrace{\frac{\partial f}{\partial z} \frac{\partial z}{\partial s}}_{\text{only modify } z}$$

If we have k variables $\{x_1, x_2, \dots, x_k\}$ we can generalize this as:

$$\frac{df}{ds} = \sum_{i=1}^k \underbrace{\frac{\partial f}{\partial x_i} \frac{\partial x_i}{\partial s}}_{x_i \text{ component}}$$

- We justify a lot of aspects of the multivariable chain rule, and the gradient, in our Gradient Descent chapter. Feel free to review there.

A.1.2 Thinking about derivatives

The typical definition of derivatives

$$\lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} \quad (\text{A.5})$$

Gives an *idea* of what sort of things we're looking for. It reminds us of one piece of information we need:

- Our derivative **depends** on the **current position** x we are taking the derivative at.

We need this because derivative are **local**: the relationship between our variables might change if we move to a different **position**.

But, the problem with vectors is that each component can act **separately**: if we have a vector, we can change in many different "directions".

$$A = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \quad B = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \quad (\text{A.6})$$

Example: Suppose we want a derivative $\partial B / \partial A$: $\Delta a_1, \Delta a_2$, and Δa_3 could each, separately, have an effect on Δb_1 and/or Δb_2 . That requires 6 different derivatives, $\partial b_i / \partial a_j$.

3 dimensions of A times
2 dimensions of B: 6
combinations.

- Every component of the input A can potentially modify **every** component of the output B .

One solution we could try is to just collect all of these derivatives into a **vector** or **matrix**.

Concept 797

For the **derivative** between two objects (scalars, vectors, matrices) A and B

$$\frac{\partial B}{\partial A}$$

We need to get the **derivatives**

$$\frac{\partial b_j}{\partial a_i}$$

between every **pair** of elements a_i, b_j : each pair of elements could have a **relationship**.

The total number of elements (or "size") is...

$$\text{Size}\left(\frac{\partial B}{\partial A}\right) = \text{Size}(B) * \text{Size}(A)$$

Collecting these values into a **matrix** will give us all the information we need.

But, how do we gather them? What should the **shape** look like? Should we **transpose** our matrix or not?



A.1.3 Derivatives: Approximation

To answer this, we need to ask ourselves *why* we care about these derivatives: we'll define them based on what we **need** them for.

- We care about the **direction of greatest decrease**: the negative of the **gradient**.
 - One application: we might want to adjust weight vector w to reduce loss \mathcal{L} .
- To find the gradient, we'll need some other **matrix** derivatives that will be useful in a **chain rule**.
 - The chain rule creates lots of intermediate derivatives, that we need to work with.

Let's focus on the first point: we want to **minimize** \mathcal{L} . Our focus is the **change** in \mathcal{L} , $\Delta\mathcal{L}$.

We want to take steps that reduce our loss \mathcal{L} .

$$\frac{\partial \mathcal{L}}{\partial w} \approx \frac{\text{Change in } \mathcal{L}}{\text{Change in } w} = \frac{\Delta \mathcal{L}}{\Delta w} \quad (\text{A.7})$$

Thus, we **solve** for $\Delta\mathcal{L}$:

All we do is multiply both sides by Δw .

$$\Delta \mathcal{L} \approx \frac{\partial \mathcal{L}}{\partial w} \Delta w \quad (\text{A.8})$$

Since this derivation was gotten using scalars, we might need a **different** type of multiplication for our **vector** and **matrix** derivatives.

Concept 798

We can use derivatives to **approximate** the change in our output based on our input:

$$\Delta \mathcal{L} \approx \frac{\partial \mathcal{L}}{\partial w} * \Delta w$$

Where the $*$ symbol represents some type of **multiplication**.

We can think of this as a **function** that takes in change in Δw , and returns an **approximation** of the loss.

We already understand **scalar** derivatives, so let's move on to the **gradient**.

A.2 Derivative: Scalar/Vector (Gradient)

Our plan is to look at every derivative combination of scalars, vectors, and matrices we can.

First, we consider:

$$\frac{\partial(\text{Scalar})}{\partial(\text{Vector})} = \frac{\partial s}{\partial v} \quad (\text{A.9})$$

We'll take s to be our scalar, and v to be our vector. So, our input is a **vector**, and our output is a **scalar**.

$$\Delta v \longrightarrow [f] \longrightarrow \Delta s \quad (\text{A.10})$$

How do we make sense of this? Well, let's write Δv_i explicitly:

$$\underbrace{\begin{bmatrix} \Delta v_1 \\ \Delta v_2 \\ \vdots \\ \Delta v_m \end{bmatrix}}_{\Delta v} \longrightarrow \Delta s \quad (\text{A.11})$$

We can see that we have m different **inputs** we can change in order to change our **one** output.

So, our derivative needs to have m different **elements**: one for each element v_i .

A.2.1 Finding the scalar/vector derivative

But how do we shape our matrix? Let's look at our **rule**.

$$\Delta s \approx \frac{\partial s}{\partial v} * \Delta v \quad (\text{A.12})$$

We can transform using our shapes:

$$\Delta s \approx \frac{\partial s}{\partial v} * \underbrace{\begin{bmatrix} \Delta v_1 \\ \Delta v_2 \\ \vdots \\ \Delta v_m \end{bmatrix}}_{\Delta v} \quad (\text{A.13})$$

How do we get Δs ? We have so many variables. Let's focus on them one at a time: breaking Δv into Δv_i , so we'll try to consider each v_i **separately**.

Last Updated: 09/03/24 03:53:41

It's usually possible to change each v_i , so we have to look at every one of them.

One problem, though: how can we treat each **derivative** separately?

- Suppose we only start by considering Δv_1 : it'll move us along one axis.
- But our derivative is based on our position! If we've just moved, all of our derivatives have changed.
- So, choosing Δv_1 first, had an effect on the other Δv_i terms: they're not really "separate".

So, what do we do?

A.2.2 Review: Planar Approximation

We'll resolve this the same way we did in chapter 3, **gradient descent**: by taking advantage of the "planar approximation".

The solution is this: assume your function is **smooth**. The **smaller** a step you take, the **less** your derivative has a chance to change.

Example: Take $f(x) = x^2$.

- If we go from $x = 1 \rightarrow 2$, then our derivative goes from $f'(x) = 2 \rightarrow 4$.
- Let's **shrink** our step. We go from $x = 1 \rightarrow 1.01$, our derivative goes from $f'(x) = 2 \rightarrow 2.02$.
 - Our derivative is almost the same!

This isn't true for big steps, but eventually, if your step is small enough, then the derivative will barely change.

If we take a small enough step Δv_i , then, if our function is **smooth**, then the derivative will hardly change!

So, if we zoom in enough (shrink the scale of change), then we can **pretend** the derivative is **constant**.

You could imagine repeatedly shrinking the size of our step, until the change in the derivatives is basically unnoticeable.

Concept 799

If you have a **smooth function**, then...

If you take sufficiently **small steps**, then you can treat the derivatives as **constant**.

This is called the **planar approximation** because that's how a smooth surface looks, at small scales:

- The derivatives of a flat plane are the same at every position: they're **constant** (we show this below).
- This is, of course, an **approximation**: we can't use it at every position, just "very nearby positions".

Clarification

This clarification is **optional**.

We can describe "sufficiently small steps" in a more mathematical way:

Our goal is for $f'(x)$ to be **basically constant**: it doesn't change much. In other words, $\Delta f'(x)$ is **small**.

Let's say we don't want it to change more than δ : this is our definition of "very small".

If you want

- $\Delta f'(x)$ to be very small ($|\Delta f'(x)| < \delta$)
- It has been proven that...
 - it's possible to take a small enough step $|\Delta x| < \epsilon$, and to get that result.

One way to describe this is to say that our function is (locally) **flat**: it looks like some kind of plane/hyperplane.

The word "locally" represents the small step size: we stay in the "local area".

Clarification 800

Why is this **true**? Because a **hyperplane** can be represented using our **linear** function

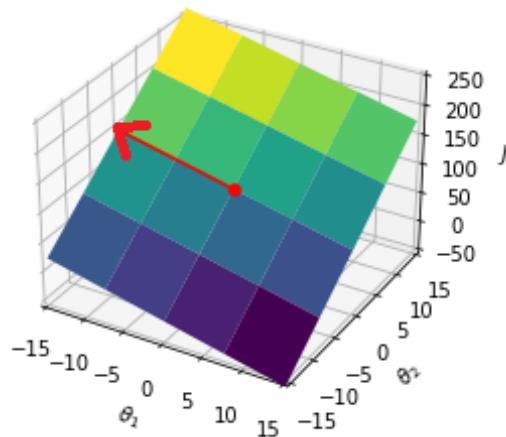
$$f(x) \approx \theta^T x + \theta_0 = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_m x_m$$

If we take a derivative:

$$\frac{\partial f}{\partial x_i} = \theta_i$$

That derivative is a **constant**! It's doesn't change based on **position**.

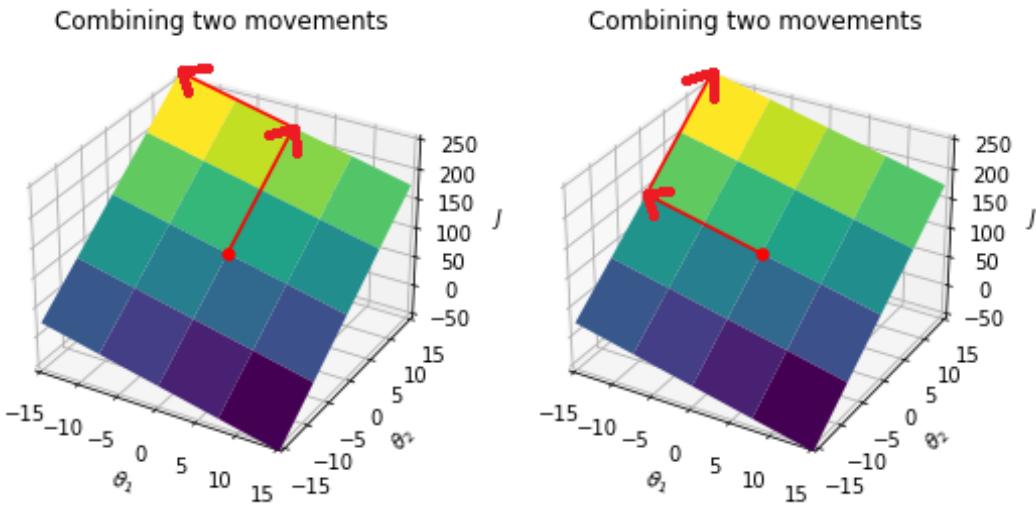
Movement in θ_1 on J



If we take very small steps, we can approximate our function as **flat**.

Why does this help? If our derivative doesn't **change**, you can take multiple steps Δv_i and the order doesn't matter.

So, you can combine your steps or separate them easily.



We can break up our big step into two smaller steps that are truly independent: order doesn't matter.

With that, we can add up all of our changes:

$$\Delta s = \Delta s_{\text{from } v_1} + \Delta s_{\text{from } v_2} + \dots + \Delta s_{\text{from } v_m} \quad (\text{A.14})$$

A.2.3 Our completed scalar/vector derivative

From this, we can get an **approximated** version of the MV chain rule.

Definition 801

The **multivariable chain rule approximation** looks similar to the multivariable chain rule, but for finite changes Δx_i .

In 3-D, we get

$$\Delta f \approx \underbrace{\frac{\partial f}{\partial x} \Delta x}_{x \text{ component}} + \underbrace{\frac{\partial f}{\partial y} \Delta y}_{y \text{ component}} + \underbrace{\frac{\partial f}{\partial z} \Delta z}_{z \text{ component}}$$

In general, we have

$$\Delta f \approx \sum_{i=1}^m \underbrace{\frac{\partial f}{\partial x_i} \Delta x_i}_{x_i \text{ component}}$$

This function lets us add up the effect each component has on our output, using **derivatives**.

This gives us what we're looking for:

$$\Delta s \approx \sum_{i=1}^m \frac{\partial s}{\partial v_i} \Delta v_i \quad (\text{A.15})$$

If we circle back around to our original approximation:

$$\sum_{i=1}^m \frac{\partial s}{\partial v_i} \Delta v_i = \frac{\partial s}{\partial v} * \begin{bmatrix} \Delta v_1 \\ \Delta v_2 \\ \vdots \\ \Delta v_m \end{bmatrix} \quad (\text{A.16})$$

When we look at the left side, we're multiplying pairs of components, and then adding them. That sounds similar to a **dot product**.

$$\sum_{i=1}^m \frac{\partial s}{\partial v_i} \Delta v_i = \begin{bmatrix} \frac{\partial s}{\partial v_1} \\ \frac{\partial s}{\partial v_2} \\ \vdots \\ \frac{\partial s}{\partial v_m} \end{bmatrix} \cdot \begin{bmatrix} \Delta v_1 \\ \Delta v_2 \\ \vdots \\ \Delta v_m \end{bmatrix} \quad (\text{A.17})$$

This gives us our derivative: it contains all of the **element-wise** derivatives we need, and in a **useful** form!

Definition 802

If s is a **scalar** and v is an $(m \times 1)$ **vector**, then we define the derivative or **gradient** $\frac{\partial s}{\partial v}$ as fulfilling:

$$\Delta s = \frac{\partial s}{\partial v} \cdot \Delta v$$

Or, equivalently,

$$\Delta s = \left(\frac{\partial s}{\partial v} \right)^T \Delta v$$

Thus, our derivative must be an $(m \times 1)$ vector

$$\frac{\partial s}{\partial v} = \begin{bmatrix} \frac{\partial s}{\partial v_1} \\ \frac{\partial s}{\partial v_2} \\ \vdots \\ \frac{\partial s}{\partial v_m} \end{bmatrix} = \begin{bmatrix} \frac{\partial s}{\partial v_1} \\ \frac{\partial s}{\partial v_2} \\ \vdots \\ \frac{\partial s}{\partial v_m} \end{bmatrix}$$

We can see the shapes work out in our matrix multiplication:

$$\underbrace{\Delta s}_{(1 \times 1)} = \underbrace{\left(\frac{\partial s}{\partial v} \right)^T}_{(1 \times m)} \underbrace{\Delta v}_{(m \times 1)} \quad (\text{A.18})$$

A.3 Derivative: Vector/Scalar

Now, we want to try the flipped version: we swap our vector and our scalar.

$$\frac{\partial(\text{Vector})}{\partial(\text{Scalar})} = \frac{\partial w}{\partial s} \quad (\text{A.19})$$

We'll take s to be our scalar, and w to be our vector. So, our input is a **scalar**, and our output is a **vector**.

$$\Delta s \rightarrow [f] \rightarrow \Delta w \quad (\text{A.20})$$

Written explicitly, like before:

$$\Delta s \rightarrow \underbrace{\begin{bmatrix} \Delta w_1 \\ \Delta w_2 \\ \vdots \\ \Delta w_n \end{bmatrix}}_{\Delta w} \quad (\text{A.21})$$

Note that we're using vector w instead of v this time: this will be helpful for our vector/vector derivative: we'll need two different symbols for "a vector".

We have 1 **input**, that can affect n different **outputs**. So, our derivative needs to have n elements.

Again, let's look at our **approximation rule**:

$$\Delta w \approx \frac{\partial w}{\partial s} \star \Delta s \quad \text{or} \quad \underbrace{\begin{bmatrix} \Delta w_1 \\ \Delta w_2 \\ \vdots \\ \Delta w_n \end{bmatrix}}_{\Delta w} \approx \frac{\partial w}{\partial s} \star \Delta s \quad (\text{A.22})$$

Here, we can't do a **dot product**: we're multiplying our derivative by a **scalar**.

A.3.1 Working with the vector derivative

How do we get each of our terms Δw_i ?

Well, each term is **separately** affected by Δs : we have our terms $\partial w_i / \partial s$.

So, if we take one of these terms **individually**, treating it as a scalar derivative, we get:

$$\Delta w_i = \frac{\partial w_i}{\partial s} \Delta s \quad (\text{A.23})$$

If you're ever confused with matrix math, thinking about individual elements is often a good way to figure it out!

Since we only have **one** input, we don't have to worry about **planar** approximations: we only take one step, in the s direction.

In our matrix, we get:

$$\mathbf{w} = \begin{bmatrix} \Delta w_1 \\ \Delta w_2 \\ \vdots \\ \Delta w_n \end{bmatrix} = \begin{bmatrix} \Delta s(\partial w_1 / \partial s) \\ \Delta s(\partial w_2 / \partial s) \\ \vdots \\ \Delta s(\partial w_n / \partial s) \end{bmatrix} \quad (\text{A.24})$$

This works out for our equation above!

It could be tempting to think of our derivative $\partial w / \partial s$ as a **column vector**: we just take w and just differentiate each element. Easy!

- In fact, this *is* a valid convention. However, this conflicts with our previous derivative: they're both column vectors!
- Not only is it **confusing**, but it also will make it harder to do our **vector/vector** derivative.

So, what do we do? We refer back to the equation we used last time:

$$\Delta w = \left(\frac{\partial w}{\partial s} \right)^T \Delta s \quad (\text{A.25})$$

We take the **transpose**! That way, one derivative is a column vector, and the other is a row vector. And, we know that this equation works out from the work we just did.

$$\Delta w = \left[\frac{\partial w_1}{\partial s}, \frac{\partial w_2}{\partial s}, \dots, \frac{\partial w_n}{\partial s} \right]^T \Delta s \quad (\text{A.26})$$

Clarification 803

We mentioned that it is a valid **convention** to have that **vector derivative** be a **column vector**, and have our **gradient** be a **row vector**.

This is **not** the convention we will use in this class - you will be confused if we try!

That means, for whatever **notation** we use here, you might see the **transposed** version elsewhere. They mean exactly the **same** thing!

$$\underbrace{\Delta w}_{(\text{orange} \times 1)} = \underbrace{\left(\frac{\partial w}{\partial s} \right)^T}_{(\text{orange} \times 1)} \underbrace{\Delta s}_{(1 \times 1)} \quad (\text{A.27})$$

As we can see, the dimensions check out.

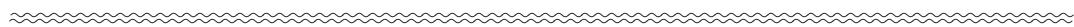
Definition 804

If s is a **scalar** and w is an $(n \times 1)$ **vector**, then we define the **vector derivative** $\partial w / \partial s$ as fulfilling:

$$\Delta w = \left(\frac{\partial w}{\partial s} \right)^T \Delta s$$

Thus, our derivative must be a $(1 \times n)$ vector

$$\frac{\partial w}{\partial s} = \left[\frac{\partial w_1}{\partial s}, \frac{\partial w_2}{\partial s}, \dots, \frac{\partial w_n}{\partial s} \right]$$



A.4 Derivative: Vector/Vector

We'll be combining our two previous derivatives:

$$\frac{\partial(\text{Vector})}{\partial(\text{Vector})} = \frac{\partial w}{\partial v} \quad (\text{A.28})$$

v and w are both **vectors**: thus, input and output are both **vectors**.

$$\Delta v \rightarrow [f] \rightarrow \Delta w \quad (\text{A.29})$$

Written out, we get:

$$\underbrace{\begin{bmatrix} \Delta v_1 \\ \Delta v_2 \\ \vdots \\ \Delta v_m \end{bmatrix}}_{\Delta v} \rightarrow \underbrace{\begin{bmatrix} \Delta w_1 \\ \Delta w_2 \\ \vdots \\ \Delta w_n \end{bmatrix}}_{\Delta w} \quad (\text{A.30})$$

Something pretty complicated! We have m inputs and n outputs. Every input can interact with every output.

So, our derivative needs to have mn different elements. That's a lot!

A.4.1 The vector/vector derivative

We return to our rule from before. We'll skip the star notation, and jump right to the equation we've gotten for both of our two previous derivatives:

$$\Delta w = \left(\frac{\partial w}{\partial v} \right)^T \Delta v \quad (\text{A.31})$$

Hopefully, since we're combining two different derivatives, we should be able to use the same rule here.

With mn different elements, this could get messy very fast. Let's see if we can focus on only **part** of our problem:

$$\begin{bmatrix} \Delta w_1 \\ \Delta w_2 \\ \vdots \\ \Delta w_n \end{bmatrix} = \left(\frac{\partial w}{\partial v} \right)^T \begin{bmatrix} \Delta v_1 \\ \Delta v_2 \\ \vdots \\ \Delta v_m \end{bmatrix} \quad (\text{A.32})$$

One input

We could try focusing on just a single **input** or a single **output**, to simplify things. Let's start with a single v_i .

$$\underbrace{\begin{bmatrix} \Delta w_1 \\ \Delta w_2 \\ \vdots \\ \Delta w_n \end{bmatrix}}_{\Delta w \text{ from } v_i} = \left(\frac{\partial w}{\partial v_i} \right)^T \Delta v_i \quad (\text{A.33})$$

We now have a simpler case: $\partial \text{Vector} / \partial \text{Scalar}$. We're familiar with this case!

$$\frac{\partial w}{\partial v_i} = \left[\frac{\partial w_1}{\partial v_i}, \frac{\partial w_2}{\partial v_i}, \dots, \frac{\partial w_n}{\partial v_i} \right] \quad (\text{A.34})$$

We get a vector. What if the **output** is a scalar instead?

One output

$$\Delta w_j = \left(\frac{\partial w_j}{\partial v} \right)^T \begin{bmatrix} \Delta v_1 \\ \Delta v_2 \\ \vdots \\ \Delta v_m \end{bmatrix} \quad (\text{A.35})$$

We have $\partial \text{Scalar} / \partial \text{Vector}$:

$$\frac{\partial w_j}{\partial v} = \begin{bmatrix} \partial w_j / \partial v_1 \\ \partial w_j / \partial v_2 \\ \vdots \\ \partial w_j / \partial v_m \end{bmatrix} \quad (\text{A.36})$$

So, our vector-vector derivative is a **generalization** of the two derivatives we did before!

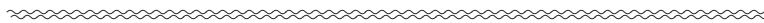
- It seems that extending along the **vertical** axis changes our v_i value, while moving along the **horizontal** axis changes our w_j value.



A.5 General derivative (Vector/Vector)

You might have a hint of what we get: one derivative stretches us along **one** axis, the other along the **second**.

But now, let's prove it to ourselves.



Our biggest problem is that we developed these tools for working with **vectors**, and now we have a **matrix**.

Rather than giving up this perspective, we'll instead take it further:

- We can think of a matrix as a "stack of vectors".

$$M = \begin{bmatrix} a_1 & b_1 & \cdots & z_1 \\ a_2 & b_2 & \cdots & z_2 \\ \vdots & \vdots & \ddots & \vdots \\ a_n & b_n & \cdots & z_n \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} & \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} & \cdots & \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} \end{bmatrix} \quad (\text{A.37})$$

- Or, a **row vector of column vectors**.

$$M = \begin{bmatrix} \vec{a} & \vec{b} & \cdots & \vec{z} \end{bmatrix} \quad (\text{A.38})$$

So, that's our plan: we first view our matrix as a row vector, hiding the column vectors inside.

- This requires treating our column vectors as if they were "scalars".

Then, we'll expand those column vectors, and see what we get.

Concept 805

One way to **simplify** our work is to treat **vectors** as **scalars**, and then convert them back into **vectors** after applying some math.

- We have to be **careful** - any operation we apply to the pretend "**scalar**", has to match how the **vector** would behave.



This is **equivalent** to when just focused on one scalar inside our vector, and then stacked all those scalars back into the vector.

Here's how we apply this to our situation.

- First, we treat each **column** as a **scalar**, and the whole object as a **vector**.
 - This will use one familiar derivative.

- Then, we'll expand each **column** into a whole **vector**. Then, we have a **matrix**.
 - This will us our other derivative.

This isn't just a cute trick: it relies on an understanding that, at its **basic** level, we're treating **scalars** and **vectors** and **matrices** as the same type of object: a structured array of numbers.

We'll get into "arrays" later.

As always, our goal is to **simplify** our work, so we can handle each piece of it.

- We treat Δv as a **scalar** so we can get the **simplified** derivative.

$$\Delta w = \left(\frac{\partial w}{\partial v} \right)^T \Delta v \quad (\text{A.39})$$

We'll only expand Δw , because that's a simpler case we know how to manage.

$$\begin{bmatrix} \Delta w_1 \\ \Delta w_2 \\ \vdots \\ \Delta w_n \end{bmatrix} = \left(\frac{\partial w}{\partial v} \right)^T \Delta v \quad (\text{A.40})$$

- Notice that we **didn't** simplify v to v_i . We aren't using only one element of v : we're pretending as if the whole vector v (including all elements) is a scalar.

We compute our derivative, based on earlier work:

$$\frac{\partial w}{\partial v} = \underbrace{\left[\frac{\partial w_1}{\partial v}, \frac{\partial w_2}{\partial v}, \dots, \frac{\partial w_n}{\partial v} \right]}_{\text{Column } j \text{ matches } w_j} \quad (\text{A.41})$$

- Our "answer" is a row vector. But, each of those derivatives is a **column** vector!

Now that we've taken care of ∂w_j (one for each column), we can **expand** our derivatives in terms of ∂v_i : each is a **column vector**!

First, for w_1 :

$$\frac{\partial \mathbf{w}}{\partial \mathbf{v}} = \left[\begin{array}{c} \frac{\partial w_1}{\partial v_1} \\ \frac{\partial w_1}{\partial v_2} \\ \vdots \\ \frac{\partial w_1}{\partial v_m} \end{array}, \quad \frac{\partial w_2}{\partial v_1}, \quad \dots \quad \frac{\partial w_n}{\partial v_1} \right] \quad \text{Column } j \text{ matches } w_j \quad \text{Row } i \text{ matches } v_i \quad (\text{A.42})$$

And again, for w_2 :

$$\frac{\partial \mathbf{w}}{\partial \mathbf{v}} = \left[\begin{array}{c} \frac{\partial w_1}{\partial v_1} \\ \frac{\partial w_1}{\partial v_2} \\ \vdots \\ \frac{\partial w_1}{\partial v_m} \end{array}, \quad \begin{array}{c} \frac{\partial w_2}{\partial v_1} \\ \frac{\partial w_2}{\partial v_2} \\ \vdots \\ \frac{\partial w_2}{\partial v_m} \end{array}, \quad \dots \quad \frac{\partial w_n}{\partial v_1} \right] \quad \text{Column } j \text{ matches } w_j \quad \text{Row } i \text{ matches } v_i \quad (\text{A.43})$$

And again, for w_n :

$$\frac{\partial \mathbf{w}}{\partial \mathbf{v}} = \left[\begin{array}{c} \frac{\partial w_1}{\partial v_1} \\ \frac{\partial w_1}{\partial v_2} \\ \vdots \\ \frac{\partial w_1}{\partial v_m} \end{array}, \quad \begin{array}{c} \frac{\partial w_2}{\partial v_1} \\ \frac{\partial w_2}{\partial v_2} \\ \vdots \\ \frac{\partial w_2}{\partial v_m} \end{array}, \quad \dots \quad \begin{array}{c} \frac{\partial w_n}{\partial v_1} \\ \frac{\partial w_n}{\partial v_2} \\ \vdots \\ \frac{\partial w_n}{\partial v_m} \end{array} \right] \quad \text{Column } j \text{ matches } w_j \quad \text{Row } i \text{ matches } v_i \quad (\text{A.44})$$

We have column vectors in our row vector... based on our new perspective, this is basically the same as a **matrix**.

Definition 806

If

- v is an $(m \times 1)$ vector
- w is an $(n \times 1)$ vector

Then we define the **vector derivative** $\partial w / \partial v$ as fulfilling:

$$\Delta w = \left(\frac{\partial w}{\partial v} \right)^T \Delta v$$

Thus, our derivative must be a $(m \times n)$ vector

$$\frac{\partial w}{\partial v} = \left[\begin{array}{cccc} \frac{\partial w_1}{\partial v_1} & \frac{\partial w_2}{\partial v_1} & \dots & \frac{\partial w_n}{\partial v_1} \\ \frac{\partial w_1}{\partial v_2} & \frac{\partial w_2}{\partial v_2} & \dots & \frac{\partial w_n}{\partial v_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial w_1}{\partial v_m} & \frac{\partial w_2}{\partial v_m} & \dots & \frac{\partial w_n}{\partial v_m} \end{array} \right] \quad \begin{array}{l} \text{Column } j \text{ matches } w_j \\ \text{Row } i \text{ matches } v_i \end{array}$$

This general form can be used for **any** of our matrix derivatives.

So, our matrix can represent any **combination** of two elements! We just assign each **row** to a v_i component, and each **column** with a w_j component.

A.5.1 More about the vector/vector derivative

Let's show a specific example: w is (3×1) , v is (2×1) .

$$\frac{\partial w}{\partial v} = \left[\begin{array}{ccc} \overbrace{\frac{\partial w_1}{\partial v_1}}^{w_1} & \overbrace{\frac{\partial w_2}{\partial v_1}}^{w_2} & \overbrace{\frac{\partial w_3}{\partial v_1}}^{w_3} \\ \overbrace{\frac{\partial w_1}{\partial v_2}}^{w_1} & \overbrace{\frac{\partial w_2}{\partial v_2}}^{w_2} & \overbrace{\frac{\partial w_3}{\partial v_2}}^{w_3} \end{array} \right] \quad \begin{array}{l} \} v_1 \\ \} v_2 \end{array} \quad (\text{A.45})$$

Another way to describe the general case:

Notation 807

Our matrix $\frac{\partial \mathbf{w}}{\partial \mathbf{v}}$ is entirely filled with **scalar derivatives**

$$\frac{\partial w_j}{\partial v_i}$$

Where any one **derivative** is stored in

- Row i
 - m rows total
- Column j
 - n columns total

We can also compress it along either axis (just like how we did to derive this result):

Notation 808

Our matrix $\frac{\partial \mathbf{w}}{\partial \mathbf{v}}$ can be written as

$$\frac{\partial \mathbf{w}}{\partial \mathbf{v}} = \left[\underbrace{\frac{\partial w_1}{\partial v_1}, \frac{\partial w_2}{\partial v_1}, \dots, \frac{\partial w_n}{\partial v_1}}_{\text{Column } j \text{ matches } w_j} \right]$$

or

$$\frac{\partial \mathbf{w}}{\partial \mathbf{v}} = \left[\begin{array}{c} \frac{\partial \mathbf{w}}{\partial v_1} \\ \frac{\partial \mathbf{w}}{\partial v_2} \\ \vdots \\ \frac{\partial \mathbf{w}}{\partial v_m} \end{array} \right] \left. \right\} \text{Row } i \text{ matches } v_i$$

These compressed forms will be useful for deriving our new and final derivatives, **matrix-scalar** pairs.

A.6 Derivative: matrix/scalar

Now, we have our general form for creating derivatives.

We'll get our derivative of the form

$$\frac{\partial(\text{Matrix})}{\partial(\text{Scalar})} = \frac{\partial \textcolor{blue}{M}}{\partial \textcolor{red}{s}} \quad (\text{A.46})$$

We have a matrix $\textcolor{blue}{M}$ in the shape $(r \times k)$ and a scalar $\textcolor{red}{s}$. Our **input** is a **scalar**, and our **output** is a **matrix**.

$$\textcolor{blue}{M} = \begin{bmatrix} \textcolor{blue}{m}_{11} & \textcolor{blue}{m}_{12} & \cdots & \textcolor{blue}{m}_{1k} \\ \textcolor{blue}{m}_{21} & \textcolor{blue}{m}_{22} & \cdots & \textcolor{blue}{m}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \textcolor{blue}{m}_{r1} & \textcolor{blue}{m}_{r2} & \cdots & \textcolor{blue}{m}_{rk} \end{bmatrix} \quad (\text{A.47})$$

This may seem concerning: before, we divided **inputs** across **rows**, and **outputs** across **columns**. But in this case, we have **no** input axes, and **two** output axes.

Well, let's try to make this work anyway.

What did we do before, when we didn't know how to handle a **new** derivative?

- We compared it to **old** versions: we built our vector/vector case using the vector/scalar case and the scalar/vector case.
- We did this by **compressing** one of our *vectors* into a *scalar* temporarily: this works, because we want to treat each of these objects the **same way**.

We don't know how to work with **Matrix/Scalar**, but what's the **closest** thing we do know? **Vector/Scalar**.

How do we accomplish that? As we saw above, a matrix is a **vector of vectors**. We could turn it into a **vector of scalars**.

Concept 809

A **matrix** can be thought of as a **column vector** of **row vectors** (or vice versa).

So, we can use our earlier technique and convert the **row vectors** into **scalars**.

We'll replace the **row vectors** in our matrix with **scalars**.

$$\textcolor{blue}{M} = \begin{bmatrix} \textcolor{blue}{M}_1 \\ \textcolor{blue}{M}_2 \\ \vdots \\ \textcolor{blue}{M}_r \end{bmatrix} \quad (\text{A.48})$$

Now, we can pretend our matrix is a vector! We've got a derivative for that:

$$\frac{\partial \mathbf{M}}{\partial s} = \begin{bmatrix} \frac{\partial M_1}{\partial s} & \frac{\partial M_2}{\partial s} & \dots & \frac{\partial M_r}{\partial s} \end{bmatrix} \quad (\text{A.49})$$

Aha - we have the same form that we did for our vector/scalar derivative! Each derivative is a column vector. Let's expand it out:

$$\frac{\partial \mathbf{M}}{\partial s} = \left[\begin{array}{c} \left[\frac{\partial m_{11}}{\partial s}, \frac{\partial m_{21}}{\partial s}, \dots, \frac{\partial m_{r1}}{\partial s} \right] \\ \vdots \\ \left[\frac{\partial m_{1k}}{\partial s}, \frac{\partial m_{2k}}{\partial s}, \dots, \frac{\partial m_{rk}}{\partial s} \right] \end{array} \right], \quad \text{Column } j \text{ matches } m_j? \quad (\text{A.50})$$

Row i matches $m_{?i}$

Definition 810

If \mathbf{M} is a matrix in the shape $(r \times k)$ and s is a scalar,

Then we define the **matrix derivative** $\partial \mathbf{M} / \partial s$ as the $(k \times r)$ matrix:

$$\frac{\partial \mathbf{M}}{\partial s} = \left[\begin{array}{cccc} \frac{\partial m_{11}}{\partial s} & \frac{\partial m_{21}}{\partial s} & \dots & \frac{\partial m_{r1}}{\partial s} \\ \frac{\partial m_{12}}{\partial s} & \frac{\partial m_{22}}{\partial s} & \dots & \frac{\partial m_{r2}}{\partial s} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial m_{1k}}{\partial s} & \frac{\partial m_{2k}}{\partial s} & \dots & \frac{\partial m_{rk}}{\partial s} \end{array} \right], \quad \text{Column } j \text{ matches } m_j? \quad (\text{A.51})$$

Row i matches $m_{?i}$

- This matrix has the shape of \mathbf{M}^T .

A.7 Derivative: scalar/matrix

We'll get our derivative of the form

$$\frac{\partial(\text{Scalar})}{\partial(\text{Matrix})} = \frac{\partial s}{\partial M} \quad (\text{A.51})$$

We have a matrix M in the shape $(r \times k)$ and a scalar s . Our **input** is a **matrix**, and our **output** is a **scalar**.

Let's do what we did last time: break it into **row vectors**.

$$M = \begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ M_r \end{bmatrix} \quad (\text{A.52})$$

The gradient for this "vector" gives us a **column vector**:

$$\frac{\partial s}{\partial M} = \begin{bmatrix} \frac{\partial s}{\partial M_1} \\ \frac{\partial s}{\partial M_2} \\ \vdots \\ \frac{\partial s}{\partial M_r} \end{bmatrix} \quad (\text{A.53})$$

This time, each derivative is a **row vector**. Let's **expand**:

$$\frac{\partial s}{\partial M} = \begin{bmatrix} \left[\frac{\partial s}{\partial m_{11}} \quad \frac{\partial s}{\partial m_{12}} \quad \cdots \quad \frac{\partial s}{\partial m_{1k}} \right] \\ \left[\frac{\partial s}{\partial m_{21}} \quad \frac{\partial s}{\partial m_{22}} \quad \cdots \quad \frac{\partial s}{\partial m_{2k}} \right] \\ \vdots \\ \left[\frac{\partial s}{\partial m_{r1}} \quad \frac{\partial s}{\partial m_{r2}} \quad \cdots \quad \frac{\partial s}{\partial m_{rk}} \right] \end{bmatrix} \quad (\text{A.54})$$

Definition 811

If \mathbf{M} is a matrix in the shape $(r \times k)$ and s is a scalar,

Then we define the **matrix derivative** $\partial s / \partial \mathbf{M}$ as the $(r \times k)$ matrix:

$$\frac{\partial s}{\partial \mathbf{M}} = \left[\begin{array}{cccc} \frac{\partial s}{\partial m_{11}} & \frac{\partial s}{\partial m_{12}} & \cdots & \frac{\partial s}{\partial m_{1k}} \\ \frac{\partial s}{\partial m_{21}} & \frac{\partial s}{\partial m_{22}} & \cdots & \frac{\partial s}{\partial m_{2k}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial s}{\partial m_{r1}} & \frac{\partial s}{\partial m_{r2}} & \cdots & \frac{\partial s}{\partial m_{rk}} \end{array} \right] \quad \begin{array}{l} \text{Column } j \text{ matches } m_{?j} \\ \text{Row } i \text{ matches } m_i? \end{array}$$

This matrix has the same shape as \mathbf{M} .

A.8 Tensors

A.8.1 Other Derivatives

After these, you might ask yourself, what about other derivative combinations?

$$\frac{\partial \textcolor{blue}{v}}{\partial \textcolor{blue}{M}}? \quad \frac{\partial \textcolor{blue}{M}}{\partial \textcolor{blue}{v}}? \quad \frac{\partial \textcolor{blue}{M}}{\partial \textcolor{blue}{M}^2}?$$
 (A.55)

There's a problem with all of these: the total number of axes is **too large**.

What do we mean by an **axis**?

Definition 812

An **axis** is one of the **indices** we can adjust to get a different scalar in our array: each index is a "direction" we can move along our object to **store** numbers.

- A **scalar** has **0 axes**: we only have one scalar, so we have no indices to adjust.
- ~~~~~
- A **vector** has **1 axis**: we can get different scalars by moving **vertically** (for column vectors): v_1, v_2, v_3, \dots

$$\left[\begin{array}{c} v_1 \\ v_2 \\ \vdots \\ v_m \end{array} \right] \quad \left. \right\} \text{Axis 1}$$

~~~~~

- A **matrix** has **2 axes**: we can move **horizontally** or **vertically**.

$$\overbrace{\left[ \begin{array}{cccc} m_{11} & m_{12} & \cdots & m_{1r} \\ m_{21} & m_{22} & \cdots & m_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ m_{k1} & m_{k2} & \cdots & m_{kr} \end{array} \right]}^{\text{Axis 2: Columns}} \quad \left. \right\} \text{Axis 1: Rows}$$

These can also be called **dimensions**.

Why does the number of **axes** matter? Remember that, so far, for our derivatives, each axis of the output represented an axis of the **input** or **output**.

Note that last bit: we're saying a vector has one dimension. Can't a vector have **multiple dimensions**? We'll clarify this in an optional following section.

$$\frac{\partial \mathbf{w}}{\partial \mathbf{v}} = \left[ \begin{array}{cccc} \frac{\partial w_1}{\partial v_1} & \frac{\partial w_2}{\partial v_1} & \cdots & \frac{\partial w_n}{\partial v_1} \\ \frac{\partial w_1}{\partial v_2} & \frac{\partial w_2}{\partial v_2} & \cdots & \frac{\partial w_n}{\partial v_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial w_1}{\partial v_m} & \frac{\partial w_2}{\partial v_m} & \cdots & \frac{\partial w_n}{\partial v_m} \end{array} \right]$$

Column j: vertical axis of  $\mathbf{w}$

Row i: vertical axis of  $\mathbf{v}$

The way we currently build derivatives, we try to get **every pair** of input-output variables: we use **one** axis for each **axis** of either the **input** or **output**.

Take some examples:

- $\partial s / \partial v$ : we need one axis to represent each term  $v_i$ .
  - 0 axis + 1 axis  $\rightarrow$  1 axis: the output is a (column) **vector**.
- $\partial v / \partial s$ : we need one axis to represent each term  $w_j$ .
  - 1 axis + 0 axis  $\rightarrow$  1 axis: the output is a (row) **vector**.
- $\partial w / \partial v$ : we need one axis to represent each term  $v_i$ , and another to represent each term  $w_j$ .
  - 1 axis + 1 axis  $\rightarrow$  2 axes: the output is a **matrix**.
- $\partial M / \partial s$ : we need one axis to represent the rows of  $M$ , and another to represent the columns of  $M$ .
  - 2 axis + 0 axis  $\rightarrow$  2 axes: the output is a **matrix**.
- $\partial s / \partial M$ : we need one axis to represent the rows of  $M$ , and another to represent the columns of  $M$ .
  - 0 axis + 2 axis  $\rightarrow$  2 axes: the output is a **matrix**.

Notice the pattern!

**Concept 813**

A **matrix derivative** needs to be able to account for each type/**index** of variable in the input **and** the output.

So, if the **input**  $x$  has  $m$  axes, and the **output**  $y$  has  $n$  axes, then the derivative needs to have the same **total** number:

$$\text{Axes}\left(\frac{\partial \textcolor{violet}{y}}{\partial \textcolor{orange}{x}}\right) = \text{Axes}(\textcolor{violet}{y}) + \text{Axes}(\textcolor{orange}{x})$$

This is where our problem comes in: if we have a vector and a matrix, we need **3 axes!** That's more than a matrix.

---

### A.8.2 Dimensions (Optional)

Here's a quick aside to clear up possible confusion from the last section: our definition of axes and "dimensions".

We said a vector has 1 axis, or "dimension" of movement. But, can't a vector have **multiple** dimensions?

### Clarification 814

We have two competing definition of **dimension**: this explains why we can say seemingly conflicting things about derivatives.

So far, by "dimension", we mean, "a separate **value** we can **adjust**".

- Under this definition, a  $(k \times 1)$  column **vector** has **k** dimensions: it contains **k** different scalars we can **adjust**.

$$\left[ \begin{array}{c} v_1 \\ v_2 \\ \vdots \\ v_k \end{array} \right] \quad \left. \right\} \text{We can adjust each of our } k \text{ scalars.}$$

- You might say a  $(k \times r)$  **matrix** has **k** dimensions, too: based on the **dimensionality** of its column vectors.
  - Since we prioritize the size of the vectors, we could say this is a very "vector-centric" definition.

In this section, by "dimension", we mean, "an **index** we can **adjust** (move along) to find another scalar.

- Under this definition, a  $(k \times 1)$  column **vector** has **1** dimension: we only have **1** axis of **movement**.
- You might say a  $(k \times r)$  **matrix** has **2** dimensions: a **horizontal** one, and a **vertical** one.
  - This **definition** is the kind we use in the following sections.

In other words:

- Vector-style dimensionality: number of separate scalars in your vector.
- Matrix(tensor)-style dimensionality: number of separate indices we can use to find scalars.

If you jumped here from X.16, feel free to follow this [link](#) back. Otherwise, continue on.

### A.8.3 Dealing with Tensors

If a vector looks like a "line" of numbers, and a matrix looks like a "rectangle" of numbers, then a **3-axis** version would look like a "box" of numbers. How do we make sense of this?

First, what is this kind of object we've been working with? Vectors, matrices, etc. This collection of numbers, organized neatly, is an **array**.

#### Definition 815

An **array** of objects is an **ordered sequence** of them, stored together.

The most typical example is a **vector**: an ordered sequence of **scalars**.

A **matrix** can be thought of as a **vector** of **vectors**. For example: it could be a row vector, where every column is a column vector.

- So, we think of a matrix as a "two-dimensional array".

We can extend this to any number of dimensions. We call this kind of generalization a **tensor**.

#### Definition 816

In machine learning, we think of a **tensor** as a "**multidimensional array**" of numbers.

- Each "dimension" is what we have been calling an "**axis**".
- A tensor with  $c$  axes is called a  **$c$ -Tensor**.

**Example:** The 3-D box we are talking about above is called a 3-Tensor. We can simply think of it as a stack of matrices.

- We can imagine stacking the following two matrices in the third dimension, with the leftmost in front, and the rightmost in the back, in a  $(2 \times 3 \times 2)$  block:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \quad \begin{bmatrix} 5 & 7 & 11 \\ 13 & 17 & 19 \end{bmatrix} \quad (\text{A.56})$$

**Clarification 817**

Note that what we call a tensor is **not** a mathematical (or physics) tensor.

- We don't make use of most of the crucial properties of a mathematical tensor.

Our tensor can be better thought of as a "**generalized matrix**", or a "multidimensional array".

This is important, because a "mathematical" tensor has properties that can be confusing from an ML perspective:

- The "tensor product" doesn't behave at all like a matrix product.
- Rather, the generalized version of the **matrix product** is something called **tensor contraction**.

Tensor contraction examples, like the einsum function in numpy ("einstein summation notation"), can get very complex for higher-dimensional tensors.

If tensors don't really behave quite like matrices, and their math is complex, how do we handle **tensors**?

Simply, we convert them into regular **matrices** in some way, and then do our usual math on them:

- If a tensor has a pattern of **zeroes**, we might be able to flatten it into a matrix.
  - For example, in this matrix, we make a vector out of the diagonal:

$$\begin{bmatrix} 3 & 0 & 0 \\ 0 & 9 & 0 \\ 0 & 0 & 4 \end{bmatrix} \rightarrow \begin{bmatrix} 3 \\ 9 \\ 4 \end{bmatrix} \quad (\text{A.57})$$

These examples aren't especially important, but you can see different variations used for different problems!

- We can also flatten it into a matrix or vector by **stacking** layers next to each other in the same dimension.

- For example, we could stack the two columns of a matrix:

This version is used when we are willing to give up our n-D structure: we lose a lot of information.

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \rightarrow \begin{bmatrix} 1 \\ 3 \\ 2 \\ 4 \end{bmatrix} \quad (\text{A.58})$$

- We cleverly "**hide**" dimensions of a matrix, as we have before: treat a scalar as a vector, etc.

This works best for high-level, conceptual stuff. We'll use it below.

- We **systematically** multiply and add our elements in a way that gives us the derivatives we want, replicating matrix multiplication.
  - This is **tensor contraction**.
  - This method is often used by softwares to implement the chain rule.

### Clarification 818

If you look into **derivatives** that would result in a **3-tensor** or higher, you'll find that there's no consistent **notation** for what these derivatives look like: shape, structure, etc.

These techniques are part of why: there are **different** approaches for how to approach these objects.

The solution tends to be directly computing the chain rule, rather than figuring out the structure of the abstract object.

As we will see in the next chapter, tensors are **very** important to machine learning.

However, because they're so troublesome to work with, we'll convert to matrices in most cases.

---

## A.9 Chapter 7 Derivatives

### A.9.1 The loss derivative

Finally, we apply all of our new knowledge to the common derivatives in section 7.5.

$$\overbrace{\frac{\partial \mathcal{L}}{\partial A^L}}^{(n^L \times 1)} \quad (\text{A.59})$$

Loss is not given, so we can't compute it. But, we can get the shape: we have a scalar/vector derivative, so the shape matches  $A^L$ .

#### Notation 819

Our derivative

$$\frac{\partial \mathcal{L}}{\partial A^L}$$

Is a scalar/vector derivative, and thus the shape  $(n^L \times 1)$ .

### A.9.2 The weight derivative

$$\overbrace{\frac{\partial \mathcal{L}}{\partial W^\ell}}^{(m^\ell \times 1) ?} \quad (\text{A.60})$$

This derivative is difficult - it's a derivative in the form vector/matrix. With **three** axes, a 3-tensor seems suitable.

But, our goal is to use this for the **chain rule**: so, we need to make matrix shapes that **match**.

$$\frac{\partial \mathcal{L}}{\partial W^\ell} = \underbrace{\frac{\partial \mathcal{L}}{\partial Z^\ell}}_{\text{Weight link}} \cdot \underbrace{\left( \frac{\partial \mathcal{L}}{\partial Z^\ell} \right)^\top}_{\text{Other layers}} \quad (\text{A.61})$$

Remember that we had to do weird reordering and transposing in chapter 7 to flip the order of the chain rule? This is why: shape consistency.

Our problem is we have **too many axes**: the easiest way to resolve this to **break up** our matrix.

$$W = [W_1 \ W_2 \ \dots \ W_n] \quad (\text{A.62})$$

Notice that, this time, we broke  $W$  into **column vectors**  $W_i$ , rather than row vectors. Each **neuron**'s weights are represented by one column vector.

So, for now, we focus on only **one neuron** at a time: it has a column vector  $W_i$ . We'll ignore everything except  $W_i$ .

For simplicity, we're gonna ignore the  $\ell$  notation: just be careful, because  $Z$  and  $A$  are from two different layers!

$$W_i = \begin{bmatrix} w_{1i} \\ w_{2i} \\ \vdots \\ w_{mi} \end{bmatrix} \quad (\text{A.63})$$

### Concept 820

When you have a derivative that has **too many dimensions**, it's easier to only focus on **one** of the elements/dimensions, and find the derivative of that first.

- In this example, rather than finding  $\frac{\partial Z}{\partial W}$ , we're finding  $\frac{\partial Z}{\partial W_i}$ .
- We're simplifying from **matrix**  $W$  to **vector**  $W_i$ .

~~~~~  
So, we're doing $\frac{\partial Z}{\partial W_i}$: we need equations relating these variables.

- W_i represents the weights of one neuron, while Z represents the pre-activation of all the neurons.

Let's focus on one pre-activation neuron at a time.

$$z_j = W_j^T A \quad (\text{A.64})$$

Notably, this equation reminds us that pre-activation z_j is only affected by weights from the same neuron, W_j .

- Weights from a different neuron have no effect on z_j .
- So, the derivative between them will be 0: we can check this using the equation above.

Concept 821

The i^{th} neuron's **weights**, W_i , have **no effect** on a different neuron's **pre-activation** z_j .

So, if the neurons don't match, then our derivative is **zero**:

- i is the neuron for **pre-activation** z_i
- j is the j^{th} weight **in neuron** k .
- k is the **neuron** for weight vector W_k

$$\frac{\partial z_i}{\partial W_{jk}} = 0 \quad \text{if } i \neq k$$

So, our only nonzero derivatives are

$$\frac{\partial z_i}{\partial W_{ji}}$$

With that out of the way, let's actually **expand** our expression from above, to compute $\frac{\partial z_i}{\partial W_{ji}}$.

$$z_i = W_i^T A \quad (\text{A.65})$$

$$z_i = \begin{bmatrix} w_{1i} & w_{2i} & \cdots & w_{mi} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} \quad (\text{A.66})$$

This matrix multiplication can be written as an easier-to-use **sum**:

$$z_i = \sum_{j=1}^n W_{ji} a_j \quad (\text{A.67})$$

Finally, we can get our derivatives, for the **non-zero** terms:

$$\frac{\partial z_i}{\partial W_{ji}} = a_j \quad (\text{A.68})$$

Now, we can combine them into a vector.

$$\frac{\partial z_i}{\partial W_i} = \begin{bmatrix} \frac{\partial z_i}{\partial W_{1i}} \\ \frac{\partial z_i}{\partial W_{2i}} \\ \vdots \\ \frac{\partial z_i}{\partial W_{mi}} \end{bmatrix} \quad (\text{A.69})$$

We plug in our new values:

$$\frac{\partial z_i}{\partial W_i} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} = A \quad (\text{A.70})$$

We get a result!

~~~~~

What if the pre-activation  $z_i$  and weights  $W_k$  don't match? We've already seen: the derivative is 0: weights from one neuron don't affect different neurons.

$$\frac{\partial z_i}{\partial W_{jk}} = 0 \quad \text{if } i \neq k \quad (\text{A.71})$$

We can combine these into a **zero vector**:

$$\frac{\partial z_i}{\partial W_k} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \vec{0} \quad \text{if } i \neq k \quad (\text{A.72})$$

So, now, we can describe all of our vector components:

$$\frac{\partial z_i}{\partial W_k} = \begin{cases} A & \text{if } i = k \\ \vec{0} & \text{if } i \neq k \end{cases}$$

(A.73)

Again, we see: "same neuron? they're related. Different neurons? They're not."

~~~~~

This derivative, despite being a vector, can be represented with a **single** symbol ($\vec{0}$ or A), for each element $\frac{\partial z_i}{\partial W_k}$.

- For the sake of trying to wrangle our 3-tensor, we'll "hide" one dimension using this fact.
- If this is the case, then we have to pretend that W_k is a "scalar": so, W is now a "vector".

We compute $\frac{\partial Z}{\partial W}$ as a "vector/vector" derivative: this is a **matrix**!

$$\frac{\partial Z}{\partial W} = \begin{bmatrix} A & \vec{0} & \cdots & \vec{0} \\ \vec{0} & A & \cdots & \vec{0} \\ \vdots & \vdots & \ddots & \vec{0} \\ \vec{0} & \vec{0} & \vec{0} & A \end{bmatrix} \quad (\text{A.74})$$

We have our result: it turns out, despite being stored in a **matrix**-like format, this is actually a **3-tensor**! Each "scalar" of our **matrix** is a **vector**: we really have 3 axes.

~~~~~  
But, we don't really... *want* a tensor. It doesn't have the right shape, and we can't do matrix multiplication.

We'll solve this by **simplifying**, without losing key information.

### Concept 822

For many of our "tensors" resulting from matrix derivatives, they contain **empty** rows or **redundant** information.

Based on this, we can **simplify** our tensor into a fewer-dimensional (fewer axes) object.

We can see two types of **redundancy** above:

- Every element **off** the diagonal is 0.
- Every element **on** the diagonal is the same.

Let's fix the first one: we'll go from a diagonal matrix to a column vector.

$$\begin{bmatrix} A & \vec{0} & \cdots & \vec{0} \\ \vec{0} & A & \cdots & \vec{0} \\ \vdots & \vdots & \ddots & \vec{0} \\ \vec{0} & \vec{0} & \vec{0} & A \end{bmatrix} \longrightarrow \begin{bmatrix} A \\ A \\ \vdots \\ A \end{bmatrix} \quad (\text{A.75})$$

Then, we'll combine all of our redundant  $A$  values.

$$\begin{bmatrix} \textcolor{orange}{A} \\ \textcolor{orange}{A} \\ \vdots \\ \textcolor{orange}{A} \end{bmatrix} \longrightarrow \textcolor{orange}{A} \quad (\text{A.76})$$

We have our final answer!

### Notation 823

Our derivative

$$\underbrace{\frac{\partial \textcolor{red}{Z}^\ell}{\partial \textcolor{blue}{W}^\ell}}_{(\textcolor{blue}{m}^\ell \times 1)} = \textcolor{orange}{A}^{\ell-1}$$

Is a vector/matrix derivative, and thus should be a 3-tensor.

But, we have turned it into the shape  $(\textcolor{blue}{m}^\ell \times 1)$ .

This is as **condensed** as we can get our information: if we compress to a scalar, we lose some of our elements.

- Even with this derivative, we still have to do some clever **reshaping** to get the result we need (transposing, changing multiplication order of the chain rule, etc.)

However, at the end, we get the right shape for our chain rule!

This is the weird order change we mentioned in chapter 7, where we reverse the order of elements.

### A.9.3 Linking Layers $\ell - 1$ and $\ell$

$$\frac{\partial \textcolor{red}{Z}^\ell}{\partial \textcolor{orange}{A}^{\ell-1}} \quad (\text{A.77})$$

This derivative is much more manageable: it's just the derivative between a vector and a vector. Let's look at our equation again:

Ignoring superscripts  $\ell$ , as before.

$$\textcolor{red}{Z} = \textcolor{blue}{W}^T \textcolor{orange}{A} \quad (\text{A.78})$$

We'll use the same approach we did last section:  $\textcolor{blue}{W}$  is a "vector", and we'll focus on  $\textcolor{blue}{W}_i$ . This will allow us to break it up **element-wise**.

- We could treat  $\textcolor{blue}{W}$  as a whole matrix, but our approach will be cleaner: otherwise, we'd have to depict every  $\textcolor{blue}{W}_i$  at **once**.

$$W = \begin{bmatrix} W_1 & W_2 & \dots & W_n \end{bmatrix} \quad W_i = \begin{bmatrix} w_{1i} \\ w_{2i} \\ \vdots \\ w_{mi} \end{bmatrix} \quad (\text{A.79})$$

Here's our equation:

$$z_i = \begin{bmatrix} w_{1i} & w_{2i} & \dots & w_{mi} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} \quad (\text{A.80})$$

We matrix multiply:

$$z_i = \sum_{j=1}^n W_{ji} a_j \quad (\text{A.81})$$

The derivative can be gotten from here -

$$\frac{\partial z_i}{\partial a_j} = W_{ji} \quad (\text{A.82})$$

We look at our whole matrix derivative:

$$\frac{\partial Z}{\partial A} = \left\{ \begin{array}{c} \text{Column } i \text{ matches } z_i \\ \left[ \begin{array}{ccc} \ddots & \vdots & \ddots \\ \cdots & \frac{\partial z_i}{\partial a_j} & \cdots \\ \ddots & \vdots & \ddots \end{array} \right] \end{array} \right\} \text{Row } j \text{ matches } a_j \quad (\text{A.83})$$

This notation looks a bit weird, but it's just a way to represent that all of our elements follow the same pattern.

Wait.

- The derivative  $\frac{\partial z_i}{\partial a_j}$  is in the  $j^{\text{th}}$  row,  $i^{\text{th}}$  column of  $\frac{\partial Z}{\partial A}$ .
- $W_{ji}$  is the element of  $W$  in the  $j^{\text{th}}$  row,  $i^{\text{th}}$  column.

$W_{ji}$  goes in the same place in both matrices! They're the same matrix:  $W = \frac{\partial Z}{\partial A}$

We get our final result:

If two matrices have exactly the same shape and elements, they're the same matrix.

**Notation 824**

Our derivative

$$\underbrace{\frac{\partial \mathbf{Z}^\ell}{\partial \mathbf{A}^{\ell-1}}}_{(m^\ell \times n^\ell)} = \mathbf{W}^\ell$$

Is a vector/vector derivative, and thus a matrix.

It takes the shape  $(m^\ell \times n^\ell)$ .

- This matches the intuition we have from the simplified, 1d version, where we have  $z = w a$ , instead of  $Z = W^T A$ .

**A.9.4 Activation Function**

$$\frac{\partial \mathbf{A}^\ell}{\partial \mathbf{Z}^\ell} \quad (\text{A.84})$$

The last derivative is less strange than its solution looks.

$$\mathbf{A}^\ell = f(\mathbf{Z}^\ell) \longrightarrow \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = f\left( \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} \right) \quad (\text{A.85})$$

We can apply our function element-wise:

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} f(z_1) \\ f(z_2) \\ \vdots \\ f(z_n) \end{bmatrix} \quad (\text{A.86})$$

As we can see, each activation is a function of only **one** pre-activation.

**Concept 825**

Each **activation** is only affected by the **pre-activation** in the **same neuron**.

So, if the **neurons** don't match, then our derivative is zero:

- $i$  is the neuron for pre-activation  $z_i$
- $j$  is the neuron for activation  $a_j$

$$\frac{\partial a_j}{\partial z_i} = 0 \quad \text{if } i \neq j$$

So, our only nonzero derivatives are

$$\frac{\partial a_i}{\partial z_i}$$

As for our **non-zero** terms, they all rely on the equation:

$$a_i = f(z_i) \tag{A.87}$$

Our derivative is:

$$\frac{\partial a_i}{\partial z_i} = f'(z_i) \tag{A.88}$$

In general, including the non-diagonals:

$$\frac{\partial a_i}{\partial z_j} = \begin{cases} f'(z_i) & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{A.89}$$

This gives us our result:

**Notation 826**

Our derivative

$$\underbrace{\frac{\partial \mathbf{A}^\ell}{\partial \mathbf{Z}^\ell}}_{(n^\ell \times n^\ell)} = \left[ \begin{array}{ccccc} f'(z_1^\ell) & 0 & 0 & \cdots & 0 \\ 0 & f'(z_2^\ell) & 0 & \cdots & 0 \\ 0 & 0 & f'(z_3^\ell) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & f'(z_n^\ell) \end{array} \right] \quad \begin{array}{l} \text{Column } j \text{ matches } a_j \\ \text{Row } i \text{ matches } z_i \end{array}$$

Is a vector/vector derivative, and thus a matrix.

It takes the shape  $(n^\ell \times n^\ell)$ .

**A.9.5 Element-wise multiplication**

Notice that, in the previous section, we could've compressed this matrix down to remove the unnecessary 0's:

$$\begin{bmatrix} f'(z_1^\ell) \\ f'(z_2^\ell) \\ \vdots \\ f'(z_n^\ell) \end{bmatrix} \quad (\text{A.90})$$

This is a valid way to interpret this matrix! The only thing we need to be careful of: if we were to use this in a chain rule, we couldn't do normal matrix multiplication.

However, because of how this matrix works, you can just do **element-wise** multiplication instead!

You can check it for yourself: each index is separately scaled.

**Concept 827**

When multiplying two vectors R and Q, if they take the form

$$R = \begin{bmatrix} r_1 & 0 & 0 & \cdots & 0 \\ 0 & r_2 & 0 & \cdots & 0 \\ 0 & 0 & r_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & r_n \end{bmatrix} \quad Q = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ \vdots \\ q_n \end{bmatrix}$$

Then we can write their product each of these ways:

$$RQ = \underbrace{R * Q}_{\text{Element-wise multiplication}} = \begin{bmatrix} r_1 q_1 \\ r_2 q_2 \\ r_3 q_3 \\ \vdots \\ r_n q_n \end{bmatrix}$$

So, we can substitute the chain rule this way.

## A.10 Terms

- Matrix/scalar derivative
- Scalar/Matrix derivative
- Axis
- Dimension (vector)
- Dimension (array)
- Array
- "Tensor" (Generalized matrix)
- $c$ -Tensor (2-tensor, 3-tensor, etc.)

# APPENDIX B

## Optimizing Neural Networks

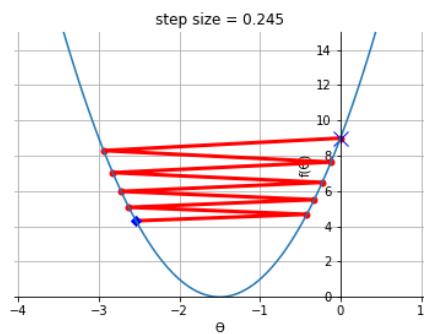
These notes carry over from the [Neural Networks](#) chapter. Here, we include topics that we previously skimmed over.

### B.1 Strategies towards adaptive step-size

#### B.1.1 Momentum

##### B.1.1.1 Solving Oscillation

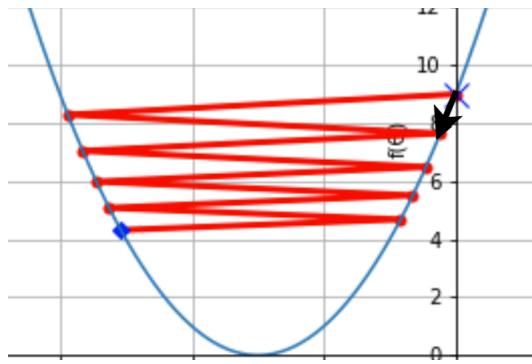
Let's look at one common problem we have with gradient descent: **oscillation**.



We overshoot our target, and then have to take another step that **undoes** most of what happened in the previous step. So, we waste a lot of time correcting the last step.

This can significantly **slow** down how quickly we converge.

For example, our first two steps land us in almost the same place we started!



The black arrow shows the combined effect of our first two steps: almost nothing!

We don't want to waste time, so we want to remove the "part" of the gradient that is likely to **cancel** out.

The **next** gradient cancels out some of the previous. Our first two steps add up, or "**average out**" to a small improvement.

If our steps are effectively "averaging", we'll speed up that process: we'll average together the gradients *before* taking our step!

This means we can take a bigger step in a direction we won't have to cancel!

### Concept 828

Since our gradient descent steps **combine** to give us our new model, we can think of them as adding, or "**averaging**" to a more accurate improvement.

When our function **oscillates**, we get the same pattern **multiple** times: past steps indicate the sort of pattern we'll see in **future** steps.

The only real difference between adding and averaging is whether we divide by the number of terms.

So, we'll average our current gradient with past gradients: that way, the **component** that gets cancelled out is **removed**, and we won't have to undo our mistake over and over.

### Concept 829

**Oscillation** causes us to move back and forth over the same region **multiple** times, where each step mostly **cancels** out the last.

One solution is to **average out** multiple of our gradients: the part that is "**cancelled out**" should be eliminated by the average.

So, we average our **past** gradients (past oscillation) with our **current** gradient, so we move in a more **efficient** direction, speeding up our algorithm.

Another way to think about it: when we **average** out our current and previous gradient, we're cancelling out what they "**disagree**" on, and keeping what they **agree** on.

So, we're taking a step in the direction that multiple gradients agree will **improve** our model!

### B.1.1.2 Weighted averages

We could naively average all of our gradients **equally**. But, this would be a bad idea:

- It doesn't give you as much control of the algorithm: what if we care more about the **present** gradient, than the previous one?
- Gradients further in the **past** are **less likely** to matter: we've moved further away from those positions.
  - We also need to **scale** down past terms, so they don't take up most of the average.

The first problem is easy to solve: we'll **weigh** each of our terms differently.

If you're averaging 100 terms, and you add one more... it's not going to change much.

#### Concept 830

A **weighted average** is used when we want some terms to affect our **average** more than others.

We represent this with **weights**: each weight represents the **proportion** of our average from that term.

$$\text{Weighted Average} = x_1 w_1 + x_2 w_2 + \dots + x_n w_n$$

**Example:** If  $w_1 = .6$ , that means **60%** of the average comes from  $x_1$ .

Note that, since we're talking about **proportions**, they need to **sum to 1**: it wouldn't make sense to have more than 100% of the average.

At each time step, we're adding one new gradient: the **present** one.

We'll simplify our average to those two terms: the **present** gradient, versus all the **past** gradients.

These weights are **separate** from the weights inside our neural network.

They do, however, represent the same type of concept: the NN weights scale the **input**, while these weights scale the **gradients**.

- We represent the importance (**weight**) of our **past** gradients using the variable  $\gamma$ .
- We want the two terms to add to 1: so, the importance of **current** gradient is  $1 - \gamma$ .

$$\overbrace{A_t}^{\text{Average}} = \underbrace{\gamma G_{t-1}}_{\text{Old gradients}} + \underbrace{(1-\gamma)g_t}_{\text{New gradient}} \quad (\text{B.1})$$

Now, we have **control** over how much the present or past gradient matters: we just have to adjust  $\gamma$ .

### B.1.1.3 Running Average

We still have some work to do: first, we haven't made it clear how we're incorporating our old gradients: we lumped them into one term.

Let's try building up from  $t = 1$ . We'll assume our previous gradients are 0, for simplicity.

$$A_0 = g_0 = 0 \quad (\text{B.2})$$

Our first step will average this with our **first** gradient:

$$A_1 = \gamma g_0 + (1 - \gamma)g_1 \quad (\text{B.3})$$

Simplifying to:

$$A_1 = (1 - \gamma)g_1 \quad (\text{B.4})$$

What about our second step?

$$A_2 = \underbrace{\gamma G_{t-1}}_{\text{Old gradients}} + (1 - \gamma)g_2 \quad (\text{B.5})$$

We *could* just plug in  $g_1$ . But,  $A_1$  contains the information about our first gradient  $g_1$ , **and** the gradient before it,  $g_0$ .

$$A_2 = \underbrace{\gamma A_1}_{\text{Contains } g_1, g_0} + (1 - \gamma)g_2 \quad (\text{B.6})$$

We can repeat this process:

$$A_3 = \underbrace{\gamma A_2}_{\text{Contains } g_2, g_1, g_0} + \underbrace{(1 - \gamma)g_3}_{\text{New gradient}} \quad (\text{B.7})$$

And so, we've created a general way to **average** as our program **runs** through different gradients.

$$A_t = \gamma A_{t-1} + (1 - \gamma)g_t \quad (\text{B.8})$$

To allow more flexibility, we'll allow  $\gamma$  to **vary in time**, as  $\gamma_t$ .

$$A_t = \gamma_t A_{t-1} + (1 - \gamma_t)g_t \quad (\text{B.9})$$

- We call this a **running average**.

**Definition 831**

A **running average** is a way to average past data with present data **smoothly**.

Our **initial** value for the average is typically zero:

$$A_0 = 0$$

Then, we begin introducing **new** data points.

- You use the parameter  $\gamma_t$  to indicate how much you want to prioritize **past data**.
- Thus,  $1 - \gamma_t$  indicates the value of **new data**.

$$\overbrace{A_t}^{\text{Average}} = \underbrace{\gamma_t A_{t-1}}_{\text{Old gradients}} + \overbrace{(1 - \gamma_t) g_t}^{\text{New gradient}}$$

- Note that instead of  $\gamma$ , we write  $\gamma_t$ : this "discount factor" can vary with time.

**Clarification 832**

This is technically only one kind of **running average**: here, we use an "**exponential moving average**".

There are different ways to average past data points, with different **weighting** schemes.

- For example, you could do a "**simple moving average**", where you average equally over the last  $n$  data points.

**B.1.1.4 Running Averages: The Distant Past**

So, how does this "running average" approach affect our different data points, further in the past? Let's find out.

For simplicity, let's assume  $\gamma_t = \gamma$ : it's a **constant**.

$$A_t = \gamma A_{t-1} + (1 - \gamma) g_t \quad (\text{B.10})$$

We can expand  $A_{t-1}$  to see further in the past:

$$A_t = \gamma(\gamma A_{t-2} + (1-\gamma)g_{t-1}) + (1-\gamma)g_t \quad (\text{B.11})$$

And even further:

$$A_t = \gamma\left(\gamma(\gamma A_{t-3} + (1-\gamma)g_{t-2}) + (1-\gamma)g_{t-1}\right) + (1-\gamma)g_t \quad (\text{B.12})$$

This is starting to get messy: don't worry if it's hard to read.

Let's rewrite this.

$$\gamma^3 A_{t-3} + \gamma^2(1-\gamma)g_{t-2} + \gamma(1-\gamma)g_{t-1} + (1-\gamma)g_t \quad (\text{B.13})$$

We see a "stacking" effect for  $\gamma$ :

- We only partly include our newest data point: we scale it by  $1 - \gamma$ , to make room for the past.

$$(1-\gamma)g_t \quad (\text{B.14})$$

- But if your gradient is 1 time unit in the past, we apply  $\gamma$  **once**, "forgetting" some more of that gradient.

$$\gamma(1-\gamma)g_{t-1} \quad (\text{B.15})$$

- But if your gradient is 2 units in the past, we apply  $\gamma$  **twice**: we've "forgotten" some of it twice.

$$\gamma^2(1-\gamma)g_{t-2} \quad (\text{B.16})$$

Each time we do an average, we scale down our older data points by  $\gamma$ . So, the further in the past, the less effect they have.

- This is exactly the kind of design we wanted!

**Concept 833**

A **running average** tends to pay less attention to data further in the **past**.

- In general, if you are  $k$  time units in the past, we apply a factor of  $\gamma^k$ .

Because  $\gamma < 1$ , this **exponentially** decays to 0.

If we want to fully expand  $A_t$ , it's easiest to use a sum:

$$A_T = \sum_{t=0}^T \gamma^{(T-t)} \cdot (1 - \gamma) \cdot g_t$$

You can compare this formulation against what we computed above.

**B.1.1.5 Momentum**

Applying a running average to our gradients gives us **momentum**.

- This analogy to **physics** represents how our point "moves" through the **weight space**, to optimize  $J$ .
- The gradient gives us a "direction" of motion. So, our **momentum** represents the direction we were "already moving": the **previous** averaged gradient.

We use  $M$  to represent the "averaged gradient" that we use to move. Our initial momentum is 0:

$$M_0 = 0 \quad (\text{B.17})$$

Gradient descent adjusts our weights: we go from one list of weights, to another. That's why we say we're moving through the "weight space".

Because our hypothesis is determined by our weights, this is also a "hypothesis space".

And we want to average our current gradient with past gradients:

$$M_t = \gamma M_{t-1} + (1 - \gamma) g_t \quad (\text{B.18})$$

What is our **past** gradient  $g_t$ ? Well, we want to use  $W$  to modify  $J$ :  $\frac{\partial J}{\partial W}$ , or  $\nabla_W J$ .

And we're moving through the **weight space**, so our input to the gradient is the previous set of weights,  $W_{t-1}$ .

$$g_t = \nabla_W J(W_{t-1}) \quad (\text{B.19})$$

And finally,  $M_t$ , our "averaged gradient", determines how we move.

$$W_t = W_{t-1} - \eta M_t \quad (\text{B.20})$$

**Definition 834**

**Momentum** is a technique for gradient descent where we do a **running average** between our current gradient, and our older gradients.

This approach reduces **oscillation**, and thus aims to improve the speed of convergence for our models.

- Our initial "momentum" (averaged gradient) is **0**.
- The amount we value new data is given by  $\gamma_t$ .
- Old data is scaled by  $1 - \gamma_t$ .

$$M_0 = 0$$

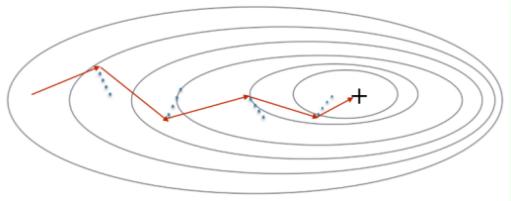
$$M_t = \underbrace{\gamma_t M_{t-1}}_{\text{Old gradients}} + \underbrace{(1 - \gamma_t) \nabla_W J(W_{t-1})}_{\text{New gradient}}$$

We use this momentum to take our step:

$$W_t = W_{t-1} - \eta M_t$$

This approach puts more emphasis on newer data points, and less on older ones.

**Example:** We can see this "damped oscillation" in action below:



The blue lines indicate the more "severe" path that we would've taken with normal gradient descent. the orange line is the line we take with momentum.

Notice that generally, the orange path is closer to correct! We still oscillate somewhat, but much less than we would have otherwise.

## B.1.2 Adadelta

### B.1.2.1 Per-weight step sizes

Earlier, we discussed creating **separate** step sizes for different weights in our network:

- Different weights could have different **sensitivities**: one weight could have a much larger/smaller impact on  $J$ , based on **structure**.
- We already have the challenge of figuring out what **step sizes** cause **slow convergence/divergence**: it would be even harder to find a step size that fit **all** of our weights, at the same time.

So, instead, we decided that each weight gets its **own step size**.

- But, we never elaborated on **how** we compute that (adjustable) step size.

#### Notation 835

Our base step size is indicated as  $\eta$ .

We'll call the **modified** step size for weight  $j$ ,  $\eta_j$ .

### B.1.2.2 Scaling

Our goal now, is to come up with a **systematic** way to **assign step sizes**. This allows us to **adjust**, rather than run our whole gradient descent with a "bad" step size.

Why do we adjust our step size? To avoid slow convergence, oscillation, or divergence.

So, we have less of a problem if we choose  $\eta$  poorly.

- We might expect **slow convergence** if the derivative is too **small**: we carefully take small steps, but we aren't having much of an impact on  $J$ .
  - Across this "flatter" region of the surface, in one direction, we might expect it to be safer to move further.
- We might expect **divergence** or **oscillation** if the derivative is too **large**.
  - We might end up "missing" possible solutions/local minima by **overshooting** them.
  - So, "steeper" regions might be riskier.

**Concept 836**

Our goal is to **scale** our step size, so that it **adapts** to the situation:

- We want to **shrink** our step for **large** gradients
- We want to **increase** our step for **small** gradients

And we're interested in the **magnitude** of these gradients.

This sort of behavior is easily captured by including a factor of  $1/\|g_t\|$ . However, this has a smoothness problem: so, we'll use  $1/\|g_t\|^2$  instead.

Though, we need to keep it separate for each of our data points.

**Notation 837**

The gradient for **weight j** at **time t** is given as

$$g_{t,j} = \nabla_{W_j} J(W_{t-1})_j$$

Note the double-subscript.

- By isolating weight  $j$ , we have a constant, not a vector.

We can now write our gradient update rule:

$$W_{t,j} = W_{t-1,j} - \eta_j \cdot g_{t,j} \quad (\text{B.21})$$

And we're currently using the step size

Don't save this equation! It isn't our final formula.

$$\eta_j = \frac{\eta}{g_{t,j}^2} \quad (\text{B.22})$$

**B.1.2.3 Averaging**

But, we don't necessarily know how "steep" our region **generally** is, based on the current gradient  $g_t$ .  $g_t$  only gives us **one point** in space.

It would be helpful to include information from the **past**: we'll be re-using the **weighted average**, once again.

$$G_t = \gamma G_{t-1} + (1-\gamma) g_t^2 \quad (\text{Maybe?})$$

Once again, it's helpful that the weighted average gradually "forgets" older information: we care less about gradients which are "further" from the present.

- Note that we're averaging the **squared** magnitude.

Technically this is still **incorrect**: we need the  $j$  notation.

$$G_{t,j} = \gamma G_{t-1,j} + (1 - \gamma) g_{t,j}^2 \quad (\text{Fixed!})$$

It's incorrect because  $g_t$  is a vector: we can't square it directly, we have to square its magnitude.

#### B.1.2.4 Division by zero

There's a problem with our weight adjustment:

$$\eta_j = \frac{\eta}{G_{t,j}} \quad (\text{B.23})$$

What happens if the denominator is near zero? It'll explode to a huge number! And at zero, it's undefined.

To solve this, we'll add a very small constant,  $\epsilon$ .

We won't prescribe any particular choice of  $\epsilon$  here.

$$\eta_j = \frac{\eta}{G_{t,j} + \epsilon} \quad (\text{B.24})$$

Now, our scaling factor will never be bigger than  $1/\epsilon$ .

#### B.1.2.5 Square root

One last concern, to wrap up: currently, we're diving by the **squared** gradient. This is actually somewhat overkill.

Remember that our goal is to do the following operation:

$$W_{t,j} = W_{t-1,j} - \underbrace{\eta_j \cdot g_{t,j}}_{\text{Update}} \quad (\text{B.25})$$

With our current formula, our update has

- a factor of  $g_{t,j}$  in the **numerator**
- from  $\eta_j$ , a factor proportional to  $g_{t,j}^2$  in the **denominator**

This is "scaling" our gradient by more than we want to. So, we'll take the **square root** of the denominator.

$$\eta_j = \frac{\eta}{\sqrt{G_{t,j} + \epsilon}} \quad (\text{B.26})$$

Our goal is to make the scales of different axes more similar, not to neglect dimensions with high gradient (high effect on loss)

This is the completed form of our **adadelta step size rule!**

### Definition 838

**Adadelta** is a technique for **adaptive step size**, which:

- **Decreases** step size in dimensions with a history of **high-magnitude** gradients
- **Increases** step size in dimensions with a history of **low-magnitude** gradients

Suppose our gradient for weight  $W_j$  at time  $t$  is represented by

$$g_{t,j} = \nabla_W J(W_{t-1})_j$$

This is accomplished by **scaling** the step size  $\eta$  to create  $\eta_j$ :

$$\eta_j = \frac{\eta}{\sqrt{G_{t,j} + \epsilon}}$$

Where  $G_{t,j}$  is a "**running average**" of the previous gradients for weight  $W_j$ .

$$\begin{aligned} G_{0,j} &= 0 \\ G_{t,j} &= \gamma G_{t-1,j} + (1 - \gamma) g_{t,j}^2 \end{aligned}$$

So, our completed gradient descent rule takes the form:

$$W_{t,j} = W_{t-1,j} - \frac{\eta}{\sqrt{G_{t,j} + \epsilon}} \cdot g_{t,j}$$

### B.1.2.6 Sparse Data

One major advantage of adadelta is its use for **sparse datasets**, where many variables only show up in a small percentage of the data.

- If a variable is much less frequent, then the weighted average  $G_t$  will be much smaller.
- So, when those data points **do** appear, the step size is much larger.

This allows our model to learn more from variables that don't show up as frequently.

**Concept 839**

**Adadelta** often works well with **sparse data**: datasets where many variables rarely show up as non-zero.

The step sizes for these variables become much larger, so our model can learn more from less-common information.

However, this can run the risk of paying attention to variables that are sparse, but **not** especially meaningful.

- It's important to choose your variables carefully, so the model doesn't "learn" from noise.

### B.1.3 Adagrad

Originally, adadelta derives from a **simpler** method, known as **adagrad**, or "adaptive gradient".

- The main difference with this approach is, rather than find the **weighted average**  $G_t$ , we simply **sum** the previous gradients.

The main problem with this approach is that, over time,  $G_t$  becomes too **large**. So, our step size becomes too small, and our algorithm slows down.

By averaging, we don't run into this problem of  $G_t$  "accumulating": older data is gradually **forgotten**.

### B.1.4 Adam

Momentum and Adadelta both bring some benefits:

- **Momentum** averages our current gradient with **previous gradients**, to reduce **oscillation** and make a more direct path to the solution.
- **Adadelta** modifies our **step sizes**: it takes smaller steps in directions of high gradient (reduce overshooting) and takes bigger steps in directions of low gradient (converge faster).

There's nothing structurally incompatible between them. So, why not incorporate both?

- This combination is called **Adam**: it has become the most popular way to handle step sizes in neural networks.

#### Concept 840

**Adam** integrates the techniques of both **momentum** and **adadelta**.

#### B.1.4.1 Momentum and Adadelta

We used two different **running averages**, both using  $\gamma$  for their "**discount factor**": how much we discount the effect of older data.

- We'll replace  $\gamma$  with  $B_1$  (momentum) and  $B_2$  (adadelta).

It's "discounting", or reducing the effect of older data, because  $\gamma < 1$ .

#### Notation 841

In the adam algorithm,  $B_1$  and  $B_2$  are **discount factors** replacing  $\gamma$  from momentum and adadelta.

We use the same notation for gradients:

$$g_{t,j} = \nabla_W J(W_{t-1})_j \quad (\text{B.27})$$

First, we'll keep track of our **averaged gradient**,  $m_{t,j}$ . This will be the **direction** we plan to move.

$$m_{t,j} = B_1 m_{t-1,j} + (1 - B_1) g_{t,j} \quad (\text{B.28})$$

Then, we'll keep track of the **average squared gradient**,  $v_{t,j}$ . This will tell us whether to scale up/down our **step size** on each variable.

$$v_{t,j} = B_2 v_{t-1,j} + (1 - B_2) g_{t,j}^2 \quad (\text{B.29})$$

We can combine these into our equation, the way you use momentum and adadelta:

$$W_{t,j} = W_{t-1,j} - \frac{\eta}{\sqrt{v_{t,j} + \epsilon}} \cdot m_{t,j} \quad (\text{B.30})$$

We're not quite done yet, though.

#### B.1.4.2 Normalization

There's one issue with our running average, that we should address.

Suppose we have a sequence of numbers:

$$[1, 1, 1] \quad (\text{B.31})$$

What's the running average of this sequence of numbers, with  $\gamma = .1$ ? You'd expect it to be 1 for all elements, right?

- We start with  $a_0 = 0$ .

$$\begin{aligned} a_1 &= .1 * 0 + .9 * 1 = .9 \\ a_2 &= .1 * .9 + .9 * 1 = .99 \\ a_3 &= .1 * .99 + .9 * 1 = .999 \end{aligned} \quad (\text{B.32})$$

It turns out not to be true! Why is that?

Because of our initial term,  $a_0$ : it has nothing to do with the data. Because it'll always be 0, it **deflates** the value of all the remaining averages.

How much does it deflate the average? Well, let's consider the general equation:

$$A_T = \sum_{t=0}^T \gamma^{(T-t)} (1 - \gamma)^t g_t \quad (\text{B.33})$$

What "fraction" of the total is missing, thanks to  $A_0 = 0$ ?

- Well, all of our weighted terms  $\gamma^{(T-t)} (1 - \gamma)^t$  should add up to 1: each one represents the "percent/fraction" of the average which comes from that  $g_t$  term.

$A_0$  matches  $t = 0$ , so we find:

$$A_T = \underbrace{A_0 \gamma^T}_{A_0=0} + \sum_{t=1}^T \gamma^{(T-t)} (1-\gamma)^t g_t \quad (B.34)$$

So, out of our total 1, or 100%, we're missing  $\gamma^T$ .

- Our new total is  $1 - \gamma^T$ .
- In order to correct for the "zeroed out" part of our average, we multiply by

$$\frac{\text{Desired total}}{\text{Real total}} = \frac{1}{1 - \gamma^T} \quad (B.35)$$

### Concept 842

We've chosen  $A_0 = 0$ : this is **unrelated** to our real data.

As a result,  $A_t$  doesn't accurately reflect our weighted average: it's slightly **smaller**.

- $A_0$  is "included" as a 0, even though it's not a **real** data point.
- So, whatever **percent** of our data is represented by  $A_0$ , is "falsely empty".

$A_0$  is scaled by  $\gamma^t$ , so that's the amount **removed** from our "true" weighted average.

The "real" weighted average, that ignores  $A_0$ , has to un-do the deflation caused by including fake, starting data:

$$\hat{A}_t = A_t \cdot \frac{1}{1 - \gamma^t}$$

We'll make this correction for our running averages:

$$\begin{aligned} \hat{m}_{t,j} &= \frac{m_{t,j}}{1 - B_1^t} \\ \hat{v}_{t,j} &= \frac{v_{t,j}}{1 - B_2^t} \end{aligned} \quad (B.36)$$

#### B.1.4.3 Adam

Now, we have our complete equation for adam:

**Definition 843**

**Adam** is a technique for improving **gradient descent** that integrates two other techniques. Our computed gradient is given as

$$g_{t,j} = \nabla_W J(W_{t-1})_j$$

- **Momentum** averages our previous gradients with our current one, to reduce oscillation and get a "averaged" view of data.  $B_1$  gives the weight of old data.

$$m_{0,j} = 0 \quad m_{t,j} = B_1 m_{t-1,j} + (1 - B_1) g_{t,j}$$

- **Adadelta** estimates how "steep" our gradient has been recently, and uses it to adjust our step size, for better convergence.  $B_2$  gives the weight of old data.

$$v_{0,j} = 0 \quad v_{t,j} = B_2 m_{t-1,j} + (1 - B_2) g_{t,j}^2$$

We include a **correction** factor for the zeroed initial conditions:  $m_0 = v_0 = 0$ .

$$\hat{m}_{t,j} = \frac{m_{t,j}}{1 - B_1^t} \quad \hat{v}_{t,j} = \frac{v_{t,j}}{1 - B_2^t}$$

Finally, our update rule takes the form:

$$W_{t,j} = W_{t-1,j} - \frac{\eta}{\sqrt{\hat{v}_{t,j} + \epsilon}} \cdot \hat{m}_{t,j}$$

Impressively, adam is not very sensitive to the initial conditions for  $\epsilon$ ,  $B_1$  and  $B_2$ .

To implement this for NNs, we just need to keep track of each value, for each weight. If you do it systematically, this is easier than it sounds.

It's able to "self-correct" its step sizes and gradient based on data.

## B.2 Batch Normalization Details

Let's review the general premise

of batch normalization:

### Definition 844

**Batch Normalization** is a process where we

- Standardize the pre-activation for each layer using the mean  $\mu_i$  and standard deviation  $\sigma_i$  (for the  $i^{\text{th}}$  dimension). Select  $\epsilon: 0 < \epsilon \ll 1$ .

$$\bar{Z}_{ij} = \frac{Z_{ij} - \mu_i}{\sigma_i + \epsilon}$$

- Choose the new mean and standard deviation for the pre-activation using  $(n \times 1)$  vectors  $G$  and  $B$

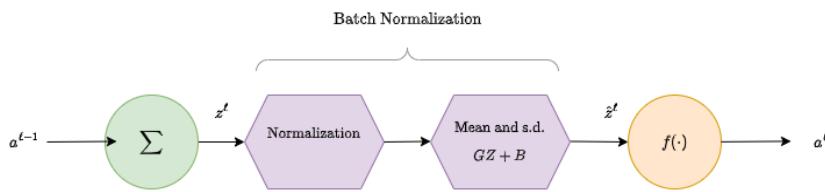
$$\hat{Z}_{ik} = G_i * \bar{Z}_{ij} + B_i$$

### B.2.1 Applying batch normalization to backprop

#### Remark (Optional)

The following section mostly deals with the details of computing batch normalization derivatives.

As we showed with our figure,



Batch normalization adds two units that **interrupt** our chain of nested functions.

That means we need to figure out how to do backprop, bridging across the new gap between  $Z$  and  $\hat{Z}$ .

So, that only leaves a couple problems:

- "Bridging the gap" between derivatives before and after BN, with  $\partial \hat{Z}^t / \partial Z^t$ .
- Gradients for  $G$  and  $B$ : they're parameters now, too, so we need to train them.

**Concept 845**

Introducing batch normalization add new functions **in between** our old ones.

- Our input data has to travel through those additional layers.
- This **changes** the relationship between our current value, and the output loss.

So, in order to do backprop correctly, we have to figure out the **derivatives** of those functions.

**B.2.1.1 Bridging the gap**

We want to connect the start and end of batch normalization:

$$\frac{\partial \hat{Z}}{\partial Z} \quad (\text{B.37})$$

As usual, with the chain rule, we'll connect them by considering any values/function **between** them.

In this case,  $Z$  is normalized ( $\bar{Z}$ ), and **then** we apply  $G$  and  $B$  ( $\hat{Z}$ ). We're missing the "normalized" step.

$$\frac{\partial \hat{Z}}{\partial Z} = \frac{\partial \hat{Z}}{\partial \bar{Z}} \cdot \frac{\partial \bar{Z}}{\partial Z} \quad (\text{B.38})$$

Let's compare any two of these :  $\hat{Z}$  and  $\bar{Z}$ , for example.

- These are both  $(n \times k)$  matrices.
- That means that each has two dimensions of variables. If we were to take the derivative between them, we would need  $2 * 2$  axes: a **4-tensor**.

That sounds terrible. Instead, we'll compute these derivatives **element-wise**.

$$\frac{\partial \hat{Z}_{ab}}{\partial Z_{ef}} = \frac{\partial \hat{Z}_{ab}}{\partial \bar{Z}_{cd}} \cdot \frac{\partial \bar{Z}_{cd}}{\partial Z_{ef}} \quad (\text{B.39})$$


---

**B.2.1.2 Indexing**

First, let's simplify these indices: only some pairs of elements matter.

$$\hat{Z}_{ik} = G_i * \bar{Z}_{ik} + B_i \quad (\text{B.40})$$

It seems that  $\hat{Z}_{ik}$  is only affected by the element with the same indices:  $\bar{Z}_{ik}$ .

**Concept 846**

$\hat{Z}_{ik}$  is only a function of the same index element  $\bar{Z}_{ik}$ .

- Any other elements from  $\bar{Z}$  has no effect.

$$(a \neq c \text{ or } b \neq d) \implies \frac{\partial \hat{Z}_{ab}}{\partial \bar{Z}_{cd}} = 0$$

So, we write our remaining derivatives as  $\partial \hat{Z}_{ik} / \partial \bar{Z}_{ik}$ .

What about the other derivative?

$$\bar{Z}_{ik} = \frac{Z_{ik} - \mu_i}{\sigma_i + \epsilon}$$

$\mu_i$  and  $\sigma_i$  include various different data points  $Z_{ik}$ , but only the  $i^{\text{th}}$  dimension.  $Z_{ik}$  requires the exact same indices.

**Concept 847**

$\bar{Z}_{ik}$  is only a function of elements in the same dimension  $i$ ,  $Z_{ij}$ .

$$c \neq e \implies \frac{\partial \bar{Z}_{cd}}{\partial Z_{ef}} = 0$$

Our remaining derivatives take the form  $\partial \bar{Z}_{ik} / \partial Z_{ij}$ .

If we boil this down, we get all of our non-zero derivatives:

$$\frac{\partial \hat{Z}_{ik}}{\partial Z_{ij}} = \frac{\partial \hat{Z}_{ik}}{\partial \bar{Z}_{ik}} \cdot \frac{\partial \bar{Z}_{ik}}{\partial Z_{ij}} \quad (\text{B.41})$$

### B.2.1.3 Computing $\partial \hat{Z}_{ik} / \partial \bar{Z}_{ik}$

We return to our previous equation:

$$\hat{Z}_{ik} = G_i * \bar{Z}_{ik} + B_i \implies \frac{\partial \hat{Z}_{ik}}{\partial \bar{Z}_{ik}} = G_i \quad (\text{B.42})$$

### B.2.1.4 Computing $\partial \bar{Z}_{ik} / \partial Z_{ij}$

For the other derivative:

$$\bar{Z}_{ik} = \frac{Z_{ik} - \mu_i}{\sigma_i + \epsilon} \quad (\text{B.43})$$

This gets a bit complicated, because  $Z_{ij}$  can affect three different terms:  $Z_{ik}$ ,  $\mu_i$ , and  $\sigma_i$ .

We'll solve this by using the multi-variable chain rule.

$$\frac{\partial \bar{Z}_{ik}}{\partial Z_{ij}} = \underbrace{\frac{\partial \bar{Z}_{ik}}{\partial Z_{ik}} \cdot \frac{dZ_{ik}}{dZ_{ij}}}_{Z_{ik}'s \text{ effect}} + \underbrace{\frac{\partial \bar{Z}_{ik}}{\partial \mu_i} \cdot \frac{d\mu_i}{dZ_{ij}}}_{\mu_i's \text{ effect}} + \underbrace{\frac{\partial \bar{Z}_{ik}}{\partial \sigma_i} \cdot \frac{d\sigma_i}{dZ_{ij}}}_{\sigma_i's \text{ effect}} \quad (\text{B.44})$$

We're linearly adding the effect due to each of these three variables, separately.

In each of these terms, we treat the other two variables as "constant".

### B.2.1.5 Lots of mini-derivatives

Let's compute the  $\bar{Z}$  derivatives.

$$\bar{Z}_{ik} = \frac{Z_{ik} - \mu_i}{\sigma_i + \epsilon} \quad (\text{B.45})$$

gives us

$$\frac{\partial \bar{Z}_{ik}}{\partial Z_{ik}} = \frac{1}{\sigma_i + \epsilon} \quad \frac{\partial \bar{Z}_{ik}}{\partial \mu_i} = \frac{-1}{\sigma_i + \epsilon} \quad \frac{\partial \bar{Z}_{ik}}{\partial \sigma_i} = -\left(\frac{Z_{ik} - \mu_i}{(\sigma_i + \epsilon)^2}\right) \quad (\text{B.46})$$

Now, let's compute the  $Z_{ij}$  derivatives.

$$\boxed{\frac{\partial Z_{ik}}{\partial Z_{ij}} = \delta_{jk}} = \mathbf{1}(j = k) = \begin{cases} 1 & j = k \\ 0 & j \neq k \end{cases} \quad (\text{B.47})$$

$$\mu_i = \frac{1}{K} \sum_{j=1}^K Z_{ij} \implies \boxed{\frac{\partial \mu_i}{\partial Z_{ij}} = \frac{1}{K}} \quad (\text{B.48})$$

$$\sigma_i^2 = \frac{1}{K} \sum_{j=1}^K (Z_{ij} - \mu_i)^2 \implies \boxed{\frac{\partial \sigma_i^2}{\partial Z_{ij}} = \frac{Z_{ij} - \mu_i}{K\sigma_i}} \quad (\text{B.49})$$

### B.2.1.6 Assembling our derivatives

Now, we plug them in.

$$\frac{\partial \bar{Z}_{ik}}{\partial Z_{ij}} = \frac{\partial \bar{Z}_{ik}}{\partial Z_{ik}} \cdot \frac{dZ_{ik}}{dZ_{ij}} + \frac{\partial \bar{Z}_{ik}}{\partial \mu_i} \cdot \frac{d\mu_i}{dZ_{ij}} + \frac{\partial \bar{Z}_{ik}}{\partial \sigma_i} \cdot \frac{d\sigma_i}{dZ_{ij}} \quad (\text{B.50})$$

First, the  $\bar{Z}$  derivatives.

$$\frac{\partial \bar{Z}_{ik}}{\partial Z_{ij}} = \left( \frac{1}{\sigma_i + \epsilon} \right) \cdot \frac{dZ_{ik}}{dZ_{ij}} - \left( \frac{1}{\sigma_i + \epsilon} \right) \cdot \frac{d\mu_i}{dZ_{ij}} - \left( \frac{Z_{ik} - \mu_i}{(\sigma_i + \epsilon)^2} \right) \cdot \frac{d\sigma_i}{dZ_{ij}} \quad (\text{B.51})$$

And now the  $Z_{ij}$  derivatives.

$$\frac{\partial \bar{Z}_{ik}}{\partial Z_{ij}} = \left( \frac{1}{\sigma_i + \epsilon} \right) \cdot \delta_{jk} - \left( \frac{1}{\sigma_i + \epsilon} \right) \cdot \left( \frac{1}{K} \right) - \left( \frac{Z_{ik} - \mu_i}{(\sigma_i + \epsilon)^2} \right) \cdot \left( \frac{Z_{ij} - \mu_i}{K\sigma_i} \right) \quad (\text{B.52})$$

### Key Equation 848

We have found the batch normalization derivatives

$$\frac{\partial \hat{Z}_{ik}}{\partial \bar{Z}_{ik}} = G_i$$

$$\frac{\partial \bar{Z}_{ik}}{\partial Z_{ij}} = \frac{1}{K(\sigma_i + \epsilon)} \left( K\delta_{jk} - 1 - \frac{(Z_{ik} - \mu_i)(Z_{ij} - \mu_i)}{\sigma_i(\sigma_i + \epsilon)} \right)$$

Which we multiply together to find  $\frac{\partial \hat{Z}_{ik}}{\partial Z_{ij}}$ .

Once we've computed these derivatives, we can use them to extend the chain of  $\partial L / \partial \hat{Z}_{ik}$ .

We're aiming to handle both of the batch normalization functions, with  $\partial L / \partial Z_{ij}$ .

- $Z_{ij}$  affects every data point  $\hat{Z}_{ik}$ , and every data point affects  $L$ .
- So, we'll have to consider every data point  $k$  using the multi-variable chain rule:

$$\frac{\partial L}{\partial Z_{ij}} = \underbrace{\sum_{k=1}^K}_{\text{MV Chain Rule}} \overbrace{\frac{\partial L}{\partial \hat{Z}_{ik}} \cdot \frac{\partial \hat{Z}_{ik}}{\partial Z_{ij}}}^{\text{Data point } k\text{'s effect}} \quad (\text{B.53})$$

We're going from  $\hat{Z}_{ik}$ , which is post-BN, to  $Z_{ij}$ , which is pre-BN.

We can use this to go further back in our layers: as far as we want, as long as we don't forget this derivative!

#### B.2.1.7 Gradients for B and G

We want  $\partial L / \partial B$  and  $\partial L / \partial G$ . Thanks to the work we did just now, we can travel backwards through numerous layers, to reach any  $B^\ell$  and  $G^\ell$ .

So, we'll assume we have  $\partial L / \partial \hat{Z}^\ell$ . Once again, we'll omit the layer notation.

- We'll focus on a single bias,  $B_i$ : this biases one dimension of  $\hat{Z}_{ik}$ .

$$\hat{Z}_{ik} = G_i * \bar{Z}_{ik} + B_i \implies \frac{\partial \hat{Z}_{ik}}{\partial B_i} = 1 \quad (\text{B.54})$$

- This bias is applied to every of our K data points: every data point affects the loss L. So, we'll have to sum them with the multi-variable chain rule.

This is exactly the same as how we did  $\partial L / \partial Z_{ij}$ .

$$\frac{\partial L}{\partial B_i} = \sum_{k=1}^K \frac{\partial L}{\partial \hat{Z}_{ik}} \cdot \frac{\partial \hat{Z}_{ik}}{\partial B_i} = \boxed{\sum_{k=1}^K \frac{\partial L}{\partial \hat{Z}_{ik}}} \quad (\text{B.55})$$

Now, we focus on a single scaling factor  $G_i$ :

$$\hat{Z}_{ik} = G_i * \bar{Z}_{ik} + B_i \implies \frac{\partial \hat{Z}_{ik}}{\partial G_i} = \bar{Z}_{ik} \quad (\text{B.56})$$

And once again, we add across different data points:

$$\frac{\partial L}{\partial G_i} = \sum_{k=1}^K \frac{\partial L}{\partial \hat{Z}_{ik}} \cdot \frac{\partial \hat{Z}_{ik}}{\partial G_i} = \boxed{\sum_{k=1}^K \frac{\partial L}{\partial \hat{Z}_{ik}} \bar{Z}_{ik}} \quad (\text{B.57})$$

We're finished.

### Key Equation 849

Here, we have the gradients for the batch scaling coefficients, B and G.

$$\frac{\partial L}{\partial B_i} = \sum_{k=1}^K \frac{\partial L}{\partial \hat{Z}_{ik}}$$

$$\frac{\partial L}{\partial G_i} = \sum_{k=1}^K \frac{\partial L}{\partial \hat{Z}_{ik}} \bar{Z}_{ik}$$

# APPENDIX C

---

## Recurrent Neural Networks

---

This chapter was originally placed immediately after the Convolutional Neural Networks Chapter.

- Additionally, this chapter relies on having read the State Machines section of the Markov Decision Process (MDP 0).

### C.0.1 Review: Neural Networks So Far

In this class, we design **models** we can **train** to handle different tasks.

All of this has culminated in the **neural network**: a model class that can handle a huge number of interesting problems.

- To create a neural network, we combine many smaller, simpler models together in a systematic, **non-linear** way.
- This creates the "**fully connected**" (FC) neural network.

We then discovered a *weakness* of FC neural networks: they don't understand **space** very well!

- **Example:** FC networks don't encode information about which pixels are **close** or **far** from each other.

Our solution was the **convolutional neural network** (CNN):

Remember that by "systematic", we mostly just mean "organized":

The parts of our model are cleanly organized, to make our math easier.

- We used **convolution** as a way to represent which elements were "near" each other in space.

### C.0.2 Time in a Neural Network

We've created models that can model *space*. We might also wonder: can we make it so they understand **time**, as well?

Right now, our neural networks have no built-in way to *represent* time: each data point stands by itself.

- As we discussed in the CNN chapter, the **order** of our input variables is mostly **ignored** by the model.

#### Concept 850

A traditional, fully connected **neural network** cannot easily use information about **time**, or the **past**.

Note that we're focused on the **finished** model:

The model changes while training, but the fully-trained model **doesn't** keep track of the past.

Previously, we added structure to NNs using **convolution**. But this *doesn't* work as well in time as it does in space:

#### Concept 851

**Time** and **space** behave differently:

- Information can be spread out over **any direction** in space.
- However, information only travels **forward**, not backward, in time.

So, we need to model them differently.

Realistically, we want model that can keep track of time, and order of past events, for plenty of purposes:

- **Example:** Stocks, weather, choosing the best plan of action, etc.
  - "It rained yesterday" and "it rained last week" have very different effects on today's weather.

## C.1 State Machines

This section is identical to the notes from MDP 0. If you have already read that section, skip to C.2.

## C.2 Recurrent Neural Networks

We've fully fleshed out our state machines above: we've developed a technique for managing time, using **memory**.

- This is similar to how, in the last chapter, we developed Convolution, to manage **space**.

Just as we did in convolution, we'll add state machines into our neural network system. This will give us the **Recurrent Neural Network (RNN)**.

### C.2.1 Building up RNNs

How do we create a "network" using our state machine system?

- Well, a traditional NN applies a **linear** operation  $z = W^T x + W_0$ , and then a **non-linear** operation,  $a = f(z)$ .
- We'll implement this process in our state machine, using  $f_s$  and  $f_o$ .

The **linear time-invariant (LTI) system** we defined at the end of the last section gives us the simplest, most reduced version of this:

$$f_s(s_{t-1}, x_t) = A s_{t-1} + B x_t \quad f_o(s_t) = C s_t$$

To replicate a neural network, we'll need to add two things.

#### Concept 852

An RNN modifies an LTI system in two ways:

- Includes **offset** terms in the linear part of  $f_o$  and  $f_s$
- Adds a **nonlinear** component after the linear component both functions.

### C.2.2 Offset

Let's add that offset term:

$$f_s(s_{t-1}, x_t) = As_{t-1} + Bx_t + D \quad f_o(s_t) = Cs_t + E$$

This is a bit cluttered. We'll rename all of these weights:

### Notation 853

For our RNN, we'll label our **weights** as  $W^{ab}$ :

- b represents the "**input**": what we're multiplying  $W^{ab}$  with.
- a represents the "**output**": what we're using  $W^{ab}$  to compute.

We'll also use subscript  $W_0$  for offset terms.

We can go through all of our weights like this.

### Notation 854

We want to rewrite  $f_s(s_{t-1}, x_t) = As_{t-1} + Bx_t + D$

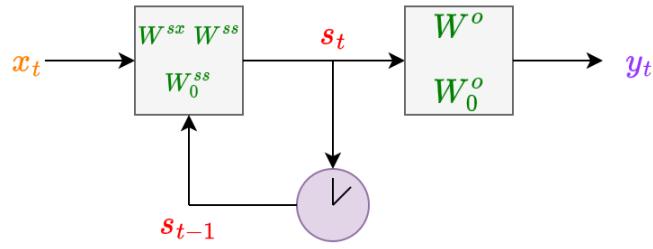
- $W^{ss}$  is combined with  $s_{t-1}$ , to compute  $s_t$ .
- $W^{sx}$  is combined with  $x_t$ , to compute  $s_t$ .
- $W_0^{ss}$  is the offset term that computes  $s_t$ .

$$s_t = W^{ss}s_{t-1} + W^{sx}x_t + W_0^{ss}$$

Now, we'll rewrite  $f_o(s_t) = Cs_t + E$ :

- $W^o$  is combined with  $s_t$  to compute the  $y_t$ .
  - Based on conventions, you could also write it as  $W^{os}$ . But, since there's no  $x$  term, this is unnecessary.
- $W_0^o$  is the offset term that computes  $y_t$ .

$$y_t = W^o s_t + W_0^o$$



### C.2.3 Activation Function

Now, we've covered the linear part of our function. We'll apply a non-linear function  $f$  and  $g$  to each:

$$s_t = f(W^{ss} s_{t-1} + W^{sx} x_t + W_0^{ss}) \quad (C.1)$$

$$y_t = g(W^o s_t + W_0^o) \quad (C.2)$$

These are our **activation functions**.

f and g are pretty vague names, so it would be more helpful to name them according to what they're being used to compute: state and output.

- So, we could say  $f = f_s$ , and  $g = f_o$ .

## Clarification 855

In previous sections, we've used  $f_s$  and  $f_o$  to indicate the **entire function** used to compute the state/output, including the **linear** part.

- However, we want to separate the linear and **non-linear** parts.

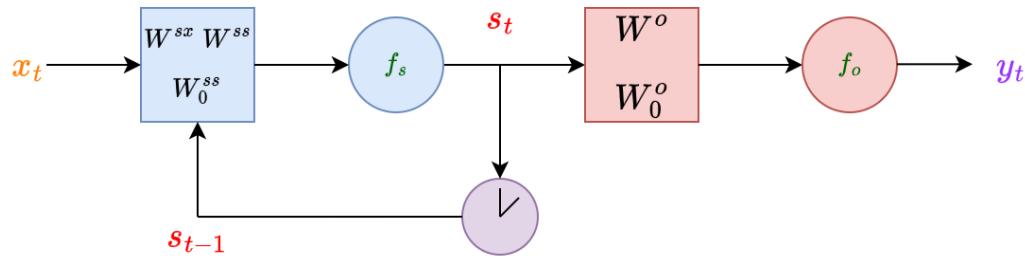
So, we'll switch conventions:  $f_s$  and  $f_o$  in sections 10.1 and 10.2 are **not the same**.

- $f_s$  and  $f_o$  now represent these **activation functions**.

$$s_t = f_s \left( W^{ss} s_{t-1} + W^{sx} x_t + W_0^{ss} \right) \quad (C.3)$$

$$y_t = f_o(W^o s_t + W_0^o) \quad (C.4)$$

We'll insert these units into our diagram:



### Concept 856

Just like in a traditional neural network, we apply our activation functions **element-wise**.

$$f \left( \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \right) = \begin{bmatrix} f(a_1) \\ f(a_2) \\ f(a_3) \end{bmatrix}$$

### C.2.4 Shape

We've the mechanics of our RNN sorted out. To do some quick book-keeping, we'll address the shapes of our various objects.

We'll keep our input, state, and output as vectors:

### Definition 857

We define the dimensions of our input, output, and state as vectors:

$$\begin{array}{lll} \mathcal{X} = \mathbb{R}^\ell & \mathbf{x}_t : (\ell \times 1) \\ \mathcal{S} = \mathbb{R}^m & \Rightarrow \quad \mathbf{s}_t : (m \times 1) \\ \mathcal{Y} = \mathbb{R}^n & \mathbf{y}_t : (n \times 1) \end{array}$$

Based on these vectors, we can derive our weight dimensions:

**Definition 858**

We define the dimensions of our RNN weights:

$$\mathbf{W}^{\text{sx}} : (\mathbf{m} \times \ell)$$

$$\mathbf{W}^{\text{ss}} : (\mathbf{m} \times \mathbf{m})$$

$$\mathbf{W}_0^{\text{ss}} : (\mathbf{m} \times 1)$$

$$\mathbf{W}^{\text{o}} : (\mathbf{n} \times \mathbf{m})$$

$$\mathbf{W}_0^{\text{o}} : (\mathbf{n} \times 1)$$

### C.2.5 Complete RNN

Now, we've done all the work we need to: We can define our RNN.

#### Definition 859

A **Recurrent Neural Network (RNN)** is a particular kind of **state machine** used as a neural network:

- We use a state machine so our network can remember and use past data, via the **state**.
- We call it "**recurrent**" because of our states. A state is created by the network to find the output, and then one timestep later, is **re-used** as a new input.

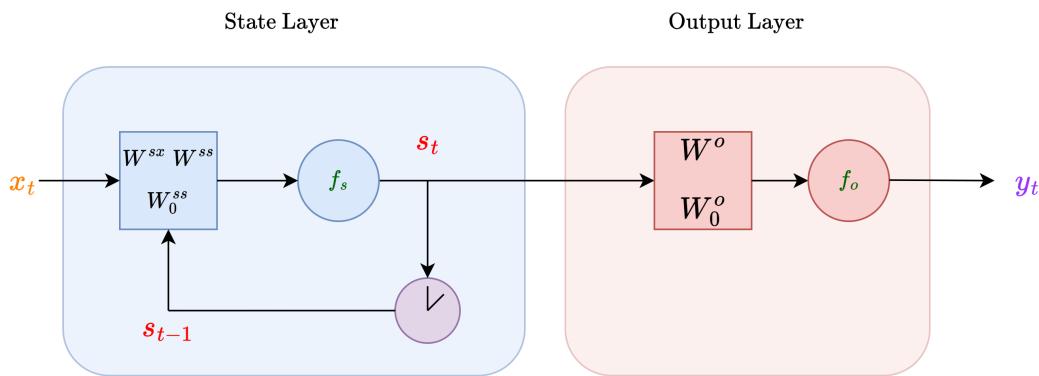
Our RNN manipulates input  $x_t \in \mathbb{R}^l$ , and state  $s_t \in \mathbb{R}^m$ , to create an output  $y_t \in \mathbb{R}^n$ .

Our state and output equations are given as:

$$s_t = f_s \left( W^{ss} s_{t-1} + W^{sx} x_t + W_0^{ss} \right)$$

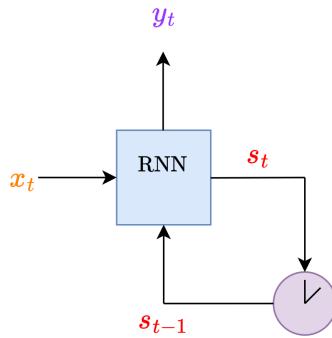
$$y_t = f_o \left( W^o s_t + W_0^o \right)$$

Where  $f_s$  and  $f_o$  are (typically non-linear) **activation functions**, and every weight  $W$  is a **matrix** with the appropriate dimensions.



We can abstract away all the math with a simpler perspective: \_\_\_\_\_

Technically, this diagram works for any state machine.



We have the basic parts we care about: input, output, and state.

Our input and output are visible from outside, while the **state** is recycled within the system.

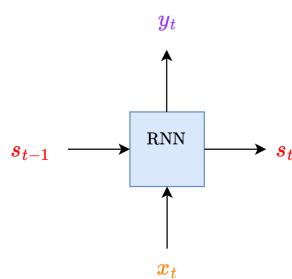
Take a moment to compare this model to the more complex one above: they're more similar than they seem.

### C.2.6 RNN as a "network"

One issue we might have with the above diagram is it doesn't look very much like a **network**: at best, it seems like a very small network.

But the simplified diagram could inspire us: currently, the RNN "feeds into itself", using the state.

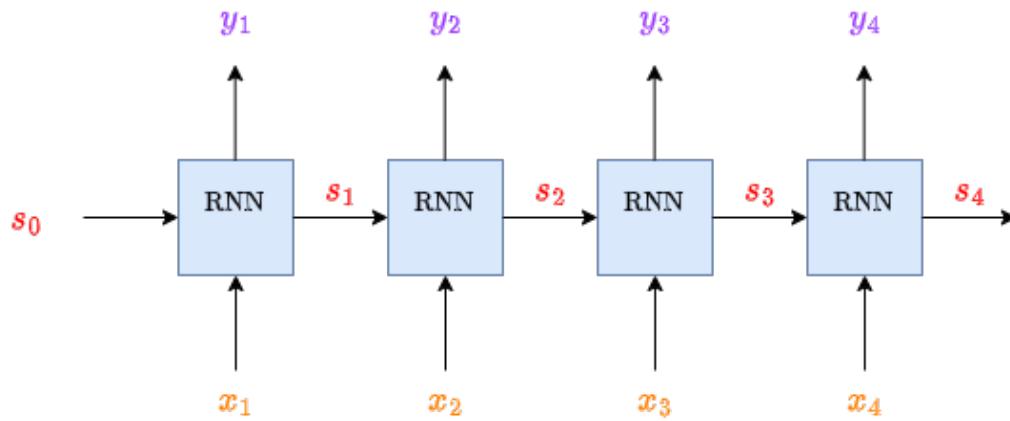
- In a different perspective, we could imagine that our first RNN unit is feeding into a second, **identical** unit.
- We'll show what we mean, by removing the clock:



We have an isolated, simple block.

Now, we can connect several of these in **series**: each one represents the input and output for the  $t^{\text{th}}$  timestep.

- The state  $s_t$  feeds into the  $(t + 1)^{\text{th}}$  block.



Each of these RNNs is the same block: we've replaced our loop by copying the same RNN multiple times.

Now, it looks more like a network! This is fascinating: an RNN is like a network where we use the **same layer** over and over again!

- With the additional caveat that each layer has its own **input** and **output**.

#### Concept 860

One perspective on RNNs is to see them as a **layered** network, where each layer is the full RNN:

- Every layer has a **distinct** input and an output.
- Every layer is structurally **identical** (same weights, activation).

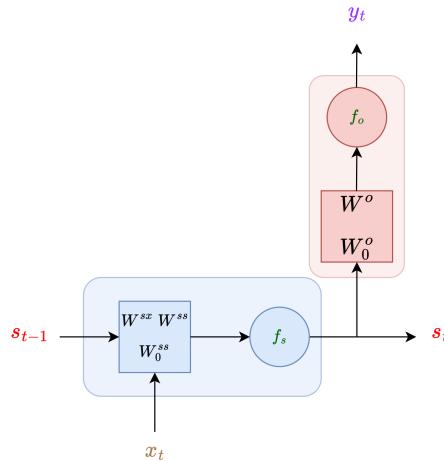
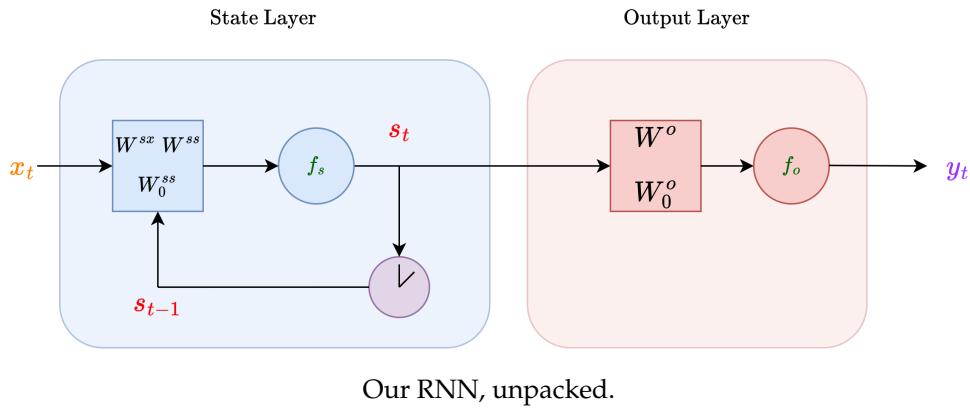
Of course, this version can be misleading: we don't actually have  $n$  copies of our RNN, we have one copy that we're using repeatedly.

- However, we could think of the  $x$ -axis as representing "time": the same RNN reused at multiple times.

#### C.2.7 RNN fully unpacked

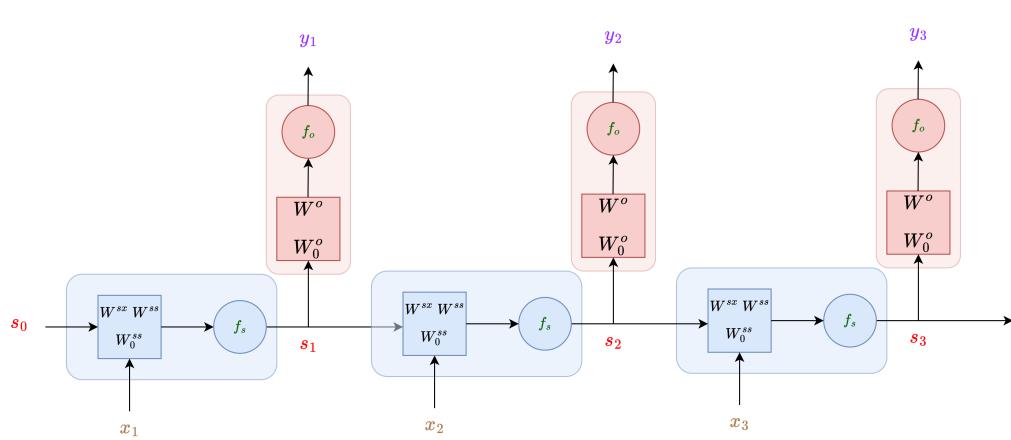
Now that we've introduced this perspective, let's use it for the "**unpacked**" version of our RNN, where we don't hide all of our inner functions. This will get a little messy.

First, we need to remove the clock:



We had to rearrange things a bit to get the effect we wanted.

But now, we can stack these into a full "network":



**Text:** Our RNN, unpacked.

This version looks complex, but it's just three copies of our previous model, side-by-side.

### C.2.8 RNN Example 1 (Optional)

Let's try a very simple example: we'll do a **weighted average** of the last 3 inputs.

This is a linear operation, so we can ignore the activation functions. Thus,  $f_s$  and  $f_o$  are the identity function:  $f_s(z) = f_o(z) = z$

$$s_t = W^{ss} s_{t-1} + W^{sx} x_t + W_0^{ss} \quad (\text{C.5})$$

$$y_t = W^o s_t + W_0^o \quad (\text{C.6})$$

Our input  $x_t$  for each turn will be a single value, stored in a  $(1 \times 1)$  matrix. Likewise, the "weighted average of 3 inputs" is a single value: another  $(1 \times 1)$ .

$$x_t = [X_t] \quad y_t = [Y_t] \quad (\text{C.7})$$

- Our state is based on the information we need to remember in order to compute the output.
- Thus, we'll store the **last three inputs**.

$$s_t = \begin{bmatrix} X_{t-2} \\ X_{t-1} \\ X_t \end{bmatrix} \quad (\text{C.8})$$

#### Concept 861

Our **state vector** is typically chosen based on what is useful for finding the **output**.

So, our goal is to get the structure we gave  $s_t$  above. We'll need to encode that in our equation.

We compute  $s_t$  from  $s_{t-1}$ ,  $x_t$ , and an offset.

- We're storing our past values, so we don't need an offset:  $W_0^{ss} = 0$ .

$$\begin{bmatrix} X_{t-2} \\ X_{t-1} \\ X_t \end{bmatrix} = W^{ss} s_{t-1} + W^{sx} x_t \implies \begin{bmatrix} X_{t-2} \\ X_{t-1} \\ X_t \end{bmatrix} = W^{ss} \begin{bmatrix} X_{t-3} \\ X_{t-2} \\ X_{t-1} \end{bmatrix} + W^{sx} \begin{bmatrix} X_t \\ X_{t-1} \end{bmatrix} \quad (\text{C.9})$$

Now, we can see that the information for  $s_t$  is spread across both terms:

- The  $s_{t-1}$  term contains  $X_{t-1}$  and  $X_{t-2}$
- The  $x_t$  term contains  $X_t$ .

So, we want to end up with:

$$\begin{bmatrix} X_{t-2} \\ X_{t-1} \\ X_t \end{bmatrix} = \underbrace{\begin{bmatrix} X_{t-2} \\ X_{t-1} \\ 0 \end{bmatrix}}_{W^{ss}s_{t-1}} + \underbrace{\begin{bmatrix} 0 \\ 0 \\ X_t \end{bmatrix}}_{W^{sx}x_t} \quad (\text{C.10})$$

We can figure out our weight matrices  $W^{ss}$  and  $W^{sx}$ , by comparing our input and output.

For starters, we showed above that the dimensions of  $W^{sx}$  depend on  $x_t$  and  $s_t$ .

$$\underbrace{\begin{bmatrix} ? & ? & ? \\ ? & ? & ? \\ ? & ? & ? \end{bmatrix}}_{W^{ss}} \begin{bmatrix} X_{t-3} \\ X_{t-2} \\ X_{t-1} \end{bmatrix} = \begin{bmatrix} X_{t-2} \\ X_{t-1} \\ 0 \end{bmatrix} \quad (\text{C.11})$$

To figure out  $W_{ss}$ , we go row-by-row, and figure out the correct values:

$$\underbrace{\begin{bmatrix} a & b & c \\ ? & ? & ? \\ ? & ? & ? \end{bmatrix}}_{W^{ss}} \begin{bmatrix} X_{t-3} \\ X_{t-2} \\ X_{t-1} \end{bmatrix} = \begin{bmatrix} X_{t-2} \\ X_{t-1} \\ 0 \end{bmatrix} \implies aX_{t-3} + bX_{t-2} + cX_{t-1} = X_{t-2} \quad (\text{C.12})$$

We find  $a = 0$ ,  $b = 1$ ,  $c = 0$ . We can repeat this process for our other rows.

Once we do the rest, we find:

$$\underbrace{\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}}_{W^{ss}} \begin{bmatrix} X_{t-3} \\ X_{t-2} \\ X_{t-1} \end{bmatrix} = \begin{bmatrix} X_{t-2} \\ X_{t-1} \\ 0 \end{bmatrix} \quad (\text{C.13})$$

We follow the same logic for the other example:

$$\underbrace{\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}}_{W^{sx}} \begin{bmatrix} X_t \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ X_t \end{bmatrix} \quad (\text{C.14})$$

### Concept 862

In order to derive our weight matrix, we can go **row-by-row**:

- For each row, we figure out which choice of **weights** gives the desired output.

Taken together, we get:

$$\begin{bmatrix} \textcolor{blue}{X_{t-2}} \\ \textcolor{blue}{X_{t-1}} \\ \textcolor{blue}{X_t} \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}}_{W^{ss}} \begin{bmatrix} X_{t-3} \\ \textcolor{red}{X_{t-2}} \\ \textcolor{red}{X_{t-1}} \end{bmatrix} + \underbrace{\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}}_{W^{sx}} \begin{bmatrix} \textcolor{brown}{X_t} \end{bmatrix} \quad (\text{C.15})$$

~~~~~

Now we can compute the weighted average. We'll pick some arbitrary numbers: 50% of X_t , 30% of X_{t-1} , and 20% of X_{t-2} .

$$Y_t = 0.5X_t + 0.3X_{t-1} + 0.2X_{t-2} \quad (\text{C.16})$$

- Again, we don't need an offset $W_0^o = 0$.

$$\begin{bmatrix} Y_t \end{bmatrix} = W^o \begin{bmatrix} \textcolor{red}{s_t} \end{bmatrix} \implies W^o \begin{bmatrix} \textcolor{blue}{X_{t-2}} \\ \textcolor{blue}{X_{t-1}} \\ \textcolor{blue}{X_t} \end{bmatrix} \quad (\text{C.17})$$

We can, again, figure out our weight matrix W^o based on the desired result.

$$\underbrace{\begin{bmatrix} 0.5 & 0.3 & 0.2 \end{bmatrix}}_{W^o} \begin{bmatrix} \textcolor{blue}{X_{t-2}} \\ \textcolor{blue}{X_{t-1}} \\ \textcolor{blue}{X_t} \end{bmatrix} = 0.5X_t + 0.3X_{t-1} + 0.2X_{t-2}$$

Remark (Optional) 863

Our final RNN comes out to:

$$f_s(z) = f_o(z) = z$$

$$W^{ss} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \quad W^{sx} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad W_0^{ss} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$W^o = \begin{bmatrix} 0.5 & 0.2 & 0.2 \end{bmatrix} \quad W_0^o = \begin{bmatrix} 0 \end{bmatrix}$$

C.2.9 RNN Example 2 (Optional)

Let's run through a more concrete example.

For simplicity, f_s and f_o are the identity function: $f_s(z) = f_o(z) = z$.

$$s_t = W^{ss} s_{t-1} + W^{sx} x_t + W_0^{ss} \quad (\text{C.18})$$

Remember that this is the activation function, not the complete function we use

$$y_t = W^o s_t + W_0^o \quad (\text{C.19})$$



- Each **input** is one number: the amount of money you earn every month.

$$x_t = \begin{bmatrix} x_t^E \end{bmatrix} \quad (\text{C.20})$$

- Our **state** will be two numbers: the money you have in the bank, and the money you've invested.

$$s_t = \begin{bmatrix} s_t^B \\ s_t^I \end{bmatrix} \quad (\text{C.21})$$

- Your **output** is your net worth: including the bank, and the invested money.

$$y_t = \begin{bmatrix} y_t^T \end{bmatrix} \quad (\text{C.22})$$

Each of these could be a vector of any length, depending on the problem.



First, we want to compute s_t . Just like we mentioned in [Example 1](#), we can go row-by-row to figure out the equation for s_t .

Let's start with your first row, s_t^B : your "savings" money.

- The money in the bank s_{t-1}^B makes no interest.
- 10% of our investing s_{t-1}^I goes into the bank.
- 80% of our earned money x_t^E goes into the bank.
- We lose \$6000 of savings every month.

We have a pretty terrible bank.

$$s_t^B = s_{t-1}^B + 0.2s_{t-1}^I + 0.8x_t^E - 6000 \quad (\text{C.23})$$

With this, we can write in vector form:

$$\begin{bmatrix} s_t^B \\ s_t^I \end{bmatrix} = \begin{bmatrix} 1 & 0.2 \end{bmatrix} \begin{bmatrix} s_{t-1}^B \\ s_{t-1}^I \end{bmatrix} + \begin{bmatrix} 0.8 \\ 0.2 \end{bmatrix} \begin{bmatrix} x_t^E \end{bmatrix} - \begin{bmatrix} 6000 \\ 0 \end{bmatrix} \quad (\text{C.24})$$

Concept 864

Just like in other neural networks, the **weights** in an RNN indicate how an **input** variable (ex: x_t^E) affects an **output** variable (ex: s_t^B in our linear system)

Now, we'll compute s_t^I , your investment money:

- The money in the bank s_{t-1}^B is not invested.
- Our invested money s_{t-1}^I grows by 1%.
- 20% of our earned money x_t^E is invested.
- **No money** is added beyond that.

$$s_t^I = 0s_{t-1}^B + 1.01s_{t-1}^I + 0.2x_t^E + 0 \quad (\text{C.25})$$

In vector form:

$$\begin{bmatrix} s_t^I \end{bmatrix} = \begin{bmatrix} 0 & 1.01 \end{bmatrix} \begin{bmatrix} s_{t-1}^B \\ s_{t-1}^I \end{bmatrix} + \begin{bmatrix} 0.2 \end{bmatrix} \begin{bmatrix} x_t^E \end{bmatrix} + \begin{bmatrix} 0 \end{bmatrix} \quad (\text{C.26})$$

Now, we can create our full representation of s_t :

$$s_t = W^{ss} s_{t-1} + W^{sx} x_t + W_0^{ss} \quad (\text{C.27})$$

Which becomes:

$$\begin{bmatrix} s_t^B \\ s_t^I \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0.2 \\ 0 & 1.01 \end{bmatrix}}_{W^{ss}} \begin{bmatrix} s_{t-1}^B \\ s_{t-1}^I \end{bmatrix} + \underbrace{\begin{bmatrix} 0.8 \\ 0.2 \end{bmatrix}}_{W^{sx}} \begin{bmatrix} x_t^E \end{bmatrix} + \underbrace{\begin{bmatrix} -6000 \\ 0 \end{bmatrix}}_{W_0^{ss}} \quad (\text{C.28})$$

We've finished the state equation of our RNN.

The output will be a bit simpler: it's just your total net worth.

$$y_t = W^o s_t + W_0^o \quad (\text{C.29})$$

- The output is the sum of your bank savings, and your investments.

- There's nothing to add beyond that.

$$\begin{bmatrix} y_t^T \end{bmatrix} = \overbrace{\begin{bmatrix} 1 & 1 \end{bmatrix}}^{W^o} \begin{bmatrix} s_t^B \\ s_t^I \end{bmatrix} + \overbrace{\begin{bmatrix} 0 \end{bmatrix}}^{W_0^o} \quad (C.30)$$

Remark (Optional) 865

Our final RNN comes out to:

$$f_s(z) = f_o(z) = z$$

$$W^{ss} = \begin{bmatrix} 1 & 0.2 \\ 0 & 1.01 \end{bmatrix} \quad W^{sx} = \begin{bmatrix} 0.8 \\ 0.2 \end{bmatrix} \quad W_0^{ss} = \begin{bmatrix} -6000 \\ 0 \end{bmatrix}$$

$$W^o = \begin{bmatrix} 1 & 1 \end{bmatrix} \quad W_0^o = \begin{bmatrix} 0 \end{bmatrix}$$

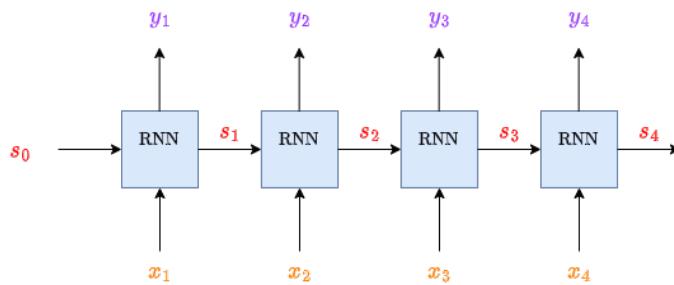
C.3 Sequence-to-sequence RNN

C.3.1 The sequence-to-sequence perspective

We've completely developed our RNN, a network designed out of a **state machine**.

- Our system takes one input, and produces one output, for each time step.

So far, we've been viewing each of these x_t and y_t terms separately. However, when we use our simplified perspective, things look a bit different:



In this view, we see a "sequence" of inputs x_t , and a "sequence" of outputs y_t .

This is why we might call the RNN problem a **sequence-to-sequence** problem.

Concept 866

Rather than seeing each x_t term as an isolated input, we could consider our input the full **sequence**

$$\mathbf{x} = [x_1 \ x_2 \ x_3 \ \cdots \ x_n]$$

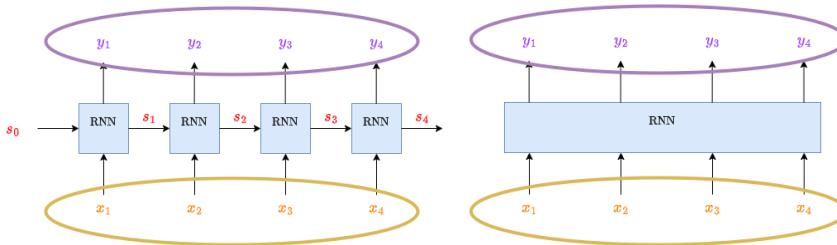
Our RNN takes the sequence \mathbf{x} and returns a paired, output sequence \mathbf{y} :

$$\mathbf{y} = [y_1 \ y_2 \ y_3 \ \cdots \ y_n]$$

In this view, we can think of our RNN as a machine that takes in one sequence, and outputs a second, **equal-length** sequence.

-
- Notice that x_t and y_t can be **vectors**: \mathbf{x} and \mathbf{y} may need to be complete **matrices**, to store all these vectors.

In this view, we "lump together" all of our inputs and outputs as a single object:



If we ignore the internal states, our RNN just turns one sequence into another.

We sometimes call this process **transduction**.

Definition 867

We say that our RNN **transduces** our **input** sequence into our **output** sequence.

Now, we can have **multiple sequences**: for example, we could have our input be, $x = [1, 2, 3, 4]$, or $x = [2, 4, 6, 7]$. Both are valid inputs.

- And in each of those sequences, you have **multiple timesteps**.

We'll need to distinguish between these two: different sequences, versus different timesteps within a sequence.

And each vector x_t can have multiple elements! We'll ignore this last bit, for our sanity.

- For this purpose, we re-use data point notation $x^{(i)}$ that we developed in earlier chapters, Regression and Classification.

Notation 868

We use x_t to distinguish between inputs in the **same sequence**.

- We'll represent a whole sequence with x .

We'll use $x^{(i)}$ to distinguish between **different sequences**.

- **Example:** $x_3^{(2)}$ is the 3rd timestep, of the 2nd sequence ("data point").

C.3.2 Sequence length

One important observation: we **need** an input x_t in order to proceed to the next output.

- So, we only have as many outputs as we have inputs.

Concept 869

The **input** and **output** sequences to an RNN will be the **same length**.

What about two different input sequences?

- Our RNN is capable of taking in sequences of **any length**: if a sequence is longer, we just run our RNN for more timesteps.

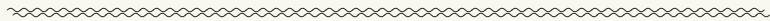
So, each sequence our RNN receives can whatever length it wants to be: they don't have to match length.

Concept 870

An RNN can receive input sequences of **any length**.

In fact, the length can be different between **different input sequences**.

- So, sequence $x^{(1)}$ and sequence $x^{(2)}$ can be used by the **same RNN**, even if they have different lengths.



This means that each data point needs a separate variable for length: the length of $x^{(i)}$ is $n^{(i)}$

However, we need to be careful of the difference between these two ideas:

Clarification 871

The output sequence $y^{(i)}$ **must** have the **same length** as its input $x^{(i)}$: they're **paired** together.

However, different inputs ($x^{(i)}$ and $x^{(j)}$) can have **different lengths**.



This also means that inputs and outputs which are **not from the same pair** ($x^{(i)}$ and $y^{(j)}$) can have different lengths.

Note that this also means that different outputs can have different lengths, as well.

C.3.3 Training data

Just like any other NN, we usually are training our RNN for a **task**. What kind of task?

- The output of our RNN is a **sequence**: so, our goal will be to take the input sequence $x^{(i)}$, and give the *desired* sequence $y^{(i)}$.

Concept 872

Training our RNN is similar to our previous model training problems, like **regression**:

- Given a particular input sequence $x^{(i)}$, we want to teach our model to produce the output sequence $y^{(i)}$.

This is similar to regression, where we want to take an input vector, and get a real number as an output.

We want to use this model for **supervised** learning: we know the sequence we want to get as an output.

Definition 873

Our RNN is trained with a **training set** with q data points:

$$\left[\begin{array}{cccc} (x^{(1)}, y^{(1)}) & (x^{(2)}, y^{(2)}) & \dots & (x^{(q)}, y^{(q)}) \end{array} \right]$$

Where, due to the behavior of RNNs (described above), we require:

- Each element $x^{(i)}$ or $y^{(j)}$ is a **sequence**.
- Elements $(x^{(i)}, y^{(i)})$ from the **same pair** must have the **same length** $n^{(i)}$.
- Different pairs** may have **different lengths**.
 - Meaning, $n^{(i)}$ and $n^{(j)}$ are allowed to be different.

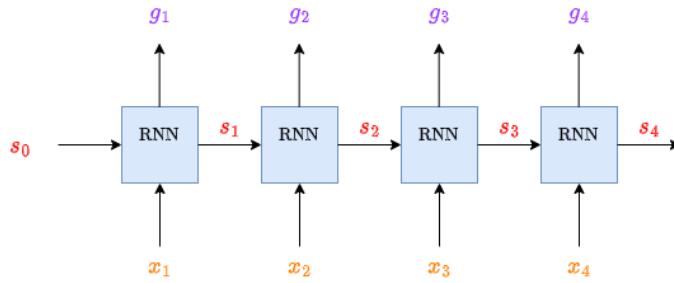
One important note: Now, we're using y to represent the **desired** output, which is not necessarily the same as what our RNN actually gives us.

- We'll need separate notation for this.

Notation 874

From this point on, we'll use y to indicate the **desired, correct** output, and g to represent the **current RNN** output.

- So, our goal is to make g and y as **similar** as possible.



Very little changes: we just replace y with g .

C.3.4 Training and Evaluation

The desired training output is $y^{(i)}$: we will **predict** it using the RNN output, $g^{(i)}$.

- For training and evaluation, we'll need a **loss function** to indicate how wrong our guess $g^{(i)}$ is.

This loss will be indicated by $\mathcal{L}_{\text{seq}}(g^{(i)}, y^{(i)})$: typically, it will tell us how **different** our sequences are.

- The easiest way to compare two sequences is to compare them **element-wise**: compare the t^{th} element of $y^{(i)}$ to the t^{th} element of $g^{(i)}$.

Definition 875

We compute the **loss** \mathcal{L}_{seq} of our sequence by adding up the loss \mathcal{L}_{elt} for each **element** in our sequence.

$$\mathcal{L}_{\text{seq}}(g^{(i)}, y^{(i)}) = \sum_{t=1}^{n^{(i)}} \mathcal{L}_{\text{elt}}(g_t^{(i)}, y_t^{(i)})$$

The choice of loss function \mathcal{L}_{elt} depends on the data type of $y^{(t)}$.

- Note that for sequence $g^{(i)}$, we have $n^{(i)}$ timesteps.

Example: If our sequence is a series of numbers, we could take the **squared error** between the elements:

$$g^{(i)} = [1 \ 4 \ 5] \quad y^{(i)} = [1 \ 2 \ 3] \quad \mathcal{L}_{\text{elt}}(a, b) = (a - b)^2 \quad (\text{C.31})$$

If compute the total loss, we get

$$\mathcal{L}_{\text{seq}}(g^{(i)}, y^{(i)}) = \sum_{t=1}^3 \left(g_t^{(i)} - y_t^{(i)} \right)^2 = (1 - 1)^2 + (4 - 2)^2 + (5 - 3)^2$$

$$\mathcal{L}_{\text{seq}}(g^{(i)}, y^{(i)}) = 8 \quad (\text{C.32})$$

Next, we'll compute the overall performance of our RNN. First, some notation:

Notation 876

We'll collectively represent all of our weights with a W :

$$W = (W^{sx}, W^{ss}, W^o, W_0^{ss}, W_0^o)$$

These are the **parameters** of our RNN – they're used for computing $g^{(i)}$. Meanwhile, $x^{(i)}$ is the **input**, so we find:

$$g^{(i)} = \text{RNN}(x^{(i)}; W)$$

Just like in other problems, we evaluate our model by taking the **average** of all of our losses:

Definition 877

The **objective function** $J(W)$ of our RNN is given by **averaging** the loss for each of our q data points:

$$J(W) = \frac{1}{q} \sum_{i=1}^q \mathcal{L}_{\text{seq}}(g^{(i)}, y^{(i)}) = \frac{1}{q} \sum_{i=1}^q \mathcal{L}_{\text{seq}}\left(\overbrace{\text{RNN}(x^{(i)}, W)}^{g^{(i)}}, y^{(i)}\right)$$

C.3.5 Activation Functions

Lastly, we want to address our activation functions, f_s and f_o .

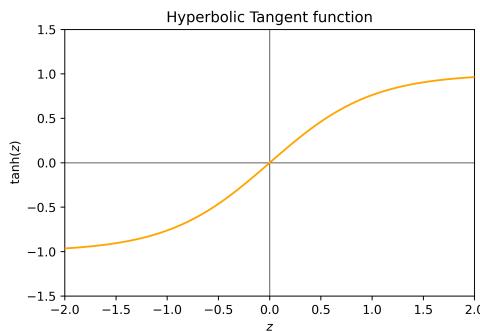
f_s is used for our **state**: because our state doesn't directly compute our output, we tend to use the same f_s for different problems:

Concept 878

Our most typical choice for f_s is the **hyperbolic tangent** function \tanh :

$$f_s(z) = \tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

A function whose output ranges over $(-1, +1)$.



A reminder of how \tanh appears: it looks relatively similar to sigmoid, with a different output range.

Meanwhile, f_o directly allows us to compute the output, so our choice of f_o depends on our problem and data type.

Concept 879

Just like in regular supervised learning, f_o is chosen based on the problem at hand, considering:

- Data type
- Range
- Sensitivity

And other properties.

C.4 RNN as a language model

Human language is written/spoken **sequentially**, with each character/word/syllable coming in a particular **order**.

Thus, we might expect RNNs, a "sequential" model, to be suited for this task.

The task in question is **predictive text**:

Not nearly as well as transformers, but we'll discuss that in the Transformers chapter.

Definition 880

In the **predictive text** problem, you're given the "past" sequence of text, and you're supposed to predict "future" text.

Example: Autocorrect is a common application: based on the words you've typed so far, your phone will predict the most likely next word.

RNNs can be trained to accomplish this type of task.

C.4.1 Tokens

Our goal is to **correctly** predict the next word in a sentence, based on the previous words in a sentence.

First, we'll break up our sentence into a sequence of **elements**: these elements will be called **tokens**.

Definition 881

A **token** is a single **unit** of our text: this might be a single letter/**character**, or a single **word**.

Example: The following sentence contains 3 tokens if a token is a **word**, and 11 tokens if a token is a **character**:

Some systems, like chat-gpt, will use several characters as a single "token": this set of characters might not line up with each word!

$$\begin{array}{ccccccc} 1 & 2 & 3 \\ \text{I} & \text{ love} & \text{ dogs} \end{array} \quad (\text{C.33})$$

$$\begin{array}{cccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ \text{I} & - & \text{l} & \text{o} & \text{v} & \text{e} & - & \text{d} & \text{o} & \text{g} & \text{s} \end{array} \quad (\text{C.34})$$

We can represent our sentence c as sequence of tokens c_i :

$$c = [c_1 \ c_2 \ \dots \ c_k] \quad (\text{C.35})$$

C.4.2 Predicting tokens

We want our RNN to predict **future** tokens in the sentence, based on **past** tokens.

- Let's start by feeding in our first token:

$$\text{RNN} \left(\begin{bmatrix} c_1 \end{bmatrix} \right) = \begin{bmatrix} G_2 \end{bmatrix} \quad (\text{C.36})$$

We want our RNN to predict the **next** character, so the **output** will be our prediction for c_2 : we'll call it G_2 .

Notation 882

Our **prediction** for the n^{th} token, c_n , is G_n .

- Thus, G_n is the token we consider **most likely** for c_n .

Alternatively, G_n could be a vector, giving the **probability** for each possible token that c_n could be.

This would be useful for evaluating our model: how sure was it of the right answer?

- Let's try our second token:

$$\text{RNN} \left(\begin{bmatrix} c_1 & c_2 \end{bmatrix} \right) = \begin{bmatrix} G_2 & G_3 \end{bmatrix} \quad (\text{C.37})$$

Note that our model gets to see the correct c_2 , **after** making its prediction, G_2 .

- That means that our RNN has only seen c_1 when it predicts G_2 : it doesn't know what the correct character, c_2 , is yet.
- Meanwhile, G_3 is generated with knowledge of c_1 and c_2 : the RNN knows what the first two characters are.

Concept 883

When our RNN **guesses** the t^{th} token, G_t , it can only see the **first $t - 1$ inputs**:

$$\begin{bmatrix} c_1 & c_2 & \dots & c_{t-1} \end{bmatrix} \xrightarrow{\text{RNN}} G_t$$

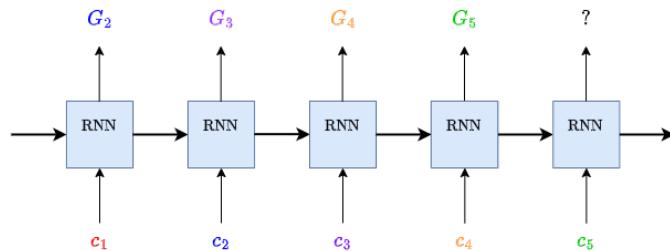
Information stored in c_t or c_{t+1} , for example, has **no effect** on G_t .

- Otherwise, our model could cheat, and predict by looking at the answer!

In this way, we can supply our entire input, and get a full vector of predictions:

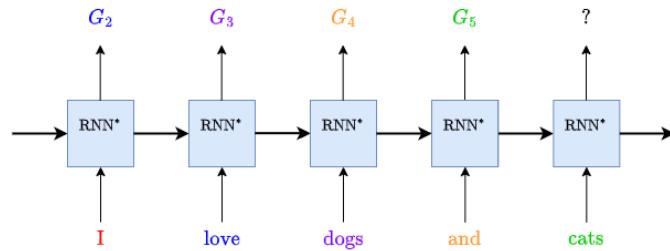
$$\text{RNN} \left(\begin{bmatrix} c_1 & c_2 & \dots & c_{k-1} & c_k \end{bmatrix} \right) = \begin{bmatrix} G_2 & \dots & G_{k-1} & G_k & ? \end{bmatrix} \quad (\text{C.38})$$

Let's show this in our simplified RNN diagram.

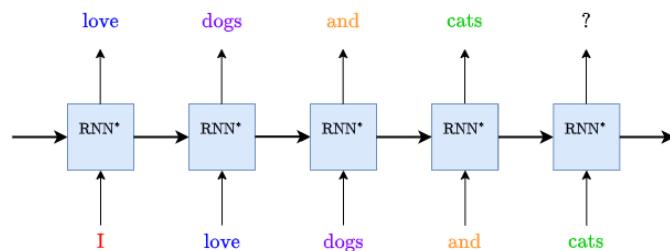


Remember, the arrows left-to-right are states: they represent all of the previous words in the sentence.

Here's an example sentence we could try to predict:



And now, here's an ideal case, where our RNN performs perfectly (we'll call it RNN^*):



Our RNN predicts the right word, right before it appears.

C.4.3 Start token and end token

There are two problems we'd like to fix.

- We'd like to try predicting the first character, c_1 , rather than skipping it.
- If we start with c_2 , there are only $k - 1$ characters we need to predict.

- We have k inputs, and therefore k outputs: we have **one more output** than we need.

Our first solution is to add a special "**START**" token to the beginning of our input:

$$x = \begin{bmatrix} \text{START} & c_1 & c_2 & \cdots & c_k \end{bmatrix} \quad (\text{C.39})$$

What does this do for us?

- During our first timestep, our RNN has nothing other than the **START** token, so it's able to spend that timestep **predicting c_1** .

$$\text{RNN}\left(\begin{bmatrix} \text{START} \end{bmatrix}\right) = \begin{bmatrix} g_1 \end{bmatrix} \quad (\text{C.40})$$

Now, our output starts with G_1 .

$$\text{RNN}\left(\begin{bmatrix} \text{START} & c_1 & c_2 & \cdots & c_{k-1} & c_k \end{bmatrix}\right) = \begin{bmatrix} G_1 & G_2 & \cdots & G_{k-1} & G_k & ? \end{bmatrix}$$

Definition 884

The **input** to our **sentence-prediction RNN** starts with the special **START** token, followed by all of the tokens in our sentence c .

$$x = \begin{bmatrix} \text{START} & c_1 & c_2 & \cdots & c_k \end{bmatrix}$$

When our RNN receives this **START** token, it has an opportunity to predict the **first word** in the sentence, with no context.

This hasn't fixed our other problem, though: now we have $k + 1$ slots, but only k outputs we need to predict.

We'll solve that by adding an **END** character at the end of the output.

$$y = \begin{bmatrix} c_1 & c_2 & \cdots & c_k & \text{END} \end{bmatrix} \quad (\text{C.41})$$

- Now, we have a desired final output: we want our RNN to predict when the sentence ends, and return **END**.

Definition 885

The **optimal output** of our **sentence-prediction RNN** starts with all the tokens in our sentence c , followed by the special **END** token.

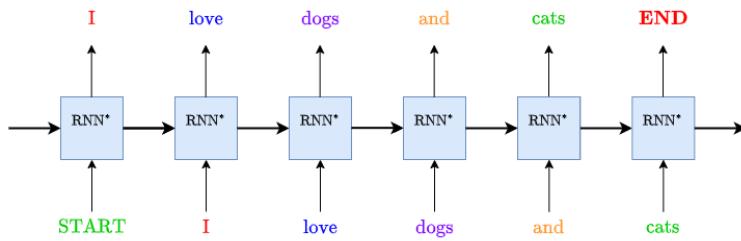
$$y = [c_1 \ c_2 \ \dots \ c_k \ \text{END}]$$

Thus, we've fully formed our problem statement:

Concept 886

The goal of our sentence-prediction RNN is to **replicate the sentence c** , token-by-token, with two caveats:

- The input starts with the special **START** token, to give your model one timestep to **predict the first token** c_1 .
- The output should terminate with the special **END** token: your model should **predict when the sentence ends**.



Here's what our system looks like, now that we've added the START and END tokens.

Let's review what happens at each timestep.

- On the bottom, we **input** one word in the sentence.
- On the top, we **output** the predicted next word.
- After an input, we update our **state**, to include new info. This is **used** by the RNN in the next timestep (left-to-right).

C.4.4 Why we might use RNNs for language

The general reason we decided to try using RNNs for language is pretty basic:

- "RNNs **output sequences**. Text is a **sequence of tokens**".

But there's a bit more to it than that: RNNs allow our model to remember the **structure** of our sentence, using our **state s_t** .

- The newest token c_t might be related to one that we saw **earlier**: for example, 5 tokens ago.
- Our state can treat words that are **closer**, differently from words which are **further away**.

This way, our **state** allows us to keep track of sentence structure: grammar, the meaning of each word, tone, etc.

Concept 887

When we use an RNN as a **language model**, we're hoping that it can distinguish between "**nearby**" words, and "**farther away**" words.

- Words which are closer/further can contribute differently to the sentence: this gives us **context**.

Convolution has a similar effect, but in a more **discrete** way: in convolution, we have a window of a **fixed size**.

- If n is our window size, but two tokens are $n + 1$ units apart, then they won't show up together in convolution.

Meanwhile, RNNs are more flexible:

Concept 888

Depending on how your **RNN state** works, it could preserve/accumulate data over **longer, non-fixed** distances than **convolution**.

- **Example:** A **running average** could factor in newer information, with older information, without completely "forgetting" that old information.

C.4.5 Why RNNs don't work (well) for language

Unfortunately, RNNs tend not to work well enough.

- Their state s_t can only store a **limited** amount of data.
- So, the longer the RNN runs, the more it forgets.

Concept 889

Over time, an RNN will "**forget**" information it learned in **earlier** timesteps.

- It's replaced by **newer** data.

Language requires this sort of longer-term memory.

- The longer the text prompt becomes, the worse the RNN performs.

In the next chapter, we'll design a model to overcome this problem: the **transformer**.

C.5 Terms

- Timestep t (Review)
- State s_t
- Input x_t
- Transition function f_s
- Output y_t
- Output function f_o
- State Machine
- Finite State Machine
- State Transition Diagram
- Linearity
- Time-Invariance
- Linear Time-Invariant System (LTI)
- Recurrent Neural Network (RNN)
- Weights $W^{ss}, W^{sx}, W_0^{ss}, W^o, W_0^o$
- Transduction
- x_t notation
- $x^{(i)}$ notation (Review)
- Sequence loss \mathcal{L}_{seq}
- Hyperbolic Tangent Function \tanh (Review)
- Predictive Text
- Token
- Start Token
- End Token

APPENDIX D

Word2vec – Skipgram Approach

D.1 Vector embeddings and tokens

In these notes, we introduce a particular way to choose "good" vector embeddings, based on the word2vec technique.

- These notes were originally spliced from the Transformers chapter, so there are some regions of overlap.

D.1.1 One-hot encoding isn't enough

First, we want to turn words into something computable, like a **vector**.

The simplest approach would be **one-hot encoding**.

It's difficult to try to do math on the word "cheddar". It's not numerical.

- **Example:** Suppose that we want to classify **furniture** as table, bed, couch, or chair.

$$\begin{bmatrix} \text{table} \\ \text{bed} \\ \text{couch} \\ \text{chair} \end{bmatrix} \quad (\text{D.1})$$

- For each class:

$$\begin{aligned} v_{\text{chair}} &= \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} & v_{\text{table}} &= \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} & v_{\text{couch}} &= \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} & v_{\text{bed}} &= \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \end{aligned} \quad (\text{D.2})$$

This approach is simple, but often, it's *too* simple.

Concept 890

One-hot encoding loses a lot of information about the objects it's representing.

- It's hard to say which words are "**similar**" to each other, for example.

Example: You probably associate the word "**sugar**" with "**sweet**", and "**salt**" with "**savory**".

- But, if you use one-hot encoding, all of these words are "**equally different**".

$$\begin{aligned} v_{\text{salt}} &= \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} & v_{\text{savory}} &= \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} & v_{\text{sugar}} &= \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} & v_{\text{sweet}} &= \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \end{aligned} \quad (\text{D.3})$$

You could **shuffle** the rows of one-hot vectors, and represent the same information.

So, we can't use the order of 1's and 0's to determine "**closeness**": the order can be **freely changed**.

In order to incorporate this information, we'll need a better way to represent words as vectors.

D.1.2 Word Embeddings: Similarity between words

Our new approach will convert each word w into a **vector** v_w of **length d**.

$$w \longrightarrow v_w \quad v_w \in \mathbb{R}^d \quad (\text{D.4})$$

Unlike one-hot encoding, we don't require that d equals the size of our vocabulary.

How do we want to convert words into vectors? Above, we mentioned that one-hot doesn't tell us how **similar** two words are.

Clarification 891

There are many ways for words to be **similar**: similar word length, similar choice of letters, etc.

But in our case, we're interested in **semantics**: the **meanings** of the words. We want to know which words have similar meanings.

- **Example:** We don't consider "sugar" and "sweet" to be similar because they both start

with "s".

- They're similar because of **meaning**: sugar tastes sweet. Sweet strawberries contain sugar.

Concept 892

We often want our **word embeddings** v_w to tell us which words are **semantically similar** to each other: which words have similar **meanings**.

$$v_a \text{ and } v_b \text{ are similar vectors} \iff a \text{ and } b \text{ are semantically similar words}$$

Our goal is to make this statement true. But we have a problem: these are *concepts*, rather than computable *numbers*.

- So, we'll have to turn each side into something computable.

D.1.3 Vector Similarity: Dot Products

First, we'll handle the left side: how do we know if vectors are **similar**?

- We've come across this problem multiple times, and we'll solve it the same way as always: using the **dot product**.

Concept 893

Review from the Classification chapter

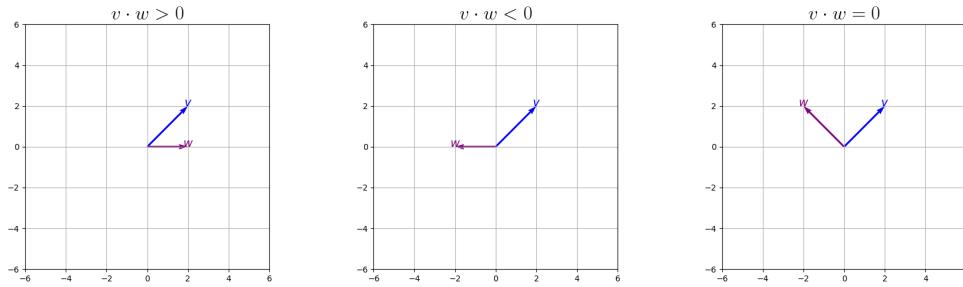
You can use the **dot product** between vectors u and v , **normalized by their magnitudes**, to measure their "**cosine similarity**".

$$S_C(u, v) = \frac{u \cdot v}{|u| \cdot |v|}$$

If two vectors are more **similar**, they have a **larger** normalized dot product.

- This function ranges from -1 (opposite vectors) to +1 (identical vectors). Perpendicular vectors receive a 0.

We call it "cosine similarity", because this is equal to the cosine of the angle α between u and v .



We can see here what we mean by "similar" or "dissimilar".

Clarification 894

You can use $S_C(u, v)$ to measure the **similarity** between two vectors, ignoring magnitude.

But for simplicity, we'll skip the **normalizing** step, and just take the **dot product**:

$$S_D(u, v) = u \cdot v = u^\top v$$

We're getting closer to a computable form:

$$\underbrace{(v_a \cdot v_b)}_{\text{Similar vectors}} \text{ is large} \iff a \text{ and } b \text{ are semantically similar words} \quad (\text{D.5})$$

D.1.4 Semantic Similarity and Word Frequency

The "right side" of our expression is a bit trickier: how do you compute which words have **similar meanings**?

We can't directly turn "meaning" into a number. But instead, we'll focus on a different concept, that might help us predict similarity:

- **Example:** Earlier, we showed that "sweet" and "sugar" were related, by referencing the fact that "**sugar tastes sweet**".
- While our machine might not understand the concept, it can see that "sugar" and "sweet" showed up **together** in a sentence.

Often, words that are related, show up in the same sentences, or paragraphs. So, we'll try to use this to our advantage:

How much do/can large language models "understand" what they're saying? Lots of very smart people continue to argue exactly how much they know.

a and b are **semantically similar words** $\xrightleftharpoons{\text{maybe?}}$ a and b **frequently** show up together

These two aren't *actually* equivalent, but we hope that we can use one to predict the other.

Concept 895

We can predict which words might be **more similar** by observing **how often** they show up **together** in a body ("corpora") of text.

- When two words occur **together** in a context, we call this **co-occurrence**.
- Thus, we're measuring **frequency of co-occurrence**.

If two words show up near each other more frequently, we predict that they might be **more similar**.

This kind of word embedding is often called "**word2vec**", named after a particular set of algorithms that use this approach.

- **Example:** The words "quantum" and "physics" go together often. So do the words "rain" and "weather".

Sometimes, "word2vec" is used to reference any technology that creates word embeddings. But this isn't always technically accurate.

Clarification 896

We **don't actually know** for certain that, if two words often show up together, they have **related meanings**.

But, in practice, we find that "**frequency of co-occurrence**" is a **surprisingly good** measure of similarity.

Because we're talking about frequency, we'll consider the **probability** of seeing both words together.

a and b are **semantically similar words** $\xrightleftharpoons{\text{maybe?}}$ $P(a \text{ and } b \text{ occur together})$ is **high**

Finally, we have something closer to math:

$$\underbrace{v_a \cdot v_b \text{ is large}}_{\text{Similar vectors}} \iff \underbrace{P(a \text{ and } b \text{ occur together}) \text{ is high}}_{\text{Similar words}}$$

Now, we have some "mathematical" concepts: we can start using these to create mathematical **objects**.

D.1.5 Clarifying our probability

In order to proceed, we need to be a little more specific.

$$\overbrace{(v_a \cdot v_b) \text{ is large}}^{\text{Similar vectors}} \iff \overbrace{P(a \text{ and } b \text{ occur together}) \text{ is high}}^{\text{Similar words}}$$

The dot product is already an equation, so the left side is fine.

The right side is all we need to clear up: "P (a and b occur together)" is a bit **vague**.

- We want to know if a and b tend to show up **together**, rather than **separately**.

Here's a concrete way to say this: "**if** we find one word, **how often** do we find the other nearby?"

One interpretation would be: "if we look at a random phrase, how often do we have words a and b?"

But we only care whether a and b are together/separate: we don't care about sentences containing neither.

Concept 897

To predict how **similar** words a and b are, we want to compute how often they **co-occur**.

- One way to phrase this: "**given** that we find a, what are the **chances** we find b nearby?"

$$P\{b \text{ nearby} \mid a \text{ found}\}$$

$$\overbrace{(v_a \cdot v_b) \text{ is large}}^{\text{Similar vectors}} \iff \overbrace{P\{b \text{ nearby} \mid a \text{ found}\} \text{ is large}}^{\text{Occur together frequently}}$$

We're getting warmer!

- We "find" a at index t:

w_t is the t^{th} word in our passage.

$$w_t = a \quad (\text{D.6})$$

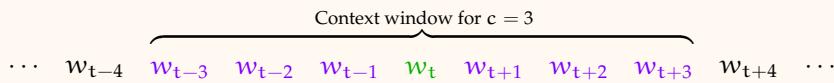
- Now, let's define what it means for b to be "**nearby**".

Definition 898

In a text, we may want to find the "context" for **center word** w_t : we want all of the words **nearby**.

- We'll use the c nearest words on either side: these are our **context words**. c is our **maximum skip distance**.

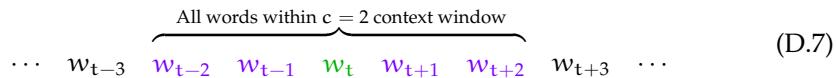
This collection of $2c + 1$ words is called our **context window**.



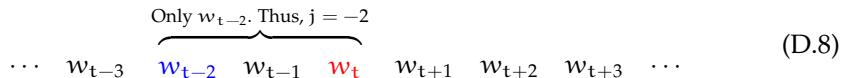
- Notice the similarity to the filter size from Convolution: we still have an idea of "locality".

So, we want to look for b in our context window. There are two ways we can turn this into a probability:

- You check all of the context words **at the same time**:



- You check **one word at a time**: j units to the right/left.



For now, it's easier to use the latter approach: **each index** has a **separate probability**.

We call c our "maximum skip distance", because it's the largest number of words we can "skip" over, starting from w_t .

We're allowed to move over by c words, in either direction.

Concept 899

We measure the **co-occurrence** of a and b by asking:

- "Given that we find a at index t ..."

$$w_t = a$$

- what are the **chances** that we find b at index $t + j$?"

$$w_{t+j} = b$$

With this, we find our result:

$$P\{w_{t+j} = b \mid w_t = a\}$$

We did it! This is a clear, explicit probability.

$$\overbrace{(v_a \cdot v_b) \text{ is large}}^{\text{Similar vectors}} \iff \overbrace{P\{w_{t+j} = b \mid w_t = a\} \text{ is large}}^{\text{Occur together frequently}}$$

Notation 900

We can make this notation a little denser:

$$P\{w_{t+j} = b \mid w_t = a\} = P\{b \mid a\}_j$$

This assumes that t **doesn't** affect our probability: it doesn't matter **where** we found a , just **how far away** b is (and on which side).

- This is a reasonable assumption for our purposes.

D.1.6 Computing predicted probabilities

How do we turn a **real number** $v_a \cdot v_b$ into a **probability** $P(b \mid a)_j$?

- $P(b \mid a)_j$ is the chance of finding b at index $t + j$, if a is at index t .

$$\cdots w_{t-3} \underbrace{w_{t-2} \quad w_{t-1} \quad w_t}_{\text{What word is at } w_{t-2} \text{? Is it } b?} \quad w_{t+1} \quad w_{t+2} \quad w_{t+3} \quad \cdots \tag{D.9}$$

- So, we need to compare b to every other word that we could find at $t + j$: this is a

multi-class problem, using the softmax function.

We have one class for each possible word we could find at $t + j$.

$$\text{Softmax}(z_k) = \frac{e^{z_k}}{\sum_i e^{z_i}} \quad (\text{D.10})$$

Let's review the concept behind "softmax":

Concept 901

Suppose that we have n possible words (n "classes"), and we want to figure out which one is **correct**.

The k^{th} class has a score, z_k , used to compute probability.

- The bigger z_k is, the **more likely** k is to be the **correct class**.

To keep it **positive**, z_k is converted to e^{z_k} : each e^{z_i} competes to see which class is more likely.

- To create a probability, we **compare** the score of class k to all of our other classes, using **softmax**.

$$\underbrace{e^{z_k}}_{\text{Class } k} \quad \text{vs} \quad \underbrace{\sum_i e^{z_i}}_{\text{All classes}} \quad \Rightarrow \quad \text{Softmax}(z_k) = \frac{e^{z_k}}{\sum_i e^{z_i}}$$

- We repeat this process for every possible word i , to get all of our predictions.

Now, the big question: what is z_k ?

$$\underbrace{(v_a \cdot v_b) \text{ is large}}_{\text{Similar vectors}} \iff \underbrace{P\{w_{t+j} = b \mid w_t = a\}}_{\text{Occur together frequently}} \text{ is large}$$

z_k and $(v_a \cdot v_b)$ serve the **same purpose**:

- Large **dot product** predicts high probability.
- Large **z_k** predicts high probability.

So, we can use our dot product as a "score" z_k :

$$z_b = v_a \cdot v_b \quad (\text{D.11})$$

Now, we can plug this into our probability equation!

Key Equation 902

The **more similar** (bigger dot product) a and b are, the **more likely** we predict to find them together.

- We use a **softmax** to compute this probability for each possible word b .

$$P\{w_{t+j} = b \mid w_t = a\} = \frac{e^{v_a \cdot v_b}}{\sum_i e^{v_a \cdot v_i}}$$

Or, in alternate notation:

$$P\{b \mid a\} = \frac{\exp(v_a \cdot v_b)}{\sum_i \exp(v_a \cdot v_i)}$$

Ta-da! We've combined two separate concepts into a single equation.

Note that, in both top and bottom, we keep v_a : we're considering every possible word for w_{t+j} , while we know $w_t = a$.

D.1.7 Skip-gram approach: Training our word2vec model

One remaining issue: this equation doesn't tell us what the "true" probabilities are: they tell us the probability that our model **predicts**.

- Now, we have to choose a **good model** (word embedding).

Clarification 903

Our equation is $P\{w_{t+j} = b \mid w_t = a\}$ is our **estimation** for the probability.

- The real probabilities could be **different**: we'll design our word embedding to give us the most **accurate probabilities**.

First: what does our model look like? How do we even **generate** word embeddings?

- Often, we rely on a neural network.

Definition 904

We have two common **models for word embedding** (θ):

- Separately assigning a **vector** to each word.
- Using a shared **neural network** to embed every word as a vector.

Our neural network uses **parameters** θ . We'll use θ to represent our embedding, that we want to **train**.

$$w \xrightarrow{\theta} v_w$$

How do we pick a good model?

- We'll **train** our embedding θ , so that our **probabilities** are as accurate as possible.

As we established, our problem is multi-class classification:

Concept 905

Review from Classification chapter

For **multi-class classification**, we use the **negative log-likelihood multiclass** (NLLM) equation to compute **loss**:

$$\mathcal{L}_{\text{NLLM}}(g, y) = - \sum_{i=1}^n y_i \log(g_i)$$

y is a one-hot vector, so all terms of the sum except the "correct" term $i = k$ cancel out to 0:

$$-y_k \log(g_k) \xrightarrow{y_k=1} -\log(g_k)$$

g_k is the probability we assigned to the correct answer.

Next, we need training data: a body of **text**.

- For an example, let's visit index t in the text: this is the center of our **context window**.
- a is replaced by whatever word we find at that index: w_t . We still want to predict w_{t+j} .

$$\dots w_{t-3} w_{t-2} w_{t-1} w_t \dots w_{t+j} w_{t+j+1} \dots \quad (\text{D.12})$$

How good is our word embedding? According to NLLM: "how likely were we to correctly

predict w_{t+j} ?"

$$\mathcal{L}_{\text{NLLM}}(g, y) = -\log \left(\overbrace{\mathbf{P}\{w_{t+j} \mid w_t\}}^{\text{How likely we thought } w_{t+j} \text{ was, based on model}} \right)$$

The correct word for index $t + j$ would be... w_{t+j} . We can read the outcome from the text, and use our model to check how **likely** we thought that outcome was.

$$\mathcal{L}_{\text{NLLM}}(g, y) = -\log \left(\overbrace{\mathbf{P}\{w_{t+j} \mid w_t\}}^{\text{How likely we thought } w_{t+j} \text{ was, based on model}} \right)$$

Note that, the higher this probability is (the more sure we are of the correct answer), the closer the loss gets to 0.

Key Equation 906

We train our **word embedding** θ by **maximizing** the probability $\mathbf{P}(w_{t+j} \mid w_t)$ of predicting the **correct word** in each spot.

In our case, we want to **minimize**

$$\mathcal{L}_{\text{NLLM}}(\theta, j) = -\log \left(\mathbf{P}\{w_{t+j} \mid w_t\} \right)$$

where

$$\mathbf{P}\{b \mid a\} = \frac{\exp(v_a \cdot v_b)}{\sum_i \exp(v_a \cdot v_i)}$$

Now, we know how to compute these odds for a **single index**, $t + j$. We want to repeat this process for the rest of our context window:

$$\cdots \quad w_{t-3} \quad \underbrace{w_{t-2} \quad w_{t-1} \quad w_t \quad w_{t+1} \quad w_{t+2} \quad w_{t+3}}_{\text{All words within } c = 2 \text{ context window}} \quad \cdots \quad (\text{D.13})$$

Key Equation 907

We can find the **total loss** of our embedding θ , over our entire **context window**, by adding up the loss from each **context word**.

This includes all of the indices $(t + j)$, going from $j = -c$ to $j = +c$. Meaning, we want

- $|j| \leq c$ (within window)
- $j \neq 0$ (don't want to compare w_t with itself)

$$\mathcal{L}_t(\theta) = - \sum_{\substack{j \neq 0 \\ |j| \leq c}} \log \left(\mathbf{P} \left\{ w_{t+j} \mid w_t \right\} \right)$$

One more modification: the loss function above only computes loss for a **single context window**.

But, for a passage of text, there are many possible context windows: all we have to do is shift our target word, w_t .

- **Example:** Below, with $c = 2$, we'll show all of our possible context windows:

Target word is red, context words are blue.

This is a sample sentence
 This is a sample sentence

(D.14)

We'll average the loss over **all context windows**.

We'll ignore negative indices, so we don't cause problems when $t = 0$ or $t = T$.

Key Equation 908

Take a body of text with T words, and a **context window** with a **max skip distance** of c . We use word embedding θ .

Our **objective function** $J(\theta)$ for the **skip-gram word2vec** algorithm, over the entire passage, is:

$$J(\theta) = \frac{1}{T} \sum_{t=1}^T \mathcal{L}_t(\theta)$$

or,

$$J(\theta) = - \frac{1}{T} \sum_{t=1}^T \left(\sum_{\substack{j \neq 0 \\ |j| \leq c}} \log \left(\mathbf{P} \left\{ w_{t+j} \mid w_t \right\} \right) \right)$$

This is the completed loss function that we can use to train our embedding.

Concept 909

We can train our **embedding** θ using our **loss function** J , via **gradient descent**.

$$\theta' = \theta - \eta \nabla_{\theta} J$$

- If we're using a **neural network** to create embeddings, we'll need **back-propagation** to train.

D.1.8 Issues with skip-gram

There are a few problems worth addressing. First, look again at our equation:

$$P\{w_{t+j} = b \mid w_t = a\} = \frac{e^{v_a \cdot v_b}}{\sum_i e^{v_a \cdot v_i}}$$

Something might strike you: our probability is **totally independent** of which index $t + j$ we want to predict.

- That means, our model would make the exact same prediction for every nearby word.
- This is more easily resolved in our transformer model, so we won't worry about it for now.

Concept 910

Our **probability** calculation in skip-gram is **independent** of the **skip distance** j between words w_t and w_{t+j} .

- All words within the **context window** have the **same probability distribution**.

Another problem: if "more **similar**" means "more likely to **co-occur**", doesn't that suggest that we would expect a word to appear with itself, really often?

- This would be true for every word: nothing can be more similar to a vector than itself, after all.
- Our solution is to just **exclude** w_t from predictions about nearby words.

Concept 911

The most similar word to w_t , is **itself!**

- So, we often **exclude** w_t from predictions.

Another problem: our objective function $J(\theta)$ includes a logarithm. To optimize θ , we'd need to compute its **derivative**.

- This becomes really expensive, especially when our vocabulary can have millions of words.
- Our solution is to "**prune**" / remove a lot of words from our probability calculation.

We can predict in advance, that some words don't need to be included.

Concept 912

Skip-gram can become expensive to train, when the vocabulary becomes too large.

- So, we prune some unlikely words in our vocabulary, to speed up our predictions.