

# The Principles of Deep Learning Theory

An Effective Theory Approach  
to Understanding Neural Networks

DANIEL A. ROBERTS

*MIT*

SHO YAIDA

*Meta AI*

*based on research in collaboration with*

BORIS HANIN

*Princeton University*



CAMBRIDGE  
UNIVERSITY PRESS

# Contents

<b>Preface</b>	<b>ix</b>
<b>0 Initialization</b>	<b>1</b>
0.1 An Effective Theory Approach . . . . .	2
0.2 The Theoretical Minimum . . . . .	3
<b>1 Pretraining</b>	<b>11</b>
1.1 Gaussian Integrals . . . . .	12
1.2 Probability, Correlation and Statistics, and All That . . . . .	21
1.3 Nearly-Gaussian Distributions . . . . .	26
<b>2 Neural Networks</b>	<b>37</b>
2.1 Function Approximation . . . . .	37
2.2 Activation Functions . . . . .	43
2.3 Ensembles . . . . .	47
<b>3 Effective Theory of Deep Linear Networks at Initialization</b>	<b>53</b>
3.1 Deep Linear Networks . . . . .	54
3.2 Criticality . . . . .	56
3.3 Fluctuations . . . . .	59
3.4 Chaos . . . . .	65
<b>4 RG Flow of Preactivations</b>	<b>71</b>
4.1 First Layer: Good-Old Gaussian . . . . .	73
4.2 Second Layer: Genesis of Non-Gaussianity . . . . .	79
4.3 Deeper Layers: Accumulation of Non-Gaussianity . . . . .	90
4.4 Marginalization Rules . . . . .	96
4.5 Subleading Corrections . . . . .	100
4.6 RG Flow and RG Flow . . . . .	103
<b>5 Effective Theory of Preactivations at Initialization</b>	<b>109</b>
5.1 Criticality Analysis of the Kernel . . . . .	110
5.2 Criticality for Scale-Invariant Activations . . . . .	123
5.3 Universality Beyond Scale-Invariant Activations . . . . .	125

5.3.1	General Strategy . . . . .	126
5.3.2	No Criticality: Sigmoid, Softplus, Nonlinear Monomials, etc. . . . .	128
5.3.3	$K^* = 0$ Universality Class: tanh, sin, etc. . . . .	130
5.3.4	Half-Stable Universality Classes: SWISH, etc. and GELU, etc. . . . .	135
5.4	Fluctuations . . . . .	137
5.4.1	Fluctuations for the Scale-Invariant Universality Class . . . . .	139
5.4.2	Fluctuations for the $K^* = 0$ Universality Class . . . . .	141
5.5	Finite-Angle Analysis for the Scale-Invariant Universality Class . . . . .	146
<b>6</b>	<b>Bayesian Learning</b>	<b>153</b>
6.1	Bayesian Probability . . . . .	154
6.2	Bayesian Inference and Neural Networks . . . . .	156
6.2.1	Bayesian Model Fitting . . . . .	157
6.2.2	Bayesian Model Comparison . . . . .	165
6.3	Bayesian Inference at Infinite Width . . . . .	169
6.3.1	The Evidence for Criticality . . . . .	169
6.3.2	Let's Not Wire Together . . . . .	173
6.3.3	Absence of Representation Learning . . . . .	178
6.4	Bayesian Inference at Finite Width . . . . .	179
6.4.1	Hebbian Learning, Inc. . . . .	179
6.4.2	Let's Wire Together . . . . .	182
6.4.3	Presence of Representation Learning . . . . .	186
<b>7</b>	<b>Gradient-Based Learning</b>	<b>191</b>
7.1	Supervised Learning . . . . .	192
7.2	Gradient Descent and Function Approximation . . . . .	194
<b>8</b>	<b>RG Flow of the Neural Tangent Kernel</b>	<b>199</b>
8.0	Forward Equation for the NTK . . . . .	200
8.1	First Layer: Deterministic NTK . . . . .	206
8.2	Second Layer: Fluctuating NTK . . . . .	207
8.3	Deeper Layers: Accumulation of NTK Fluctuations . . . . .	211
8.3.0	<i>Interlude: Interlayer Correlations</i> . . . . .	211
8.3.1	NTK Mean . . . . .	215
8.3.2	NTK–Preactivation Cross Correlations . . . . .	216
8.3.3	NTK Variance . . . . .	221
<b>9</b>	<b>Effective Theory of the NTK at Initialization</b>	<b>227</b>
9.1	Criticality Analysis of the NTK . . . . .	228
9.2	Scale-Invariant Universality Class . . . . .	233
9.3	$K^* = 0$ Universality Class . . . . .	236
9.4	Criticality, Exploding and Vanishing Problems, and None of That . . . . .	241

<b>10 Kernel Learning</b>	<b>247</b>
10.1 A Small Step . . . . .	248
10.1.1 No Wiring . . . . .	250
10.1.2 No Representation Learning . . . . .	250
10.2 A Giant Leap . . . . .	252
10.2.1 Newton’s Method . . . . .	253
10.2.2 Algorithm Independence . . . . .	257
10.2.3 <i>Aside</i> : Cross-Entropy Loss . . . . .	259
10.2.4 Kernel Prediction . . . . .	261
10.3 Generalization . . . . .	264
10.3.1 Bias–Variance Tradeoff and Criticality . . . . .	267
10.3.2 Interpolation and Extrapolation . . . . .	277
10.4 Linear Models and Kernel Methods . . . . .	282
10.4.1 Linear Models . . . . .	282
10.4.2 Kernel Methods . . . . .	284
10.4.3 Infinite-Width Networks as Linear Models . . . . .	287
<b>11 Representation Learning</b>	<b>291</b>
11.1 Differential of the Neural Tangent Kernel . . . . .	293
11.2 RG Flow of the dNTK . . . . .	296
11.2.0 Forward Equation for the dNTK . . . . .	297
11.2.1 First Layer: Zero dNTK . . . . .	299
11.2.2 Second Layer: Nonzero dNTK . . . . .	300
11.2.3 Deeper Layers: Growing dNTK . . . . .	301
11.3 Effective Theory of the dNTK at Initialization . . . . .	310
11.3.1 Scale-Invariant Universality Class . . . . .	312
11.3.2 $K^* = 0$ Universality Class . . . . .	314
11.4 Nonlinear Models and Nearly-Kernel Methods . . . . .	317
11.4.1 Nonlinear Models . . . . .	318
11.4.2 Nearly-Kernel Methods . . . . .	324
11.4.3 Finite-Width Networks as Nonlinear Models . . . . .	330
<b><math>\infty</math> The End of Training</b>	<b>335</b>
$\infty.1$ Two More Differentials . . . . .	337
$\infty.2$ Training at Finite Width . . . . .	347
$\infty.2.1$ A Small Step Following a Giant Leap . . . . .	351
$\infty.2.2$ Many Many Steps of Gradient Descent . . . . .	358
$\infty.2.3$ Prediction at Finite Width . . . . .	373
$\infty.3$ RG Flow of the ddNTKs: The Full Expressions . . . . .	384
<b><math>\varepsilon</math> Epilogue: Model Complexity from the Macroscopic Perspective</b>	<b>389</b>

<b>A</b>	<b>Information in Deep Learning</b>	<b>399</b>
A.1	Entropy and Mutual Information . . . . .	400
A.2	Information at Infinite Width: Criticality . . . . .	409
A.3	Information at Finite Width: Optimal Aspect Ratio . . . . .	411
<b>B</b>	<b>Residual Learning</b>	<b>425</b>
B.1	Residual Multilayer Perceptrons . . . . .	428
B.2	Residual Infinite Width: Criticality Analysis . . . . .	429
B.3	Residual Finite Width: Optimal Aspect Ratio . . . . .	431
B.4	Residual Building Blocks . . . . .	436
	<b>References</b>	<b>439</b>
	<b>Index</b>	<b>445</b>