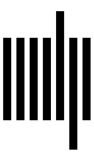
# Gradient Expectations

## Structure, Origins, and Synthesis of Predictive Neural Networks

By: Keith L. Downing

## Citation:

*Gradient Expectations: Structure, Origins, and Synthesis of Predictive Neural Networks*
**By:** Keith L. Downing
**DOI:** 10.7551/mitpress/14723.001.0001
**ISBN (electronic):** 9780262374675
**Publisher:** The MIT Press
**Published:** 2023



The MIT Press

# Preface

The main purpose of the brain, any brain, of any organism, is prediction—or so we are told by many prominent neuro- and cognitive scientists. I first encountered that claim in 2002, then again in 2005, and then again and again and again. By around 2007, I figured it was worth the time and effort to explore this fascinating hypothesis, so I started digging into the neuroscience literature—not an easy chore for a computer scientist. What I found were many areas of the brain whose structure and (purported) function made a lot of sense when viewed through *predictive glasses*, which I kept securely in place through many long hours of reading.

That personal journey led me to write two journal articles on systems neuroscience, with the brain's predictive machinery as a key focus. Unfortunately, each brain region appeared to be predicting in different ways, through complex interactions between a host of diverse neuron types and network motifs. As an artificial intelligence (AI) researcher, I found this particularly frustrating, since the primary AI interest in neuroscience has a very reductionist tint: we are a looking for *a few good neural principles* that mechanistically explain cognition and lend themselves to computer implementations (that ideally can plug-and-play in larger systems, such as self-driving cars, smart-home consoles, and so on). Neuroscience seemed to offer no such cheat sheet for intelligence, so I moved on to other endeavors.

In 2016, Andrew Clark (a popular philosopher among those of us who promote artificial life (ALife) approaches to AI) published *Surfing Uncertainty*, an enlightening account of the predictive mind. After digesting Clark's compact theory and the many examples that it convincingly explained, I knew that my return to the topic was inevitable. So in 2020, I began another lengthy investigation into prediction, but this time with the explicit goal of striking a healthy balance between neuroscience and connectionism, which, I quickly learned, had plenty to say about prediction, but only if you were willing to spend many hours absorbing the mathematics.

I invested those hours, and along the way it became clear that there was indeed a set of primitive computational mechanisms that both realized prediction in silico and, intuitively, presented perfect candidates for the building blocks of biology's expectation-producing strategies. Chief among these are *gradients*: relationships among factors that summarize the basic effect that a modification to one factor has on the other. From math class, we know these as *derivatives*: the change in Y divided by the change in X. Combining gradients with

a few other simple operations, such as summing, averaging, and comparing, yields a small kit of versatile tools that support computational predictions, many of which can also be realized by simple combinations of neural units.

Not coincidentally, gradients also play a large role in the field of deep learning (DL), which has come to dominate AI over the past decade. They represent precisely the same concept in DL as in math class, but now their complexity puts to shame those triple-starred exercises in the final only-for-students-destined-for-a-Cal-Tech-PhD chapter of your calculus book. As a college professor, I spend several lecture hours explaining them to my (very bright) students. These gradients are the heart of DL; in fact, a common moniker for DL techniques is *gradient-based methods*. If you skip all the gradient calculations, you will never have more than a superficial understanding of DL. They are that pivotal.

The title of this book has multiple meanings. First of all, the blatant knock-off of a Dickens' classic seemed more appealing than *Predictions from Derivatives: Finally, Something Useful from Those Math Classes You Slept Through*. With *expectations* from *gradients*, the same idea should come to mind. So *gradients* have clearly earned a spot on the marquee, as have *expectations*, as a synonym for *predictions*.

This word combination has a second, equally important, connotation for this book: the expectations for DL's gradient-based methods are enormous. So much of the hype and hyperbole surrounding AI stems from the legitimate successes of DL, but most of us AI folks bristle at the mention of the many utopian and dystopian visions of a future dominated by our AI tools, many of which are envisioned as basic extrapolations of contemporary DL achievements. Although this book is certainly not an attack on DL—other authors have seized that gauntlet—it does call some of these expectations into question, particularly those concerning an *artificial general intelligence* (AGI) based on gradient methods.

Despite this skepticism to a continued domination of AI by DL, my foray into connectionism revealed many neural network designs from the late twentieth century that clearly accentuated the role of predictive machinery in cognition. Invented by some of the same people who drive the DL revolution in the 2020s, these older networks have not gone quietly into the night; their principles continually reemerge in nascent systems that try mightily to replace the complex, biologically unrealistic derivatives of DL with simpler, local (and thus biologically plausible) gradients. In so doing, these networks manifest the tight synergy between recognition (of sensory patterns) and prediction (of future patterns) that many view as fundamental to actual understanding. After all, we exhibit some of the deepest levels of comprehension by marrying our concept-identifying faculties with examplar-generating skills. Nobody has ever seen a mastadon on a putting green, but we can predict what such a scenario might entail; and by doing so, we reveal considerable deep knowledge of prehistoric mammals and golf courses.

The three components of this book's subtitle, *structure, origins*, and *synthesis*, refer to three primary subject goals of this work. First, the primitive predictive mechanisms can combine in many ways to produce predictive neural structures, as seen in both brains and artificial neural networks. The various primitives and resulting structures form the basis of the first three chapters of the book, with high-level conceptual explanations in chapter 1, more of the mathematical flesh and bones in chapter 2, and then the various neural implementations in chapter 3. Additional neural network structures appear throughout the book, including chapter 4, which delves into the older connectionist models mentioned above.

Chapter 5 focuses on *predictive coding*, a well-known principle dating all the way back to the 1950s, but still very prevalent in most discussions of predictive neural networks, several of which appear in this chapter.

The second goal, *origins*, takes center stage in chapter 6, which paints a picture of how predictive networks may have evolved, beginning with the simplest organisms and continuing on up through the mammalian brain and several of its subdivisions. This is far from a complete phylogenetic tree of predictive progress, but it helps set the stage for chapter 7, where the third goal, *synthesis* becomes the primary theme. My main interest and belief lies in evolutionary approaches to neural network design—a view that is anathema to many DL experts. Chapter 7 addresses the competition and cooperation between gradient-based and evolutionary routes to synthetic intelligence, before using some of the concepts from chapter 6 as the conceptual basis for an emergent predictive-network system that includes the three key adaptive mechanisms of this (and my earlier) book: evolution, development, and learning. The topic of synthesis arises throughout the book as discussions (both general and specific) concerning how various predictive techniques have been or could be implemented, but synthetic issues reach a head in chapter 7. Finally, chapter 8 summarizes the earlier chapters, derives several generalizations from them, sketches a few predictions of its own for the future of AI, and then boldly refuses to pick sides.

This book's target audience is anyone with a deep interest in intelligence and how neural structures might achieve it. This, quite naturally, pertains to college students who study psychology, neuroscience, or AI; but the main concepts should be accessible to anyone with an interest in cognition and the patience to follow my virtual dissections of neural networks, both simple and sophisticated, biological and artificial.

Although mathematics appears throughout the book, only chapter 4 burrows into it very deeply, as I try to do more than just hand-wave at the ties among crucial concepts such as prediction, recognition, information, surprise, and free energy. However, for those who trust a hand wave, the many gray boxes filled with derivations can easily be skipped with no repercussions in later chapters.

Prediction is a popular topic, of which many books have been written. My original thought was, essentially, *Who needs another one?* But as I read through many of them, I found myself wanting more concrete mechanisms, more of the dirty little details that make things resonate in my own mind. So although this book is hardly the definitive manual of a prediction machine, I hope it gives you a feel for how the high level expectations of mice, macaws, and men ground out in patterns of neural activity . . . and how scrutinizing these vertical slices, from the coarse spatiotemporal scales of ethology down to the microns and milliseconds of neuroscience, might help us improve the intelligence of our machines.


The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Times New Roman by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Downing, Keith L., author.
Title: Gradient expectations : structure, origins, and synthesis of predictive neural networks / Keith L. Downing.
Description: [Cambridge, Massachusetts] : The MIT Press, [2023] | Includes bibliographical references and index.
Identifiers: LCCN 2022037237 (print) | LCCN 2022037238 (ebook) | ISBN 9780262545617 (paperback) | ISBN 9780262374682 (epub) | ISBN 9780262374675 (pdf)
Subjects: LCSH: Deep learning (Machine learning) | Neural networks (Computer science) | Conjugate gradient methods.
Classification: LCC Q325.73 .D88 2023 (print) | LCC Q325.73 (ebook) | DDC 006.3/2—dc23/eng20230302
LC record available at https://lccn.loc.gov/2022037237
LC ebook record available at https://lccn.loc.gov/2022037238