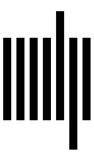
Gradient Expectations

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

8 Conclusion

8.1 Schrodinger's Frozen Duck

Turning left off of E6, Norway's main interstate highway (a two-laner with a decent shoulder), we begin the final 30 kilometers (on a two-laner with no shoulder and sporadic midline markers) of our yearly Christmas trek to Mormor's (Grandma's) house. At this exact point in the journey, the excitement mounts, not so much for the anticipation of presents under the tree, but for the start of yet another ridiculous family contest: predicting the temperature at Mormor's house, located in one of Norway's coldest towns, Folldal.

At the turnoff, my wife and I read aloud the current outdoor temperature, often around −15C; then everyone has to pick a unique integer. The backseat is alive with calculational cacophony as each child promotes their number with a geographic/meteorologic justification. There is little elevation change from the turnoff to Folldal center, but from there up to Mormor's house is a steep climb with anywhere from a 5 to 15 degree temperature swing, which is often positive, since cold air sinks.

The arguments rage for a kilometer or two before we force everyone to lock in their prediction. There are the occasional, illicit attempts to text ahead to Mormor to get a mercury reading, but the kids tend to self-police such blatant cheating attempts. Elevated vigilance levels pervade the tense atmosphere as the minute-by-minute temperature readings allow everyone to evaluate their winning prospects. Gradients of temperature change per kilometer also enter the picture: *The temperature dropped 3 degrees since the turnoff, and we're now halfway to Folldal, so . . .*

Nothing is decided until we actually pull in the driveway and park the car, since the temperature has been known to change several degrees in the last 100 meters of the trip, during the final ascent. The winner gets bragging rights for the frigid holiday season, of which they tend to take full advantage.

Wrapped inside this half-hour contest are a host of smaller-scale predictive acts, such as the numerous expectations that the driver employs to safely navigate treacherous backroads across a frozen landscape. Flashes of light when entering sharp turns reveal oncoming vehicles, while reflections from the pavement provide invaluable warnings of upcoming black ice, and odd roadside background shifts may be the only clue one gets of a crossing moose. In the backseat, any suspicious roll toward the door may be a dead giveaway of cell-phone cheating.

Beneath these moment-by-moment predictions are split-second brain activities involving expectations, some explicit (conscious and verbalized) and others merely implicit, but all supporting the decision making of the competitors and driver. And all involve predictions more about future brain states than about future states of the world as a whole, since, for the most part, the temperature at Mormor's when we reach the turnoff is the same as when we reach her driveway. It's our awareness of that reading that lies in the future, and for the sake of the contest, that is all that matters.

This subjective nature of *future* becomes evident in another consequence of my geographical situation: delayed awareness of American sporting outcomes. As an avid fan of the University of Oregon Ducks, I am frequently plagued by the nine-hour time difference between Scandinavia and Oregon. I retire on Saturday evenings in heightened anticipation of an Oregon football or basketball game that will start and end during my slumber, and my dreams often include wild predictions of win and loss scenarios. Could the team mascot score the winning points? Could the pounding Oregon rain suddenly freeze and turn a football game into a hockey match? I awaken on Sunday morning into *Schrodinger moments*: until I open my digital tablet, the ducks are both victorious and defeated *where it counts*, in my mind.

Of course, my state of mind means nothing to Las Vegas bookies; when the final whistle blows on the field or court, speculations of future outcomes become objective exchanges of cold hard cash, and the case closes for all intents and purposes. In contrast, I hope that this book has given readers an appreciation for prediction as a very personal internal phenomenon, one that is rich and expansive in scope, yet grounded in simple, ubiquitous, neural mechanisms.

This book began with that famous quote by Yogi Berra: It's tough to make predictions, *especially about the future.* After a few hundred pages, the oxymoron in that statement may have lost some vigor. When expectations involve projections about a subjective information state concerning present and past world states, the bond between prediction and future attenuates. Through that weakening, the concept expands in scope, which I hope has encouraged readers to see the broader perspective of prediction.

8.2 Expectations Great and Small

Underlying any overt prediction, such as that of tomorrow's soybean futures, run a host of declarative and procedural expectations, many of events as simple as the upcoming activation levels of particular neurons. At the end of chapter 7, I used the term *functional fractalization* to highlight the similarity of the overt and low-level functions / purposes: they are all predictions. However, akin with most functional decompositions, they clearly involve different mechanisms and scales.. Although a stack of cortical columns presents a nice image of modular predictive coders, and possibly a productive computational model for certain tasks, it omits many key differences between peripheral and internal neural processing. These discrepancies are vital to a neuroscientist but anathema to AI researchers looking for a short list of basic principles that can explain as much of intelligence as possible. So those of us in sciences of the artificial will jump at opportunities to flatten our learning curves with a juicy serving of the finest abstractions. Prediction is one such meal.

This need not paint a reductionist picture of intelligence as *prediction all the way down*. Rather, the moral of this and other books (Llinas 2001; Clark 2016; Buzsaki 2019) might be closer to this:

Everything is not prediction, but prediction is everywhere.

The predictive primitives described in these chapters (delays, gradients, averages, and so on) are ubiquitous in neural networks (both natural and artificial), but they do not always combine and interact in support of prediction. Still, their omnipresence within a diverse collection of neural circuits makes the emergence of predictive motifs nearly unavoidable. Facilitated variation, development, and learning provide means, and selection provides myriad motives.

Proficient predictors surely have a selective advantage over organisms whose mental world is confined to the present. A good deal of intelligent behavior involves supplementing the present with enough relevant memories of the past to imagine the future. Of course, fully knowing the future is impossible, but having biases that correlate well with it can only increase one's odds of survival. And, as detailed earlier, the temporal disparities between fast motor actions and slow sensory processing in most species rewards any mechanisms that can provide reasonably accurate hints as to near-future sensor readings. At the cellular level, basic constraints such as energy and information efficiency would favor predictive-coding schemes that can reduce neural firing as much as possible while still maintaining the essential information coupling needed to keep a body running smoothly, both inside and out.

As overviews of the cerebellum, hippocampus, basal ganglia, and neocortex should indicate, predictive circuits can take many forms; and the function of expectation generation can peacefully coexist and cooperatively dovetail with vital faculties such as memory, perception, and action selection. Unfortunately, each of these neural structures exhibits a level of cellular heterogeneity and behavioral complexity that precludes any quest for basic principles of neural prediction. These are hardly the frictionless planes of neuroscience. However, the myriad species- (or class-) specific predictive topologies further highlight a selective pressure to develop something (anything) that can bring an organism's visions of tomorrow a little closer to today.

8.3 As Expected

The search for general predictive principles may also lead deep into development, where cooperative and competitive interactions among neurons and neural groups (fortified by Hebbian STDP) may have led to the (nearly inevitable) emergence of predictive motifs consisting of top-down signals, delays, inhibitors, comparators, and bottom-up error signals. In short, prediction and control may have arisen as natural consequences of facilitated variation's modularity and weak linkage (Kirschner and Gerhart 2005), neural Darwinism's *survival of the best networkers* (Edelman 1987), and displacement theory (Deacon 1998).

Since the natural world as *we* experience it puts a premium on predictive competence, the fact that evolution found respectable forecasting techniques should come as no surprise. However, a sloth lives under different constraints than a human; its actions are very slow. One can imagine many worlds, based on carbon or silicon, where sensing and acting have more similar timescales—or where vision, working, after all, with the fastest known entity (light),

fully eclipses motricity—and thus the obvious need for prediction diminishes. Of course, without fiber-optic brains, we still need to differentiate the speed of waves and particles from the speed at which our default machinery can interpret them. Robots may circumvent this problem: a well-designed humanoid may have very high-frequency sensing but slower motricity, particularly if it should safely interact with humans. Thus, from the sensorimotor perspective, a robot may have less need for prediction than a human.

However, any truly intelligent agent, real or artificial, requires lookahead: the ability to envision and evaluate future options, only some of which will actually come to pass. Unless it can physically backtrack (i.e., run) at the speed of light, an agent must occasionally commit to irreversible choices whose consequences will be known only in the future. For example, a chieftain preparing to visit an unfamiliar new tribe may have to choose between spears and body armor, or gifts and festive garb, with serious ramifications for either mismatch. The contents of a fisherman's boat as he heads out to sea can have extreme repercussions later in the day, with no second chances for repacking. Carefully designed experiments show that even ravens can plan ahead by choosing the proper tool for a task that they will encounter minutes or hours in the future (Kabadayi and Osvath 2017). A good deal of higher intelligence would simply not be possible without prediction.

In *Intelligence Emerging*, I wrote a lot about search and its essential contribution to the emergence of intelligence, across multiple spatiotemporal scales. Predictive machinery is one of the golden nuggets that all that search eventually found. It props up very high points in the fitness landscape. A revised theory of facilitated variation (Kirschner and Gerhart 2005) might one day include predictive coders in its list of *core components* found by evolution and then repeatedly exploited to ratchet up complexity and intelligence.

Thus, it seems that predictive machinery has eventually emerged, *as expected*, given the world in which we live and evolve. Any crystal-ball mystique surrounding prediction has hopefully been dispelled by these chapters, which both broaden the range of predictive activity and also give indications of how the overt predictions of our daily lives have reasonable neuroscientific explanations, many of which involve the brain's own version of forecasting. Expectations are a natural part of our evolutionary past and present, and the future prospects for automated prediction seem bright.

8.4 Gradient Expectations

The recent, wild successes of deep learning (DL) have created overwhelming expectations for both itself and AI as a whole. Though many of the predictions, such as fully autonomous automobiles, have had *reality checks* as researchers understand the difficulties of the remaining (peripheral but essential) aspects of the problem (i.e., recognizing pedestrian intentions), these have not derailed DL. There are too many triumphs to simply discard these techniques as *science fiction*[1] and scurry back to the safety of more traditional science and engineering approaches, those whose results can be verified mathematically, or empirically, in the lab.

As of around 2012, DL has proven that gradient-based methods perform expertly in a wide range of domains, when given enough data and computing power, both of which have become abundant in the past decade. Long-distance gradients have withstood stellar challenges from the population-based search methods of evolutionary computation, and they have partnered with the trial-and-error search techniques of reinforcement learning (RL)

to form the DRL colossus, which has brought the world's chess and go champions to their knees—and all in the absence of expert domain knowledge and human-game data cases. DRL success now requires only a good programmer and powerful machines. DRL workers have masterfully weaned themselves from the data-as-oil that many companies have leveraged to attract AI partners. For better or worse, AI progress can continue quite independently in those areas where the basic principles and rules are freely accessible (e.g., in textbooks) and available computation permits extensive self-investigation / self-play by a digital agent.

The looming question is whether DL and its gradients can get us all the way to artificial general intelligence (AGI). Skepticism abounds (Mitchell 2019; Larson 2021; Hawkins 2021), as does the push for a return to biology for more hints and inspiration (Hiesinger 2021; Soltoggio, Stanley, and Risi 2018; Miller 2021). At the same time, nascent Hebbian learning methods, based on predictive coding (Whittington and Bogacz 2017), can preserve many of the powers of backpropagation. These local mechanisms, along with some of the other predictive circuitry discussed in this book, could carry neural networks beyond the jumbo-gradient era and into a more versatile, biologically realistic future.

A broader philosophical view of AGI and its origins further weakens arguments about the supreme importance of DL and super-sized gradients. A common artificial life (ALife) perspective on AGI is that it can only arise via the interactions of agents with other agents and their environment over time periods spanning many generations. As (most) early AI researchers eventually realized, you cannot simply pound AGI into a machine as logical rules for understanding and behaving in the world; and as (some) contemporary DL adherents would probably admit, you cannot feed millions of examples into a neural network and expect it to induce all salient generalities from those images, text, sounds, and the like. The value of actually exploring the world and generating data oneself has been grossly underestimated by much of AI, although self-play in DRL systems such as AlphaZero has surely opened many eyes and minds to the possibilities and advantages.

As discussed earlier, prediction is a nice trick for creating personal data sets for training ML systems to generate expectations from current states. Thus, supervised learning can surely play an important role in AGI. However, our brains probably do a lot more unsupervised and reinforced learning than supervised, as implied by the proposed functional breakdowns of various brain regions, such as neocortex, hippocampus, and cerebellum, where only the latter exhibits anything close to truly supervised learning (Doya 1999). We cannot employ semantic deceptions such as calling DL's autoencoders unsupervised learning just because the input and target are the same; the training algorithm is still backpropagation. Of course, this book argues for a different semantic stretch: viewing one level as producing a target value for the predictions of a neighbor layer, with the ensuing prediction error providing all the feedback needed for effective learning. Neither can the billions of local gradients in this extensive population of PID controllers pave the whole road to AGI on its own.

The expectations generated by gradient-based methods seem overblown and unrealistic to anyone who takes seriously the connections between nature and engineering. As I argued extensively in *Intelligence Emerging*, the role of persistent but relatively random search in all phases of evolution, development, and learning seem absolutely fundamental to the design of cognitive machinery. Readers of this book have hopefully gained an appreciation for a wider array of gradients, how they enable prediction, and how the mechanisms for handling local gradients fit nicely into various accounts of the evolution of intelligence. In this

way, local gradients and prediction constitute important bio-inspirations for the continuing pursuit of AGI.

8.5 Expecting the Unexpected

Despite many years in the field, I am probably no better a predictor of AI's future than, say, Yogi Berra. If my technological forecasts over the years had been wagers, I would surely be penniless. The predilections formed by several decades of deep technical immersion (in the neurons of the brains of the bugs in the trees) often cloud one's view of the forest. My only antidote to this myopia is reading higher-level accounts written for the general public; and in the past decade, a flood of such books, on AI, have hit the market.

Some paint a very grim picture of a denuded planet ruled by robots, with us as their slaves. Others strike a more reasonable balance of power in which the human-machine cooperative reaches highly productive and socially and emotionally pleasing levels. It's hard to avoid a sense of AI awe when reading a nearly perfect DL translation of English to French; or viewing microscopic images of tumors that machines, but not humans, could detect; or watching a robot hurdle obstacles like an Olympic champion. But, giving pause to even the most optimistic futurist are tales of racist bots that actually know nothing of race or suffering, medical ML systems having no understanding of basic biology, image classifiers thrown off by a little piece of tape on a stop sign, and "autonomous" robots controlled by an engineer hidden behind a shopping-mall palm tree. I am often tempted to view the current (very hot) AI summer as little more than prelude to another AI winter, as the roller-coaster of hype, hope, and disillusionment rumbles on.

Ray Kurzweil (2012, 2005) is one futurist whose visions resonate well with my own interests and experiences. One of his key speculations concerning the *singularity* (a term coined by Vernor Vinge in the 1980s) is an accelerating feedback relationship between AI and neuroscience. As AI, robotics, and hardware improve, so does our ability to measure and interpret neural data, and thus our ability to understand the mind. This enhanced comprehension of the brain then cycles back to fortify our AI systems. As the hardware and algorithms improve, so too does the frequency of this cycle, eventually leading to a state in which the AI bots are intelligent enough to design their own neural theories, and the robots have the dexterity to physically test them. Humans are then politely ushered out of the loop, and into the cheap seats in the back of the arena, and progress proceeds at lightning speeds, yielding systems that dwarf human skill and intelligence. Stop the level-5 autonomous Lamborghini; I want to get out and walk!

Putting dystopian drama aside, the positive feedback between AI and neuroscience at the core of Kurzweil's theory has very positive overtones of mutual progress for both fields. Although the ratio of intersection to union of these two disciplines is small, there are a sufficient number of researchers worldwide with the cross-disciplinary fortitude to investigate intelligence from both the natural and artificial perspectives. Still, a great many pilgrimages by AI workers to nature's holy ground result in little satisfaction, spiritual or otherwise. The bug of biological inspiration bites many, but the victim grows weak of trying to turn the fascination into a competitive algorithm. But just as we continue to trawl the vast rainforests for miracle cures, there is no reason to discontinue our trips across campus to the neuroscience department.

AI's involvement in another, quite different, positive feedback has decidedly negative consequences: the viral spread of outrageous hyperboles, vicious insults, and bald-faced lies. We live in very unnerving times, when democracies have begun a spiraling decline fueled by ever-growing fears and animosities, for which AI has played no small part. So many of these incredulous fantasies and conspiratorial chimera tear through layers of cyberspace unhindered, never meeting reality at sorely needed comparators, but instead feeding back into and magnified by the overheated matrix. It is imperative that humanity, in tandem with our technologies, actively addresses the dangers that have been so widely and loudly anticipated, before all red lights are on, and all bets are off.

The simple act of juxtaposing fiction (no matter how innocent) with fact, identifying the differences, and using them to improve the quality of our collective reporting (to legions of followers) would go a long way. Error and failure are, after all, the catalysts of learning and improvement, but only when recognized and admitted. In *Great Expectations*, Dickens (1861) addresses this adaptivity that often emerges from hardship:

I have been bent and broken, but—I hope—into a better shape.

Unfortunately, Yogi Berra paints a different picture:

The future ain't what it used to be.

It may be just as enlightening to let AI predict its own future. When given the first five italicized words, DeepAI's text-generation system completes the thought (and artfully segues into acknowledgments):

In the future, artificial intelligence will be able to learn what you thought you knew, but humans will just have to figure out the difference from what you actually had. I want to thank the team for working so hard to find a way to make this possible and thank my backers for being involved.

Whether the classics, the comical, or the artificial should serve as our guide to the future is anyone's guess. But most readings of the technological gradients indicate that AI will play an oversized role. Ours is to pay extremely close attention.

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