The Boundedness Illusion: Projections from early evolution can underestimate evolutionary potential

Michael J Wiser[1], Emily L Dolson[1], Anya Vostinar[2], Richard E Lenski[1], and Charles Ofria[1]

[1]Michigan State University, East Lansing, MI 48824
[2]Grinnell College, Grinnell, IA 50112
mwiser@msu.edu

Abstract

Open-ended evolution researchers seek to create systems that continually produce "new" evolutionary outcomes, attempting to mimic the power and diversity of evolution in nature. The specific metrics used (novelty, complexity, diversity, etc) vary by researcher, but the holy grail would be a system where any of these can accumulate indefinitely. Of course, one challenge that we face in reaching this goal is even recognizing if we have succeeded. To determine the evolutionary potential of a system, we must conduct finite experiments; based on their results we can predict how we would expect evolution to progress were the run to have continued. Here we begin to explore how such predictions might be made and how accurate they might be. In this initial study, we focus on predicting fitness; this metric can be easy to calculate, and often correlated with increases in traits like novelty and complexity. We find the best fit to measured values of fitness in a simple digital evolution experiment, and demonstrate that projecting this fit out usually predicts that fitness will be constrained by an asymptote. Extending the experiments, however, we see that fitness often shoots past this asymptote, belying the boundedness that it implies. Extending past a premature end point allows us to see through this "boundedness illusion".

Introduction

Evolution has produced an astounding degree of diversity in biological life. Current evidence suggests that this has been going on for 3.8 billion years on Earth (Mojzsis et al., 1996; Rosing, 1999), and yet evolutionary innovation is far from over. New strains of respiratory diseases circulate yearly (Grenfell et al., 2004; Nelson et al., 2007). Populations of predators continuously race against prey in co-evolutionary cycles, where changes in one population alter selection pressures on the other, and those changes in the second population reflect back as new selection pressures on the first (Abrams, 2000). Even a long-term laboratory study of *E. coli* in a simple glucose-rich environment provides no evidence of evolution reaching an endpoint (Wiser et al., 2013). Natural systems are thus characterized as undergoing open-ended evolution, showing continual generation of change, novelty, complexity, and diversity (Taylor et al., 2016).

Unlike in natural systems, computational evolutionary systems often seem to show rapid change early followed by stagnation (Lampinen and Zelinka, 2000; Piotrowski, 2014). After a sufficient time of little to no change, researchers conclude that the system has reached an endpoint. However, there are some potential pitfalls to this conclusion. For one, not all curves that decelerate are bounded. Power law relationships, which are common in both biology (Clauset et al., 2009; May, 2006) and physics (Adams et al., 1993; Newman, 2005), show rapid changes early on, followed by decelerating changes. But power law relationships are not asymptotic[1]; there is no upper limit to the function the way there is is in, say, a rectangular hyperbola. Secondly, and perhaps more importantly, the experiments we conduct are, by nature, finite; they run for a specified length of time, and stop. As a field, we often use these experiments to make predictions about what will happen over much longer, sometimes indefinite time scales (Sajjad et al., 2016), but these predictions are, by their very nature, extrapolations. Portions of a curve that look like stagnation when viewed at one scale can be part of a steep increase when viewed over a larger interval. Further, certain features of an evolutionary trajectory – for example, a rare event of large magnitude in the midst of more common events of relatively small magnitude, as predicted by notions of punctuated equilibrium (Eldregde and Gould, 1972) – may cause a trajectory to *appear* bounded, even when the general pattern is *unbounded*. We term this discrepancy the "boundedness illusion".

Selection can act upon many different features to optimize individuals within a population. Ultimately, these selected features influence Darwinian fitness, which is a measure of how much genetic information an individual contributes to future generations. Indeed, biologists measure fitness in terms of contribution of genetic material to future gene pools, whether directly through competitions over generations (Wiser and Lenski, 2015) or through features that are components of fitness but are easier to measure than an

---

[1] Here we consider increasing power law functions, where the exponent is positive. When the exponent is negative, the function will decrease and may have an asymptote (depending on where the independent variable is relative to the exponent), as the function cannot cross 0.

integrated fitness measure in some systems, such as maximum growth rate in a population (Gerstein and Otto, 2011), number of young fledged (Velmala et al., 2015), or seeds set (Remold, 2002). In any population subject to selection pressures stronger than neutral drift, the expectation is that, on average, fitness will increase from one generation to the next (Orr, 2009). How fitness changes over time is tightly linked with whether the evolution in a system is open-ended, or is itself bounded by an upper value. By focusing our analysis on fitness, we expect to address a fundamental driver of other aspects of open ended evolution as well.

Many researchers see fitness in existing computational systems as inherently bounded (reviewed in (Taylor et al., 2016)). In one sense, there must be a global fitness optimum within a computational system. Theoretically there are a finite number of possible organisms, due to limits of computer memory, if nothing else (Davies, 2004). As such, every individual that could exist has an explicit fitness value for a give computational environment, with the highest such value being a fundamental bound. However, the eventual theoretical bound on a population's fitness may not translate into a practical bound. The potential search space in non-trivial computational evolutionary systems is vast. If the time it would take a system to reach its upper boundary of fitness, on average, is longer than what remains before entropic heat death of the universe for example, that boundary is likely to be irrelevant over experimental time scales. Shorter time scales – perhaps merely trillions of years – may show unbounded increases in fitness, regardless of any theoretical upper bound in the system. Similarly, while the whole genomic search space of individuals may be technically ergodic – any state is theoretically reachable from any other state – how long it would take to transition from one particular state to another may make the space practically non-ergodic, even if it theoretically is so.

To make sense of the vast genetic search space available in multi-objective optimization, researchers often turn to the concept of adaptive landscapes. First described by Sewall Wright in 1932 (Wright, 1932), adaptive landscapes are a heuristic tool for how traits contribute to the fitness of an individual. Every point along the landscape represents a combination of trait values. When researchers assign a fitness value to each one of those points, a multidimensional landscape emerges, with peaks and plateaus of high fitness, and valleys or basins of low fitness. As a population reaches a fitness peak, its adaptation slows and eventually stops if that peak is the highest one in the landscape. The fitness landscapes associated with non-trivial problems will themselves be complex, and typically exist in far more than three dimensions. For reasons directly analogous to how improvements in fitness may be theoretically bounded but practically unbounded, so too may a fitness landscape be treated as infinite even when it is theoretically finite. Beyond even these limits, natural systems rarely, if ever, exist in perfectly static environments; changes in resources, competing populations, and predators or prey introduce additional variation in the form of dynamic landscapes, which are beyond the scope of this paper.

Background

Open-ended evolution

Is there a fundamental difference between evolutionary dynamics in populations of DNA-based organisms and evolutionary dynamics in today's artificial life systems? If so, this would be important to understand when drawing inferences from computational systems. Moreover, figuring out the cause of such a difference would provide insight into the conditions that are necessary and/or sufficient to produce evolution akin to what we observe in nature. These topics are the focus of research on open-ended evolution (Taylor et al., 2016).

In order to address these questions, we must first identify the dynamics that we expect an open-ended evolutionary system to exhibit. Previously, we proposed that these dynamics can be lumped into five categories: the potential for continuous meaningful change, the potential for the continuous production of meaningful novelty, the potential for unbounded growth in the complexity of ecological communities, the potential for unbounded growth in the complexity of individual organisms, and the potential for major evolutionary transitions in individuality (Taylor et al., 2016). We specify that these dynamics must be "meaningful" to clarify that it is not sufficient to continuously produce novel individuals via deleterious mutations that will be rapidly purged from the population; there must be new genetic information to count as novel.

Note that "unbounded growth in fitness" is not one of our categories. It is, however, inextricably interwoven with all of them. Over the time frames and population sizes commonly used in artificial life experiments, unbounded growth in fitness is sufficient to continuously produce change and novelty, as it requires at least an occasional introduction of new, fitter genotypes. Moreover, most ways of filtering for "meaningful" evolutionary dynamics rely in some way on fitness. Previously, we suggested an approach in which genomes are simplified down to sites that affect fitness (*i.e.* informative sites), and populations are filtered to only include genotypes that survive for a (Taylor et al., 2016). Unbounded growth in fitness should facilitate the evolution of new genotypes that pass these filters and are thus able to contribute to metrics of open-endedness.

Other approaches to quantifying open-ended evolution are also closely tied to fitness and boundedness. For example, in Bedau et al.'s Evolutionary Activity Statistics, the population is split into taxonomic "components" (Bedau et al., 1998). The evolutionary activity of these components is measured by the length of time they persist in the population. Evolution in the system being measured can then be placed into a class of open-endedness based on whether the diversity and cumulative evolutionary activity of these components are bounded or unbounded, and whether new components keep being generated. As before, when these statistics are being used on a computational system, it is necessary to filter out noise in some way. This is traditionally done by comparing runs of the system being measured to runs of a "shadow" version of that system in which all fitnesses are equal. Evolutionary activity in the shadow run is then subtracted out from that observed in the real run to get a measurement of non-trivial evolutionary dynamics.

Unbounded growth in fitness is sufficient to create a scenario in which novel components keep appearing and the cumulative activity of components is unbounded. However, it is not sufficient to guarantee unbounded growth in the coexisting diversity of components. There is some disagreement over how to classify this scenario (Channon, 2003; Kitto, 2006), but it is generally agreed to be at least somewhat open-ended. Thus, unbounded fitness growth dynamics also have clear implications for a system's open-endedness under this framework as well. Additionally our observations about boundedness have the potential to apply to boundedness in the context of evolutionary activity statistics as well.

Ultimately, in this work we focus on open-endedness in terms of whether populations stop adapting because they reach optima on the adaptive landscape. We argue that this is a useful lens through which to examine open-ended evolution because it facilitates comparison to experimental evolution in a laboratory setting while simultaneously having implications for conventional artificial life conceptions of open-ended evolution. By bridging these two bodies of research, we can refine our expectations for the behavior of all evolving systems.

Boundedness

Much of the debate about open-ended evolution concerns whether evolutionary trends within study systems will continue forever, or whether there is some eventual limit (Lampinen and Zelinka, 2000; Piotrowski, 2014). Biological evolution is commonly assumed to be open-ended, and therefore unbounded, while most researchers argue that computational systems are inherently closed because of an inevitable limit to evolutionary potential (Taylor et al., 2016).

It is important to note that there is a difference between whether a system is theoretically unbounded versus practically unbounded. The number of positive integers is theoretically unbounded: there are an infinite number of positive integers. The number of positive integers any human will ever say aloud is theoretically bounded: given that the universe will eventually end, and that it takes non-zero time to say a number, there will be numbers never said. But the number of positive integers some human will ever say may be practically unbounded at the moment, because we are so far away from that theoretical upper limit that the existence of that limit has no discernible impact on the present. Anything we wish to measure would eventually run into physical constraints based on physics the amount of matter and energy in the universe may be incomprehensibly vast, but it is still eventually finite, and that places an upper bound on information, and therefore complexity (Davies, 2004). Those eventual universal physical limits, though, do not always matter in a practical sense; if we do not expect to come close to the theoretical limit, does it really matter that such a limit exists?

Adaptive landscapes

The adaptive landscape is a conceptual tool imported from evolutionary biology (Wright, 1932). In an extreme simplification, imagine a three-dimensional landscape (see Fig Conceptual Fitness Landscape?). The x- and y- axes represent two traits of relevance to evaluating an individual. In an organic biological system, these could be traits such as limb length and maximum running speed; in a computational system they could be traits such as total resources collected and number of messages sent to other individuals. The z-axis represents the fitness value associated with that specific combination of x and y values. When fitness values are viewed across a range of the x and y values, high fitness regions rise as hills, while low fitness regions sink as valleys, resulting in the landscape of the metaphor. Of course, true fitness landscapes are unlikely to be only 3-dimensional; it is exceedingly rare for only two traits to influence fitness (though see (Ganco and Hoetker, 2009) about the use of NK-landscapes in research). Nevertheless, discussions of the adaptive landscape are pervasive in the literature in both evolutionary computation (Vassilev and Miller, 2000; Yu and Miller, 2001; Islam et al., 2012) and evolutionary biology (Page and Nowak, 2002; Mahler et al., 2013; Martin and Wainwright, 2013; de Visser and Krug, 2014), with a great deal of attention paid to how populations can move from a local optimum to a better optimum despite a region of poor fitness in between (Covert et al., 2013; Chou et al., 2014; Kvitek and Sherlock, 2011).

In toy problems, researchers can map out the adaptive landscape in its entirety. This quickly becomes impractical as we move from the toy problems on which many genetic algorithms are tested to the much more complex landscapes of either non-trivial engineering problems, or the vast search spaces of biological populations. Unlike in biological systems, it is often plausible in a digital system to quickly gather information about the local adaptive landscape by systematically measuring the impact of all possible one- or two-step mutations from a particular organism (Covert et al., 2013). Yet even this local knowledge of the landscape does not necessarily provide information about the landscape as a whole; different regions of the landscape may have different properties. From an artificial life perspective, then, it is less the case that we can use the adaptive landscape to make predictions about how evolution will proceed than that we can use information about how evolution has proceeded – and, in particular, from different evolutionary run starting at the same point – to infer the shape of the adaptive landscape.

A further complication of adaptive landscapes is that while there may be a single global optimum, it is not necessarily the case that an evolving population will settle at that peak. Imagine two different regions of a fitness landscape: one has a high but narrow peak, with even points a single mutation away from this peak having very low fitness; the other has a plateau, where the highest point is lower than the highest point of the narrow peak, but there is a fairly wide neutral network at this high-but-not-highest value. When there is a non-zero mutation rate, populations will consist of a cloud of points near each other, even at equilibrium. The most fit individuals – in the first case, sitting exactly on the top of the high, narrow peak – will, on average, produce the most offspring, which will either be on that same point (clonal offspring), or within a short distance of it (mutant offspring), tending to center the population on at least a local fitness peak during periods of equilibrium. The fitness of the population, therefore, will be based on the aggregate value in this cloud of points. If a peak is sufficiently narrow, and the difference in height between a narrow on a broad peak is small enough, population-level fitness can be optimized at the shorter, broader peak (Wilke et al., 2001).

Methods

Study System

Avida is a digital evolution software platform (Ofria and Wilke, 2004; Ofria et al., 2009). In this software, organisms are represented by individual programs, written in an assembly-like Turing-complete language. These organisms reproduce themselves by execution of their instructions. This reproduction, however, is not perfect; the user defines rates of mutations, such as instructions being replaced by other instructions, new instructions being added, or existing instructions being deleted. These mutations produce variation within the population of organisms. The user also creates an environment for these organisms, defining whether certain behaviors are rewarded with additional CPU cycles, and what the value of this reward is. Organisms have associated fitness values, which correspond to their expected rate of reproduction. With mutations as a source of variation, heredity due to the self-replication of organisms, and selection imposed by the environment, Avida represents an instance of evolution, rather than a simulation of it (Pennock, 2007).

Experimental Design

We evolved ten populations of digital organisms in an environment called logic-77, in which organisms were rewarded for performing any of 77 distinct 1-, 2-, or 3-input logic tasks (excluding the simplest 1-input task, Echo, where organisms output the same number they received as input). Organisms could perform any combination of these tasks, repeating any individual task up to 10 times. We focused on the logic-77 environment for two main reasons. First, the logic-77 environment is a relatively complex environment, and thus more likely to have a rugged adaptive landscape – where different high fitness regions are separated by substantial valleys of low fitness intermediaries – than a simpler environment. Secondly, in earlier work we have found evolution in the logic-77 environment to show substantial variation across replicate runs (Wiser, 2015), which is an important element for our questions. We then allowed these initial ten populations to evolve for 200,000 generations. We subsequently selected populations whose evolution appeared to have plateaued, extracted the most common genotype, and allowed it to evolve for another 200,00 generations to examine their long-term dynamics. For these second rounds of evolution, we ran ten replicate instances of Avida from each of the ten evolved starting points to assess the variability in potential outcomes.

Statistical Methods

We fit our data to two distinct mathematical models of fitness change over time. One of the models is a rectangular hyperbola, of the form

$$w = \frac{a * t}{t + b} + 1 \tag{1}$$

where w is relative fitness (measured fitness divided by ancestral fitness), a and b are model parameters, and t is time, measured here in generations. The other model is a power law, of the form

$$w = (b * t + 1)^{a} \tag{2}$$

where w is relative fitness (measured fitness divided by ancestral fitness), a and b are model parameters, and t is time, measured here in generations. We fit our models with the nls() command in R version 3.4.0 (R Core Team, 2016). (Technically, all our models were fit on log(base 2)-transformed fitness.) We compared our model fits by the difference in BIC value; note that because both of our models have the same number of parameters, and are fit on the same underlying data, the difference in BIC value is formally equivalent to the difference in AIC value in this case. We used these differences in BIC values to compare model fits, as outlined in (Raftery, 1995).

Results and Discussion

When separate evolutionary runs start from the same ancestor, they experience different mutations. These mutations make it such that the resulting organisms in different replicates will differ from each other. Sometimes, this variation has no effect on the phenotype of the organism; at other times, it does. In this work, we can see that ten replicate trials evolving from the same ancestor reach ten different fitness values after 200,000 generations, and they do so through different trajectories (see Fig.1)

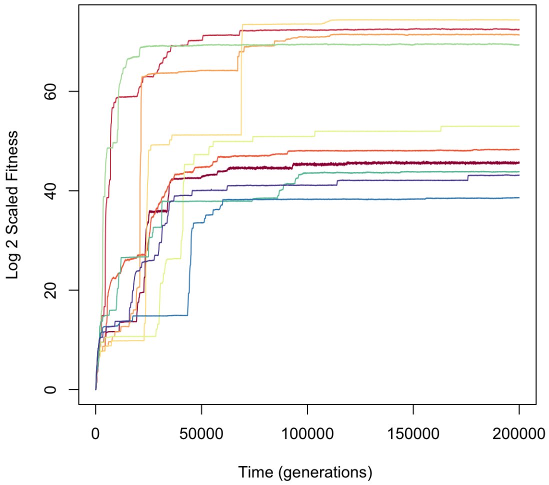

Figure 1: Evolution from original ancestor. Each line represents one individual evolutionary replicate.

For each replicate, we considered two different models to explain the fitness trajectory. One of them is a bounded model (Equation 1); in it, fitness will approach, though never quite reach, an upper limit. In this case, as t approaches infinity, the equation approaches a. The other model (Equation 2) is unbounded; in it, fitness increases decelerate, but do not have an upper limit. Both of these models have two parameters, so are of the same complexity as each other. The specific models considered are taken from a similar analysis of bacterial data (Wiser et al., 2013); the model fits for this data set are presented in Table 1. We are not here making claims that these specific models are the best fit of all possible models, but merely that they represent examples of a fundamental difference between a model that is bounded and a model that is not.

In ten of ten cases, the data is better fit by the bounded model than by the unbounded model. At first glance, this would argue that the fitness in these runs is approaching a maximum value that it will be unable to exceed. However, that cannot be the case. For one, each organism in each of the populations is descended from the same original ancestor through an unbroken chain of viable organisms. As such, there is a mutation path from any organism in one population to any organism in another population that passes entirely through organisms capable of survival and reproduction. This means that, in theory, any organism has the potential to mutate into any other organism in this experiment in a stepwise fashion. The probability of certain paths will be very low, as they go through extended sections of poor fitness relative to the existing population, but they are still potential, viable paths. As such, any population which appears as if it has an asymptote of fitness lower than the highest population observed must necessarily be capable of reaching the same fitness as that highest observed population, and thus is not truly bounded by that asymptote.

Seed	BIC (Power)	BIC (Hyper)	BIC Difference
1001	62532	54223	8310
1002	64908	47684	17224
1003	57051	42340	14711
1004	74087	67613	6474
1005	73477	67538	5939
1006	69025	64012	5013
1007	64835	49042	15792
1008	48405	36917	11488
1009	60901	57751	3150
1010	57096	46967	10129

Table 1: Model comparison between the bounded (Hyper) and unbounded (Power) models for 200,000 generations of evolution from the original ancestor. The BIC Difference column is the BIC value for the unbounded model minus the BIC value for the bounded model; positive values indicate a better fit for the bounded model, while negative values indicate a better fit for the unbounded model. BIC differences $<$10 are considered very strong support (Raftery, 1995). Note that all values reported on the table are rounded to the nearest integer, leading to rounding differences in the BIC Difference column

We further chose to test whether predictions of upper bounds of fitness were accurate by taking advantage of a convenient aspect of digital evolution: we were able to take organisms from the end of the initial 200,000 generations of evolution and used them as the ancestors for a subsequent 200,000 generations of evolution (see Fig. 2). For each of the ten intermediate ancestors, we conducted ten replicate evolutionary runs, for a total of 100 new evolutionary runs. If the predictions of an asymptote were accurate, for any given initial run which was better fit by the bounded model, we would expect none of the subsequent evolutionary runs from its final organism to exceed that asymptote. This is not the case. Table 2 shows the predicted asymptote for any initial run better fit by the bounded model, and the fitness attained in the subsequent evolution, which exceeds the predicted maximum in 17 of 100 cases.

The results stemming from these intermediate ancestors fall into several broad categories. Broadly speaking, these are cases where 1) the replicates consistently demonstrate lower fitness than either model predicts; 2) the actual data as a whole is more consistent with the bounded model, but indi-

1^{st} Seed	Pred. Asymptote	Highest Fit	# Exceed
1001	52.45	46.39	0
1002	75.73	74.80	0
1003	53.21	59.04	3
1004	82.56	75.26	0
1005	97.02	74.47	0
1006	68.55	70.37	2
1007	72.04	74.52	2
1008	47.22	56.23	4
1009	48.46	72.96	6
1010	47.98	45.18	0

Table 2: Comparisons of predicted asymptotes with highest realized fitnesses. Each listed asymptote is the base 2 log of the scaled fitness. Highest Fit. denotes the highest fitness measured in this subsequent 200,000 generations from an ancestor corresponding to an individual from the end of the first 200,000 generations of evolution. # Exceed indicates how many of the ten replicates started from that ancestor exceeded this predicted asymptote in an additional 200,000 generations of evolution.

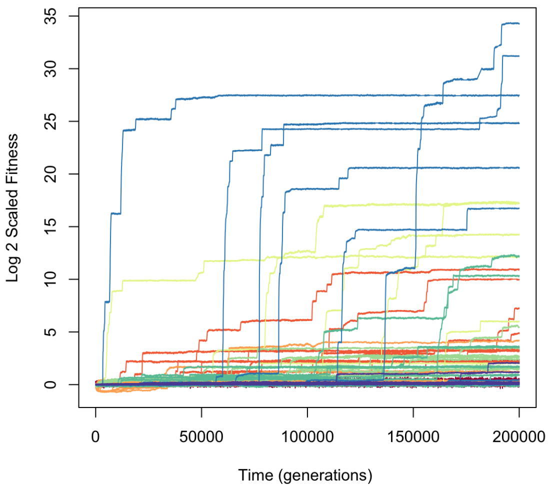

Figure 2: Evolution from evolved ancestors. Each line represents one individual evolutionary replicate. Different colors represent different ancestors. The colors match those from 1; the run from 1 that produced the red line there is used as the ancestor for all the lines with the same shade of red in this figure. Note that in this figure, each ancestor is scaled to its own starting point. Because both this and 1 have Log 2 y axes, this scaling simply moves each trajectory down horizontally to start at 0 in this Figure; the magnitude of increase is unaffected

vidual replicates are highly variable; 3) a substantial fraction of the replicates exceed even the unbounded model's predictions. Thhe third category is the least common of these – only one of the ten initial populations fall into this category – but the data are evenly split between categories 1 (five replicates) and 2 (four replicates).

For replicates in category 1, such as in Fig. 3 A, even the bounded model from the first 200,000 generations overestimates the fitness trajectory in the second 200,000 generations. In cases like this, the actual populations exhibit fewer significant increases in fitness in this second phase, and/or these increases are smaller than expected. The lower or less frequent increases can sometimes be explained directly by the increases themselves being small or rare in the second phase. At other times, the increases in the second phase can still be substantial, but particularly large or early gains in the first phase can lead to an *expectation* of large and/or frequent increases, beyond what the subsequent data demonstrate. Particularly in the latter of these cases, later portions of the evolutionary trajectory appear to better fit the bounded model than the unbounded model, even though the fitness of populations is increasing notably; these increases are, however, less than predicted from the large, early gains, and thus can give the illusion of unboundedness.

In other cases, such as in Fig. 3 B, the set of the subsequent runs as a whole are better fit by the bounded model, but many individual runs are not. Instead, some of the subsequent runs achieve notably higher fitness than the bounded model predicts, while others underperform compared to model expectations. Indeed, in some cases, such as in Fig. 3 D, some of the individual replicates actually better fit with the unbounded model, even occasionally exceeding this model's predictions. This category, then, reflects circumstances where the average across all replicates from this intermediate ancestor appears to be better described by a bounded model, but individual replicates may have very different results.

In the final category, represented here in Fig. 3 C, not only do many of the replicates from the intermediate ancestor exceed the theoretical asymptote from the bounded model, they even exceed the predictions of the particular unbounded model used. This particular case is a good example of how happenstance can play a large role in the accuracy of long-term predictions. The initial run leading to the intermediate ancestor had only modest changes in fitness between 80,000 and 200,000 generations, leading to the appearance of a population having reached a fitness plateau. Yet in only a very short time after the start of evolution from the intermediate ancestor, one of the replicates experienced a rapid growth in fitness, rising from a $\log_2$ fitness of slightly less than 40, to one greater than 60 in a few, rapid increases. Had this rapid increase happened slightly before the end of the initial phase, rather than slightly after the start of the second phase, it would have had a profound impact on the pre-

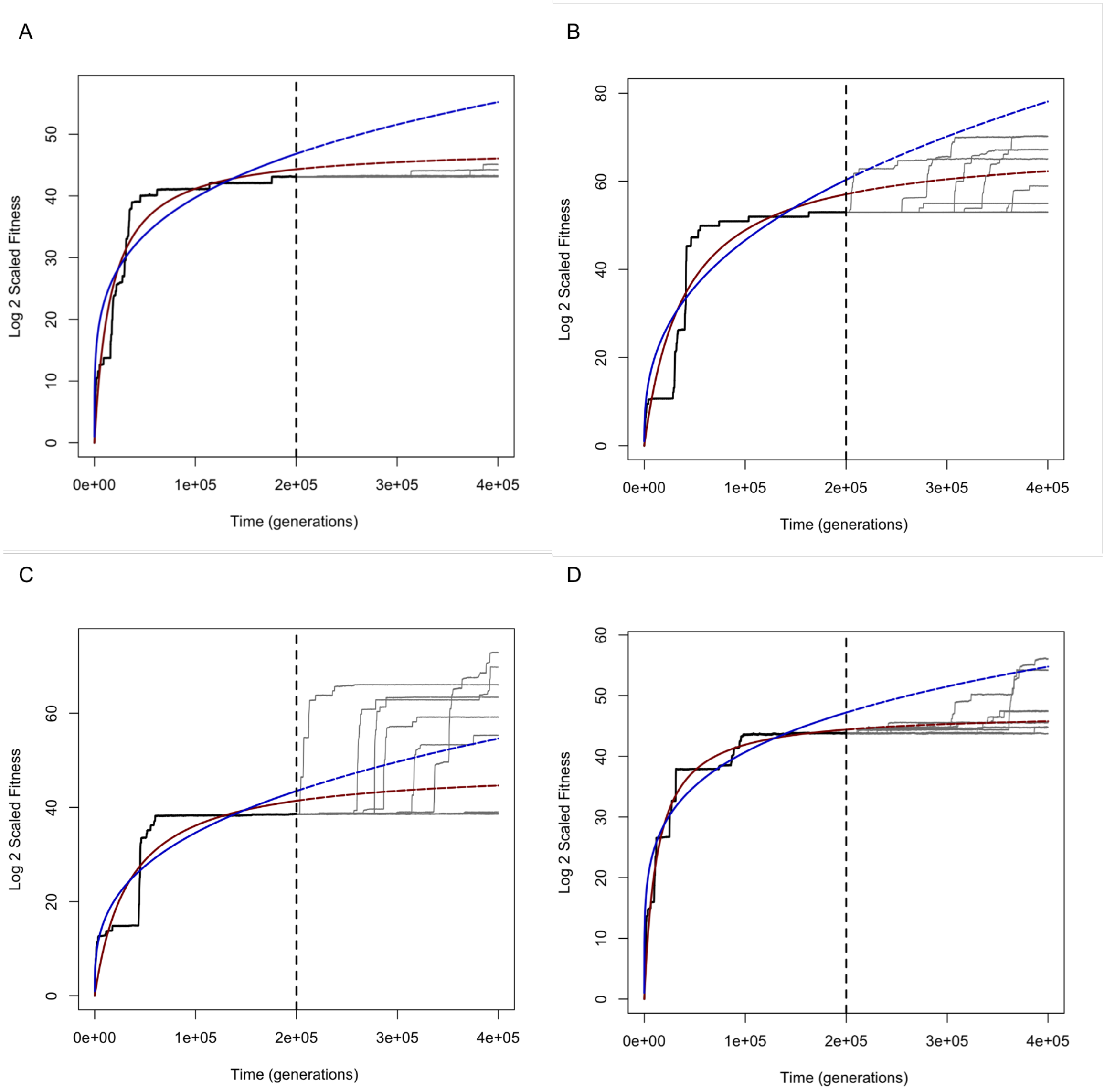

Figure 3: Example comparisons of model projections to actual data. For each case, the Hyperbolic model is shown in dark red; the power law model is shown in medium blue. The models are fit through the first 200,000 generation (solid lines), and then projected for the next 200,000 generations (dashed lines). The thin gray lines show the ten replicates started from the endpoint of the first phase. Note that the different panels have different y-axes, as the different initial populations reach notably different fitness levels.

dictions made, and the subsequent analysis of which populations and replicates exceeded the predictions of one or both models.

Across the whole set of experiments, many of the subsequent evolutionary runs – 83 of 100 – fail to exceed the theoretical asymptote predicted by the initial 200,000 generations of evolution. At first glance, this might seem to be support these asymptote predictions as being largely correct. That, however, would be a mistake. Firstly, these predictions are of the maximum fitness the relevant populations could attain; an example of exceeding this limit is sufficient to prove that the theoretical limit was incorrect. Secondly, these asymptotes are predictions of the maximum fitness these populations would ever attain, given infinite time, if the pattern of fitness change in the first 200,000 generations continued. Instead, we are testing these only against an additional 200,000 generations; this is a large number of generations, to be sure, but it is a trivial number compared to the history of microbes on Earth so far. Thirdly, we have examples of populations exceeding these theoretical maximum from five of the ten intermediate ancestors examined. This is not just a case of a small percentage of predicted limits being routinely exceeded, but instead a case where fully half of the predicted limits being exceeded at least once – in a set of only ten replicates each running, as just mentioned, a large but non-exhaustive number of generations.

What features of these fitness trajectories may drive the boundedness illusion? One striking feature is the large stepwise increases in fitness available in this environment. Any time a new task is acquired, it has the potential to immediately double fitness; moreover, because tasks can be rewarded up to ten times per organism, a task performed inside of a loop could be rewarded up to 2^{10} essentially instantly. These very large rewards will be far more visible in the fitness trajectory than the more frequent but smaller improvements that come from more efficient replication, which can lead to the *appearance* of fitness being unchanging, even while it is increasing, because it isn't increasing on the scale of these rare but large jumps. For example, in Fig. 3 C – the most extreme case of subsequent evolution exceeding the theoretical asymptote of the bounded model – one of the ten replicate extensions underwent a series of very rapid increases in fitness almost immediately, resulting in a more than 2^{20} gain in fitness in this second phase of evolution. Yet even so, it is eventually superseded by several other replicates from this same intermediate ancestor that took slower paths towards even higher fitness regions of the landscape. For any given replicate, there are points at which it seems that the population has plateaued – this ancestral population spent more than 100,000 generations at a fairly consistent fitness – but these times of apparent stasis do not signify a lack of future change.

Further, the scale of other changes can obscure real evolutionary improvements. In Fig. 3 C, four of the replicates started from the intermediate ancestor look as if they have flat trajectories in the second phase. In reality, these four range from one that declines 1.6% to one that increases 36.6% over the second 200,000 generations. Changes of this magnitude can be meaningful in biological systems – populations in the aforementioned *E. coli* study increased in fitness roughly 60-70% over 50,000 generations of adaptation – but are invisible in this graph because they are dwarfed by the rare but exceptionally large effects of task acquisition.

Conclusion

For systems that lack feedback cycles and frequency-dependence, a global optimal genotype must exist. That optimum, however, may be so far a randomly chosen starting position that an upper bound does not have an appreciable impact on evolutionary dynamics within the system. Deceleration in fitness gain can give a strong impression of fitness growth being bounded by an upper limit, lending statistical support to models which have an asymptotic limit. However, subsequent evolution frequently exceeds these predicted limits, thereby calling into question the utility of such models. Further, the variation in predicted limits across different replicates starting from the same point reduces the likelihood of many of those limits being real. Instead, we should be careful to provide evidence of effective limits to evolution, rather than assume they are present without adequately demonstrating it.

Acknowledgements

We thank Caroline Turner, Alita Burmeister, and Michael Blazanin for helpful discussion and feedback on this manuscript. This research was supported in part by the BEACON Center for the Study of Evolution in Action (NSF Cooperative Agreement No. DBI-0939454), NSF Award No. DEB-1655715, and by Michigan State University through the computational resources provided by the Institute for Cyber-Enabled Research. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Abrams, P. A. (2000). The Evolution of Predator-Prey Interactions: Theory and Evidence. *Annual Review of Ecology and Systematics*, 31(1):79–105.

Adams, F. C., Bond, J. R., Freese, K., Frieman, J. A., and Olinto, A. V. (1993). Natural inflation: Particle physics models, power-law spectra for large-scale structure, and constraints from the Cosmic Background Explorer. *Physical Review D*, 47(2):426–455.

Bedau, M. A., Snyder, E., and Packard, N. H. (1998). A classification of long-term evolutionary dynamics. In *Artificial life VI*, pages 228–237. Cambridge: MIT Press.

Channon, A. (2003). Improving and still passing the ALife test: Component-normalised activity statistics classify evolution in Geb as unbounded. In Standish, R. K., Bedau, M. A., and Abbass, H. A., editors, *Proceedings of the Eighth International Conference on Artificial Life*, pages 173–181. MIT Press, Cambridge, MA.

Chou, H.-H., Delaney, N. F., Draghi, J. A., and Marx, C. J. (2014). Mapping the Fitness Landscape of Gene Expression Uncovers the Cause of Antagonism and Sign Epistasis between Adaptive Mutations. *PLOS Genetics*, 10(2):e1004149.

Clauset, A., Shalizi, C., and Newman, M. (2009). Power-Law Distributions in Empirical Data. *SIAM Review*, 51(4):661–703.

Covert, A. W., Lenski, R. E., Wilke, C. O., and Ofria, C. (2013). Experiments on the role of deleterious mutations as stepping stones in adaptive evolution. *Proceedings of the National Academy of Sciences*, 110(34):E3171–E3178.

Davies, P. C. W. (2004). Emergent biological principles and the computational properties of the universe: Explaining it or explaining it away. *Complexity*, 10(2):11–15.

de Visser, J. A. G. and Krug, J. (2014). Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics*, 15:480.

Eldregde, N. and Gould, S. J. (1972). Punctuated equilibria: an alternative to phyletic gradualism. In Schopf, T. J. M., editor, *Models in Paleobiology*, pages 82–115. Freeman, Cooper, San Francisco.

Ganco, M. and Hoetker, G. (2009). NK modeling methodology in the strategy literature: Bounded search on a rugged landscape. In *Research Methodology in Strategy and Management*, volume 5 of *Research Methodology in Strategy and Management*, pages 237–268. Emerald Group Publishing Limited.

Gerstein, A. C. and Otto, S. P. (2011). Cryptic Fitness Advantage: Diploids Invade Haploid Populations Despite Lacking Any Apparent Advantage as Measured by Standard Fitness Assays. *PLOS ONE*, 6(12):e26599.

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L. N., Daly, J. M., Mumford, J. A., and Holmes, E. C. (2004). Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science*, 303(5656):327.

Islam, S. M., Das, S., Ghosh, S., Roy, S., and Suganthan, P. N. (2012). An Adaptive Differential Evolution Algorithm With Novel Mutation and Crossover Strategies for Global Numerical Optimization. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(2):482–500.

Kitto, K. J. (2006). *Modelling and generating complex emergent behaviour*. PhD thesis, Flinders University, Adelaide, New South Wales, Australia.

Kvitek, D. J. and Sherlock, G. (2011). Reciprocal Sign Epistasis between Frequently Experimentally Evolved Adaptive Mutations Causes a Rugged Fitness Landscape. *PLOS Genetics*, 7(4):e1002056.

Lampinen, J. and Zelinka, I. (2000). On stagnation of the differential evolution algorithm. In Osmera, P., editor, *Proceedings of MENDEL 2000*, pages 76–83, Brno, Czech Republic.

Mahler, D. L., Ingram, T., Revell, L. J., and Losos, J. B. (2013). Exceptional Convergence on the Macroevolutionary Landscape in Island Lizard Radiations. *Science*, 341(6143):292.

Martin, C. H. and Wainwright, P. C. (2013). Multiple Fitness Peaks on the Adaptive Landscape Drive Adaptive Radiation in the Wild. *Science*, 339(6116):208.

May, R. M. (2006). Network structure and the biology of populations. *Twenty years of TREE - part 2*, 21(7):394–399.

Mojzsis, S. J., Arrhenius, G., McKeegan, K. D., Harrison, T. M., Nutman, A. P., and Friend, C. R. L. (1996). Evidence for life on Earth before 3,800 million years ago. *Nature*, 384:55.

Nelson, M. I., Simonsen, L., Viboud, C., Miller, M. A., and Holmes, E. C. (2007). Phylogenetic Analysis Reveals the Global Migration of Seasonal Influenza A Viruses. *PLOS Pathogens*, 3(9):e131.

Newman, M. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323–351.

Ofria, C., Bryson, D. M., and Wilke, C. O. (2009). Avida. In *Artificial Life Models in Software*, pages 3–35. Springer London.

Ofria, C. and Wilke, C. O. (2004). Avida: A Software Platform for Research in Computational Evolutionary Biology. *Artificial Life*, 10(2):191–229.

Orr, H. A. (2009). Fitness and its role in evolutionary genetics. *Nature Reviews Genetics*, 10:531.

Page, K. M. and Nowak, M. A. (2002). Unifying Evolutionary Dynamics. *Journal of Theoretical Biology*, 219(1):93–98.

Pennock, R. T. (2007). Models, simulations, instantiations, and evidence: the case of digital evolution. *Journal of Experimental & Theoretical Artificial Intelligence*, 19(1):29–42.

Piotrowski, A. P. (2014). Differential Evolution algorithms applied to Neural Network training suffer from stagnation. *Applied Soft Computing*, 21:382–406.

R Core Team (2016). R: A language and environment for statistical computing.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, 25:111–164.

Remold, S. K. (2002). Unapparent virus infection and host fitness in three weedy grass species. *Journal of Ecology*, 90(6):967–977.

Rosing, M. T. (1999). 13c-Depleted Carbon Microparticles in >3700-Ma Sea-Floor Sedimentary Rocks from West Greenland. *Science*, 283(5402):674.

Sajjad, M., Singh, K., Paik, E., and Ahn, C.-W. (2016). A data-driven approach for agent-based modeling: simulating the dynamics of family formation. *Journal of Artificial Societies and Social Simulation*, 19(1):9.

Taylor, T., Bedau, M., Ackley, D., Channon, A., Banzhaf, W., Beslon, G., Dolson, E., Froese, T., Hickinbotham, S., Ikegami, T., McMullin, B., Packard, N., Rasmussen, S., Virgo, N., Agmon, E., Clark, E., McGregor, S., Ofria, C., Ropella, G., Spector, L., Stanley, K. O., Stanton, A., Timperley, C., Vostinar, A., and Wiser, M. (2016). Open-Ended Evolution: Perspectives from the OEE Workshop in York. *Artificial Life*, 22(3):408–423.

Vassilev, V. K. and Miller, J. F. (2000). The Advantages of Landscape Neutrality in Digital Circuit Evolution. In Miller, J., Thompson, A., Thomson, P., and Fogarty, T. C., editors, *Evolvable Systems: From Biology to Hardware*, pages 252–263. Springer Berlin Heidelberg.

Velmala, W., Helle, S., Ahola, M. P., Klaassen, M., Lehikoinen, E., Rainio, K., Sirki, P. M., and Laaksonen, T. (2015). Natural selection for earlier male arrival to breeding grounds through direct and indirect effects in a migratory songbird. *Ecology and Evolution*, 5(6):1205–1213.

Wilke, C. O., Wang, J. L., Ofria, C., Lenski, R. E., and Adami, C. (2001). Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844):331–333.

Wiser, M. J. (2015). *An Analysis of Fitness in Long-Term Asexual Evolution Experiments.* PhD thesis, Michigan State University.

Wiser, M. J. and Lenski, R. E. (2015). A Comparison of Methods to Measure Fitness in Escherichia coli. *PLoS ONE*, 10(5):e0126210.

Wiser, M. J., Ribeck, N., and Lenski, R. E. (2013). Long-term dynamics of adaptation in asexual populations. *Science*, 342(6164):1364–1367.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In *Proceedings of the 6th International Congress of Genetics*, pages 356–366.

Yu, T. and Miller, J. (2001). Neutrality and the Evolvability of Boolean Function Landscape. In Miller, J., Tomassini, M., Lanzi, P. L., Ryan, C., Tettamanzi, A. G. B., and Langdon, W. B., editors, *Genetic Programming*, pages 204–217. Springer Berlin Heidelberg.