Kludged*

Jeffrey C. Ely†

October 4, 2007

Abstract

Is there reason to believe that our brains have evolved to make efficient decisions so that the details of the internal process by which these decisions get made are irrelevant? Or can we understand the persistence of behavioral anomalies as the consequence of specific imperfections in the decision-making circuitry that remain despite evolutionary pressure? I develop a formal model which illustrates a fundamental limitation of adaptive processes: improvements tend to come in the form of *kludges*. A kludge is a marginal adaptation that compensates for, but does not eliminate fundamental design inefficiencies. When kludges accumulate the result can be perpetually suboptimal behavior. This is true even in a model of evolution in which arbitrarily large innovations occur infinitely often with probability 1. This has implications for traditional defenses of both positive and normative methodology and provides a foundation for behavioral theories built on the methodology of constrained-optimal design.

Keywords: kludge.

*Preliminary (cite anyway.) This paper was not previously circulated under the titles "Bad Adaptation" or "Ex-Post Regrettable Organism Design."

†Department of Economics, Northwestern University. jeffely@northwestern.edu. I thank Larry Samuelson and Jeroen Swinkels and Marcin Peski for early conversations on this subject and advice which steered me in a good direction. Conversations with Sandeep Baliga, George Mailath, Bill Zame, Ilya Segal, Joel Sobel, Alvaro Sandroni, Daron Acemoglu and Tomasz Strzalecki were also influential. Ben Handel was a valuable research assistant at an early stage. For some of the technical results, I took inspiration from a paper by Sandholm and Pauzner (1998). I accept the blame.

1 Introduction

In July of 2004, Microsoft announced that the release of Vista, the next generation of the Windows operating system, would be delayed until late 2006. Jim Allchin famously walked into the office of Bill Gates and proclaimed, "It's not going to work." Development of Windows had become unmanageable and Allchin decided that Vista would have to be re-written essentially from scratch.

> Mr Allchin's reforms address a problem dating to Microsoft's beginnings. . . . PC users wanted cool and useful features quickly. they tolerated – or didn't notice – the bugs riddling the software. Problems could always be patched over. With each patch and enhancement, it became harder to strap new features onto the software since new code could affect everything else in unpredictable ways.[1]

The Alternative Minimum Tax was introduced by the Tax Reform Act of 1969. It was intended to prevent taxpayers with very high incomes from exploiting numerous tax exemptions and paying little or no tax at all. Over time, the shortcomings of the AMT as a solution to the proliferation of exemptions have begun to appear. However, over this same time, the federal tax and budgeting system has come to depend on the AMT to the point that many observers think that changing the AMT, without complicated accompanying adjustments elsewhere, would be worse than leaving it as is.

Flat fish inhabit the sea floor, but for many this was not the original habitat. When their ancestors moved to the sea floor, they adapted by changing their orientation from swimming "upright" to on their sides. Given their existing bone structure, this was the only way to become "flat", but it rendered one eye useless. So, by a further adaptation many of today's species of flatfish migrate one eye to the opposite side of their body during development.[2]

As beautifully documented the film *The March of the Penguins*, emperor penguins spend a nearly 9 month breeding and nurturing cycle which involves walking up to 100 KM away from any food source in order to avoid predators. The problem for penguins is that they are birds, and hence lay eggs; but they are flightless birds, so they find it inconvenient to move to

[1] "Code Red: Battling Google, Microsoft Changes How it Builds Software." *The Wall Street Journal,* Robert Guth, September 2005.

[2] For a vivid account of the evolution of bony flat fish, see Dawkins (1986).

areas where the eggs can be easily protected. They adapted not by rectifying either of these two basic problems,[3] but instead by compensating for them by an extremely costly and risky behavior.

Each of these examples represents a *kludge*: an improvement upon a highly complex system that solves an inefficiency but in a piecemeal fashion and without addressing the deep-rooted underlying problem. There are three ingredients to a kludge. First the system must be increasing in complexity so that new problems arise that present challenges to the internal workings of the system. Second, a kludge addresses the problem by patching up any mis-coordination between the inherited infrastructure and the new demands. Third, the kludge itself– because it makes sense only in the presence of the disease it is there to treat– intensifies the internal inefficiency, necessitating either further kludges in the future or else eventually a complete revolution.[4]

Microsoft Windows is a complex system whose evolution is guided by a forward-looking dynamic optimizer. It is not surprising therefore that, after two decades worth of kludges that accompanied the expansion from DOS to Windows to 32 bit and evenutally 64 bit architecture, revolution was the final solution. In the case of the US Tax Code, or for that matter any sufficiently complex body of contracts that govern interactions among diverse interests, while the evolution may be influenced by forward-looking considerations, full dynamic optimization is more tenuous as a model of the long-run trade-offs.

But the story is very different for flat fish and penguins, and, to come to the point, for human brains, whether we are considering the evolution of the brain across generations or the development of the decision-making apparatus within the life a single individual. Here, progress is *adaptive.* An adaptive process is not forward-looking and certainly not governed by dynamic optimization. An adaptive process inherits its raw material from the past, occasionally modifies it by chance (mutation or experimentation), and selects among variants according to success *today.*

Nevertheless there is the possibility, not completely fanciful, that an adaptive process can produce complex systems that perform as well today as those that were designed by an optimizer given the same set of raw materials. Indeed, there is a tradition in economics that accepts the dis-

[3]Incidentally, it has happened in evolutionary history that oviparous (egg-laying) species have adapted to vivipary (giving birth to live offspring.) Some species of sharks are important examples. Vivipary enables a long internal gestation so that the developing offspring is protected and nourished within the body of the mother.

[4]See wikipedia for the history, usage, and pronunciation of the word kludge.

tinction between adaptation and optimization, but rationalizes a positive methodology based on unfettered optimization by an appeal to this unwritten proposition.[5]

In this paper I present a model intended to suggest that this hope was a longshot at best. I analyze a simple single-person decision problem. An *organism* is a procedure for solving this problem. I parameterize a family of such algorithms which includes the optimal algorithm in addition to algorithms that perform less well. An adaptive process alters the organism over time, favoring improvements. I show conditions under which no matter how long the adaptive process proceeds, an engineer, at any point in time, working only with the raw materials that presently make up the organism, could eliminate a persistent structural inefficiency and produce a significant improvement. In the model, kludges arise naturally and are the typical adaptations that improve the organism. A kludge always improves the organism at the margin, but also increases both its complexity and its internal complementarity and as a by-product makes it harder and harder for adaptation to undo these inefficiencies in the future.

In the model, a resource is available at a randomly determined location. The organism evolves a procedure for collecting and processing information about the location. Two trade-offs govern the design of the optimal organism. First, a fixed number of computational steps must be allocated between estimation of the location and exploitation of the resource. More precise estimates come at the expense of reduced intensity of exploitation. Second, the organism must evolve the optimal protocol for processing the information. The pitfall is that the organism may adapt an inefficient protocol which requires too many processing steps to achieve a given precision. The cost is reduced intensity. However, once this inefficient protocol is in place, future evolution (modeled as expansion of computational power) continues to "invest" in it making it increasingly difficult to re-optimize.

The problem in the model is not due to "local optima." The model admits arbitrarily large mutations with positive probability, so they occur infinitely often. Given enough time, the process would escape any non-global static optimum. Indeed I present a benchmark model (see Proposition 1) in which there is an artificial upper bound on the complexity of the organism. In this model the optimally adapted organism eventually appears with probability 1. Also, the effect is not due to altered evolutionary incentives that come from strategic interactions with other agents. The model analyzes the performance of a single agent solving an isolated deci-

[5]The classic defense is Friedman (1966).

sion problem.

Structurally inefficient decision-makers present a problem not just for positive methodology, but normative as well. Much of welfare economics is founded on revealed preference and agent sovereignty. These principles presume that the choices we observe reveal what benefits the agent. But when the adaptive process creates a wedge between observed behavior and the underlying objective the agent is designed to satisfy, there is a corresponding wedge between revealed preference and true preference. Put differently, if we grant that there is some underlying objective that guides the adaptive process, then at best we can view the organism as an agent whose efforts at achieving that objective are the result of a second-best solution designed by nature, the principal. We can no better infer that underlying objective from the choice behavior of the organism than we can identify the distorted choices made by an incentivized agent with the principal's first-best solution.

1.1 Organism Design

Indeed, this principal-agent metaphor is the basis of an increasingly popular methodology for behavioral economics. For example, Robson (2001) studies the biological rationale for hedonic utility. His model shows that utility can be understood as an optimal compensation scheme for an agent who has private information about the fitness consequences of various consumption bundles. Implicit is the interpretation that natural selection can be equivalently regarded as a fitness-maximizing principal with a freedom to design the agent's preferences limited only by asymmetric information. The bottom-line of such a model is an agent whose revealed-preference exactly coincides with nature's first-best.

By contrast, interesting non-standard preferences can be generated by a similar model in which metaphorical nature is assumed to face additional constraints. For example, Samuelson and Swinkels (2006) consider a design problem in which the agent necessarily makes errors in information processing and nature's incentive scheme must trade off the value of incorporating the agent's private information about the local environment against the risks of granting too much leeway to imperfectly formed beliefs. Constrained to use the blunt instrument of utility to provide incentives to the agent, nature's optimal design necessarily induces behavioral biases, including self-control problems, menu-dependence, and present-bias. In Rayo and Becker (2007) nature is constrained to use outcome-contingent rewards to motivate the agent and there are limits to the granularity of this

"happiness" instrument. The resulting optimal incentive scheme is equivalent to reference-dependent preference. In dynamic decision problems, certain behavioral responses can substitute for expanded memory and in Baliga and Ely (2007), nature's design economizes on memory capacity by utilizing this trade-off. The result is observationally equivalent to a sunk-cost bias.

In each of these models, the conclusions are driven by the particular constraint imposed on nature's representative, the principal. It follows that the same evolutionary argument can turn each one of these conclusions on its head. Nature is appropriately modeled as an optimizing principal only if natural selection can be assumed to operate long enough to reach an optimum. But then there should also have been ample time for nature to relax these constraints. In the language of incentives, because there is no intrinsic conflict of interest between principal and agent, in the long run nature simply "sells the firm" to the organism. Whatever residual effect of the constraints persists should have negligible costs.[6] Equivalently, observable behavior should be arbitrarily close to the first-best, and hence free from (costly) behavioral anomolies.

The present paper provides a defense of this methodology against such arguments. On the one hand, the model yields arbitrarily large innovations in the design of the organism. It follows that each "component" of the organism is optimally designed taking as given the existence of, and interactions with, other components. That is, the organism is optimal subject to certain design constraints. And on the other hand, these constraints needl not be eliminated despite the fact that arbirarily large innovations occur infinitely often. Indeed, their presence can have non-vanishing shadow costs even in the long run.

2 The Model

An organism is designed to solve a fixed decision problem, instances of which are presented to the organism repeatedly over time. The decision problem has the following interpretation. A resource is available at a certain location. The location is realized independently in each period. Signals

[6]Of course there are physical constraints which could not be eliminated no matter how long nature is left to act. But the appropriate comparison here is between the outcome of the evolutionary process as modeled by a designer subject to physical and non-physical constraints and the design that is optimal subject only to the physical constraints. The argument here is that all residual internal design inefficiencies should, in the long run, have negligible bottom-line consequences.

which reveal the location of the resource are available to the organism. The problem for the organism is to input these signals, interpret them, and then choose a location in attempt to exploit the resource. The fitness of the organism is determined by the distance between the actual location of the resource and the location chosen.

The organism is described by an algorithm for inputing and processing signals. The components of this algorithm adapt over time according to a general evolutionary process which selects for improvements in overall fitness. We describe the long run behavior of this evolutionary process.

The Decision Problem A resource is hidden at a location $\theta \in [-1, 1]$ which is drawn from the uniform distribution. The organism will choose a location a and search intensity i. The payoffs are

$$u(a, i, \theta) = i\left[2a\theta - a^2\right]. \tag{1}$$

In the environment there is a collection of signals σ available to the organism which convey information about the current location of the resource. A new location, and a new set of signals are selected independently across each of an infinite sequence of periods. The problem for the organism is to evolve an efficient process to input, interpret, and aggregate the signals in order to optimally exploit the resource.

In each period, there is an unlimited number of signals that the organism could potentially use to locate the resource, but the organism is limited by the number of signals it can collect. Formally, $\sigma = \sigma_1, \sigma_2, \ldots$ is an infinite sequence from $\{-1, +1\}^\infty$ and the organism is able to collect a finite sample of size k. The parameter k is referred to as the *precision* of the organsim. Over time, increasing precision is one of the ways in which the organism will evolve.

The organsim must also learn how different signals σ_j correlate differently with the location of the resource. This correlation structure is fixed over time and represents the *environment* to which the organism must adapt. At any point in time, the organism's current scheme for translating a sample $\sigma_1, \ldots, \sigma_k$ into a choice of location is characterized by another finite sequence $\pi = \pi_0, \pi_1, \ldots, \pi_k \in \{-1, +1\}^k$. When the organism observes the sample $\sigma_1, \ldots, \sigma_k$ it applies the following formula

$$a(\sigma_1, \ldots, \sigma_k | \pi) = \frac{\pi_0}{k+2} \sum_{j=1}^{k} \pi_j \sigma_j. \tag{2}$$

to select a location $a(\sigma_1, \ldots, \sigma_k)$ to search. To interpret this, note that the sign of $\pi_0 \pi_j$ expresses whether the organism acts as though signal σ_j is positively or negatively correlated with the location θ. The sequence $\pi_0, \ldots, \pi_k$ can be viewed as the genetic code of the organism. Over time, the organsim will evolve by increasing the length and tuning the configuration of this genetic code.

The specific functional form and the special role played by the parameter π_0 will be explained next.

The Environment Now I describe the probability distributions from which the location θ and the signals σ are drawn independently in each period. One aspect of the environment is fixed throughout. An infinite sequence $\lambda = \lambda_1, \lambda_2, \ldots \in \{-1, +1\}^\infty$ is determined at the beginning of time according to an i.i.d. process with $\text{Prob}(\lambda_j = 1) = l > 1/2$. We will refer to λ as the *environment*.

The environment determines the correlation between signals and the location of the resource. Specifically, each signal σ_j is chosen independently according to the distribution

$$\text{Prob}(\sigma_j = \lambda_j) = \frac{\theta + 1}{2}. \tag{3}$$

To understand this structure, first consider a signal j for which $\lambda_j = 1$. In this case, observing $\sigma_j = 1$ indicates that the resource is likely to be located further to the right, whereas $\sigma_j = -1$ indicates that the resource is likely to be located further to the left. However, when $\lambda_j = -1$, these inferences are reversed. For this reason, if $\lambda_j = -1$, then we say that the jth signal is *inverted*.

The optimal estimator It is instructive to begin by considering the first-best algorithm for estimating the location of the resource, given prior knowledge of the structure of the environment. Suppose that λ is known. It follows from standard properties of binomial sampling that the posterior expected value of θ conditional on observing the sample of signals $\sigma_1, \ldots, \sigma_k$

is given by[7]

$$\bar{\theta}_k := \mathbf{E}(\theta|\sigma_1, \ldots, \sigma_k) = \frac{1}{k+2} \sum_{j=1}^{k} \lambda_j \sigma_j. \tag{4}$$

Therefore, given the structure of payoffs (see Equation 1), the organism should hunt for the resource at location

$$a^*(\sigma_1, \ldots, \sigma_k) = \frac{1}{k+2} \sum_{j=1}^{k} \lambda_j \sigma_j. \tag{5}$$

Alignment Now, returning to the organism, we can see that there are two types of organisms which implement this optimal strategy. Compare Equation 2 and Equation 5. A *positively-aligned* organism is one with $\pi_0 = +1$ and $\pi_j = \lambda_j$ for $j = 1, \ldots, k$. A *negatively-aligned* organism is one with $\pi_0 = -1$ and $\pi_j = -\lambda_j$ for $j = 1, \ldots, k$. Both types of organism select the conditional expected fitness maximizing location $a^*(\sigma_1, \ldots, \sigma_k)$ given a sample size of k. Any other organism of equal precision chooses an inferior location.

Computation and Complexity We view the organism as an algorithm for locating and exploiting the resource. The organism is limited by the number of computational steps in can perform in this process. This number x will be called the *complexity* of the organism. Each use of the following operations requires a single step: collecting an additional signal σ_j, multiplying by -1 (henceforth referred to as a *pre-processing step*), and increasing the search intensity. Therefore, an organism of complexity x which uses l steps to calculate its location will exploit the resource at that location with an intensity equal to $x - l$. As the organism evolves, it will improve by increasing x and adjusting the allocation of these computational steps..

While positive and negatively aligned organisms of the same precision select the same location $a^*(\sigma_1, \ldots, \sigma_k)$, they typically require a different number of steps to do it and therefore they will differ in the intensity i with which they are able to exploit the resource. This means that, for a given

[7]Write $\hat{\theta} = (\theta + 1)/2$. Then $\hat{\theta}$ is distributed uniformly on $[0, 1]$ and observation of $\{\lambda_j \sigma_j\}_{j=1}^{k}$ is a binomial sampling process from $\{-1, 1\}$ with unknown probability $\hat{\theta}$ that $\lambda_j \sigma_j = 1$. It is a standard result that in this case the posterior distribution of $\hat{\theta}$ is a Beta distribution with parameters (ζ_1, ζ_2) where ζ_1 is one plus the number of j for which $\lambda_j \sigma_j = 1$ and ζ_2 is one plus the number of j for which $\lambda_j \sigma_j = -1$. The expectation of the Beta distribution is $\frac{\zeta_1}{\zeta_1 + \zeta_2}$. This yields Equation 4 after some algebra.

total complexity x, only one of these two types of organism will achieve the maximum fitness. The gene π_0 adds flexibility to the design of the organism potentially allowing it to economize on computational complexity.

The diagrams in Figure 1 illustrate the optimal organism for a fixed complexity x. The "budget" lines capture the tradeoff between intensity and precision for positively- (dashed) and negatively- (solid) aligned organisms respectively. Adding the jth unit of precision requires a sacrifice of one or two units of intensity, depending on the alignment and the value of λ_j. This yields the following budget equations

$$x = i + k\left(\frac{3}{2} - \frac{1}{k}\sum_{j=1}^{k}\frac{\lambda_j}{2}\right)$$

for positive alignment and

$$x = i + k\left(\frac{3}{2} + \frac{1}{k}\sum_{j=1}^{k}\frac{\lambda_j}{2}\right)$$

for negative alignment. The "indifference curve" is the set of pairs (i,k) which achieve the same fitness.

Figure 1(a) shows a case in which the optimal organism is negatively aligned. As the organism increases in complexity, the budget lines shift up, potentially switching the alignment of the optimal organism. This is illustrated in Figure 1(b). Indeed, the optimal alignment depends on the sign of the moving average

$$L(k) := \frac{1}{k}\sum_{j=1}^{k}\lambda_j > 0.$$

If it is positive, then the fraction of inverted signals up to k is greater than $1/2$, and the optimal organism will be positively aligned. The negatively aligned organism is optimal in the alternative case.

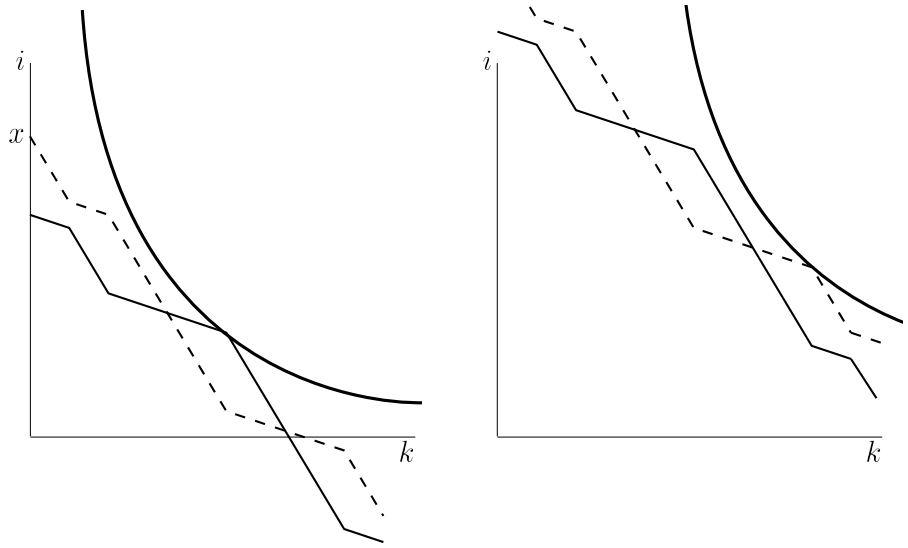
Recall that we have assumed that $l > 1/2$. This implies that for sufficiently complex organisms, positive alignment is optimal. A convenient way to visualize this is to consider k sufficiently large so that $L(k) \approx 2l - 1$ and the two budget lines are approximately

$$x \approx i + k\,(2 - l)$$

for positive alignment and

$$x \approx i + k\,(1 + l)$$

for negative alignment. This is illustrated in Figure 2.



(a) Low x. Negative alignment (solid line) is optimal.

(b) Higher x. Budget lines shift upward and now positive alignment is optimal.

Figure 1: Optimal organism for a fixed level of complexity x.

Kludge Note that for sufficiently complex organisms, positive alignment yields a greater budget. Once this is the case, any negatively aligned organism is attempting to implement the optimal decision rule via an inefficient protocol. For this reason and reasons developed further below, we refer to such an organism as a *kludge.*

Definition 1. *Suppose that the fraction of inverted signals up to k exceeds* $1/2$*, i.e.*

$$\frac{1}{k}\sum_{j=1}^{k}\lambda_j > 0.$$

Then we say that an optimal negatively aligned organism with precision k is a kludge.

We can quantify the inefficiency of a kludge of complexity x. A switch to positive alignment would produce an organism of the same precision

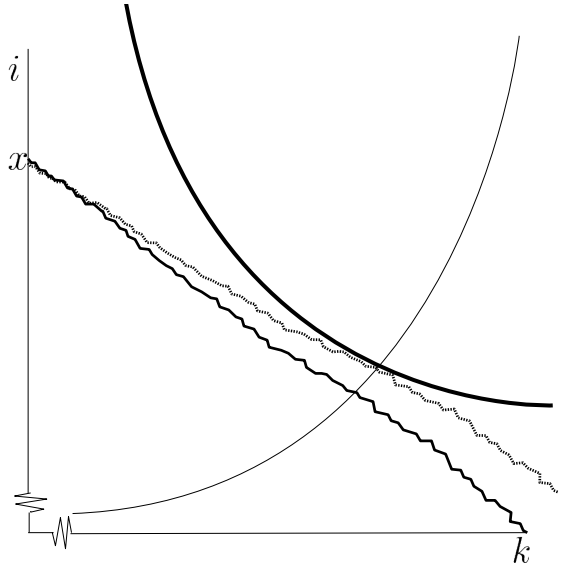


Figure 2: Optimal alignment for large k.

but strictly higher intensity. Indeed the intensity and therefore the fitness can be increased by a number which (on average) increases linearly in k.

However, this measure may be hard to interpret as it depends on a cardinal interpretation of payoffs. As an alternative, let us define the following ordinal concept of inefficiency of an organism. Say that the organism is *asymptotically structurally inefficient* if there is a given component of the organism (here, a subset of tokens) such that at point in time, and forever thereafter, this component should be altered as a part of some improvement to the organism, but nevertheless the component remains fixed forever.[8]

[8] A virtue of this definition is that it excludes "marginal inefficiencies" where at any point in time some inefficiencies are present, but every inefficiency, once it appears, is eventually eliminated. For example, we may imagine that the most recently developed features of the organism might begin in an inefficient state, but eventually as the organism matures, these features are improved to their optimal state and align optimally with the rest of the organism. By contrast, asymptotic structural inefficiency identifies persistent mis-alignments. It would be desirable to sharpen the definition even further by considering dynamic efficiency issues. Without going into the details of such a definition, I note that the kludges in this paper represent static as well as dynamic inefficiencies. In addition to outperforming a kludge at each point in time, positively aligned organisms also grow in intensity and precision faster than kludges.

The Adaptive Process The final ingredient in the model is a description of the process by which the organism evolves. I adopt a simple model of mutation and natural selection designed to capture the effects of a general class of adaptive processes. The specific assumptions are chosen mostly for expositional and analytical convenience.

Each period t, the organism $\mathcal{O}_t$ is evaluated according to its overall fitness. The realized payoff of the organism in a period when the resource is located at θ and the signal is σ is given by

$$i\left[2\theta a(\sigma_1,\ldots,\sigma_k|\pi) - a(\sigma_1,\ldots,\sigma_k|\pi)^2\right].$$

The overal fitness $V(\mathcal{O})$ of the organism $\mathcal{O}$ is defined as the expected value of this payoff with respect to the distributions of θ and σ.

With positive probability, a variation occurs which results in a new version $\mathcal{O}'$ of the organism. If the variant $\mathcal{O}'$ is more fit, i.e. $V(\mathcal{O}') > V(\mathcal{O}_t)$, then the variant replaces the existing version and survives to date $t+1$, that is $\mathcal{O}_{t+1} = \mathcal{O}'$. If not, then the existing version survives, i.e. $\mathcal{O}_{t+1} = \mathcal{O}_t$.

Two types of variations can occur. First, with probability q, the organism increases in complexity. It keeps the analysis simple to assume that when complexity increases it it increases by two, and the two additional computational steps are allocated optimally taking as given the existing allocation. On the other hand, with probability $(1-q)$ the organism does not increase complexity, but undergoes a mutation in which some (possibly empty) subset of existing computational steps are re-allocated.

We model mutations as follows. The genetic code of the organism is represented by a sequence of symbols $\{-, 1, *\}$. The initial string of length $x - i$ consists of symbols in $\{-, 1\}$ and represents π. The remaining symbols, numbering i, are all $*$. An example is illustrated in Figure 3.

$$1\ -1\ -1\ 1\ -1\ \ldots\ 1\ *\ *\ \ldots\ *$$

Figure 3: An encoding of an organism.

When a mutation occurs, a subset of the $-$ and $*$ symbols is selected at random. Each of the symbols in the selected set is then randomly re-assigned. When a symobl is re-assigned it may mutate into a different symbol and it may change position in the sequence, subject to the following

restrictions. A new 1 symbol (which represents an increase in precision) is added at the end of the π substring, and a new $*$ symbol (representing an increase in intensity) is added at the end of the sequence. A new $-$ symbol (representing a new signal pre-processing step) can be placed in front of any 1 symbol which does not already have a $-$ in front of it.

It remains to specify the probabilities with which various sets of symbols are selected and re-assigned. One simple and natural model would be as follows. There is a fixed mutation probability $\mu > 0$ and each gene is subject to mutation with independent probability μ. When a gene mutates, it is assigned a new symbol and location within the sequence with each possible combination being equally likely. This *independent mutation* model is useful for building intuition but imposes more structure than is needed for the results. All that is necessary is the following assumption on the asymptotic probabilities with which large subsets are selected.

Definition 2. *Let M_n be a probability distribution over subsets of $\{1, \ldots, n\}$. We say that the family of distributions $\{M_n\}_{n \in \mathbb{N}}$ satisfies a* large-deviation condition *if there exists $\mu \in (0,1)$ and a function $\delta : (\mu, 1] \times \mathbb{N} \to (0,1)$ such that if T is any subset of $\{1, \ldots, n\}$, and $m \geq \mu$, then the probability under M_n of selecting a mutation set which includes more than a fraction m of elements from T is no greater than $\delta_m(|T|)$ and*

$$\limsup_N \frac{\delta_m(N+1)}{\delta_m(N)} \leq \beta(m) < 1.$$

This is a large-deviation property which limits the probability of selecting a large fraction m (greater than μ) of genes from any given large subset. This probability must shrink to zero at a rate which is asymptotically faster than some $\beta(m) < 1$. Note that by a standard result from large-deviation theory, the independent mutation model is a special case. Also note that it places no restrictions at all on how the selected set is re-assigned. [9]

[9] So for example it accomodates a model in which the selected set is re-assigned to maximize fitness given the fixed structure of the remainder. This would represent an optimizing (but not too patient) designer to whom ideas about how to improve the organism arrive randomly. Ideas which involve coordinated changes of larger and larger components of the organism have smaller and smaller probability. And in this context, if we think of the organism as a *theory* of the environment, the model can be interpreted to say something about the evolution of scientific theories.

3 Analysis

First, we consider an instructive benchmark case in which $q = 0$. In this case, the complexity of the organism is fixed and cannot increase. Then, because the mutation probabilities are strictly positive, with probability 1 the organism will be optimally adapted after some finite timespan.

Benchmark with $q = 0$ Consider an arbitrary organism $\mathcal{O}$ of complexity x. Let $\mathcal{O}^*$ be an optimal organism of the same complexity. There is a lower bound on the probability of a mutation large enough to transform any such $\mathcal{O}$ into $\mathcal{O}^*$. In the worst case, a change to the entire genetic structure will be required and the probability of such a large mutation is strictly positive by assumption. When $q = 0$, the organism will never increase in complexity and so this remains forever a lower bound on the probability of reaching an optimally adapted organism in a single step. It follows that with probability 1 the optimal organism will appear eventually. Moreover an optimal organism can never be replaced if the complexity of the organism cannot increase.

Proposition 1. *When $q = 0$, with probability 1 the organism is eventually optimally adapted, regardless of the initial complexity.*

The proposition shows that any asymptotic inefficiency that arises when $q > 0$ is not due to a simple problem of local optima. The model allows for arbitrarily large mutations at any point in time. Thus, any improvement, of any fixed size, if available for sufficiently long, will eventually be realized. On the other hand, this argument does not apply to improvements which require larger and larger mutations. Potentially, the organism can gradually improve at the margin by increasing in complexity, all the while intensifying the complementarity among its components. This would mean that substantial improvements decline in probability. Whether such improvements will be realized will depend on the rate at which this probability declines.

The main result of the paper concerns the case of $q > 0$ and asymptotic structural inefficiency.

Theorem 1. *Suppose $\mu < 1/6$. When $q > 0$ there is a positive probability that the organism will be forever kludged and thus asymptotically structurally inefficient.*

3.1 Proof of Theorem 1

Recall that we have assumed that $l > 1/2$. The parameter l determines the probability that each $\lambda_j = +1$. As discussed above, what matters for the optimal design of the organism is the sign of the moving average

$$L(k) = \frac{1}{k} \sum_{j=1}^{k} \lambda_j > 0.$$

Because $l > 1/2$, by the strong law of large numbers, there is probability one on the set of environments in which there exists some $\bar{k}$ such that $L(k)$ is positive for all $k > \bar{k}$. Throughout the proof, we fix such an environment λ and integer $\bar{k}$ and consider the stochastic process of evolution in that environment.

Definition 3. *Let $\mathcal{O}$ be a kludge. We say that a* drastic mutation *occurs if a mutation produces an organism $\mathcal{O}'$ with the same complexity but strictly larger fitness.*

Let η_x be the probability that an kludge of complexity x will undergo a drastic mutation. Consider the following simplified stochastic process. Let the states of the process correspond to the levels of overall complexity x of the organism. At each state, three transitions are possible. With probability q, the value of x increases by two. With probability $(1-q)\eta_x$, the process terminates. Finally, with the remaining probability, the process continues and the value of x is unchaged. Figure 4 illustrates. The process begins in state $\bar{x}$, defined as the smallest level of complexity such that the optimal organism has precision $\bar{k}$. We refer to this as the *stochastic exit process.*

Lemma 1. *The probability that a negatively aligned organism with precision $\bar{k}$ remains asymptotically structurally inefficient is bounded below by the probability that the stochastic exit process never terminates. This probability is positive if and only if*

$$\sum_{x} \eta_x < \infty.$$

Proof. A standard tool from the theory of countable-state Markov chains[10] indicates an analysis of the following system of equations in unknowns

[10] See (Billingsley, 1995, Theorems 8.4 and 8.5)

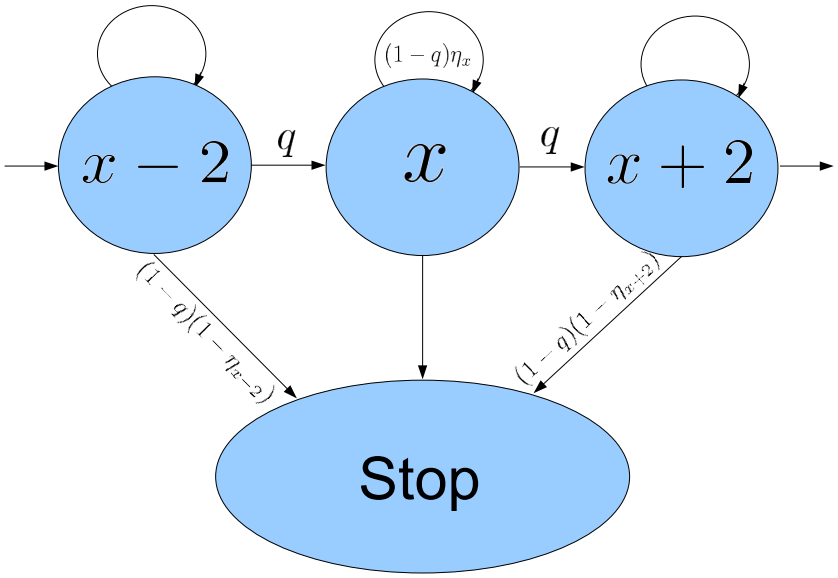


Figure 4: Stochastic Termination Process.

$Z_0, Z_2, \ldots$.

$$
\begin{aligned}
Z_0 &= qZ_2 + (1-q)(1-\eta_{\bar{x}})Z_0 \\
Z_2 &= qZ_4 + (1-q)(1-\eta_{\bar{x}+2})Z_2 \\
&\vdots \\
Z_x &= qZ_{x+2} + (1-q)(1-\eta_{\bar{x}+x})Z_x \\
&\vdots
\end{aligned}
$$

If the system can be solved by a bounded sequence Z_x, then the probability is strictly positive that the process will never terminate.

We can set $Z_0 = 1$ and then solve the system recursively, first writing

$$Z_{x+2} = \left(\frac{1-(1-q)(1-\eta_{\bar{x}+x})}{q} \right) Z_x$$

for each x, or

$$Z_{x+2} = \left(1 - \frac{(1-q)\eta_{\bar{x}+x}}{q}\right) Z_x$$

then recursively substituting to obtain

$$Z_{x+2} = \prod_{n=2}^{x} \left[1 - \frac{(1-q)\eta_{\bar{x}+n}}{q}\right].$$

We wish to show that $\lim Z_x < \infty$ which is equivalent to the convergence of the following series.[11]

$$\sum_{n=\bar{x}+2}^{\infty} \frac{(1-q)\eta_n}{q}.$$

which is convergent iff $\sum \eta_n < \infty$. □

Lemma 2. *There exists a function $M(k)$ such that the probability that any kludge with precision k has a drastic mutation is bounded above by $M(k)$ and*

$$\limsup_k \frac{M(k+1)}{M(k)} < 1. \tag{6}$$

Lemma 3. *There exists a function $C(k)$ such that there are at most $C(k)$ values of x such that a kludge of complexity x can have precision k, and*

$$\lim_k \frac{C(k+1)}{C(k)} = 1. \tag{7}$$

Combining lemmas 2 and 3 enables us to conclude the proof of Theorem 1 as they establish the bound

$$\sum_{x=\bar{x}}^{\infty} \eta_x \leq \sum_{k=\bar{k}}^{\infty} C(k)M(k) < \infty$$

which by Lemma 1 is enough to prove the theorem. □

We now turn to the proofs of Lemma 2 and Lemma 3. Each makes use of the following statistical lemma, whose proof is in Appendix A.

[11] Note that for any sequence of positive numbers R_n, $1 + \sum_1^x R_n \leq \prod_1^x (1 + R_n) \leq \exp(\sum_1^x R_n)$.

Lemma 4. *For any level of precision k,*

$$\operatorname{var}(\bar{\theta}_k) - \operatorname{var}(\bar{\theta}_{k-1}) = \frac{1 - \operatorname{var}(\bar{\theta}_{k-1})}{(k+2)^2}$$

and

$$\operatorname{var}(\theta) - \operatorname{var}(\bar{\theta}_k) < \frac{1}{k+3}$$

Proof of Lemma 2. (Preliminary. The proof covers the case of $l = 1$.)

Let $\mathcal{O}^*$ be a kludge with precision k^* and intensity i^*. The precision and intensity satisfy the "first-order condition"

$$i^* \cdot \left[\operatorname{var} \bar{\theta}_{k^*+1} - \operatorname{var} \bar{\theta}_{k^*}\right] < 2 \operatorname{var} \bar{\theta}_{k^*}.$$

Applying Lemma 4 and rearranging,

$$i^* < 2 \left[\frac{\operatorname{var} \bar{\theta}_{k^*}}{1 - \operatorname{var} \bar{\theta}_{k^*}}\right] (k^* + 3)^2 .$$

Let us define α by the following equation. It gives the maximum amount by which intensity can be reduced and still produce a drastic muation.

$$\begin{aligned} (i^* - \alpha k^*) \operatorname{var} \theta &= i^* \operatorname{var} \bar{\theta}_{k^*} \\ \alpha k^* \operatorname{var} \theta &< 2 \left[\frac{\operatorname{var} \bar{\theta}_{k^*}}{1 - \operatorname{var} \bar{\theta}_{k^*}}\right] (k^* + 3)^2 \left(\operatorname{var} \theta - \operatorname{var} \bar{\theta}_{k^*}\right) \end{aligned}$$

Applying lemma 4,

$$\alpha < \frac{2}{1 - \operatorname{var} \theta} \frac{k^* + 3}{k^*}. \tag{8}$$

Figure 5 illustrates the situation. The kludge $\mathcal{O}^*$ achieves the maximum fitness among all points on the budget-line for negative alignment. The horizontal axis is now the effective fitness of an organism. A necessary condition for an organism to achieve a higher fitness than $\mathcal{O}^*$ is for the (effective fitness, intensity) pair to lie above the budget line. It is convenient to normalize the axes by dividing by k^*, yielding Figure 5.

By definition of a kludge, a drastic mutation requires that π_0 change sign. This change by itself however cannot improve because all of the inputs will be misaligned and the organism will be choosing the negative of

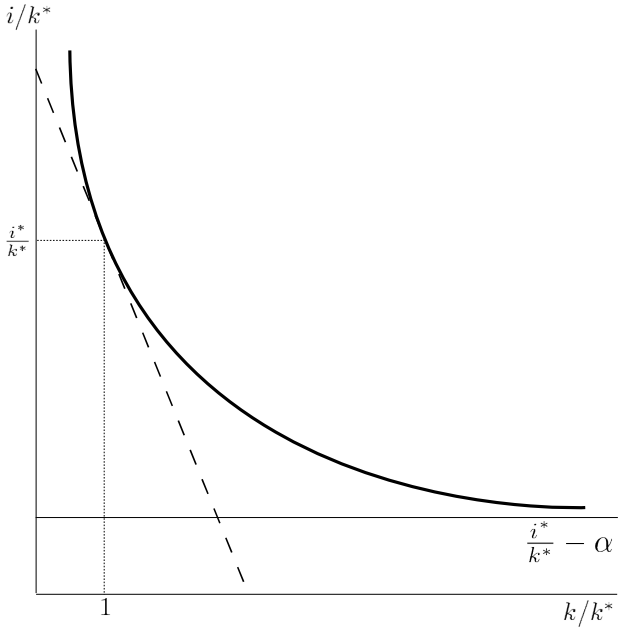


Figure 5: The optimal kludge $\mathcal{O}^*$. The axes have been rescaled. The asymptote represents the minimum intensity of any organism which achieves at least the fitness of $\mathcal{O}^*$.

the optimal location. So a drastic mutation requires accompanying changes elsewhere. To identify the necessary changes it is useful to consider a measure of the *effective precision* of an organism, defined as follows.

$$\tilde{k}(\mathcal{O}) = \pi_0 \sum_{j=1}^{k} \pi_j \lambda_j$$

Notice that the change in sign of π_0, by itself, reduces the effective precision to $-k^*$.

The fitness of any organism with effective precision $\tilde{k}$ is no greater than that of a kludge with precision equal to $\tilde{k}$and with the same intensity. To see this, note first that for any even number z the estimator

$$\frac{1}{\tilde{k} + z + 2} \sum_{j=1}^{\tilde{k}} \lambda_j \sigma_j$$

is strictly worse than the optimal estimator from a sample of size $\tilde{k}$ (see Equation 4). And when $z = 2$, we can show that the above estimator is strictly better than the estimate produced by an organism of precision $\tilde{k} + 2$ and effective precision $\tilde{k}$. The difference between the two estimators is that the latter incorporates two additional inputs, one of which is misaligned. When the signals from these two inputs have the same sign, the two estimators produce identical estimates. When the signals from these two inputs have opposite signs, the displayed estimator produces the optimal estimate while the latter does not. Now by induction, we can show that the displayed estimator is strictly better for any even number z.

It follows that a necessary condition for a drastic mutation is that the resulting organism have an (intensity/effective-precision) pair that is above the budget line in Figure 5.

We search for paths to improvement which combine the three types of mutations that can increase effective precision.[12]

1. Change a $*$ to a 1.
2. Change a $-$ to a 1.
3. Change a $-$ to a $*$.

Refer to Figure 6 below. We first show that mutations involving only changes of type 1 will not improve upon $\mathcal{O}^*$. Each change of type 1 reduces intensity by one unit and increases effective precision by one unit. Mutations of type 1 move along the solid line with slope -1. It improves upon $\mathcal{O}^*$ only if it moves past the intersection point with the dashed budget line.

Noting that the slope of the solid line is -1 and the slope of the budget line is -2, the vertical coordinate of their intersection is $\frac{i^*}{k^*} - 4$. When $k^* > 14$, this is below the horizontal asymptote and hence no organism with such a low intensity could achieve a fitness higher than that of $\mathcal{O}^*$. It follows that when $k^* > 14$, no mutation consisting only of changes of type 1 can improve.

Next we can rule out paths that involve mutations of type 3. Any improvement must move from the solid line to the right of the budget line. Each type 3 step moves two units to the right and one unit upward. By

[12]Here I am using the fact that $l = 1$ When $l = 1$, there is no gain to adding addtional $-$ symbols once π_0 switches from -1 to 1.

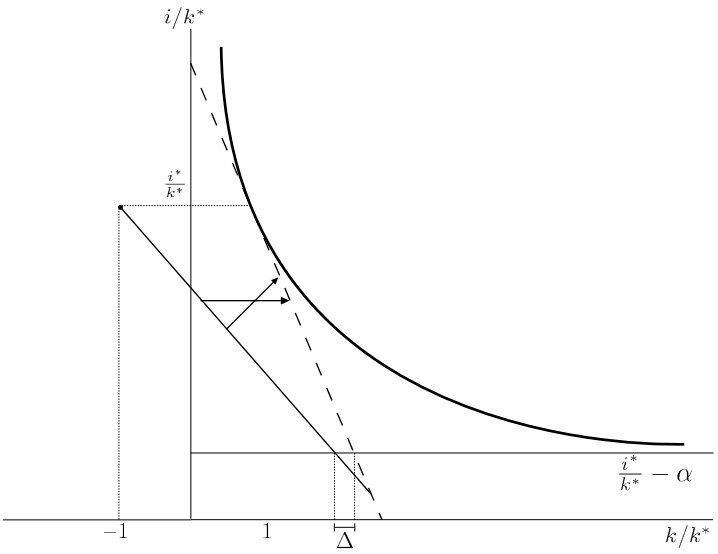


Figure 6: Paths to improvement. After the change in alignment, the (normalized) effective precision is -1. The downward sloping, horizontal, and upward sloping paths represent the three types of mutations.

comparison, each type 2 step moves three steps to the right. Because of the relative slopes of the two lines, type 2 steps close the gap more quickly.

Thus, in trying to find an improvement we can confine to paths that involve only mutations of type 1 and 2. Among these paths, we calculate the minimum number of type 2 steps required for an improvement. Any improvement must reach the budget line before falling below the horizontal asymptote of the indifference curve. Thus, the horizontal distance is bounded below by the minimum among points above the asymptote of the distance between the solid line (which is where the initial type 1 steps land) and the budget line. Because of the relative slopes of these lines, this minimum is obtained at the asymptote where the horizontal distance, denoted Δ in the figure, is equal to $\frac{4-\alpha}{2}$. (Moving a distance $4-\alpha$ upward from the intersection point puts this much distance between the two lines because of their relative slopes.) Applying Equation 8, Δ is at least

$$\frac{1}{2}\left[4 - \left(\frac{2}{1-\operatorname{var}\theta}\right)\left(\frac{k^*+3}{k^*}\right)\right] = 2 - \left(\frac{1}{1-\operatorname{var}\theta}\right)\left(\frac{k^*+3}{k^*}\right)$$

which, multiplied by k^* gives the total increase in effective precision result-

ing from the horizontal type 2 steps. Recalling that $\operatorname{var}\theta = 1/3$ and each type 2 step increases effective precision by 3, the total number of type 2 steps required is

$$\frac{k^*}{3}\left[2 - \frac{3}{2}\left(\frac{k^*+3}{k^*}\right)\right].$$

Using this result, for all $k > 14$ we can bound $M(k)$ by the probability that a large fraction of the pre-processor tokens are selected for mutation. We apply the large-deviation property (recall Definition 2.) Pick m to satisfy

$$\mu < m < 1/6.$$

Then since the number of $-$ symbols is k^* , the probability that a fraction m of these are selected for mutation is no greater than $\delta_m(k)$. Set $M(k) = \delta_m(k)$ where $\delta_m(k)$ is given by the large-deviation property. We have shown that the probability of a drastic mutation is bounded by $M(k)$ and since $\mu < m$, the large deviation property implies

$$\limsup_N \frac{M(k+1)}{M(k)} \leq \beta(m) < 1,$$

concluding the proof of the lemma. □

Proof of Lemma 3. Suppose that an organism has intensity i and precision k. Then if the organism is a kludge the following inequality must be satisfied.

$$i\left[\operatorname{var}(\bar{\theta}_k) - \operatorname{var}(\bar{\theta}_{k-1})\right] > \operatorname{var}(\bar{\theta}_{k-1}).$$

The left-hand side is the marginal increment to fitness from an increase in precision, while the right-hand side is the marginal increment to fitness from instead increasing intensity. [13] As complexity increases, a kludge with precision k will use the additional computational steps to increase intensity until it reaches the smallest level i which satisfies the corresponding inequality. It follows that the organism will increase its the level of precision from $k-1$ to k as soon as

$$i = \left(1 + \mathbb{1}_{\lambda_j=-1}\right)\frac{\operatorname{var}(\bar{\theta}_{k-1})}{\operatorname{var}(\bar{\theta}_k) - \operatorname{var}(\bar{\theta}_{k-1})}$$

(up to an integer,) and thus will spend at most

$$C(k) \leq \frac{2\operatorname{var}(\bar{\theta}_k)}{\operatorname{var}(\bar{\theta}_{k+1}) - \operatorname{var}(\bar{\theta}_k)} - \frac{\operatorname{var}(\bar{\theta}_{k-1})}{\operatorname{var}(\bar{\theta}_k) - \operatorname{var}(\bar{\theta}_{k-1})}$$

[13] In fact, when $\lambda_j = -1$, two tokens are required to increase precision, so in that case the left-hand side must exceed twice the right-hand side.

steps of the stochastic exit process with precision k.

Applying Lemma 4, we have

$$\begin{aligned} C(k) \leq C(k) &:= \frac{2\,\mathrm{var}(\bar{\theta}_k)}{1-\mathrm{var}(\bar{\theta}_k)}(k+3)^2 \\ &< \frac{2\,\mathrm{var}(\theta)}{1-\mathrm{var}(\theta)}(k+3)^2 \end{aligned}$$

so that

$$\lim \frac{C(k+1)}{C(k)} = \lim \frac{(k+4)^2}{(k+3)^2} = 1.$$

□

A Proof of Lemma 4

Define $\tau_j = \lambda_j \sigma_j$. Note that the $\mathbf{E}\tau_j = 0$ and so by Equation 4, $\mathbf{E}\bar{\theta}_k = 0$ and hence $\mathrm{var}(\bar{\theta}_k) = \mathbf{E}_\sigma\left(\bar{\theta}_k^2\right)$. We calculate

$$\mathbf{E}\left(\tau_j | \tau_1, \ldots, \tau_{j-1}\right) = \mathrm{Prob}(\tau_j = 1 | \tau_1, \ldots, \tau_{j-1}) - \mathrm{Prob}(\tau_j = -1 | \tau_1, \ldots, \tau_{j-1})$$

and by Equation 3 and the law of total probability,

$$\begin{aligned} &= \mathbf{E}\Big(\frac{\theta+1}{2} | \tau_1, \ldots, \tau_{j-1}\Big) - \mathbf{E}\Big(\frac{1-\theta}{2} | \tau_1, \ldots, \tau_{j-1}\Big) \\ &= \frac{\bar{\theta}_{k-1}+1}{2} - \frac{1-\bar{\theta}_{k-1}}{2} \\ &= \bar{\theta}_{k-1}. \end{aligned} \tag{9}$$

From Equation 4,

$$\bar{\theta}_k = \left(\frac{k+1}{k+2}\right)\bar{\theta}_{k-1} + \frac{\tau_j}{k+2},$$

so

$$\begin{aligned} \mathrm{var}(\bar{\theta}_k) &= \mathbf{E}\left[\left(\frac{k+1}{k+2}\right)^2 \bar{\theta}_{k-1}^2 + 2\left(\frac{\tau_j}{k+2}\right)\left(\frac{k+1}{k+2}\right)\bar{\theta}_{k-1} + \frac{1}{(k+2)^2}\right] \\ &= \mathbf{E}_{\tau_1,\ldots,\tau_{j-1}}\mathbf{E}\left[\left(\frac{k+1}{k+2}\right)^2 \bar{\theta}_{k-1}^2 + 2\left(\frac{\tau_j(k+1)}{(k+2)^2}\right)\bar{\theta}_{k-1} + \frac{1}{(k+2)^2} | \tau_1, \ldots, \tau_{j-1}\right] \end{aligned}$$

Applying Equation 9

$$= \mathbf{E}_{\tau_1,\ldots,\tau_{j-1}}\left[\frac{1}{(k+2)^2} + \frac{2(k+1)+(k+1)^2}{(k+2)^2}\bar{\theta}_{k-1}^2\right]$$

and we have

$$\begin{aligned}
\mathrm{var}(\bar{\theta}_k) - \mathrm{var}(\bar{\theta}_{k-1}) &= \mathbf{E}\left[\frac{1}{(k+2)^2} - \left(1 - \frac{2(k+1)+(k+1)^2}{(k+2)^2}\right)\bar{\theta}_{k-1}^2\right] \\
&= \mathbf{E}\left[\frac{1}{(k+2)^2} - \frac{1}{(k+2)^2}\bar{\theta}_{k-1}^2\right] \\
&= \frac{1 - \mathrm{var}(\bar{\theta}_{k-1})}{(k+2)^2}
\end{aligned}$$

establishing the first part of the lemma. To show the second part, note that

$$\begin{aligned}
\mathrm{var}(\theta) - \mathrm{var}(\bar{\theta}_k) &= \sum_{j=k}^{\infty} \mathrm{var}(\bar{\theta}_{j+1}) - \mathrm{var}(\bar{\theta}_j) \\
&= \sum_{j=k}^{\infty} \frac{1 - \mathrm{var}(\bar{\theta}_j)}{(j+3)^2} \\
&< \sum_{j=k+3}^{\infty} \left(\frac{1}{j}\right)^2 \\
&< \frac{1}{k+3}
\end{aligned}$$

□

References

BALIGA, S., AND J. C. ELY (2007): "Limited Memory and Sunk Cost Bias," coming soon.

BILLINGSLEY, P. (1995): *Probability and Measure*. Wiley and Sons.

DAWKINS, R. (1986): *The Blind Watchmaker*. W.W. Norton.

FRIEDMAN, M. (1966): "The Methodology of Positive Economics," in *Essays in Positive Economics*. University of Chicago Press.

RAYO, L., AND G. BECKER (2007): "Evolutionary Efficiency and Happiness," *Journal of Political Economy*, 115(2), 302–337.

ROBSON, A. (2001): "Why Would Nature Give Individuals Utility Functions?," *Journal of Political Economy*, 109, 900–914.

SAMUELSON, L., AND J. SWINKELS (2006): "Information, Evolution and Utility," *Theoretical Economics*, 1(1), 119–142.

SANDHOLM, W., AND A. PAUZNER (1998): "Evolution, Population Growth, and History Dependence," *Games and Economic Behavior*, 22, 84–120.