Proposal Technical Three-Pager

Nicholas Adams Judge, PhD

1 Motivations

1.1 TradFi

For two years I have been building a research environment for algorithmic trading, architected to grow into a system of artificial intelligence. An indirect outgrowth of my dissertation, finished in 2013, the remaining roadmap is 4-6 months long. Funding is appropriate at this point, as efficacy is related to scale.

Definitions of AI in contradistinction to machine learning read more like the product of engineers' convenience than of rich scientific theory. One arguable condition for AI in a market context: *Feature **and** model selection must be governed by some optimization routine, not human decision-making.*

Adapting the code base to this condition alters the character of research:

1. Assessment of fit becomes orders of magnitude more efficient
2. Domain-specific improvements are pushed throughout the system; e.g. if a model of 10 year Treasury notes is added, its forecasts are automatically given the opportunity to enhance models of ETH prices
3. Algorithms adapt more organically to structural breaks

There are of course other conditions that must be met to characterize a system in as grandiose a fashion as 'artificially intelligent.' But the above gives a sense of how the right architecture bends progress from the linear to the exponential, giving software the opportunity to outpace market efficiency.

1.2 DeFi

Section 2 outlines one way to meet the above criterion: an enhanced binary genetic algorithm (GA). The way GAs are typically used in-house at advanced quant funds is altered to take in data and algorithmic design features at scale.

That scale is sufficient not just for a fund, but for a decentralized network of scientists and quant developers. The AI system organized as a traditional fund could thus become the progenitor of, and participant in, a decentralized systematic trading fund capable of outpacing its centralized peers.

1.2.1 Defi and Black Box Quantitative Strategies

Black box quant funds are the tip of the TradFi spear: The most successful fund in history is effectively a software stack for hundreds of PhDs, and the high volume and salaries typical of the top quant funds structure their power

relationships with banks and other institutions. *If DeFi is going to rework finance for a more democratic future, the biggest body blow it could land would be a break with the existing institutional economics of expertise.*

1.2.2 Previous Gestures Toward Decentralization

Numerai is a fund built on crowdsourced forecasts of US equities. It offers its own token but has struggled to grow its AUM. This was predictable from its design, which worsened instead of improved the incentive structure of experts engaged in highly profitable activity. In a similar vein, Worldquant's 'Quant Championship' and other such forecasting tournaments appeal to students and data scientists with an amateur interest in market forecasts.

In reality, the multi-year development cycle of systematic strategies prevents their crowdsourcing. Section 3 explains how a properly-designed GA can lower the labor cost of participation: By providing a software stack that 'puts the pieces together' and protects IP, a decentralized system can attract more 'bite size' contributions - namely data and algorithmic design - from actual experts.

2 Genetic Algorithms

Genetic algorithms (GAs) are a stochastic optimization routine. They are inspired by natural selection and are therefore fairly intuitive.

A set of parameter values in a model is referred to interchangeably as a 'person,' 'individual' or 'chromosome.' Any (typically scalar) measure of fit can be used to assess individuals' fitness.

The first iteration - generation - starts with a population of individuals. Everyone's fitness is assessed. The fitter individuals are more likely to mate. This mating, or crossover, function is the main source of randomness: Which genes come from which parent is subject to chance. There is also a mutation function that introduces some randomness at the gene level. After the crossover and mutation steps are complete, a new generation is born. The algorithm repeats across a large number of generations. For example, in,

$$Y = f(X, \Theta) \tag{1}$$

Θ is a k-length vector, X a matrix with T rows and k columns, and Y a T-length vector. The fitness function, $g(Y, \hat{Y})$, can be the out-of-sample fit, a penalized likelihood, etc. Each $\theta_{1...k}$ is a gene, and each Θ an individual. After a max N iterations or some convergence criteria is met, the algorithm stops.

GAs are not computationally efficient, nor do they necessarily hone in on exact solutions. These two traits make them subideal in most scientific settings.

The weaknesses of GAs are the source of their strengths: They are relatively robust against local maxima and deal well with non-smooth spaces. Indeed,

contemplating the properties of genetic algorithms vis-à-vis other optimization routines helps one understand why organic matter organizes itself as it does.

3 Feature Selection & Binary Genetic Algorithms

With market data, feature selection is particularly important: variable lags, differences, and so on can create very large combinations of potential features.

In a binary GA, a wrapper is placed around (1). It takes in a k-length vector of zeros and ones representing the elements of Θ and columns of X. If zero, the element is left out of the model, and if one, it is included. E.g. in,

$$Y = h(\Omega, f(X, \Theta)) \tag{2}$$

Ω, a k-length vector of zeros and ones, is the subject of GA stochastic variation and selects subsets of Θ and X. $h()$ returns the same fitness measure $g(Y, \hat{Y})$.

3.1 A Binary GA for Complex Research Environments

(2) is a powerful step for one-off projects, helping the researcher analyze a larger set of models than could be run by hand. For a research environment to be channeled towards exponential progress, however, an alteration is needed:

$$Y = h(\Omega, f_{1...p}(X, \Theta)) \tag{3}$$

Now Ω is a $p + k$-length vector that chooses features and which version(s) of $f()$ to run. Genes $1...p$ can represent a choice to pursue a simple regression, a complex neural network, and so on.

4 Conclusion

The scientific research process can be thought of as a feedback loop wherein each finding increases the probability of the next. At some point, however, complexity overwhelms, causing so much information to leak out of that loop that predictive capacity stops growing or even shrinks.

Viewed in this manner, evolution can be thought of as a set of harshly-designed expiriments whose information aggregation scales beyond analytical limits.

Whether for an in-house software stack or a similar process with decentralized inputs, (3) by design forces contributions to be made in the same manner: piece-by-piece and tested redundantly across many situations.

With this core of scientific decision-making, it is a simple matter to design an incentive structure that (a) is tailored to sophisticated quant funds, (b) rewards early contributions, (c) protects IP by obfuscating participants' contributions, and (d) harnesses the resulting economies of scale to out-perform markets.