Referee Report: The Stability of Misperceptions under Mutations (Fudenberg and Lanzani, 2020)

Nicholas Wu

Fall 2020

Summary

In this paper, Fudenberg and Lanzani expand the parametric learning models from Esponda and Pouzo (2016) to consider whether misspecified learning is resistant to shocks in the parametric space.

The paper starts by considering the single-agent misspecified learning problem and uses much of the same formalization and assumptions as per prior work. The paper restates the solution concepts of Berk-Nash equilibrium, and what it means for a Berk-Nash equilibrium to be pure, unitary, (uniformly) strict, and self-confirming.

After reiterating the basic solution concept of Berk-Nash equilibrium, the paper then introduces the aggregate system, with a continuum of agents, and describe evolutionary dynamics for changes in parameter models within the population of agents. The general framework assumes that there is a distinct generation of agents for each time period, and the share of agents with a particular subjective model changes over time, where the share of agents either increases or decreases according to the payoff of the model in the previous period; that is, intuitively, the next generation at each time consists of higher fractions of the more successful models, which are Bayesian updated with the results of the previous generation.

In the main section, the paper introduces several models of model mutation, allowing agents to potentially change their misspecified worldview, and investigating whether such mutations can result in changes to steady-state beliefs in the aggregate system. Fudenberg/Lanzani define a Berk-Nash equilibrium (Θ, ψ) as resisting invasion from Θ' if for small enough ε, if an ε fraction of the agents change their parameter model to Θ', there exists a path such that as time goes to infinity, the system reverts back to the pre-mutation signal-action profile ψ. Fudenberg and Lanzani also argue that a necessary condition for a monomorphic steady state $\delta_\Theta \times \psi$ to be susceptible to invasion by Θ' (or equivalently, the inverse of a sufficient condition to be resistant to invasion) as explanation-improving; i.e. something in the new parameter space Θ' strictly reduces the expected KL divergence of the true signal distribution and the signal distributions in the model under parameters in Θ', as opposed to the original parameter space Θ.

The first type of mutation the paper considers are local mutations; specifically, everything in Θ' is at most some ϵ distance from something in Θ (under the L2 norm), and Fudenberg/Lanzani prove that every uniformly strict Berk-Nash equilibrium resists local mutations. Further, the paper introduces the notion of a *most informative ray*, or vector-parameter pair that characterizes the direction of enlargement of Θ that allows for the greatest reduction of KL divergence with the true distribution. Then Fudenberg/Lanzani prove

that if some optimal strategy along a most informative ray results in strictly greater expected utility, the steady-state is not resistant to local mutations, and if no such strategy exists in the support of the previous steady state strategy ψ, then the steady state is resistant to local mutation. Finally, with regard to local mutations, they show that in the directional environment, a Berk-Nash equilibrium that is resistant to local mutations must be pure or self-confirming or both.

The second main type of mutation the paper considers are one-hypothesis mutations. This starts with the presupposition that the parameter space of the subjective model, Θ, is defined as the feasible set under a set of constraints, and considers what happens when the feasible set is expanded under the removal of a single constraint. Similarly to local mutations, the paper shows that when an ε fraction of agents experience a one-hypothesis mutation, if some strategy in the set of best responses under the one-hypothesis mutation yields a strictly higher utility, the steady state is not resistant to one-hypothesis mutations. Further, a sufficient condition for resistance to one-hypothesis mutations is that every best response strategy π under the parameter space mutation yields strictly less utility than the previous steady state strategy ψ, and $\pi \in$ support(ψ) or ψ is uniformly strict.

The paper briefly discusses third model of mutation considered are attention-channeled mutations. The paper is light in this section; the paper defines attention partitions as the partition of outcomes which distinguish parameters with distinct best-reply strategies, and refines the notion of explanation-improving mutations to attention-improving mutations by restricting the the KL divergence sum over the attention partitions.

Lastly, they show a foundational result that establishes as the number of observations tends to infinity, the Bayesian updating process per generation has essentially full information of the previous generation's outcome distribution and hence the strategies taken by players are optimal.

Commentary

Technical Extensions

There are a some potential avenues for extending the work in this paper to slightly more mathematically general settings. Specifically, in the finite data learning foundation establishment, the authors make an explicit assumption that Θ, the parameter space for the subjective model, is finite. The paper conjectures that the same foundational result as their Proposition 6 holds even when Θ is non-finite, but does not prove this. One potential follow-up would be proving that their result still holds under non-finite Θ.

One of the main questions the paper does not address is the recovery rate of misspecified models that are stable under mutation. Specifically, Fudenberg/Lanzani concern themselves with whether the population state returns to ψ as the number of generations tends to infinity. One potential further investigation is how quickly populations that experience a resistable ϵ-mutation shock revert to their original state, and further how the rate of reversion to the original state depends on the size of the population shock ϵ under small-value approximation of ϵ.

The paper's relatively light content on attention partitions also could use further extension; Fudenberg/Lanzani argue that missperceptions are more stable under attention-channeled mutations. This argument is relatively straightforward, and it doesn't feel like anything particularly substantial is offered in regard

to handling the limited attention restriction. It might be interesting to characterize perhaps the equivalence class of mutations for which misspecified models Θ are resistant to attention-channeled mutations to Θ' if and only if the non-attention-filtered mutation to Θ' is resistant, or provide sufficient conditions on the learning problem for this equivalence.

Population Modeling

One of the interesting contributions of the paper is the evolutionary model for dynamics of a population's subjective model. Fudenberg/Lanzani propose that each generation of agents's subjective model is inherited from the previous generation, where a higher fraction of agents inherit their model from the more successful subjective models (in terms of expected utility of the outcome realization in the previous period under each subjective model, and assuming there are sufficiently many agents that this is observable).

The modeling of evolutionary dynamics seems to come from the necessity of modeling subjective model heterogeneity within the population, and allowing the degree of heterogeneity to change depending on the utility of the model (which appears naturally analogous to the biological concept of "fitness"). I am curious about the actual convergence rates of the evolutionary process. The paper generically assumes the payoff monotonicity assumption of Samuelson and Zhang (1992) as governing the evolutionary dynamics. Specifically, it might be interesting to specify the evolutionary function T they introduce and explore how the rate of reconvergence of populations under local mutations depends on parameters of T.

The model presented in this paper assumes that all agents share an identical utility function across the population. One direction of further research I might be interested in is exploring utility heterogeneity. One real-world example this might be useful in might be polarizing speech in public forums, where we suppose agents can choose to talk or not in a public forum, and their observed outcome depends on the average political sentiment of the forum, and agents are misspecified about the fraction of bad-faith actors in the forum. Specifically, it would be interesting to apply the similar analysis to consider the stability of this misspecified steady-state under the shock of some random ϵ-fraction of individuals suddenly expanded their belief-space of the average political sentiment.

Mutation Modeling

There are many potential further ideas for characterizing model mutation beyond the one-shot mutation models presented in this paper. The mutation dynamics considered for this paper mostly suppose that at time 0, some ε fraction of the population of agents tweak their subjective model for a little bit, and ask whether the new model can propagate among the population under the evolutionary dynamics of their model.

Additional Types of Mutation

All the models for mutation presented in this paper assume that the mutated parameter space Θ' is a strict superset of the original parameter model Θ. This helps model ideas of continuous changes in misspecification, and may be reasonable if assuming agents are conformist in a sense that mutated agents always have the option of resorting to the original parametric space. However, there might be interesting evolutionary dynamics if we introduce a small fraction ϵ of radically disrupting agents with an alternative parameter space Θ' disjoint from Θ, i.e. $\Theta \cap \Theta' = \emptyset$. I believe most of the proof of Proposition 1 in the paper follows

readily; that is, it seems fairly straightforwared to show that the paper's notion of "explanation-improving" in Θ' is necessary to propagate the mutated parametric model Θ'. One potentially interesting result of allowing such dramatic paradigm shift is that a misspecification may be locally stable at Θ under a certain learning structure, but not locally stable at Θ'. For example, the optimal strategy along the most informative ray for Θ does not yield strictly greater expected utility, and hence by the paper's Proposition 3, the subjective model Θ is resistant to local mutation, but perhaps under Θ', there does exist a most informative ray along which the optimal strategy does lead to model expansion towards a more correctly specified model. If we want to consider interactions between multiple mutations or sequences of mutations, we may want to consider the generalized notion of mutation rather than the locally-specific mutations considered in the paper.

Repeated, Stochastic Mutation after Unexpected Outcomes

One of the questions that arose for me throughout many of the lectures on misspecified learning was whether agents might change their model after seeing contrarian results that are inexplicable under their worldview. I think a long-term direction building off of the mutation work from Fudenberg/Lanzani would be to treat mutation as a stochastic process depending on the signal-outcome realizations. I think one avenue for exploration of this process would be to suppose agents not only perform a Bayesian update of their belief after each time period, but also generate a Bayesian credible interval prediction for the outcome realization in the next period (with some probability level $1-\alpha$). We could then model unexpected outcome realizations as catalyzing events for model mutation; for example, one potential model direction might be to suggest that when the outcome realization in a period lies outside the $(1-\alpha)$ Bayesian-credible interval predicted in the previous period, the agent takes a random ϵ-local mutation of their parameter space (i.e., the agent begins doubting their model and decides to expand their model after an unlikely event).

Note that after infinite time, even a correctly specified agents will almost surely doubt their own model at some point, but I think we can use the results from this paper to show that even though this happens, the population in the long term will retain their original parameter model. Preliminarily, I think that due to the mutation-stability results argued in this paper, it should not be very difficult to show that a correctly specified model is resistant to all mutations. I think it could be very interesting to characterize when misspecified models can correct themselves, and how a resistance parameter α dictating the agents' tolerances for seeing unusual results affects the rate at which a population's subjective model corrects itself. Further, I am curious what potential heterogeneity in α in a population might be able to do to impact the overall population's willingness to overcome misspecification.

One last idea in the same vein as the previous thoughts would be to consider instead of generating prediction intervals for the outcome realizations, supposing agent are sophisticated enough to test their own model by tracking a long-term statistic of the outcome realizations, and begin mutating their model if they fail to meet the expected convergence of their statistic. For example, one example might be to track the mean of some function of the outcome and signal, and mutating their model if this sample statistic fails to meet the expected convergence rate (i.e. falls outside a confidence interval predicted by the model).