A Simple Theory of Scientific Learning[*]

E. Glen Weyl[†]

September, 2007

Abstract

Scientists use diverse evidence to learn about the relative validity of various broad theories. Given the lack of statistical structure in this *scientific learning* problem, techniques of model selection and meta-analysis are not directly useful as quantitative guides. I use five simplifying assumptions to make the problem tractable by standard statistical methods. Combining Bayesian and frequentist approaches, I derive simple, intuitive rules for updating beliefs. The theory incorporates trade-offs among seemingly incomparable dimensions often used to judge models: ex-ante plausibility, precision, empirical accuracy and general applicability. I establish necessary and sufficient conditions for the consistency of the learning procedure which provide easy robustness checks for applied analysis and a simple algorithm for choosing a robustly consistent trade-off between precision and accuracy. I develop the theory in the context of a motivating application to social preference data collected by Charness and Rabin (2002). In contrast to the authors' analysis, I find (for a wide range of prior beliefs and parameter values) that after taking into account its greater precision, Selfishness is the best model of choice in the simple games they consider.

[†]Bendheim Center for Finance, Department of Economics, Princeton University, 26 Prospect Avenue, Princeton, NJ 08540: eweyl@princeton.edu.

> *In so far as a scientific statement speaks about reality, it must be falsifiable; and in so far as it is not falsifiable, it does not speak about reality.*

– Karl R. Popper, "The Logic of Scientific Discovery"

1   Introduction

Scientists, particularly social scientists, often use their knowledge of general theories to make predictions or prescribe policies in situations, like the design of international institutions, where collecting new data about the particular problem is prohibitively expensive or impossible. Careful comparison of the merits of different theories is therefore crucial. To aid such comparative analysis, scientists draw on diverse sources of evidence to learn about the relative validity of different theories. Because of the lack of statistical structure, no formal rules currently exist to provide quantitative guidance to this inference.

In this paper, I begin an attempt at developing such rules. First I provide a mathematical formulation of this *scientific learning* problem. A scientist uses her observation of the outcomes of a experiments, which have a finite number of possible outcomes, to learn about the relative validity of different theories which make (not necessarily unique) predictions about the outcomes of the experiments. I discuss why the problem cannot be directly solved through existing methods. I then make five simplifying assumptions that provide statistical structure to the problem:

1. *Bayesian inference*: the scientist believes one theory is "true" in the sense of describing the stochastic process generating experimental outcomes and learns, by Bayesian inference, about the probability that different theories are true, starting from a prior distribution over the truth of various theories.

2. *Independence*: each experiment is independent (conditional on the identity of the true theory).

3. *Theoretical imperfection*: even the true theory may make errors with probability independent of the identity of the theory.

4. *Uniformity*: conditional on an error occurring, the outcome of the experiment is uniformly distributed on the set of all possible outcomes of that experiment.

5. *Minimal theoretical structure*: conditional on no error, the outcome of the experiment is uniformly distributed on the outcomes predicted by the true theory.

I then use standard statistical techniques to provide a solution. First, I use Bayesian learning theory to derive a simple rule for updating beliefs. The resulting rule formalizes the trade-off between commonly cited criteria for judging theories: ex-ante plausibility, empirical accuracy, precision of predictions and broadness of applicability. Second, I apply frequentist methods (such as maximum likelihood estimation) to calibrate the crucial parameter of this learning rule, the rate at which errors occur, which controls the trade-off between precision and accuracy. I then fully embed the Bayesian learning within a classical framework to provide necessary and sufficient conditions for long-term consistency of the learning process in terms of the choice of this parameter. As long as the error rate is not chosen by the scientist very far from its true value, the learning process is consistent. These results provide a simple algorithm for choosing an error parameter that is consistent given a range of possible values for this error parameter.

The simplifying assumptions I make limit the applicability of the theory to contexts where there is no natural notion of distance between outcomes of experiments. In the appendix, I outline a strategy for relaxing these assumptions (and thereby complicating the theory) to make it more broadly applicable. The theory developed here may be viewed primarily as providing intuition for a more complete theory I am working to develop. Nonetheless, the theory in its current form is useful for some important applications. To make the exposition more concrete, I develop it in the context of a simple motivating application that conforms well to the assumptions of the theory.

I consider data collected by Charness and Rabin (2002), who ran a set simple experiments to determine which of several theories of social preferences were best able to explain behavior of subjects in a wide range of simple two-player, sequential binary choice games. Because many of the theories are consistent with both choices for one or both players in a particular game, the authors have no natural strategy for comparing the performance of the theories. Is a theory that makes a unique prediction 90% of the time but only explains 50% of behavior better or worse than a theory that makes a unique prediction 50% of the time but explains 90% of observed behavior? My focus on the trade-off between precision and accuracy makes their data a particularly natural application of my theory. When applied to this problem, the update rule can rationalize either an interpretation of their data as supporting Selfishness or Social Welfare (i.e. altruistic) preferences as the best theory of choice in the simple games they consider. The theory's primary contribution is to supply quantitative estimates of the range of prior beliefs and error rate choices supporting each interpretation. I find that the prior beliefs and parameter values needed to support the authors' interpretation of the data as supporting Social Welfare over Selfishness are extreme and probably implausible to most economists.

The paper following this introduction is organized into eight sections and four appendices. Section 2 discusses existing statistical approaches related to the scientific learning problem. Section 3 outlines the Charness-Rabin data and the problems with the way it was analyzed. This section motivates the development of the theory which follows. Section 4 formally develops the scientific learning problem. It then explains and motivates the assumptions I use to simplify the problem in the context of the Charness-Rabin application. It then derives and interprets basic formulae that can be used to apply the model. Section 5 is therefore dedicated to techniques for choosing the error parameter. I discuss a rule of thumb estimator loosely related to the method of moments, a maximum likelihood approach to estimation and a pseudo-Bayesian technique. Section 6 presents the results of applying the theory to analyze the Charness-Rabin data.

Because the error parameter is of such crucial importance, one may be concerned about the robustness of analysis to misspecification of this parameter. Section 7 therefore develops the core theoretical results of the paper by imbedding the subjective learning of the scientist in an objective probability framework. This yields conditions under which the learning rule implied by the theory is consistent despite misspecification of the error parameter. The learning rule always consistently learns the truth when alternatives are overly precise (subsets), but may be inconsistent if misspecified when alternatives are vaguer than (supersets of) the true theory. The error parameter conditions the trade-off between speed of consistency against subsets and inconsistency against supersets. This in turn provides a simple algorithm for choosing a robustly consistent error parameter, namely choosing the smallest error parameter that is guaranteed to be consistent given a range of possible values for the parameter. I show that the values of the error parameter

2

which support the Charness-Rabin interpretation of their data has substantial risk of inconsistency under reasonable conditions.

Section 8 returns to the criteria of plausibility, accuracy, precision and general applicability, discussing their importance in detail. These four criteria were drawn from the intersection of a recent methodology paper by Gabaix and Laibson (2007) and a classic treatise on induction by Carnap (1950) and I therefore refer to these criteria as the Carnap-Gabaix-Laibson Criteria. I argue that any reasonable theory of scientific learning should, as the theory developed here does, incorporate trade-offs among these criteria. Section 9 concludes by discussing directions for further research.

Appendix A addresses the fact that Charness and Rabin may have formulated versions of the social preference theories that were overly vague, as the analytic techniques they used did not penalize such vagueness. I therefore consider a simple "precision-augmentation" of the theories that makes them more falsifiable and therefore put them on a more even playing field with Selfishness. This check confirms my initial results. Appendix B provides a proof for one lemma in Section 7. Appendix C provides a proof of the primary theorem in Section 7. Appendix D, as discussed above, begins work towards extending the theory to make it more broadly applicable.

2 Related literatures

Three broad literatures in statistics and economics address aspects of the problem of scientific learning. The first, on model selection, shares with my analysis an emphasis on quantitative standards for selecting the best models of certain phenomena. The second, on meta-analysis, shares my emphasis on combining evidence from disparate sources. The third, on the game theory of empirical tests, shares my focus on formalizing standards from the philosophy of science. This section briefly discusses each of these literatures, as well as a paper from epidemiology that has a similar aim to this one, with an emphasis on their relation to my theory and the reasons why they are not directly applicable to scientific learning. Section 4, which formulates the problem formally, further reinforces this distinction.

The two approaches to model selection most related to my theory here are the Bayesian approach and the Statistical Learning Theory of Vapnik and Chervonekis. The Bayesian approach, summarized well in Gelman and Rubin (1995), compares theories which embed a full probability distribution over outcomes that could be realized. For reasons discussed below, this full Bayesian approach is cumbersome and overly demanding on applied analysts without substantial simplifying assumptions. Nonetheless the spirit of the model here is basically Bayesian and can be seen as supplying those simplifying structure necessary to make the Bayesian approach operational.

The Statistical Learning Theory of Vladimir Vapnik and Alexey Chervonenkis , summarized nicely in Vapnik (1995), is much more oriented than the Bayesian approach towards handling theories which specify only predictions, not full probability distributions over possible outcomes. It is also designed to be computationally tractable. However, it provides very loose bounds, rather than quantitatively precise posteriors, about the merits of various models. It therefore is not useful, particular in the small samples I consider, for providing precise estimates of how much weight show be put on various theories in a small set. Nonetheless, its focus on falsifiability and its rejection of simplistic notions of degrees of freedom in favor of measures of parsimony were a major inspiration for this work.

Other less theoretical model selection techniques are much farther from being applicable to the scientific

learning problem. For example, re-sampling techniques such as the bootstrap, jackknife and cross-validation described in Efron (1982) require a much more structured setting than afforded by the scientific learning problem to be valid. Strict degree-of-freedom based approaches, such as the AIC of Akaike (1974), do not speak the problem of theories that are inherently vague and, even when they do, can be very misleading when degrees of freedom capture very different things in different models being compared. While likelihood clearly plays a role in the model below, as in any Bayesian model, scientific learning often faces situations where the amount of independent information is far too small to ignore residual uncertainty about the correct theory. Therefore simple maximum likelihood estimation, while trivial given the model developed below, is usually inappropriate. Techniques based on information theoretic criteria, most prominently those that choose models with minimum description length such that suggested by Rissanen (1978), are widely believed to only be applicable in situations, usually in computer science, where description length is easily understood, measured and is closely related to the plausibility of a model. In economics and science more generally none of these criteria seem to hold.

The oft-cited Popperian view, as developed in Popper (1959), that among equally accurate models, the most falsifiable one should be favored has simply never been formalized and does not provide guidance on how falsifiability should be traded off against empirical accuracy. My theory can be viewed as an attempt to formalize Popper's persuasive philosophical reasoning.

All currently existing meta-analysis approaches are suited to inference for simple, clearly defined parameters and therefore require much more structure than the scientific learning problem contains. However, as I develop the theory further, these ideas from meta-analysis are likely to be very useful in correctly weighting across different experiments and different dimensions within a particular experiment. My independence assumption and general use of constant error rates abstracts away from such problems; however in including errors driven by statistical uncertainty in my application to the Charness-Rabin data, I am implicitly drawing on standard meta-analytic techniques. As I begin to develop the theory in the context of metric experiment spaces (see Appendix D), the connection to meta-analysis will become more explicit. For a survey of classical meta-analytic and quantitative review techniques see Rosenthal (1984) and for a more Bayesian approach see Eddy and Shachter (1990).

The literature in economic theory proper, most related to this paper is on the empirical testing of strategic experts. It considers whether an expert's claim to empirical knowledge can be verified if that expert is strategic. A brief survey of this work is provided in a recent paper by ?. They, like I, are interested in various strategies for empirically testing broad, vague theories based on ideas from the philosophy of science and methodological theory. However, they assume that theories are produced strategically. While this creates an important and interesting game theoretic problem, I think that my (classical) approach of seeing alternative theories as coherent hypotheses which should be tested on their merits, rather than as manipulable ploys for prestige, fits better with the way most scientists do and should interpret theories. Furthermore, none of the "tests" in the literature incorporate trade-offs among all of the criteria listed above, none of them formalize degrees of precision and none lead to posterior distributions over beliefs in various theories. Therefore, their models are not useful for applied meta-analysis of different theories. They are game theory fundamentally, not statistics.

A final paper closely related to mine develops a quite different methodology for addressing a similar question to the scientific learning problem I formulate here. However, Katz and Singer (2007) take an

entirely qualitative approach to the problem and their analysis is therefore more suited to situations where quantification is much more difficult than those my theory is appropriate for. Furthermore they primarily focus on the issue of the relevance and possible manipulation of data, rather than on the trade-off between precision and accuracy. While I think these issues are quite important, particularly in applying the more serious metric version of the theory, my primary contribution is in the trade-offs among the four listed criteria, not in how various evidence should be weighted. I therefore view their paper, like meta-analysis, as an eventual compliment to my work, rather than as a substitute.

3 Charness and Rabin (2002) data and analysis

In order to make the basic assumptions and mechanism of the theory as transparent as possible, I develop it in the context of a motivating example. Because of the simple structure of the Charness-Rabin data it is a particularly good fit with the assumptions of my theory and is easy to explain. Furthermore, it exemplifies in a simple way the problem of precision-accuracy trade-off that the theory addresses.

In their 2002 paper "Understanding Social Preferences with Simple Tests", Gary Charness and Matthew Rabin tried to overcome confounds existing in many experiments to test which models of social preferences are most consistent with a "broad range" of experimental data. Towards this end they devised twenty nine simple games designed to distinguish sharply between various theories. The paper addresses a number of issues, including the role of reciprocity and the dynamics of a few three-player games the authors consider. However, a primary focus is on using results from the twenty seven two player, binary choice games to distinguish among four theories of social preferences. All of these games had a very simple structure:

1. First player A chooses between a known payoff profile and giving player B a chance to play.

2. Conditional on being allowed by player A to play, player B faces the choice between two known payoff profiles.

3. In a few of the games, player A has no choice but to let player B play.

The authors provide simple parameterizations of the four theories of social preferences:

1. Standard selfishness: individuals act to maximize their payoff and are indifferent among alternatives which all achieve maximal payoff.

2. Difference aversion: individuals put (strictly) positive weight on the other individual's welfare when the other individual is earning a lower payoff (than oneself), but a negative weight on the other individual's welfare when the other individual is earning a higher payoff.

3. Social welfare: individuals put positive weight on the other individual's payoff always, but puts higher weight on the other player's payoff when the other player is behind than when the other is ahead.

4. Competitive: individuals always put negative weight on the other individual's welfare, but put higher negative weight on the other's payoff when the other is ahead than when she is behind.

The authors then used two basic strategies to analyze the data. The first was to assume all subjects have the same preferences, which they implement with errors, the rate of which is proportional to the utility cost of this error. The authors then calibrate the various different models and compare how well they organize the data. Such an analysis suffers from two major problems. First, if preferences are heterogeneous in the population so that different behaviors are not driven errors but by different utility functions then this analytical strategy is unlikely to give sensible results. A causal survey of their data does not seem to indicate that the stakes of experiments are strongly negatively related to the rate of errors off of theories with good fit, as would be necessary to support this procedure. Second, it seems to me unlikely that readers of the paper, or even the authors, take very seriously the estimated parameter values coming out the calibration or believe that these particular parameter values represent a reasonable alterative model to selfishness. Finally, this procedure has no way to quantify or take into account the greater degrees of freedom the social preference theories have.

The second strategy, upon which my analysis builds, is to calculate which actions in each experiment are consistent with each preference model *for some parameter value*. Because this sometimes depends on the expected actions of the other player, Charness and Rabin have two versions of this analysis, one allowing players to hold any beliefs about the other players likely actions and the other assuming players have correct beliefs. Because this "correct beliefs" version offers much sharper analytical possibilities, I focus on it here[1].

Charness and Rabin go on to calculate two measures of the "fit" of a theory. First, they ask what fraction of decisions are consistent with the predictions of the theory for some parameter value. Second, they ask what fraction of decisions are consistent with the predictions of the theory *when those predictions are unique* (i.e. independent of the parameter value). Both of these approaches suffer from the same problem: they entirely ignore the value of a theory being falsifiable. Suppose that one theory, call it the "vague theory", predicts a unique outcome in one experiment and in all other experiments is consistent with any outcome. Further imagine that in the one experiment where it predicts uniquely, 90% of subjects act according to its prediction. Consider a second theory, call it the "precise theory", that predicts uniquely on every experiment and about 70% of subjects act in accordance with its predictions. It would seem clear that the second model should be judged better: it gives precise predictions that consistently are not falsified, a performance that would be extremely unlikely if there were not "something to" this model. On the other hand the first model might well have gotten that one experiment right by chance. Yet in both of Charness and Rabin's measures, the first theory would perform better. In fact, their analysis exactly follows this path: they conclude the "social welfare" model out-preforms selifshness[2]. However, selfishness is falsifiable (makes a unique prediction) much more often in their experiments than social welfare preferences, so it is difficult to interpret their conclusions.

This situation is one concrete and relatively simple example of a general problem: how should a scientist combine many tests of different, potentially vague theories to learn about their general validity? In the

---

[1]Thus, in this context, the assumptions "all subjects are perfectly rational" and "all subjects have rational expectations about the proportion of partners choosing each strategy" can be seen as identifying assumptions for my empirical analysis. The issue of additional identifying (maintained) assumptions used to allow the model to be applied is an important one.

[2]The authors begin their paper with the claim that "Participants in experiments frequently choose actions that do not maximize their material payoffs when their actions affect others' payoffs." They later state, when considering the fraction of behavior explained by each theory, that "the proportion of explained by social-welfare preferences is significantly higher the proportions explained by the other three theories." Finally, in comparing the performance of theories when their predictions are unique the authors again argue "we see that social-welfare preferences substantially outperform the other models."

section that follows I formulate this problem formally, discuss why it cannot directly be solving through existing statistical methods and propose five assumptions that make the problem tractable.

4 The problem of scientific learning and a simple solution

4.1 The problem

Charness and Rabin collected data about choices in a set of simple two player, binary choice games. While one could use every choice made by every participant, I will focus on the choices made by a majority of participants in any particular role, in particular game and view this as being the outcome of the experiment. From this perspective, the behavior of a majority of participants in any role in a game may have two outcomes. Therefore, it is certainly reasonable in this setting (and many other settings) to assume that outcomes of experiments are discrete.

Definition 1. *An* experiment *is a finite set* Θ.

The experiments they chose to run are elements of several, increasingly large sets: the set of all two player binary choice games, the set of all two player discrete choice games, the set of all experimental games, the set of all human strategic interactions, etc. To interpret or learn from the Charness-Rabin data, it is useful to identify a set of situations from which the games they consider might reasonably be randomly drawn. That is, I will view their experiments of being representative of some class of situations[3]. In this case, a reasonable class would be the set of all two player, binary choice games where the first player chooses either to take a certain payoff or to allow the second player to move and, conditional on moving, the second player faces a binary choice with known payoffs.

Definition 2. *The* studied phenomenon *is a (generally infinite) set of experiments* Γ.

Of course, Charness and Rabin do not observe the outcomes of all such games; rather they observe the outcomes of only the finite number of experiments that they ran.

Definition 3. *The* observation Ξ *is a finite subset of* Γ.

The purpose Charness and Rabin had in mind when collecting their data was to test various different theories of social preferences. These theories are not full probabilistic models of the set of all possible outcomes, but rather are ways of generating predictions, in any given experiment, as to what the outcome of that experiment will be. These predictions are not necessarily unique: sometimes, for example, Social Welfare preferences are consistent with both choices a participant might make. More generally, a theory might be consistent with some strict but non-singleton subset of the possible outcomes of an experiment.

Definition 4. *A* theory *is a mapping* $\Theta \mapsto \lambda(\Theta) \subseteq \Theta$, *called the set of* outcomes predicted by λ *for* Θ. *Let the* theory space Λ *be the set of all possible theories satisfying this definition.*

In order to learn from an experiment, one of course needs to observe its outcome as well as the set of possible outcomes. In the Charness-Rabin application, I identify the outcome of any experiment as the choice in that role made by the majority of subjects.

---

[3]The performance of various theories in binary choice situations should be viewed itself as one of many "experiments" for learning about human behavior.

Definition 5. *A* world *is a mapping* $\omega : \Gamma \mapsto \omega(\Theta) \in \Theta$. *Let* Ω *be the set of all possible worlds and some* $\omega^\star \in \Omega$ *be called the* true world. $\omega^\star(\Theta)$ *is called the* outcome *of* Θ.

These definitions together provide a formulation of the *problem of scientific learning*. A scientist observes the outcomes of all experiments in the observation $\{\omega^\star(\Theta)\}_{\Theta \in \Xi}$ and wants to learn which theory in Λ is the best theory (call it $\lambda^\star$). This formulation allows a more precise statement of what makes the scientific learning problem different from other statistical problems. There are four differences, which I list in increasing order of importance.

1. First, we naturally think of the sample size $|\Xi|$ as being quite small. This means that asymptotic approximations (if any were valid) would be unappealing. But more importantly it means that, realistically, even after observing Ξ different scientists will still have substantial disagreement about which theory is best. Therefore simple hypothesis testing is not likely to be appealing. This of course is not a problem for Bayesian methods.

2. More significant, the space Γ from which the experiments are drawn has very little structure. There are no "natural" parametric distributions over it. This poses a particular problem, again, for frequentist approaches.

3. Even more substantially, the "parameter space" in this problem, Λ is enormously largely and more importantly has very little structure. It is therefore difficult to think about how a scientist could have a prior distribution that is non-dogmatic over this space. Confidence sets in Λ are probably not terribly appealing either.

4. Finally and most importantly, there is no "obvious" way to think about the joint distribution of $(\lambda^\star, \omega^\star)$. Theories in this model do not specify a distribution over Ω, but rather make deterministic predictions about the value of $\omega^\star$ at particular points in Γ. Therefore work is needed in order to transform this inference problem into a proper statistical question.

Because of these problems, none of the traditional statistical approaches discussed in Section 2 can be applied directly to the problem. In order to provide the structure necessary to make the problem tractable by traditional statistical methods, I will make a series of five assumptions, motivating and interpreting them in the context of the Charness-Rabin application.

4.2   Five simplifying assumptions

To begin, one must ask what is meant by learning which theory is best. A way to formally understand this is to view one theory as being "true", in the sense of providing a good description of the stochastic process generating observed data. Then the scientist's goal is to learn about which of the theories tested is this true[4] theory. From a Bayesian perspective, such learning requires that the learning agent, henceforth the *scientist*, has a prior distribution over which theory is true. While Bayesians would argue this assumption is always (axiomatically) justified, as a cautious pseudo-Bayesian I would argue that it is particularly reasonable here as scientists usually have a fairly good sense (devote a lot of time to thinking about) the *a*

---

[4]In section 7, I discuss how theories close to the true theory will be consistently learned using the learning rule developed below. Thus we might interpret this sentence as reading "which of the theories test is closest to this true theory."

priori plausibility of different theories. However, note that these priors will usually be dogmatic[5], for the reasons discussed above; in fact, the scientist will usually only have a few theories in the support of her priors. For example, in the Charness-Rabin application there are only four theories considered.

Assumption 1. *The scientist believes that one theory* $\lambda^\star \in \Lambda$ *is the* true theory *but is uncertain as to its identity. The* plausibility distribution $\pi : \lambda \mapsto \pi(\lambda) \in [0,1]$ *is a probability distribution over* Λ *that represents the scientist's priors over the different theories being the true theory.*

What the scientist learns from the observation about the true theory, as well as the true world, will depends on her beliefs about the joint distribution of the true theory and the true world. One way to substantially simplify this problem is to assume that the scientist believes that each experiment is independent, conditional on the identity true theory. This assumption is somewhat unrealistic as some experiments may be very similar to one another so that even conditional on the value of the true theory their outcomes may be correlated. While this is an important concern and an enriched version of the theory[6] would provide a means for handling this problem, as a baseline I will assume independence. Letting P be the probability distribution of the scientist, this assumption can be written formally as:

Assumption 2. $P\Big[\omega^\star(\Theta_1) = \theta_1, \ldots, \omega^\star(\Theta_N) = \theta_N \big| \lambda^\star = \lambda\Big] = \prod_{i=1}^{N} P\Big[\omega^\star(\Theta_i) = \theta_i \big| \lambda^\star = \lambda\Big]$ *for all* $\{\Theta_i\}_{i=1}^{N} \subseteq \Gamma$ *and all* $N \in \mathbb{N}$ *such that* $\Theta_i \neq \Theta_j$ *for* $i \neq j$.

While this simplifies calculations across experiments somewhat, it still does not specify what it means for a theory to be the "true theory". One might be tempted to think that if a theory is true, it should always be consistent with realized outcomes. This seems somewhat unrealistic and overly ambitious, however, at least in the social sciences. Even good economic models (and scientific theories more generally) are not consistent with all available evidence and even in the cases when they are they usually provide only good approximations of the empirical outcomes. Furthermore, assuming that the true theory is always consistent with available evidence would lead to a strong preference for vague theories which fit available evidence by luck. A simple way to capture the inherent limits on the accuracy of a theory is to assume that, in any experiment, there is some chance that the true theory will be make an error.

Assumption 3. *The scientist believes that even the true theory* λ *may be* imperfect *in the sense that it makes errors with some probability* $\Theta \mapsto \sigma(\Theta) \in [0,1]$ *that may depend on the experiment but is independent of the identity of the true theory.*

The scientist might believe that the probability of such an error depends on the identity of the true theory. By assuming away such dependence, I focus the theory below on learning about the identity of the true theory for scientists that can only have a rough sense of the correct error rate, rather than focusing on theory-conditional error rates precisely and then learning the correct theory[7].

---

[5]Therefore they should not be viewed as the scientist's "true" or full priors over theories. Rather they represent a simple way of giving statistical structure to the problem of comparing theories.

[6]Again, one has to trade off getting things "right" against allowing too much flexibility or demanding too much from applied modelers. I will later discuss where, in this application, this assumption becomes strained.

[7]An alternative approach, which builds on suggestions I received from Paul Milgrom and Wolfgang Pesendorfer, would be to obtain an estimate of the error parameter conditional on each theory through maximum likelihood estimation (as I do below) and then update beliefs with each theory using its "most competitive" (MLE) estimate of the error parameter. A nice feature of this approach is that it theoretically simple and very internally consistent. It should be easy to show that in the long-run the theory-specific estimate by MLE for the true theory converges to the true theory-specific error rate and that,

Now I need to define exactly what I mean by the true theory "making an error". A simple assumption, very loosely justified by Laplacian maximum entropy arguments[8], is that, when the true theory makes an error, the outcome is uniformly distributed on the set of all possible outcomes[9].

Assumption 4. *The scientist believes that with probability $\sigma(\Theta)$, regardless of the true theory, $\omega^\star(\Theta)$ is uniformly distributed on Θ.*

The crucial substance of this assumption is that all outcomes are equally close to one another. Suppose that, rather than discrete choice experiments, we were analyzing models that predict inflation. Conditional on a model that predicts 1% inflation being correct we might think it reasonable that 1.5% inflation would be realized. But we would not think 7% inflation would be likely. On the other hand, in a some discrete settings where any pair of outcomes has roughly equal similarity (vacuously true in binary choice), it is a reasonable approximation. In Appendix D, discuss how the basic structure of the model here can be extended to more realistic spaces with metrics.

Finally, if the theory does not make an error, which occurs with probability $1 - \sigma(\Theta)$, then we should expect the outcome to be among those predicted by the theory. But this does not immediately imply a distribution over outcomes predicted by the theory, unless the theory makes a unique prediction. Again, the simplest way deal with this is to assume, in the spirit of maximum entropy, that the entire content of the theory is its predictions and therefore that the distribution of the outcome if the true theory does not make an error is the restriction of the distribution under error to the set of predictions of the theory. Given assumption 5, this implies that if there is no error by the true theory on a particular experiment, then the outcome of that experiment is uniformly distributed on the outcomes predicted by the true theory.

Assumption 5. *The scientist believes that with probability $1 - \sigma(\Theta)$, $\omega^\star(\Theta)$ is distributed uniformly on $\lambda^\star(\Theta)$.*

Assumptions 3 and 5 together imply that

---

given this correct estimate, the scientist will consistently learn the identity of the true theory. Furthermore, given that the procedure I roughly envision being one of learning about the identity of the true theory and then learning about the error rate given that theory, this approach would seem somewhat more internally consistent than the one advocated here.

However, this approach seems somewhat less satisfying in application than in theory. First it would be very difficult to preform robustness checks, like those below, on this procedure. Given that two stages of long-run convergence are needed to justify its consistency, I am doubtful that its small sample properties would be desirable. Furthermore, I think it takes too seriously the the notion that the scientist is trying to find the theory that most closely matches the stochastic process generating outcomes. The goal of the scientist learning the true theory may not be to literally use it as a probabilistic model to predict on the next experiment. Instead, she may use standard Bayesian model selection criteria and use the output of scientific learning to obtain priors. In this setting, theory-specific error rates seem odd. However, given the simplicity of this alterative approach it offers an easy way to develop a working theory in the metric experiment context I begin to develop in Appendix D. Thus if the mathematical challenges of robustness results become overly difficult there, this is an alternative that is probably preferable to the approach developed here without robustness results.

[8] For classic philosophical defence see Carnap (1950), Laplace (1814) orLaplace (1812). For the classic modern treatment of the maximum entropy approach see Jaynes (1957) and for a more recent summary and defense of the methodology see Jaynes (1982).

[9] This corresponds roughly to the idea of a uninform "prior" in experiments: in the absence of knowing which theory is correct, every outcome of the experiment is equally plausible. In the Charness-Rabin application, this assumption is quite compelling as it is not clear what, other than theories of choice, would make one outcome more likely than another. In other applications, some outcomes may be more plausible than others for reasons orthogonal to which theory is correct. In such instances, it is easy to modify the model to put such an "ex-ante shape" on the experiment space. However, one should be cautious about introducing this extra degree of freedom, as it may somewhat increase the demands on analysts and reduce transparency.

$$P\Big[\omega^\star(\Theta) = \theta | \lambda^\star = \lambda\Big] = \begin{cases} \frac{\sigma(\Theta)}{|\Theta|} & \theta \notin \lambda(\Theta) \\ \frac{1-\sigma(\Theta)}{|\lambda(\Theta)|} + \frac{\sigma(\Theta)}{|\Theta|} & \theta \in \lambda(\Theta) \end{cases} \tag{1}$$

Together with Assumptions 1 and 2 these completely define the scientists joint probability distribution over $\omega^\star$ and $\lambda^\star$.

4.3 Basic solution

Having set up the basic probabilistic structure of the model, I can now solve for Bayesian update rule (Bayes Factors). To motivate the learning rule, consider the scientist's posterior beliefs after observing the outcome of all experiments in Ξ. By assumption 1 and Bayes's Rule:

$$P\Big[\lambda^\star = \lambda | \big\{\omega^\star(\Theta)\big\}_{\Theta \in \Xi}\Big] = \frac{P\Big[\big\{\omega^\star(\Theta)\big\}_{\Theta \in \Xi}\Big| \lambda^\star = \lambda\Big] \pi[\lambda]}{P\Big[\big\{\omega^\star(\Theta)\big\}_{\Theta \in \Xi}\Big]}$$

From now on I will use the short hand that $P[\cdot|\Xi] \equiv P[\cdot|\big\{\omega^\star(\Theta)\big\}_{\Theta \in \Xi}]$ and $P[\Xi|\cdot] \equiv P[\big\{\omega^\star(\Theta)\big\}_{\Theta \in \Xi}|\cdot]$. For any two theories $\lambda, \lambda' \in \Lambda$:

$$\frac{P\big[\lambda^\star = \lambda | \Xi\big]}{P\big[\lambda^\star = \lambda' | \Xi\big]} = \frac{P\big[\Xi \big| \lambda^\star = \lambda\big]}{P\big[\Xi \big| \lambda^\star = \lambda'\big]} \frac{\pi(\lambda)}{\pi(\lambda')}$$

Assumption 1 takes plausibility (priors) as given, so the interesting term above is the Bayes Factor, which I will sometime refer to as the *learning rule*. Using assumption 2 and equation 1 allows us to re-express the learning rule in its *error form*:

$$\frac{P\big[\Xi \big| \lambda^\star = \lambda\big]}{P\big[\Xi \big| \lambda^\star = \lambda'\big]} = \prod_{\Theta \in \Xi} \frac{P\big[\omega^\star(\Theta) | \lambda^\star = \lambda\big]}{P\big[\omega^\star(\Theta) | \lambda^\star = \lambda'\big]} = \prod_{\Theta \in \Xi} \frac{\frac{\sigma(\Theta)}{|\Theta|} + \frac{1-\sigma(\Theta)}{|\lambda(\Theta)|} 1_{\omega^\star(\Theta) \in \lambda(\Theta)}}{\frac{\sigma(\Theta)}{|\Theta|} + \frac{1-\sigma(\Theta)}{|\lambda'(\Theta)|} 1_{\omega^\star(\Theta) \in \lambda'(\Theta)}} \tag{2}$$

If we let the *informativeness* of an experiment $\mu : \Gamma \to \mathbb{R}_+ \cup \{\infty\}$ be defined by $\mu(\Theta) = \frac{1-\sigma(\Theta)}{\sigma(\Theta)}$, then we can rewrite expression 2 in its *informativeness form*:

$$\prod_{\Theta \in \Xi} \frac{1 + \mu(\Theta) \frac{|\Theta|}{|\lambda(\Theta)|} 1_{\omega^\star(\Theta) \in \lambda(\Theta)}}{1 + \mu(\Theta) \frac{|\Theta|}{|\lambda'(\Theta)|} 1_{\omega^\star(\Theta) \in \lambda'(\Theta)}} \tag{3}$$

4.4 Interpretation

To better understand the learning rule, I will focus on the case when $\sigma(\Theta) \equiv \bar{\sigma}$ or equivalently $\mu(\Theta) \equiv \bar{\mu}$. To interpret the preceding expressions, it is useful to consider the extreme cases when $\bar{\mu} \to 0$ ($\bar{\sigma} \to 1$) and $\bar{\mu} \to \infty$ ($\bar{\sigma} \to 0$).

In the first case, note learning will occur towards λ over λ' if and only if:

$$\prod_{\Theta \in \Xi} \frac{1 + \bar{\mu} \frac{|\Theta|}{|\lambda(\Theta)|} 1_{\omega^\star(\Theta) \in \lambda(\Theta)}}{1 + \bar{\mu} \frac{|\Theta|}{|\lambda'(\Theta)|} 1_{\omega^\star(\Theta) \in \lambda'(\Theta)}} > 1$$

This is equivalent to

$$\sum_{\Theta \in \Xi} \text{Log}\left[1 + \bar{\mu}\frac{|\Theta|}{|\lambda(\Theta)|}1_{\omega^\star(\Theta) \in \lambda(\Theta)}\right] - \sum_{\Theta \in \Xi} \text{Log}\left[1 + \bar{\mu}\frac{|\Theta|}{|\lambda'(\Theta)|}1_{\omega^\star(\Theta) \in \lambda'(\Theta)}\right] > 0$$

In the limit as $\bar{\mu} \to 0$ a first-order Taylor approximation of $\text{Log}[1+x]$ about $x = 0$ allows this expression to be simplified to

$$\bar{\mu}\sum_{\Theta \in \Xi}\frac{|\Theta|}{|\lambda(\Theta)|}1_{\omega^\star(\Theta) \in \lambda(\Theta)} - \bar{\mu}\sum_{\Theta \in \Xi}\frac{|\Theta|}{|\lambda'(\Theta)|}1_{\omega^\star(\Theta) \in \lambda'(\Theta)} > 0$$

We can define *precision*[10] of a theory in an experiment to be $\rho(\lambda, \Theta) \equiv \frac{|\Theta|}{|\lambda(\Theta)|}$. This seems like a reasonable definition, as this is the inverse of the fraction of possible outcomes that the theory predicts may occur and is therefore a measure of how precise the theory's prediction is in that experiment. Then the above expression becomes:

$$\sum_{\Theta \in \Xi}\rho(\lambda, \Theta)1_{\omega^\star(\Theta) \in \lambda(\Theta)} > \sum_{\Theta \in \Xi}\rho(\lambda', \Theta)1_{\omega^\star(\Theta) \in \lambda'(\Theta)}$$

We can call this expression the *sum of correct precisions* of a theory in an observation $\nu(\lambda, \Xi) \equiv \sum_{\Theta \in \Xi}\rho(\lambda, \Theta)1_{\omega^\star(\Theta) \in \lambda(\Theta)}$. Thus for $\bar{\mu} \to 0$, the scientist learns towards whichever theory has the greatest sum of correct precisions. This can be seen as placing the maximal weight possible on precision over accuracy: even theories which are highly empirically inaccurate may be updated in favor of, as long as they occasionally produce a correct prediction that is highly precise. $\bar{\mu} \to 0$ is the same as $\bar{\sigma} \to 1$, so this result is intuitive: if one believes that even the best theories are likely to fail to predict well often, then accuracy will not be very important.

On the other hand, consider the case when $\bar{\mu} \to \infty$ (or $\bar{\sigma} \to 0$). In this case the learning rule becomes

$$\prod_{\Theta \in \Xi}\frac{\rho(\lambda, \Theta)1_{\omega^\star(\Theta) \in \lambda(\Theta)}}{\rho(\lambda', \Theta)1_{\omega^\star(\Theta) \in \lambda'(\Theta)}}$$

so that a single inaccuracy causes the probability that a theory is the true theory to go to 0. Conditional on both theories being perfectly accurate, the rule updates in favor of the one with the greater product of precisions. But for a rule that disqualifies entirely any theory that is not perfectly accurate, precision is obviously of secondary importance. Thus as $\bar{\mu} \to \infty$ maximal weight is placed on accuracy. If one lives in a world where the correct theory never makes mistakes, accuracy should be crucially important in judging which theory is correct[11]. $\bar{\mu}$ can thus be viewed as a parameter governing the accuracy-precision trade-off, as well as the speed of learning from any particular experiment. Greater values of $\bar{\mu}$ imply that accuracy is more important and quicker learning; lower values of $\bar{\mu}$ place a greater weight on precision and slower learning.

---

[10]Note that if we relax assumption 5 (that is, even when an error is made the distribution over outcomes is not flat) and maintain all other assumptions, then the relevant notion of precision will be the inverse ex-ante probability of the theory's predictions. Thus perhaps falsifiability is a better word than precision, as really what matters is the likelihood with which, under the "ex-ante" error distribution, the theory would be falsified.

[11]Perhaps this should be viewed as true in the physical sciences, where as Einstein famously argued, a single experiment is sufficient to disprove a theory.

5   Choice of the error rate

Having developed the basic structure of the model, I can now apply it to the Charness-Rabin data that motivated it. To operationalize the model I still need a strategy for choosing the error function. The constant error rate assumption seems quite reasonable in this context and has the advantage, as I will highlight below, of allowing the data to tell us something about an appropriate value for this parameter. However, in the Charness-Rabin application it has one important weakness. I interpret the outcome of an binary choice game to be the action taken by a majority of participants playing that game. However, this majority is a sample, not a population, majority. Therefore some correction for statistical errors (in addition to fundamental errors of the true theory) is important.

In particular suppose we observe that x subjects make choice 1 in some experiment and $y \leq x$ subjects make choice 2. The standard statistical asymptotic estimate of the probability this outcome would have been obtained if it is in fact the case that the majority of the population would tend to choose choice 2 is:

$$\tilde{p}(x, y) \equiv 1 - \Phi \left(\frac{(x + y)\left(\frac{x}{x+y} - \frac{1}{2}\right)}{\sqrt{(x + y - 1)xy}} \right)$$

where Φ is the standard normal CDF. Thus if we assume that the outcome of the experiment is the *population* majority, then if there is a constant error of $\bar{\sigma}$ the observed error rate will be (well estimated by):

$$\bar{\sigma} + 2(1 - \bar{\sigma})\tilde{p}(\Theta) \tag{4}$$

where $p(\Theta) \equiv p(x, y)$ for the appropriate x and y corresponding to experiment Θ. This model of the error rate finds reasonable support in the data: the correlation between errors made by the selfishness model (for example) and the value of $\tilde{p}$ is .56.

Nonetheless, one is still left with the problem of choosing the base error rate $\bar{\sigma}$. If the number of experiments in the observation is very small, one is essentially forced to choose an arbitrary value for this parameter that seems reasonable. However, when, as in the Charness-Rabin data, the observation has at least moderate size, 47 to be precise, one can learn something from the data not only about the identity of the true theory, but also about what may be reasonable values of $\bar{\sigma}$.

There are several strategies one might use for this purpose. All of these should be regarded as hacks with limited theoretical justification[12]. Finding more satisfactory ways of choosing this parameter is an important area for future research. In order to make this section applicable beyond the Charness-Rabin example, I assume in most of the development that $\bar{\sigma}$ is a true constant. I then show how each technique, modified slightly, can be applied to the Charness-Rabin example.

5.1   A rule of thumb

Suppose that one knew which theory is true and was simply attempting to learn about the value of $\bar{\sigma}$. Suppose that the true theory $\lambda^\star$ makes an error on experiment Θ. Conditional on making an error, there

---

[12]Because $\bar{\sigma}$ can be viewed roughly as a nuisance parameter similar to the population variance in standard statistical models, these approaches have a close resemblance to empirical Bayes methods. These techniques use classical methods to choose values for nuisance parameters in models which are primarily Bayesian in their style of inference. For an overview of empirical Bayes methods, see Chapter 4 of Lehmann and Casella (1998).

is a probability $\frac{|\Theta|-|\lambda^\star(\Theta)|}{|\Theta|}$ that the realized outcome will be outside the theory's predictions. So as long as $\Theta \neq \lambda^\star(\Theta)$ a simple, natural and unbiased estimator[13] is $\frac{1_{\omega^\star(\Theta)\notin\lambda^\star(\Theta)}|\Theta|}{|\Theta|-|\lambda^\star(\Theta)|}$.

This estimate has the important problem that it is always 0 or greater than 1. However, if the size of the observation is relatively large, one can to use an estimate like this and an appeal to the law of large numbers to obtain an estimator of $\bar{\sigma}$. In addition, one can invokes standard arguments for generalized least squares[14] by weighting each term inversely by its standard deviation which is proportional[15] to $\sqrt{\frac{|\Theta|}{|\Theta|-|\lambda^\star(\Theta)|}-\bar{\sigma}}$. Of course this depends on the value of $\bar{\sigma}$, but that problem can easily be overcome by using a constant prior estimate in the place of a true value of $\bar{\sigma}$, approximating $\bar{\sigma}$ by 1 or 0 or using a two-stage procedure where one of the preceding hacks is used to estimate $\bar{\sigma}$ and then the estimate of sigma is used to estimate the variance of each experiment's contribution to the estimation of $\bar{\sigma}$. Regardless one needs to use some $\tilde{\sigma}$ to estimate the error rate. The theory-specific rule of thumb estimator is then given by:

$$\hat{\sigma}_{RoT,\lambda^\star} \equiv \min\left(\frac{1}{\sum_{\Theta\in\Xi}\sqrt{\frac{|\Theta|-|\lambda^\star(\Theta)|}{|\Theta|-\left[|\Theta|-|\lambda^\star(\Theta)|\right]\tilde{\sigma}}}}\sum_{\Theta\in\Xi}\frac{1_{\omega^\star(\Theta)\notin\lambda^\star(\Theta)}|\Theta|}{\sqrt{\left[|\Theta|-|\lambda^\star(\Theta)|\right]\left(|\Theta|-\left[|\Theta|-|\lambda^\star(\Theta)|\right]\tilde{\sigma}\right)}},1\right) \quad (5)$$

where the minimum operator is necessary to prevent estimate being above one, though this obviously introduced some bias into the estimator. A larger problem is that this estimator, by construction, assumes one already knows the true theory. To obtain a useful estimator, one can take a weighted average of this $\hat{\sigma}_{\text{RoT},\lambda}$'s obtained for different theories. A natural way to do this is to weight according to the scientist's (prior) plausibility distribution:

$$\hat{\sigma}_{RoT} \equiv \sum_{\lambda\in\text{supp}(\pi)}\hat{\sigma}_{\text{RoT},\lambda}\pi(\lambda) \quad (6)$$

where $\text{supp}(\pi)$ denotes the support of the scientist's plausibility distribution. Another version[16] of this estimator might be derived from a two-step procedure where the scientist first calculates $\hat{\sigma}_{RoT}$ using equation 6, updates using the resultant learning rule and then recalculates $\hat{\sigma}_{RoT}$ using in place of π her posterior distribution over the true theory.

This procedure must be modified slightly[17] to take into account non-constant error rates before it can

---

[13]A more consistent and satisfying approach here would be a method of moments estimator.

[14]Which are not particularly well-justified here given non-normality. But for a rule of thumb they likely produce a better estimate than giving equal weight to all observations.

[15]To see this, note that the expectation of the estimate is $\bar{\sigma}$. The second moment of the estimate is:

$$\left(\frac{|\Theta|}{|\Theta|-|\lambda^\star(\Theta)|}\right)^2\frac{|\Theta|-|\lambda^\star(\Theta)|}{|\Theta|}\bar{\sigma}$$

So the variance is $\bar{\sigma}\left(\frac{|\Theta|}{|\Theta|-|\lambda^\star(\Theta)|}-\bar{\sigma}\right)$. Because the $\bar{\sigma}$ term appears in the weighting of all observations, it can be dropped, yielding the expression in the text.

[16]I have not thought about whether there is any theoretical justification at all for either of these procedures. It would be interesting to know whether any consistency, approximate unbiasedness, etc. properties hold.

[17]The error rate in this context is not $\bar{\sigma}$ but, for a particular experiment, $\bar{\sigma}+2(1-\bar{\sigma})\tilde{p}(\Theta)$. The unbiased, theory-specific estimator or $\bar{\sigma}$ based on a single experiment derived above was $s(\Theta)\equiv\frac{1_{\omega^\star(\Theta)\notin\lambda^\star(\Theta)}}{1-\frac{1}{\rho(\lambda^\star,\Theta)}}$. Solving $\bar{\sigma}+2(1-\bar{\sigma})\tilde{p}(\Theta)=s(\Theta)$ for $\bar{\sigma}$, the equivalent unbiased, theory-specific estimator based on a single experiment is $s'(\Theta)\equiv\frac{s(\Theta)-2\tilde{p}(\Theta)}{1-2\tilde{p}(\Theta)}$. The variance is proportional to $\frac{1}{\left[1-2\tilde{p}(\Theta)\right]^2}$. Here whenever $\Theta\neq\lambda^\star(\Theta)$, $|\Theta|=2$ and $|\lambda^\star(\Theta)|=1$, so the only inverse standard deviation weighting comes

be applied to the Charness-Rabin data. The theory specific rule of thumb estimator for the Charness-Rabin application is:

$$\hat{\sigma}_{\text{RoT},\lambda^\star}^{CR} \equiv \max \left[\min \left(\frac{2}{\sum_{\Theta \in \Xi : \Theta \neq \lambda^\star(\Theta)} \left[1 - 2\tilde{p}(\Theta)\right]} \sum_{\Theta \in \Xi : \Theta \neq \lambda^\star(\Theta)} 1_{\omega^\star(\Theta) \notin \lambda^\star(\Theta)} - \tilde{p}(\Theta), 1 \right), 0 \right]$$

Rule of thumb estimates based on each of the Charness-Rabin theories are collected in Table 2 at the end of this section.

5.2 Maximum likelihood estimation

Another alternative, with perhaps a bit potential theoretical justification, is maximum likelihood estimation. In this simple context, direct maximum likelihood estimation is computationally trivial. Again, beginning with the assumption of constant error rate $\bar{\sigma}$, the theory-specific maximum likelihood estimator is simply:

$$\hat{\sigma}_{MLE,\lambda^\star} \equiv \text{argmax}_{\sigma \in [0,1]} \sum_{\Theta \in \Xi} \log \left(\frac{\sigma}{|\Theta|} + \frac{(1-\sigma)1_{\omega^\star(\Theta) \in \lambda^\star(\Theta)}}{|\lambda^\star(\Theta)|} \right)$$

To combine across theories we can use one of the simple weighting procedures discussed above:

$$\hat{\sigma}_{MLE} \equiv \sum_{\lambda \in \text{supp}(\pi)} \left[\text{argmax}_{\sigma \in [0,1]} \sum_{\Theta \in \Xi} \log \left(\frac{\sigma}{|\Theta|} + \frac{(1-\sigma)1_{\omega^\star(\Theta) \in \lambda(\Theta)}}{|\lambda(\Theta)|} \right) \right] \pi(\lambda) \tag{7}$$

Adapting this estimator to the Charness-Rabin context is even more straightforward than for the rule of thumb estimator. The theory-specific estimator is given by:

$$\hat{\sigma}_{MLE,\lambda^\star}^{CR} \equiv \text{argmax}_{\sigma \in [0,1]} \sum_{\Theta \in \Xi} \log \left(\frac{\sigma + (1-\sigma)\tilde{p}(\Theta)}{|\Theta|} + \frac{(1-\sigma)\left[1 - \tilde{p}(\Theta)\right]1_{\omega^\star(\Theta) \in \lambda^\star(\Theta)}}{|\lambda^\star(\Theta)|} \right)$$

Again, see Table 2 at the end of the section at the end of this section for numerical estimates in the Charness-Rabin application. An alternative, richer maximum likelihood estimator, which I call the *full* maximum likelihood estimator, explicitly incorporates uncertainty about the true theory into the likelihood maximization.

$$\hat{\sigma}_{FMLE} \equiv \text{argmax}_{\sigma \in [0,1]} \sum_{\lambda \in \text{supp}(\pi)} \left(\prod_{\Theta \in \Xi} \frac{\sigma}{|\Theta|} + \frac{(1-\sigma)1_{\omega^\star(\Theta) \in \lambda(\Theta)}}{|\lambda(\Theta)|} \right) \pi(\lambda) \tag{8}$$

While theoretically somewhat more satisfying, this estimator has quite terrible computational properties as it destroys the product structure generated by independence conditional on the true theory. However if the observation is fairly small it is not computationally intractable. In particular, we can again modify this estimator slightly and apply it to the Charness-Rabin context.

---

from the noise introduced by statistical sampling error. Also, it allows me to simplify the general expression for $s(\Theta)$ to $2 \cdot 1_{\omega^\star(\Theta) \notin \lambda^\star(\Theta)}$. Also, the maximum operator added here takes care of the possibility (needed to adapt the estimator to the Charness-Rabin context) that the (unconstrained) estimator may go below 0. Again this may introduce bias.

Plausibility vector (S, SW, DA, C)	$\hat{\sigma}_{FMLE}^{CR}$
(.25,.25,.25,.25)	.12
(.8,.1,.1,0)	.13
(.15,.5,.25,.1)	.1
(.1,.1,.70,.1)	.059
(.05,.05,.85,.05)	0

Table 1: Full maximum likelihood estimate of $\bar{\sigma}$ for Charness-Rabin data given various priors

$$\hat{\sigma}_{FMLE}^{CR} \equiv \mathrm{argmax}_{\sigma \in [0,1]} \sum_{\lambda \in \mathrm{supp}(\pi)} \left(\prod_{\Theta \in \Xi} \frac{\sigma + (1-\sigma)\tilde{p}(\Theta)}{|\Theta|} + \frac{(1-\sigma)\big[1 - \tilde{p}(\Theta)\big]1_{\omega^\star(\Theta) \in \lambda(\Theta)}}{|\lambda(\Theta)|} \right) \pi(\lambda)$$

Table 1 shows the full maximum likelihood estimate of the error rate for various values of the plausibility function.

5.3 Pseudo-Bayesian learning

A third alternative, which seems to me complex and likely unproductive, is to embed the uncertainty about the informativeness directly into the model, by endowing the scientist with priors over the the value of this parameter and solving for learning process jointly over theories and errors. I think there are three basic problems with this apparently elegant solution:

1. Intractability: I suspect that the expressions that will emerge for updating on theories will become sufficiently complicated/computationally intensive so as to deter applied analysis, particular when the model is extended to metric experiment spaces as outline in Appendix D.

2. Transparency: The introduction of a full Bayesian model is likely to obscure the analyst's choices and how they affect the analysis, rather than illuminate them and so should probably be avoided. Importantly, the robustness checks (based on choice of the error parameter) discussed in the following section are inapplicable to the full Bayesian approach.

3. Wrong focus: I fear that building a more complete Bayesian model for the choice of the error parameter is likely to obscure the model's focus on learning about *theories* taking the error rate as a parameter and shift the focus to learning about error rates, which (while an interesting problem) is simply not the scientific learning problem as developed above.

A somewhat more palatable alternative than full Bayesian learning which avoids some of these problems is to imagine that Bayesian learning is taking place for the error rate separately from that taking place for the beliefs about the true theory. One could have the scientist learn in a Bayesian manner about the value of the error rate conditional on knowing the true theory[18]. Then an expectation[19] could be taken over the

---

[18]It might even be feasible, if one wishes to take a further step towards full Bayesianism, to have the scientist do full Bayesian learning and use this to get an estimate for the error parameter, but then revert to the "known-error parameter" formulas derived above for final estimates update rule.

[19]An alternative to simply taking an expectation of the error parameter, which corresponds to a Bayesian point estimate based on a mean-squared error loss function, would be to allow for a more general loss function in constructing a Bayesian

Theory	$\hat{\sigma}^{CR}_{\mathrm{RoT},\lambda^\star}$	$\hat{\sigma}^{CR}_{MLE,\lambda^\star}$	$\hat{\sigma}^{CR}_{PB,\lambda^\star}$
Selfishness	.18	.13	.18
Social Welfare	0	0	.1
Difference Aversion	.38	.37	.43
Competitive	.59	.49	.52

Table 2: Theory-specific estimates of $\bar{\sigma}$ by various methods for the Charness-Rabin data

value of the error in the resultant posterior distribution and the expected error value would be used as a theory-specific estimate of the error rate. Formally, this theory-specific estimator for a uniform prior over error rates is given by

$$\hat{\sigma}_{PB,\lambda^\star} \equiv \frac{\int_0^1 \sigma \prod_{\Theta \in \Xi} \frac{\sigma}{|\Theta|} + \frac{(1-\sigma)1_{\omega^\star(\Theta) \in \lambda^\star(\Theta)}}{|\lambda^\star(\Theta)|} d\sigma}{\int_0^1 \prod_{\Theta \in \Xi} \frac{\sigma}{|\Theta|} + \frac{(1-\sigma)1_{\omega^\star(\Theta) \in \lambda^\star(\Theta)}}{|\lambda^\star(\Theta)|} d\sigma} \tag{9}$$

Again this can be slightly modified to apply to the Charness-Rabin context:

$$\hat{\sigma}^{CR}_{PB,\lambda^\star} \equiv \frac{\int_0^1 \sigma \prod_{\Theta \in \Xi} \frac{\sigma+(1-\sigma)\tilde{p}(\Theta)}{|\Theta|} + \frac{(1-\sigma)[1-\tilde{p}(\Theta)]1_{\omega^\star(\Theta) \in \lambda^\star(\Theta)}}{|\lambda^\star(\Theta)|} d\sigma}{\int_0^1 \prod_{\Theta \in \Xi} \frac{\sigma+(1-\sigma)\tilde{p}(\Theta)}{|\Theta|} + \frac{(1-\sigma)[1-\tilde{p}(\Theta)]1_{\omega^\star(\Theta) \in \lambda^\star(\Theta)}}{|\lambda^\star(\Theta)|} d\sigma}$$

Table 2 summarizes all of the theory-specific estimators. The three estimators yield fairly similar results to one another for each theory. Selfishness seems to yield an error estimate of somewhere in the mid-teens, Social Welfare yields an estimate of (or near) 0, Difference Aversion yields an estimate near forty and Competitive preferences yield a an estimate in the fifties. Unsurprisingly, the pseudo-Bayesian method appears to tilt the estimate away from extreme (low) values relative to the maximum likelihood estimator. These estimates provide some evidence that, at least in this context, the weighting across theories to choose combine these theory-specific estimates, rather than the method of estimation, makes a more important difference in determining an appropriate error parameter value. This should provide at least some reassurance about robustness problems across methods of estimation. More difficult are robustness problems across various methods of weighting different theory-specific estimates. These problems are the focus of section 7.

6 Some results

Before discussing in detail issues of robustness, it is useful to consider the results of the learning rule in this context for various values of the error parameter in the range shown in Table 2. Table 3 shows the relative value of the Bayes Factors for the theories, given different choices of $\bar{\sigma}$. Selfishness is normalized to 1. The results reveal both something substantive about how the Charness-Rabin data should be interpreted and methodologically about the properties of the model.

On the substantive side, Table 6 shows the sense in which the data generally supports selfishness as the best model, but also shows the range of parameter values for which the data may be interpreted as

point estimate of $\bar{\sigma}$. For example, one might try to explicitly take into account the trade off between consistency against supersets and speed of consistency against subsets discussed below. I will not attempt to develop this approach here, as my goal in this paper is more to pose and provide some simple ideas than to solve the problem of correct error parameter choice.

$\bar{\sigma}$	$\bar{\mu}$	Selfish	Social welfare	Difference aversion	Competitive
.001	999	1	39	$1.5 \cdot 10^{-7}$	$4.1 \cdot 10^{-12}$
.01	99	1	2.9	$2.9 \cdot 10^{-7}$	$1.2 \cdot 10^{-10}$
.022	45	1	1	$5.2 \cdot 10^{-7}$	$7.2 \cdot 10^{-10}$
.059	16	1	.23	$1.8 \cdot 10^{-6}$	$1.6 \cdot 10^{-8}$
.1	9	1	.11	$4.7 \cdot 10^{-6}$	$1.3 \cdot 10^{-7}$
.13	6.7	1	.075	$8.4 \cdot 10^{-6}$	$4 \cdot 10^{-7}$
.18	5.3	1	.058	$1.4 \cdot 10^{-5}$	$1.1 \cdot 10^{-6}$
.21	3.8	1	.044	$3.0 \cdot 10^{-5}$	$4.2 \cdot 10^{-6}$
.25	3	1	.04	$5.4 \cdot 10^{-5}$	$1.1 \cdot 10^{-5}$
.37	1.7	1	.037	$2.5 \cdot 10^{-4}$	$1.2 \cdot 10^{-4}$
.43	1.3	1	.04	$5.3 \cdot 10^{-4}$	$3.3 \cdot 10^{-4}$
.49	1.1	1	.046	$1.1 \cdot 10^{-3}$	$8.5 \cdot 10^{-4}$
.52	.92	1	.05	.0016	.0013
.59	.69	1	.065	.0037	.0037

Table 3: Relative values of updates in favor of various theories, selfishness normalized to 1

supporting social welfare preferences. For an error rate of .25 the model updates twenty-five times as heavily in favor of selfishness as social welfare preferences, meaning that a flat prior would imply posterior probabilities of approximately .96 on selfishness, .038 on social welfare preferences and a negligible weight on the two other models. In fact, for all values of $\bar{\sigma}_C$ above .022, the model interprets the data as providing evidence in favor of the selfishness model. On the other error rates significantly below .02, such as .001, the model views the data providing strong evidence in favor of social welfare preferences. This should be intuitive, given the discussion above about $\bar{\sigma}$ as a measure of the accuracy-precision trade-off. Social welfare preferences are never falsified in the data, while selfishness is on several occasions. Therefore placing high value on accuracy over precision by choosing a low value of the error parameter makes the evidence weigh against selfishness. The model therefore shows how the Charness-Rabin interpretation of the data is internally consistent. If one believes for prior reasons that a good model of social preferences should virtually always be consistent with behavior, even if this requires some vagueness, or that social welfare preferences are a good model (and therefore should form the basis of the error calibration), then the data can be viewed as a confirming both views.

However, I suspect many economists, who continued to use selfishness as their primary model of choice even in simple experiments, interpreted the data effectively using a parameter value above .02, as supporting selfishness. This interpretation seems reasonable (given a wide range of prior beliefs) for a few reasons. First note that in this context an error parameter of .02 or below corresponds to the belief that a good model of binary choices like these would only incorrectly predict the population average behavior one in a hundred times. Particularly given that the data used makes the somewhat heroic assumption that subjects correctly anticipate the distribution of their partner's play, this seems a very high degree of accuracy to demand. Second recall that for the weighted maximum likelihood estimation procedure a prior of at least .85 on social welfare preferences is required to achieve an error parameter below .02. For full maximum likelihood estimation a weight of at least .7 is necessary. While there is nothing wrong with these (or any other) priors, I doubt it is a prior held widely in the economics community.

On a methodological level, the results reveal several things about the model. First, perhaps the the-

ory's most attractive feature is that it connects the appropriate weight in the precision-accuracy trade-off (something quite nebulous) quantitatively to the rate at which good theories make mistakes, which seems a bit easier to judge and calibrate, in the ways discussed in the previous section. Second, the theory yields some non-obvious insights which are not overly sensitive to the choice of $\bar{\sigma}$.

1. Difference Aversion clearly does worse than Selfishness regardless of the value of $\bar{\sigma}$, where it preformed roughly even with selfishness in Charness and Rabin's measures. This effect arises because Difference Aversion, like Social Welfare preferences, is vaguer than Selfishness.

2. Both Difference Aversion and Competitive preferences are overwhelmingly rejected, regardless of the value of $\bar{\sigma}$ chosen.

3. For values of $\bar{\sigma}$ at or above .13, the relative learning in favor of Selfishness over Social Welfare preferences is not terribly sensitive to the value of $\bar{\sigma}$, ranging only within a factor of two.

However, this exercise also shows some important weaknesses in the model. First the magnitude of the update rate differences seems perhaps too large. Even if the observed phenomena we are trying to model are choices by individuals in simple experiments, the model may imply too much learning from this small data set. This is fundamentally driven by my assumption that each experiment conveys a piece of independent information. All of the experiments involved binary choices and many were qualitatively similar to one another. They therefore should not be viewed as fully independent pieces of information. Finding a more satisfactory way to deal with this problem is an important avenue for future research.

Perhaps more importantly, the analysis demonstrates some dimensions along which the results of an analysis based on the theory are quite sensitive to the choice of the error parameter. In particular, as discussed above, two sorts of interpretations of the data are possible for different values of the error parameter. This leads to a natural question of whether the learning procedure outlined above is robust to incorrect choice of the error parameter. Will a scientist living in a world where the true value of the error parameter is $\bar{\sigma}^\star$ consistently learn the identity of the true theory if she assumes that the value of the error parameter is $\bar{\sigma}$? Is the answer the same for all values of $\bar{\sigma}$? On a somewhat different note, what if (as seems likely) none of the theories considered by Charness and Rabin is the "true theory"? Then will the learning rule lead the scientist to believe in a model which is "close" to the true model in some reasonable sense? Understanding such robustness properties would make the results and differentia interpretations discussed above easier to understand.

7   Robustness

Questions about the potential consequences of "incorrect" choices of the error parameter cannot be addressed within a subjective Bayesian probability framework, because a Bayesian scientists cannot consider the possibility that he is incorrect. Because I find answers to such questions informative[20] about the model

---

[20]From a Bayesian perspective, the objective probability distribution may be viewed as the beliefs of another, skeptical scientist. Therefore all the results of this section have a Bayesian interpretation in terms of the beliefs of different scientists about the convergence of beliefs. Some additional work would have to be done in order to take into account the fact that the other scientist is herself uncertain about the identity of the true theory and to deal with the selection of the observation from the studied phenomenon, but the basic results should carry over into this Bayesian interpretation.

developed above, I devote the core of this section to proving some results within an objective probability setting in which the scientist's subjective learning is embedded. To do this, a fair amount of technical machinery is necessary and much of this will not be relevant to most readers. Therefore this section is divided into two subsections. In the first, I provide a brief summary of the results, showing how they can easily be used for robustness checks and error parameter selection in applied analysis and in particular how they are useful in the Charness-Rabin application. In the second, I develop a formal objective probability framework and prove one of the results referred to in the first subsection. The remaining proofs appear in the appendices.

7.1   A non-technical summary

The primary goal of this section is to ask whether a scientist who incorrectly chooses the constant error rate parameter $\bar{\sigma}$ will consistently learn the identity of the true theory. I answer this question in three steps:

1. First in Lemma 1, I show that, regardless of the error parameter chosen by the scientist, the learning rule is consistent against theories that are subsets of the true theory, in the sense that their predictions are always subsets of the predictions of the true theory. I also show that the smaller $\bar{\sigma}$ is, the faster is the asymptotic rate of consistency.
   Intuitively a scientist can always distinguish the truth from a subset theory as the subset theory will make errors at a higher rate than the true theory. Given that the scientist will eventually learn the error rate estimate given any theory, the simple rule "reject any theory that does not have the lowest estimated error rate" is consistent against subsets. Therefore it is not surprising that the Bayesian update rule is also consistent against subsets. Furthermore the lower the error rate used by the scientist, the more inaccuracy is penalized. Given that subset theories will always be less accurate, it is not surprising that using smaller $\bar{\sigma}$ leads to faster consistency against subset theories.

2. Second in Lemma 2, I show that so long as $\bar{\sigma}$ is not too small relative to the true error rate $\bar{\sigma}^{\star}$ the learning rule is consistent against supersets of the true theory. I discuss what I mean quantitatively by "too small" below; however, the intuition behind this result is simple. Suppose that on a particular experiment with say five possible outcomes the true theory predicts uniquely and an alternative superset theory predicts all five outcomes. If the scientist believes the error rate is near zero, then she will view the true theory as predicting its unique outcome with probability near one and view the alternative as predicting that the outcome is distributed uniformly on the five possible outcomes. However, if the true error rate is close to 1, then the true distribution over outcomes given the true theory is close to uniform over all possible outcomes. Thus, given the scientist's overly aggressive choice of the error parameter, the distribution over outcomes expected by the scientist if the alternative is true will be more similar to the actual distribution of outcomes than will the distribution over outcomes the scientist expects given that the true theory is true. Thus a scientist who chooses too small of an error parameter, relative to the true error rate, will be inconsistent against superset theories.

3. Finally in Theorem 1, I argue that supersets and subsets are the hardest theories to distinguish from the truth. Intuitively, any alternative that is not a superset or subset of the true theory can be

20

changed to make it a superset or subset in a way that will only make it more likely that the scientist learns this theory. Combining this with the first two results establishes that the learning rule is consistent if and only if $\bar{\sigma}$ is not too small relative to $\bar{\sigma}^\star$. This provides an answer to the question of the robustness of learning with an incorrect error parameter.

These results reinforce the notion that $\bar{\sigma}$ mediates the trade-off between precision and accuracy. Low values of $\bar{\sigma}$ risk inconsistency if there are vague alternative theories, favoring accuracy over precision. High values of $\bar{\sigma}$ lead to only slow consistency against subset theories, favoring precision over accuracy. The result also has a bit of the flavor of the distinction between Type I and Type II error. Choosing $\bar{\sigma}$ too low may lead to Type I error as it may causes the scientist to reject theories, even if they are true, in favor of vague alternatives. Choosing $\bar{\sigma}$ too high leads to Type II error, as the scientist fails to quickly reject inaccurate subset theories.

In addition to its applied uses, which I will discuss below, the results supply some general insight in learning about competing theories, at least in this model. It shows that there is, in some sense, greater danger that scientists with incorrect parameters will reject true theories that are very precise than those that are vague. If the true theory is very vague, its competitors will mostly be (approximate) subsets. Because these subsets theories will consistently make too many errors, it will be easy to learn, in the long run, that they are false. On the other hand if a true theory is very precise, its competitors will tend to be (approximate) supersets. Because discriminating against superset theories requires a careful weighing of precision and accuracy, rather than a simple comparison of empirical consistency, scientists who interpret data assuming that true theories must be very accurate will be at risk of learning a false theory[21].

In applications, what matters most is quantitatively how strongly the possibility of inconsistency against supersets bites. How much smaller can $\bar{\sigma}$ be than the true error rate $\bar{\sigma}^\star$ while the scientist is still consistent? The most general answer is given by the somewhat messy expression 13 in the next subsection. However it turns out[22] that the hardest superset theory to distinguish from the truth is the maximally vague theory and that this is hardest to distinguish from the truth when the truth becomes maximally vague. Thus if one cannot put ex-ante lower bounds on the precision of theories, the expression upper-bounding $\bar{\sigma}^\star$ given $\bar{\sigma}$ that ensures consistency simplifies to

$$\bar{\sigma}^\star < -\frac{1-\bar{\sigma}}{\log(\bar{\sigma})} \equiv \varphi(\bar{\sigma}) \tag{10}$$

In practice, this expression is quite close to the value of the expression 13 derived in the following subsection, as demonstrated in Table 2. Figure 1 shows expression as a function of $\bar{\sigma}$ for several ranges of values of $\bar{\sigma}$. It shows that for most values of $\bar{\sigma}$, particularly very small values, expression is much larger than $\bar{\sigma}$. For example, expression is greater than .1 for values of $\bar{\sigma}$ as low as .00005. This provides some

---

[21]Note, however, that by construction the distribution over outcomes induced by the theory learned will be closer in relative entropy (given the scientist's error rate) to the true theory given the correct error rate. While this might seem to make the results above seem a bit silly, note that the procedure outlined in this paper involves some learning about both the theory and the error rate and much of the learning about the error rate depends on the choice of the true theory. Furthermore, the procedure here should not be interpreted too literally as learning the full probability distribution over future outcomes, even though this is the literal content of Assumption 1. Often social scientist use theories to identify a reasonable set of potential outcomes or to interpret data. Some scientists applying a model may not have a clear sense of the estimated error parameter and only know which theory has been found most probably correct. Thus learning which theory is true remains of central importance.

[22]The proof of this result depend on an inequality proved by Mihai Manea, as noted in Appendix B
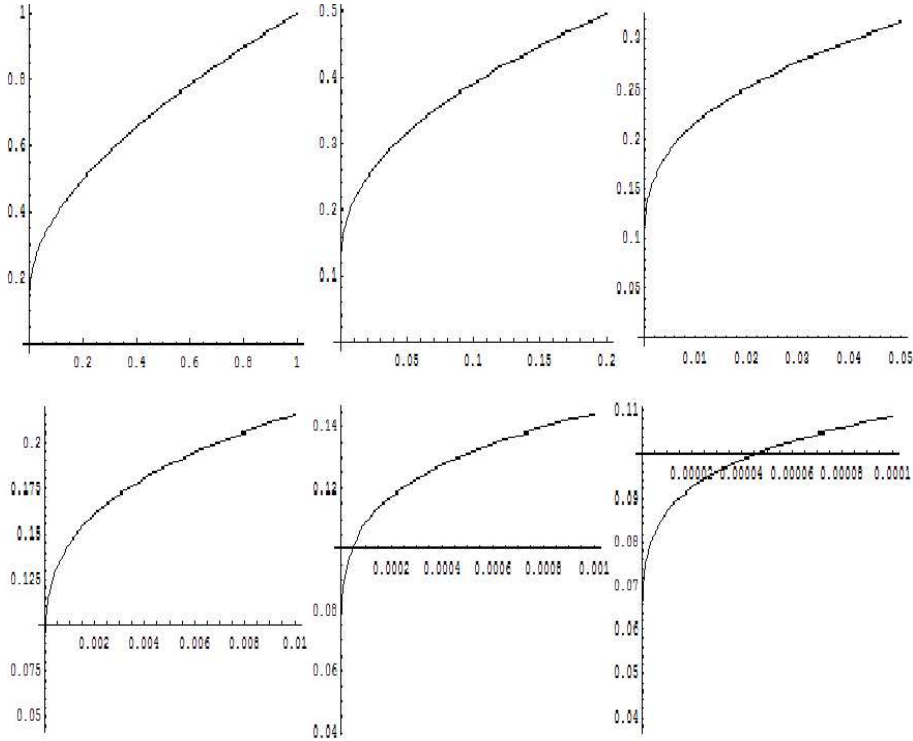
Figure 1: Values of expression 7.1 for different ranges of $\bar{\sigma}$.

reassurance that, even using this conservative bound, the risk of inconsistency is not too severe.

Returning to the Charness-Rabin data shows how Theorem 1 can be useful in applied analysis. The motivation I gave above for developing the robustness result embodied in Theorem 1 was that it might help determine why different values of $\bar{\sigma}$ lead to learning in favor of different theories. In the Charness-Rabin data natural lower bounds exist for the precision of theories. In particular, given that all of their data is from binary choices, the only way a theory can be a super set of another on a particular experiment is that one theory has a precision of 2 and the super set has a precision of 1 on that experiment. Using the formula 13 in the following subsection shows that in the Charness-Rabin context the relevant bound is

$$\bar{\sigma}^\star < 2\left(1 - \frac{\log(1+\bar{\mu})}{\log(1+\bar{2}\mu)}\right) \equiv \varphi_{CR}(\bar{\mu}) \tag{11}$$

Table 2 reproduces the results displayed in Table 1, but adds the value of expression 11 and 7.1, to show that they don't differ dramatically. Recall that the theory-specific estimates of the error parameter based on Selfishness ranged from .13 to .18, for Social Welfare preferences from 0 to .1, for Difference Aversion from .37 to .43 and for Competitive preferences from .49 to .59. Table 2 therefore indicates that values of the error parameter much below .25 are likely to lead to learning rules inconsistent against estimates of that parameter derived from at least one of the theories considered. Values of the error parameter below .022 (the level of this parameter at which the pro-Social Welfare preferences interpretation of the data begins to become valid) are inconsistent against the parameter values estimated based on both Competitive and Difference Averse preferences. They even get close to being inconsistent against the estimates based on Selfishness. This does not provide a fully satisfactory answer to the question of why these two interpretations exist for the values of the error parameter where they do. One would expect,

22

$\bar{\sigma}$	$\bar{\mu}$	$\varphi_{CR}(\bar{\mu})$	$\varphi(\bar{\sigma})$	Selfish	Social welfare	Difference aversion	Competitive
.001	999	.18	.14	1	39	$1.5 \cdot 10^{-7}$	$4.1 \cdot 10^{-12}$
.01	99	.26	.21	1	2.9	$2.9 \cdot 10^{-7}$	$1.2 \cdot 10^{-10}$
.022	45	.3	.26	1	1	$5.2 \cdot 10^{-7}$	$7.2 \cdot 10^{-10}$
.059	16	.4	.33	1	.23	$1.8 \cdot 10^{-6}$	$1.6 \cdot 10^{-8}$
.1	9	.44	.39	1	.11	$4.7 \cdot 10^{-6}$	$1.3 \cdot 10^{-7}$
.13	6.7	.47	.43	1	.075	$8.4 \cdot 10^{-6}$	$4 \cdot 10^{-7}$
.18	5.3	.52	.48	1	.058	$1.4 \cdot 10^{-5}$	$1.1 \cdot 10^{-6}$
.21	3.8	.54	.51	1	.044	$3.0 \cdot 10^{-5}$	$4.2 \cdot 10^{-6}$
.25	3	.58	.54	1	.04	$5.4 \cdot 10^{-5}$	$1.1 \cdot 10^{-5}$
.37	1.7	.66	.63	1	.037	$2.5 \cdot 10^{-4}$	$1.2 \cdot 10^{-4}$
.43	1.3	.7	.68	1	.04	$5.3 \cdot 10^{-4}$	$3.3 \cdot 10^{-4}$
.49	1.1	.73	.71	1	.046	$1.1 \cdot 10^{-3}$	$8.5 \cdot 10^{-4}$
.52	.92	.75	.73	1	.05	.0016	.0013
.59	.69	.79	.78	1	.065	.0037	.0037

Table 4: Relative values of updates in favor of various theories, selfishness normalized to 1; values of expressions 11 and 7.1 are added

from a theoretical perspective[23], that if Selfishness is the true theory then the values of the error parameter leading to the learning in favor Social Welfare preferences should be inconsistent if the true value of the error parameter is that from estimates based on Selfishness being the true theory. However, I think Table 2 does show that the values of $\bar{\sigma}$ giving rise to the pro-Social Welfare preferences interpretation are quite low and lead to substantial risk of inconsistency against reasonable values of the error parameter. Given that Selfishness is a much more precise theory than Social Welfare preferences in the current context, and therefore if Selfishness is the true theory, Social Welfare preferences may be seen roughly as a superset theory, this provides at least a partial explanation of the two possible interpretations of the data that are possible in the learning model developed here. Furthermore, Table 2 shows that the more precise value of φ_{CR} is in practice quite similar to that of φ, despite the strong ex-ante lower bounds on precision possible here.

The analysis here suggests a general principle to help ensure the robustness of analysis using the theory developed here. Suppose that one settles on using some estimate of $\hat{\sigma}$. And suppose that, after the resultant analysis, updating occurs that most favors some theory $\tilde{\lambda}$. Suppose, too, that there is some other theory, λ', which, while perhaps not a strict subset of $\tilde{\lambda}$, does tend to be more precise than $\tilde{\lambda}$ in most experiments in the observation. Then it seems reasonable to consider whether $\hat{\sigma}$ is consistent against reasonable $\hat{\sigma}_{\cdot,\lambda'}$'s. If not, one might worry that the updating that occurred was a pathological result of an incorrectly chosen estimate of the error parameter. While this should not automatically lead one to change the chosen parameter, it provides a potential general way to use Theorem 1 to preform robustness checks[24] in applied analysis.

---

[23]Two answers to this puzzle immediately suggest themselves. The first is that Selfishness is likely not the true theory, but simply the theory closest in relative entropy terms (given the true error parameter) to the true theory, along the lines of Theorem 2. It might well be that the risk of inconsistency is greater when the theory closest to the true theory, rather than learning the true theory itself. Another possibility is that there is simply measurement error in the estimates of $\bar{\sigma}^{\star}$ based on Selfishness being the true theory.

[24]There is another potentially useful check. Suppose is there is a vaguer theory than the one updated in favor of, call the vaguer theory λ' and the one updated in favor of λ. And suppose its $\hat{\sigma}_{\cdot,\lambda'}$ is such that even if the true informativeness $\hat{\sigma}_{\cdot,\lambda}$ the

For analysis on the fly, this robustness check can be reduced to a quick and easy procedure[25] for choosing a reasonable value of $\bar{\sigma}$. Suppose one arrives, say through calibrations as above, at some range of reasonable values for $\bar{\sigma}$, $[\bar{\sigma}', \bar{\sigma}'']$. Then using $\bar{\sigma} = \hat{\bar{\sigma}}_{Robust} \equiv \varphi^{-1}(\bar{\sigma}'')$ has the nice property that it is the smallest value of $\bar{\sigma}$ which ensures consistency so long as $\bar{\sigma}^\star \in [\bar{\sigma}', \bar{\sigma}'']$. If one considers the reasonable range of values of $\bar{\sigma}$ to be values arising from some calibration, given some theory considered by Charness and Rabin, then the relevant range is $[0, .52]$. In this case $\hat{\bar{\sigma}}_{Robust}$ using φ is approximately .23 and using φ^{CR} is approximately .18. If one excludes the values based on Competitive preferences then the relevant range is $[0, .43]$ and the values using φ and φ^{CR} are approximately .13 and .1 respectively. All of these values are in the range that strongly supports the pro-Selfishness interpretation of the data.

One problem in interpreting the model laid out here is that it seems unlikely that the true theory has yet been identified by scientists. New economic models are constantly proposed and most economists see themselves as learning about the relative merit of theories that are increasingly good approximations to the truth. This naturally leads to the question of whether a scientist will learn something sensible even if her priors (plausibility function) are not absolutely continuous with respect to the true theory, so long as she uses the correct error parameter. In what follows, I also show that in this case the scientist will always consistently learn a theory (or theories if there are many equally close) that are maximally close in expected relative entropy to the truth.

7.2 Formal results

In this subsection set up the necessary formal framework and prove the results outlined above.

Assumption 6. *There (objectively) exists a true measure Q, a true theory $\lambda^{\star\star} \in \Lambda$ and a true error parameter $\bar{\sigma}^\star$. Under Q the distribution of $\omega^\star$ is the distribution of $\omega^\star$ under P for a scientist placing plausibility of 1 on $\lambda^\star$ and having error parameter $\bar{\sigma}^\star$.*

Q corresponds to the true probability generating process that the scientist believes she is learning. In what follows I will consider what a scientist who has the wrong error parameter value or whose plausibility is not absolutely continuous with respect to the true theory.

To prove results about consistency, it is useful[26] to make assumptions not just about the distribution of experimental outcomes, but also about how the observation is drawn from the studied phenomenon. I will assume that the observation is drawn independently and identically from the studied phenomenon under some probability distribution. It is because of this assumption that it was important, as discussed in section 2, that one identifies the studied phenomenon with a class of situations of which the observed data is at least roughly representative.

Definition 6. *An* increasing infinite sequence of observations *(IISO) is a sequence of observations $\{\Xi_n\}_{n=1}^\infty$ such that there exists a sequence of experiments $\{\Theta_i\}_{i=1}^\infty$ for which $\Xi_n = \bigcup_{i=1}^n \Theta_i$.*

Assumption 7. *Γ is uncountably infinite. There exists a non-atomic probability distribution ν over Γ such that under Q the observation Ξ is drawn* iid *from ν in the sense that $\Xi \in \{\Xi_n\}_{n=1}^\infty$, an IISO, for*

---

update rule implied by $\hat{\sigma}_{\cdot,\lambda'}$ is still consistent against distinguishable super sets. Then it is probably worth checkin if instead of the chosen informativeness, $\hat{\mu}_{RoT,\lambda'}$ is used whether the basic conclusions of the analysis change much.

[25]This algorithm is based in a conversation I had with Xavier Gabaix and I am grateful to him for it.

[26]It may not be necessary if one places enough regularity conditions on the observation.

which the corresponding sequence of experiments is drawn independently and identically from ν. ν and Γ are such that all expectations taken below exist[27].

The assumption that ν is non-atomic means that the probability of the same experiment being drawn twice from Γ is 0, eliminating the problem of drawing the same experiment twice. I will now define formally the concepts referred to in the summary above.

Definition 7. *Two theories, λ and λ', are said to be* distinguishable *(written $\lambda \nsim \lambda'$) under Q if $\nu\big(\{\Theta \in \Gamma : \lambda(\Theta) \neq \lambda'(\Theta)\}\big) > 0$.*

Definition 8. *$\lambda \subseteq (\supseteq)\lambda'$, read λ is a subset(superset) of λ', if $\lambda(\Theta) \subseteq (\supseteq)\lambda'(\Theta), \forall \Theta \in \Gamma$.*

In the argument that follows I will make heavy use of a version of a well known theorem in information theory due to Kelly (1956).

Theorem (Kelly 1956). *Let $\{v_i\}_{i=1}^{\infty}$ and $\{u_i\}_{i=1}^{\infty}$ be infinite sequences of i.i.d. (within, not across, sequences) random variables under some measure Q. Suppose that $E_Q\big[\log(v_i)\big] > E_Q\big[\log(u_i)\big]$. Then for any $r \in \mathbb{R}, \lim_{n \to \infty} Q\big[\prod_{i=1}^{n} \frac{v_i}{u_i} > r\big] = 1$.*

Intuitively, if the expected logarithmic growth rate of one random variable is greater than the expected logarithmic growth rate of another, the law of large numbers ensures that the first will, with probability 1, become arbitrarily large relative to the second.

Definition 9. *An error parameter $\bar{\sigma}$ is said to be* consistent *against a set of theories $\Lambda' \subseteq \Lambda \backslash \{\lambda^{\star\star}\}$ if*

1. *For any measure P defined by obedience of Assumptions 1-6, use of parameter $\bar{\sigma}'$ and a plausibility function whose support is $\Lambda' \cup \lambda^{\star}$*

2. *For an IISO $\{\Xi_n\}_{n=1}^{\infty}$ that is drawn i.i.d. under Q*

3. *For any $\epsilon > 0$*

$$\lim_{n \to \infty} Q\big[P[\lambda^{\star} = \lambda^{\star\star}|\Xi_n] \leq 1 - \epsilon\big] = 0$$

This is a natural definition of consistency of an error parameter in this context against a set of theories Λ': the parameter is consistent against a set of false theories if any scientist using that parameter will become arbitrarily confident of the true theory if she only considers competing theories in Λ'. The following definition provides the equivalent natural definition of the rate of consistency.

Definition 10. *Consider two error parameters $\bar{\sigma}$ and $\bar{\sigma}'$ that are consistent against some Λ'. $\bar{\sigma}$ is said to be* consistent more quickly *than $\bar{\sigma}'$ against Λ' if*

1. *For any measure P and P' defined by obedience of Assumptions 1-6, use of parameter $\bar{\sigma}$ and $\bar{\sigma}'$ respectively and a plausibility function with support $\Lambda' \cup \lambda^{\star}$*

2. *For an IISO $\{\Xi_n\}_{n=1}^{\infty}$ that is drawn i.i.d. under Q*

---

[27]$|\Theta| < M$ for all $\Theta \in \Gamma$ and some $M \in \mathbb{R}$ is sufficient to insure this, for example.

3. For any $r \in \mathbb{R}$

$$\lim_{n \to \infty} Q\left[\frac{P'[\lambda^\star \in \Lambda'|\Xi_n]}{P[\lambda^\star \in \Lambda'|\Xi_n]} > r\right] = 1$$

One error parameter is consistent more quickly than another against a set of alternatives Λ' if, asymptotically, the probability assigned by a scientist using the first parameter to all theories in Λ' becomes arbitrarily small relative to probability assigned by a scientist using the second parameter to these theories. Note that this definition of rate of consistency is very much a partial (or even null) ordering even for a given Λ'. With these definitions, I can now state formally the first result mentioned above.

Lemma 1. *If under $\nu, \Lambda' \subseteq \{\lambda \in \Lambda : \lambda \nsim \lambda^{\star\star} \wedge \lambda \subseteq \lambda^{\star\star}\}$ is a finite set of distinguishable subsets of the true theory then*

1. Any $\bar{\sigma} \in [0, 1)$ is consistent under ν against Λ'

2. If $0 \leq \bar{\sigma} < \bar{\sigma}' < 1$ then $\bar{\sigma}$ is consistent more quickly under ν than $\bar{\sigma}'$ against Λ'

Proof. The proof has two steps.

1. First, I show that in any experiment the expected logarithm of the scientist's Bayes Factor for the true theory is greater than the scientist's Bayes Factor for any strict subset and that the expected logarithmic growth difference between these two Bayes Factors is larger for a scientist using $\bar{\sigma}$ than for one using $\bar{\sigma}'$.

2. Second, I use this result, in conjunction with Kelly's Theorem and the fact that all elements of Λ' are distinguishable subsets of $\lambda^\star$, to establish the two parts of the lemma.

First consider the expected logarithm of the (informativeness form of the) Bayes Factor of the true theory for a scientist using $\bar{\sigma}$ on an experiment where $\lambda^{\star\star}(\Theta) \neq \lambda'(\Theta)$

$$\left(1 - \left[1 - \frac{1}{\rho(\lambda^\star, \Theta)}\right]\bar{\sigma}^\star\right) \log\left(1 + \bar{\mu}\rho(\lambda^{\star\star}, \Theta)\right)$$

where $\bar{\mu} \equiv \frac{1-\bar{\sigma}}{\bar{\sigma}}$. The expected logarithm of the scientist's Bayes Factor for a subset theory λ' is

$$\left(\frac{\rho(\lambda^{\star\star}, \Theta)}{\rho(\lambda', \Theta)} - \left[\frac{\rho(\lambda^{\star\star}, \Theta) - 1}{\rho(\lambda', \Theta)}\right]\bar{\sigma}^\star\right) \log\left(1 + \bar{\mu}\rho(\lambda', \Theta)\right)$$

Simplifying the difference between these two expressions yields

$$\left(\frac{\bar{\sigma}^\star}{|\Theta|} + \frac{1 - \bar{\sigma}^\star}{|\lambda^{\star\star}(\Theta)|}\right)\left[|\lambda^{\star\star}(\Theta)| \log\left(1 + \bar{\mu}\frac{|\Theta|}{|\lambda^{\star\star}(\Theta)|}\right) - |\lambda'(\Theta)| \log\left(1 + \bar{\mu}\frac{|\Theta|}{|\lambda'(\Theta)|}\right)\right]$$

Suppose that $x > x' > 0$ and $y > 0$. Then note that

$$x \log\left(1 + \frac{y}{x}\right) > x' \log\left(1 + \frac{y}{x'}\right)$$

26

because

$$\log\left(1 + \frac{y}{x}\right) = \log\left(\frac{x'}{x}\left[1 + \frac{y}{x'}\right] + \frac{x' - x}{x}\right) > \frac{x'}{x}\log\left(1 + \frac{y}{x'}\right) + \frac{x - x'}{x}\log(1)$$

by concavity of the logarithm and Jensen's Inequality. Thus for positive y and x, $x\log\left(1 + \frac{y}{x}\right)$ is increasing in x and therefore, given that $\bar{\mu} > 0$ because $\bar{\sigma} < 1$,

$$\left(\frac{\bar{\sigma}^\star}{|\Theta|} + \frac{1 - \bar{\sigma}^\star}{|\lambda^{\star\star}(\Theta)|}\right)\left[|\lambda^{\star\star}(\Theta)|\log\left(1 + \bar{\mu}\frac{|\Theta|}{|\lambda^{\star\star}(\Theta)|}\right) - |\lambda'(\Theta)|\log\left(1 + \bar{\mu}\frac{|\Theta|}{|\lambda'(\Theta)|}\right)\right] > 0$$

as $|\lambda^{\star\star}(\Theta)| > |\lambda'(\Theta)|$ by the assumption that $\lambda' \subset \lambda^{\star\star}$ and $\lambda^{\star\star}(\Theta) \neq \lambda'(\Theta)$.

If $x > x'$, then $x\log\left(1 + \frac{y}{x}\right) - x'\log\left(1 + \frac{y}{x'}\right)$ is increasing in y as its derivative

$$\frac{1}{1 + \frac{y}{x}} - \frac{1}{1 + \frac{y}{x'}} > 0$$

Intuitively, Jensen's inequality bites stronger for larger y as the "risk" between $1 + \frac{y}{x'}$ and 1 grows with y. Thus

$$\left(\frac{\bar{\sigma}^\star}{|\Theta|} + \frac{1 - \bar{\sigma}^\star}{|\lambda^{\star\star}(\Theta)|}\right)\left[|\lambda^{\star\star}(\Theta)|\log\left(1 + \bar{\mu}\frac{|\Theta|}{|\lambda^{\star\star}(\Theta)|}\right) - |\lambda'(\Theta)|\log\left(1 + \bar{\mu}\frac{|\Theta|}{|\lambda'(\Theta)|}\right)\right]$$

is strictly increasing in $\bar{\mu}$ and therefore strictly decreasing in $\bar{\sigma}$ as $\lambda'(\Theta) \subseteq (\subset)\lambda^{\star\star}(\Theta)$. This completes step 1.

Now note that the scientist's Bayes Factor for the true theory is the product of its Bayes factor on each of the experiments. Given that each of the experiments in the observation is drawn i.i.d. from ν under Q, all we need show in order to apply Kelly's Theorem is that the expected logarithm of the scientist's Bayes Factor for the true theory is greater than the expected logarithm of the Bayes Factor for any $\lambda' \in \Lambda'$. Let $\beta(\lambda, \bar{\sigma}, \Theta)$ be the Bayes Factor of a scientist with error parameter $\bar{\sigma}$ for theory λ on experiment Θ. Then

$$E_Q\left(\log\left[\beta(\lambda^{\star\star}, \bar{\sigma}, \Theta)\right]\right) > E_Q\left(\log\left[\beta(\lambda', \bar{\sigma}, \Theta)\right]\right)$$

because I showed above that $\log\left[\beta(\lambda^{\star\star}, \bar{\sigma}, \Theta)\right] \geq \log\left[\beta(\lambda', \bar{\sigma}, \Theta)\right]$ (as anytime $\lambda^{\star\star}(\Theta) = \lambda'(\Theta)$ clearly the expected logarithm of the Bayes Factors are equal) with strict inequality on a set of positive measure under ν. So by Kelly's Theorem for any $\lambda' \in \Lambda$, any $r \in \mathbb{R}$ and any IISO $\{\Xi_n\}_{n=1}^{\infty}$ drawn i.i.d. from ν

$$\lim_{n \to \infty} Q\left[\prod_{\Theta \in \Xi_n} \frac{\beta(\lambda^{\star\star}, \bar{\sigma}, \Theta)}{\beta(\lambda', \bar{\sigma}, \Theta)} > r\right] = 1 \tag{12}$$

This, together with the finiteness of Λ', implies that

$$\forall r \in \mathbb{R}, \forall \epsilon > 0, \exists N^\star \in \mathbb{N} : \forall \lambda' \in \Lambda', \forall N > N^\star, Q\left[\prod_{\Theta \in \Xi_N} \frac{\beta(\lambda^{\star\star}, \bar{\sigma}, \Theta)}{\beta(\lambda', \bar{\sigma}, \Theta)} > r\right] > 1 - \epsilon$$

Now note that

$$\frac{P[\lambda^\star = \lambda^{\star\star}|\Xi]}{P[\lambda^\star \neq \lambda^{\star\star}|\Xi]} = \frac{\pi(\lambda^{\star\star}) \prod_{\Theta \in \Xi} \beta(\lambda^{\star\star}, \bar{\sigma}, \Theta)}{\sum_{\lambda' \in \Lambda'} \pi(\lambda') \prod_{\Theta \in \Xi} \beta(\lambda', \bar{\sigma}, \Theta)}$$

where π is the scientist's plausibility function. By the assumption that $\lambda^{\star\star} \in \text{supp}(\pi)$ we have that $\frac{\pi(\lambda^{\star\star})}{\sum_{\lambda \in \Lambda'} \pi(\lambda)} > 0$. So for any $\epsilon > 0$, if $r^\star \equiv \frac{(1-\epsilon)\pi(\lambda^{\star\star})}{\epsilon \sum_{\lambda \in \Lambda'} \pi(\lambda)}$ then if $\forall \lambda' \in \Lambda', \prod_{\Theta \in \Xi} \frac{\beta(\lambda^{\star\star}, \bar{\sigma}, \Theta)}{\beta(\lambda', \bar{\sigma}, \Theta)} > r^\star$, then

$$\frac{P[\lambda^\star = \lambda^{\star\star}|\Xi]}{P[\lambda^\star \neq \lambda^{\star\star}|\Xi]} = \frac{\pi(\lambda^{\star\star}) \prod_{\Theta \in \Xi} \beta(\lambda^{\star\star}, \bar{\sigma}, \Theta)}{\sum_{\lambda' \in \Lambda'} \pi(\lambda') \prod_{\Theta \in \Xi} \beta(\lambda', \bar{\sigma}, \Theta)} > \frac{1 - \epsilon}{\epsilon}$$

and so $P[\lambda^\star = \lambda^{\star\star}|\Xi] > 1 - \epsilon$. Plugging $r^\star$ in for r in 12 establishes consistency. Applying the reasoning in step 2 in precisely the same way, starting with the ratio of the two different scientist's plausibility values on the false theories, proves the second half of the lemma.

□

The proof relies crucially on Jensen's Inequality. Subset theories concentrate their probability more tightly than the true theory does. Because the logarithm is concave, this means that the expected logarithm of their Bayes Factor is smaller than the expected logarithm of the Bayes Factor of the true theory, regardless of the error parameter chosen. However, the smaller the error parameter chosen, the greater the stakes (informativeness) on any particular experiment, so the more harmful it is for a subset theory to concentrate its probability too densely.

Now I need to establish the second results discussed above, namely conditions under which an error rate $\bar{\sigma}$ is consistent against supersets of the true theory. The next Lemma supplies these conditions.

Lemma 2. *If*

1. *Under $\nu, \Lambda' \subseteq \{\lambda \in \Lambda : \lambda \sim \lambda^{\star\star} \wedge \lambda \supseteq \lambda^{\star\star}\}$ is a finite set of distinguishable supersets of the true theory*

2. *$\exists \rho^{\star\star} > 1, \underline{\rho}' \geq 1 : \forall \lambda' \in \Lambda'$ and for all but finitely many $\Theta \in \Gamma : \lambda'(\Theta) \neq \lambda^{\star\star}(\Theta), \rho(\lambda', \Theta) \geq \rho' \wedge \rho(\lambda^{\star\star}, \Theta) \geq \rho^{\star\star}$*

3.
$$\bar{\sigma}^\star < \frac{\log(1 + \bar{\mu}\rho^{\star\star}) - \log(1 + \bar{\mu}\rho')}{\left(1 - \frac{1}{\rho^{\star\star}}\right) \log(1 + \bar{\mu}\rho^{\star\star}) - \left(1 - \frac{1}{\rho'}\right) \log(1 + \bar{\mu}\rho')} \tag{13}$$

then $\bar{\sigma}$ is consistent against Λ'.
Conversely if

1. *$\exists \lambda' \in \Lambda'$ that is a superset of the true theory*

2. *$\exists$ infinitely many $\Theta \in \Gamma : \bar{\sigma}^\star > \frac{\log\left[1 + \bar{\mu}\rho(\lambda^{\star\star}, \Theta)\right] - \log\left[1 + \bar{\mu}\rho(\lambda', \Theta)\right]}{\left(1 - \frac{1}{\rho(\lambda^{\star\star}, \Theta)}\right) \log\left[1 + \bar{\mu}\rho(\lambda^{\star\star}, \Theta)\right] - \left(1 - \frac{1}{\rho(\lambda', \Theta)}\right) \log\left[1 + \bar{\mu}\rho(\lambda', \Theta)\right]}$*

then for some ν under which λ' is distinguishable from $\lambda^{\star\star}$, $\bar{\sigma}$ is inconsistent against Λ'.

Proof. See Appendix B.
□

Lemma 2 is a bit cumbersome and deserves some interpretation. As is proved in Appendix B, $\frac{\log(1+\bar{\mu}\rho^{\star\star}) - \log(1+\bar{\mu}\rho')}{\left(1 - \frac{1}{\rho^{\star\star}}\right) \log(1+\bar{\mu}\rho^{\star\star}) - \left(1 - \frac{1}{\rho'}\right) \log(1+\bar{\mu}\rho')}$ is decreasing in $\bar{\mu}, \rho'$ and $\rho^{\star\star}$. Therefore Lemma 2 states that, in order

to ensure consistency, $\bar{\mu}$ must be chosen sufficiently small ($\bar{\sigma}$ chosen sufficiently large) relative to the true error rate $\bar{\sigma}$ and the vagueness of both the true theory and its competitors. If no lower bound can be placed on the precision of the true and competitor theories except the trivial lower bound of 1, then expression 13 simplifies to expression 11 as proved in Appendix B.

Combining the two lemmas with an argument that superset and subset theories are the strongest competitors with the true theory proves the following theorem, which is the primary theoretical result of this paper.

Theorem 1. *If*

1. *Under ν, Λ' is a finite set of theories distinguishable from the true theory*

2. $\exists \underline{\rho}^{\star\star} > 1, \underline{\rho}' \geq 1 : \forall \lambda' \in \Lambda'$ *and for all but finitely many* $\Theta \in \Gamma : |\lambda'(\Theta)| > |\lambda^{\star\star}(\Theta)|, \rho(\lambda', \Theta) \geq \underline{\rho}' \wedge \rho(\lambda^{\star\star}, \Theta) \geq \underline{\rho}^{\star\star}$

3. $\bar{\sigma}^{\star} < \dfrac{\log(1+\bar{\mu}\underline{\rho}^{\star\star}) - \log(1+\bar{\mu}\underline{\rho}')}{\left(1 - \frac{1}{\underline{\rho}^{\star\star}}\right)\log(1+\bar{\mu}\underline{\rho}^{\star\star}) - \left(1 - \frac{1}{\underline{\rho}'}\right)\log(1+\bar{\mu}\underline{\rho}')}$

then $\bar{\sigma}$ is consistent against Λ'.

As an approximate converse, if

1. $\exists \lambda' \in \Lambda'$ *that is a superset of the true theory*

2. $\exists$ *infinitely many* $\Theta \in \Gamma : \bar{\sigma}^{\star} > \dfrac{\log\left[1+\bar{\mu}\rho(\lambda^{\star\star},\Theta)\right] - \log\left[1+\bar{\mu}\rho(\lambda',\Theta)\right]}{\left(1 - \frac{1}{\rho(\lambda^{\star\star},\Theta)}\right)\log\left[1+\bar{\mu}\rho(\lambda^{\star\star},\Theta)\right] - \left(1 - \frac{1}{\rho(\lambda',\Theta)}\right)\log\left[1+\bar{\mu}\rho(\lambda',\Theta)\right]}$

then for some ν for which λ' is distinguishable from $\lambda^{\star\star}$, $\bar{\sigma}$ is inconsistent against Λ'.

Proof. See Appendix C. □

The basic idea behind the proof is simple. Supersets and subsets are the theories that are most difficult to distinguish from the truth. Any theory which is not a superset or a subset of the truth will be dominated by theory of the same cardinality which is (or by the truth itself). Thus the theorem can be seen as a corollary of Lemmas 1 and 2.

Now I turn to the question of what a scientist will learn if her priors are not absolutely continuous with respect to the true theory. To answer this question I will assume[28] that the scientist uses the correct error rate, but for generality I will now allow $\sigma^{\star}(\Theta)$ to depend on the experiment Θ. Note that this is a slight modification of Assumption 7.

Assumption 7'. *There (objectively) exists a true measure Q, a true theory $\lambda^{\star\star} \in \Lambda$ and an error function $\sigma^{\star} : \Gamma \to (0,1)$. Under Q the distribution of $\omega^{\star}$ is the distribution of $\omega^{\star}$ under P for a scientist placing plausibility of 1 on $\lambda^{\star}$ and having error function $\sigma^{\star}(\cdot)$. The scientist uses $\sigma^{\star}(\cdot)$.*

The theorem below depends heavily on the notion of relative entropy (also known as Kullback-Leibler divergence):

---

[28]Understanding what results when both the scientist has the wrong error parameter and her plausibility is not absolutely continuous with respect to the truth would be interesting, but is not something I have yet found a simple way of formulating, much less solving.

Definition 11. *The* relative entropy *of a theory λ' from another theory λ in an experiment Θ is*

$$RE_\Theta(\lambda'||\lambda) \equiv \sum_{\theta \in \Theta} P\big[\omega^\star(\Theta) = \theta | \lambda^\star = \lambda\big] \log \left(\frac{P\big[\omega^\star(\Theta) = \theta | \lambda^\star = \lambda'\big]}{P\big[\omega^\star(\Theta) = \theta | \lambda^\star = \lambda\big]} \right)$$

Definition 12. *The* average relative entropy *of a theory λ' from another theory λ is*

$$\overline{RE}_\Theta(\lambda'||\lambda) \equiv E_Q \left[\sum_{\theta \in \Theta} P\big[\omega^\star(\Theta) = \theta | \lambda^\star = \lambda\big] \log \left(\frac{P\big[\omega^\star(\Theta) = \theta | \lambda^\star = \lambda'\big]}{P\big[\omega^\star(\Theta) = \theta | \lambda^\star = \lambda\big]} \right) \right]$$

Theorem 2. *For any measure P defined by obedience of Assumptions 1-7 and use of a finite support π as the plausibility function, any IISO $\{\Xi_n\}_{n=1}^\infty$ that is drawn i.i.d. under Q and any $\epsilon > 0$*

$$\lim_{n \to \infty} Q\Big(P\big[\lambda^\star \in \Lambda^\star(\pi)|\Xi_n\big] \leq 1 - \epsilon \Big) = 0$$

where

$$\Lambda^\star(\pi) \equiv argmin_{\lambda \in supp(\pi)} \overline{RE}_\Theta(\lambda||\lambda^{\star\star})$$

In words, the scientist will always consistently learn a theory (or theories if there are many equally close) that are maximally close in expected relative entropy to the truth[29].

Proof. By construction the expected logarithm of the Bayes Factor of any theory in $\Lambda^\star(\pi)$ is greater than the expected logarithm of the Bayes Factor of any theory outside $\Lambda^\star(\pi)$. Thus the result follows from the reasoning in Lemma 1.

$\square$

The primary value of this result is that it allows the theory above to be interpreted as process of learning the theory closest, in some sense, to the true theory, rather than as a process of learning the actually true theory. An important weakness of the result, however, is that it provides little intuitive sense of what closeness in terms of expected relative entropy means. For example, is the relative entropy of a theory that is a subset missing one outcome greater than that of a superset with one extra outcome? Some result addressing this problem would be a useful direction for further research.

8 Carnap-Gabaix-Laibson criteria

In a recent paper, "Seven Properties of a Good Model", Gabaix and Laibson (2007) compiled a list of principles that philosophers of science, as well as economists and other scientists, generally agree are desirable properties of models. These criteria for judging models can roughly be divided into two categories, with some of the criteria having aspects in both categories. Some of the properties primarily address the

---

[29]Note that in this setting if the scientist's priors only support one theory, the scientist will learn that theory. If the the support of the scientist's priors contain several theories, the scientist will learn the best of these, in the sense discussed here. Thus, at least in a vague, qualitative manner my results accord with those from the literature on strategic testing of experts; **?** establish that when experts (theories) are tested comparatively, the tester (scientist) can eventually distinguish true experts from false experts. This contrasts with earlier results by Sandroni (2003) and Olszewski and Sandroni (2007) that showed the general impossibility of determining whether a single potential expert did in fact have knowledge of an underlying stochastic process.

usefulness of a model: how easy is it to work with, does it speak to relevant issues and is it helpful in guiding decisions. Other properties are primarily about the truth value of a model: is it likely to provide a correct description of the stochastic process, as best as we can know it, underlying events we observe. It is this second group of properties that I focus on understanding and trading-off here.

I have therefore refined[30] the Gabaix-Laibson properties into a set of four criteria. These criteria also coincide with four "requirements" of a good "explicatum" (theory) identified by the great philosopher of science Rudolph Carnap in his classic work *The Logical Foundations of Probability*[31]. I therefore refer to the properties as the Carnap-Gabaix-Laibson criteria.

Because these properties seem intuitively to do a good job capturing the dimensions along which scientists (should) judge the truth value of a model, any reasonable theory of scientific learning should incorporate them. After discussing briefly each criterion, I will argue that the theory developed above incorporates it and that it makes precise trade-offs among them.

Criterion 1. Ex-ante plausibility[32]: *Do the assumptions and mechanisms through which the model generates its predictions seem plausible from an ex-ante perspective, before observing how well they explain particular phenomena? While this is certainly the most subjective[33] of the criteria, it is an essential one: priors will always be an input into[34] a reasonable model of belief formation.*

Clearly, the ex-ante plausibility of various models figures directly into the theory developed above through the plausibility function.

Criterion 2. Precision[35]: *Does the model make predictions that can be falsified? Are the predictions*

---

[30]The principles of a good model that they include, which I omit, are "Tractability" and "Conceptual Insighftulness". I also compress their notions of "Falsifiability" and "Predictive Precision" into a single criterion, as I see the distinction between these two as being one between positivity and degree of positivity. This distinction does not play an important role judging models.

[31]Interestingly, Carnap (1950), like I, advocates a view of probability in understanding scientific models that relies heavily on uniform distributions. He advocates this view on the basis of a Laplacian (maximum entropy) argument about symmetry. While I find this to be a reasonable justification for my assumptions above, I do not think it should be taken literally, given the problems with this view, and thus do not include it in the main text of the paper. In some ways, however, this paper can be seen as working out many of Carnap's ideas in more detail, as well as Popper's. His book offers a sort of philosophical defence of the approach taken here.

[32]Gabaix and Laibson refer to this as *parsimony*. I refer to it as plausibility, despite the substantial equivalence of the two, as it fits better with the probabilistic interpretation I make. Carnap describes it as "The explicatum should be as simple as possible."

[33]It certainly seems necessary that any theory of scientific learning will take priors as an input somewhere. However, it is worth noting that, in some sense not formally expressible with models we currently have, our priors about the validity of certain models are not always well-formed. For example, the primary value of the Von Neumann-Morganstern Theorem was to make it easier to judge the plausibility of the assumptions on which expected utility rests. What exactly this means, and what exactly the role of theoretical papers that "transform models" to make their plausibility "easier to judge", is an interesting topic for future meta-theoretical research.

Another interesting issue along these lines is that of "parsimony". Many models are explicitly (by calibration of some parameter) or implicitly (by making one or more of a class of well-defined possible assumptions) drawn from a broader class of potential models. It is often easy to judge the plausibility of this broader class of models, but difficult to directly judge the plausibility of the particular sub-model. A simple approach to dealing with this difficulty is to imagine that the prior on sub-models is uniformly distributed across all sub-models and therefore that the prior of the sub-model is $\frac{1}{T}$ times the prior of the super-model, where T is the number of sub-models. While this simple solution is likely reasonable in cases when the number of possible sub-models are finite and when all sub-models are (like experimental outcomes below) "equally close to one another". However, this is probably not a very satisfactory approach when there is a continuum of possible models or when some models are "very similar" while others are "very different". In this case, there may be a way of "clustering together" several sub-theories that are similar into groups and then using this as a method for constructing sub-models over which the uniform distribution hack is more persuasive. I hope to work out this idea in a future paper.

[34]Or output from, if one uses the theory to back priors out, beginning with posteriors.

[35]Carnap writes, "The characterization of the explicatum, that is, the rules of its use (for instance, in the form of a definition),

31

sufficiently precise that falsification of the model is not just theoretically possible but actually probable if the theory is incorrect? Theories that are falsifiable and make precise predictions are likely to be disproved if they are not reasonable models of reality. Therefore a precise model may be a good one, even if it is often incorrect, as the few instances in which it is born out provide strong evidence in its favor.

Precision is clearly defined in the theory and plays a central role in the learning rules derived.

Criterion 3. General applicability[36]*: Does the model make predictions about a broad range of phenomena? The more distinct phenomena a model makes predictions about the more independent opportunities there are for testing the theory.*

If we view a theory that is not applicable to a particular situation as making no prediction and therefore being maximally vague, then the generally applicability criterion naturally arises in the model. Suppose that one model makes predictions in a much broader range of settings than another theory. Then if we consider the union of all settings in which the two theories make predictions to the observation, then a theory which makes predictions more often will be much more precise than a theory which fails to do so. In a sense, general applicability can be seen, therefore, as a form of precision (or both can be seen as ways of achieving falsifiability).

Criterion 4. Empirical accuracy[37]*: Are the predictions of the model broadly born out by empirical study? Models, however plausible, precise or general, that are inconsistent with data do not provide accurate descriptions of reality and should be rejected.*

Empirical accuracy clearly plays a key role in the theory developed above.

Any model of scientific learning should incorporate, in some reasonable fashion, all of these elements. Clearly the theory developed above passes this basic test. Furthermore the model should make explicit it how it treats trade-offs among these various desirable feature. If possible it should provide a strategy that is not entirely subjective for determining the way in which these trade-offs are made and where this is not possible it should at least provide a method for checking the robustness of the models conclusions to the relative "weight" put on these various criteria. The trade-off between precision and accuracy in the model is set by the choice of the error parameter, which in turn I provide some techniques for choosing and checking the robustness of. However, the theory largely fails with respect to independently choosing the speed of updating, which conditions the trade-off between priors and the other attributes, and with respect to the weight given to general applicability. To adequately address the question of how fast updating should occur, a better model would be necessary of correlations across experiments. I hope that a similar analysis might also clarify the value of general applicability and how this should be traded off against other criteria. For example, some economists might argue that neuro-science data is very "correlated" in terms of errors, as are more economic questions, so that the performance of economic models on economic data is likely to provide more information about performance on future economic questions than is neuro-scientific data.

---

is to be given in an exact form, so as to introduce the explicatum into a well-connected system of scientific concepts."

[36]Carnap writes, "The explicatum is to be a fruitful concept, that is, useful for the formulation of many universal statements (empirical laws in the case of a nonlogical concept, logical theorems in the case of a logical concept)."

[37]Carnap writes, " The explicatum is to be similar to the explicandum in such a way that, in most cases in which the explicandum has so far been used, the explicatum can be used; however, close similarity is not required, and considerable differences are permitted." Note that Carnap's emphasis on allowing imperfect accuracy is closely tied to my notion of theoretical imperfection.

Trying to find a reasonable way of addressing such concerns would be an important step in improving the realism of the model here.

A theory of scientific learning should also provide baseline standards that put relatively few demands on analyst using the model to provide non-obvious probabilistic structure. Along the same lines, analysis using the model should be transparent and difficult to manipulate. The simplifying assumptions needed to achieve the last two goals will doubtless increase the "variance" of analysis using the model around the true conclusions that should be drawn. Nonetheless, I would argue that this added variance is worthwhile to reduce or eliminate incentives for analysts to strain the interpretation of models in non-transparent ways so as to arrive at results supporting their favorite model, as well as to limit debate about non-substantive issues. Throughout the development of the theory I have emphasized my efforts to achieve this.

Finally, a good theory of scientific learning should be tractable for applied researchers, easy enough to be used, as standard econometrics is, for common analysis, not merely in abstract theoretical inquiry or debate. By developing the theory in the context of a simple application, I have tried to show how it can easily be applied.

9   Conclusion

In this paper I provide (as far as I know) the first mathematical formulation of the general problem of scientific learning. I make five assumptions that give structure to the problem, allowing me to derive a simple, easily applicable solution. I apply the theory to analyze data on social preferences collected by Charness and Rabin (2002) and show that, in contrast to their conclusions, my theory suggests that their data supports Selfishness as the best theory of choice in simple games. I also prove some consistency results about the learning procedure inherent to the theory and showed how these results can be used as robustness checks, and as an algorithm for error parameter selection, in applied analysis. Finally, I discussed the ways in which the theory succeeds in incorporating some features that any to the scientific learning problem should.

The work here has numerous limitations that leave large areas for further inquiry. On the theoretical side many more robustness questions about the properties of the learning rule would be useful. The techniques I specify for choosing the error parameter are not fully satisfactory and improving on these would is important. Work on relaxing some of the models assumptions, particularly Independence and Uniformity would be helpful. The most important direction for future research, which I begin to explore in Appendix D, is the development of a theory, that is applicable to metric experiment spaces. In some sense the theory presented here can be seen as building intuition in a simple context for a more complex and mathematically dense metric theory to come. However, the current model likely has some applications to discrete settings beyond the Charness-Rabin data I discuss. In fact, because the Charness-Rabin example is so simple (every experiment is binary choice) a more sophisticated application would be important not just for the potential substantive insights it could yield, but also because it might act as a better test of the validity of the theory developed above.

$\bar{\sigma}_C$	$\bar{\mu}_C$	Consistent if $\bar{\sigma}_A <$	Selfish	Social welfare	Difference aversion	Competitive
.13	6.7	.47	1	$1.5 \cdot 10^{-7}$	$5.4 \cdot 10^{-6}$	$4 \cdot 10^{-7}$
.2	4	.54	1	$1.2 \cdot 10^{-6}$	$2.7 \cdot 10^{-5}$	$3.3 \cdot 10^{-6}$
.39	1.6	.67	1	$8.5 \cdot 10^{-5}$	$6 \cdot 10^{-4}$	$1.2 \cdot 10^{-4}$
.44	1.3	.7	1	$2.1 \cdot 10^{-4}$	.0012	$3.9 \cdot 10^{-4}$
.49	1	.73	1	$5.1 \cdot 10^{-4}$	.0023	$8.4 \cdot 10^{-4}$
.53	.89	.76	1	$9.9 \cdot 10^{-4}$	.0037	.0016

Table 5: Relative values of updates in favor of various (precision-augmented) theories, selfishness normalized to 1.

A   Precision-augmented Charness-Rabin analysis

In the analysis above of the Charness-Rabin data, one might worry that Charness and Rabin's formulation of the social preference models was "calibrated" to an analytical frame that disregarded precision, so that to provide a fair test of these theories, one should formulate more precise versions of the theories and run these against selfishness. It is to this task that this appendix is devoted.

I formulate the more precise version of each theory as allowing the weight on non-selfish effects to range from being one quarter as important as the player's own payoff to being two-thirds as important. I also preserve the inequalities in preference parameters Charness and Rabin assume. In the Social Welfare model this corresponds to putting a relative weight of between $\frac{1}{4}$ and $\frac{2}{3}$ on the other player's payoff, allowing this weight to differ between the cases when the other player is earning a higher or lower payoff and assuming that the weight is higher when the other player is behind than when she is ahead. In the Difference Aversion model this corresponds to putting a relative (negative) weight of between $\frac{1}{4}$ and $\frac{2}{3}$ on difference between the two players' payoffs, allowing this weight to differ between cases when my partner is ahead and I am ahead and insisting that the weight be a larger negative number when my partner is ahead than when I am ahead. In the Competitive model, it corresponds to putting a relative (negative) weight of between $\frac{1}{4}$ and $\frac{2}{3}$ on the other player's payoff, allowing this weight to differ between the case when I or my partner is ahead, insisting that the negative weight be greater when I am behind and adjusting scale to avoid discontinuities at the point when our payoffs are equal. I use Selfishness without any changes.

After adding this additional structure, all of the theories of social preferences are much more precise, but also much less accurate. Theory-specific maximum likelihood estimates of error parameters are now .13, .53, .41, .49 for, respectively, the selfishness, social welfare, difference aversion and competitive models. Thus the uniform prior, "focal" value of the error parameter is .39. Table 8 provides information analogous to tables 7 for this case.

Table 8 reveals a few interesting things. First, making Social Welfare preference more precise makes it much less accurate, causing it to do worse for all parameter values shown. Second, Difference Aversion does slightly better, especially for high error rates and now consistently out preforms Social Welfare preferences, even though it still substantially lags selfishness. This may provide some rationale for why Difference Aversion, which Charness and Rabin largely dismiss, was a popular model of social preferences: for a "reasonable range" of parameter values it captures social preferences in a reasonably precise and not entirely accurate manner. However, the analysis indicates that the data supports Selfishness strongly as the best model.

B   Proof of Lemma 2

First I establish the forward direction of the Lemma and then I turn to the (approximate) converse. I prove in three steps:

1. First, I show that if, in a particular experiment,

$$\bar{\sigma}^\star < \frac{\log\left[1 + \bar{\mu}\rho(\lambda^{\star\star}, \Theta)\right] - \log\left[1 + \bar{\mu}\rho(\lambda', \Theta)\right]}{\left(1 - \frac{1}{\rho(\lambda^{\star\star}, \Theta)}\right)\log\left[1 + \bar{\mu}\rho(\lambda^{\star\star}, \Theta)\right] - \left(1 - \frac{1}{\rho(\lambda', \Theta)}\right)\log\left[1 + \bar{\mu}\rho(\lambda', \Theta)\right]}$$

   and $\lambda'(\Theta) \neq \lambda^{\star\star}(\Theta)$ then $E_Q\left(\log\left[\beta(\lambda^{\star\star}, \Theta)|\Theta\right]\right) > E_Q\left(\log\left[\beta(\lambda', \Theta)|\Theta\right]\right)$.

2. Second, I show that this bound (the RHS of the above expression) is increasing in both $\rho(\lambda^{\star\star}, \Theta)$ and $\rho(\lambda', \Theta)$, so that if the hypotheses of the lemma (which lower bound precision) hold, then, unconditionally, because $\lambda' \approx \lambda^{\star\star}$, $E_Q\left(\log\left[\beta(\lambda^{\star\star}, \Theta)\right]\right) > E_Q\left(\log\left[\beta(\lambda', \Theta)\right]\right)$.

3. Finally, I invoke the argument from the proof of Lemma 1 to use this to establish consistency.

Recall from the proof of Lemma 1 that the expected logarithm of the Bayes Factor of the true theory on a particular experiment Θ is

$$\left(1 - \bar{\sigma}^\star + \frac{\bar{\sigma}^\star}{\rho(\lambda^\star, \Theta)}\right)\log\left[1 + \bar{\mu}\rho(\lambda^{\star\star}, \Theta)\right]$$

For λ' satisfying $\lambda'(\Theta) \supset \lambda^{\star\star}$ this is given by

$$\left(1 - \bar{\sigma}^\star + \frac{\bar{\sigma}^\star}{\rho(\lambda', \Theta)}\right)\log\left[1 + \bar{\mu}\rho(\lambda', \Theta)\right]$$

The first expression is greater than the second iff

$$(1-\bar{\sigma}^\star)\left(\log\left[1+\bar{\mu}\rho(\lambda^{\star\star}, \Theta)\right] - \log\left[1+\bar{\mu}\rho(\lambda', \Theta)\right]\right) > \bar{\sigma}^\star\left(\frac{1}{\rho(\lambda', \Theta)}\log\left[1+\bar{\mu}\rho(\lambda', \Theta)\right] - \frac{1}{\rho(\lambda^\star, \Theta)}\log\left[1+\bar{\mu}\rho(\lambda^{\star\star}, \Theta)\right]\right) \iff$$

$$\log\left[1+\bar{\mu}\rho(\lambda^{\star\star}, \Theta)\right] - \log\left[1+\bar{\mu}\rho(\lambda', \Theta)\right] > \bar{\sigma}^\star\left[\left(1-\frac{1}{\rho(\lambda^{\star\star}, \Theta)}\right)\log\left[1+\bar{\mu}\rho(\lambda^{\star\star}, \Theta)\right] - \left(1-\frac{1}{\rho(\lambda', \Theta)}\right)\log\left[1+\bar{\mu}\rho(\lambda', \Theta)\right]\right]$$

Note that by the assumption that $\lambda'(\Theta) \supset \lambda^{\star\star}(\Theta)$ we have that $\rho(\lambda^{\star\star}, \Theta) > \rho(\lambda', \Theta)$ so that the above expression will hold iff and only if

$$\bar{\sigma}^\star < \frac{\log\left[1 + \bar{\mu}\rho(\lambda^{\star\star}, \Theta)\right] - \log\left[1 + \bar{\mu}\rho(\lambda', \Theta)\right]}{\left(1 - \frac{1}{\rho(\lambda^{\star\star}, \Theta)}\right)\log\left[1 + \bar{\mu}\rho(\lambda^{\star\star}, \Theta)\right] - \left(1 - \frac{1}{\rho(\lambda', \Theta)}\right)\log\left[1 + \bar{\mu}\rho(\lambda', \Theta)\right]} \tag{14}$$

which establishes step 1.

Next I want to show that the RHS of inequality 14 is increasing in both $\rho(\lambda^{\star\star}, \Theta)$ and $\rho(\lambda', \Theta)$ given that $\rho(\lambda^{\star\star}, \Theta) > \rho(\lambda', \Theta)$. For aesthetic reasons I will abbreviate $\rho(\lambda^{\star\star}, \Theta)$ as ρ_1 and $\rho(\lambda', \Theta)$ as ρ_2. To see that the right hand side of the expression above in ρ_1, take the derivative which yields

$$\frac{\rho_2\Big(\bar{\mu}\rho_1[\rho_1+\rho_2-2\rho_1\rho_2]\log[1+\bar{\mu}\rho_2]+\rho_2\Big[2\bar{\mu}\rho_1(\rho_1-1)+(1+\bar{\mu}\rho_1)\big(\log[1+\bar{\mu}\rho_1]-\log[1+\bar{\mu}\rho_2]\big)\Big]\log[1+\bar{\mu}\rho_1]\Big)}{(1+\bar{\mu}\rho_1)\big[(\rho_1-1)\rho_2\log(1+\bar{\mu}\rho_1)-\rho_1(\rho_2-1)\log(1+\bar{\mu}\rho_2)\big]^2}$$

Recalling that $\rho_1>\rho_2\geq 1$ and $\bar{\mu}>0$, clearly $(1+\bar{\mu}\rho_1)$ and ρ_2 are positive, so the sign of the expression is the same as the sign of

$$(1+\bar{\mu}\rho_1)\rho_2\log(1+\bar{\mu}\rho_1)\big(\log[1+\bar{\mu}\rho_2]-\log[1+\bar{\mu}\rho_1]\big)+\bar{\mu}\rho_1(\rho_1-\rho_2)\log(1+\bar{\mu}\rho_2)$$

Clearly this expression is 0 when $\rho_1=\rho_2$. I now want to show that for $\rho_1\neq\rho_2$, $\rho_1,\rho_2\geq 1$ we have this expression strictly positive. I am extremely grateful to Mihai Manea for the proof of this fact, which I now present. Let $e^a\equiv 1+\bar{\mu}\rho_1$ and $e^b\equiv 1+\bar{\mu}\rho_2$. Then proving that the above expression is positive for $\rho_1\neq\rho_2$, $\rho_1,\rho_2\geq 1$ is equivalent to proving that the following expression is strictly positive when $a\neq b$ and $a,b\geq 0$:

$$e^a\big(e^b-1\big)a(b-a)+\big(e^a-1\big)\big(e^a-e^b\big)b$$

I show this in two cases. First suppose that $\rho_2>\rho_1$. Then the above expression being strictly positive is equivalent to the following inequality:

$$\frac{e^{b-a}-1}{b-a}\cdot\frac{e^a-1}{a}<\frac{e^b-1}{b}$$

or equivalently

$$\ln\left(\frac{e^{b-a}-1}{b-a}\right)+\ln\left(\frac{e^a-1}{a}\right)<\ln\left(\frac{e^b-1}{b}\right)$$

Now let $f(x)\equiv\ln\left(\frac{e^x-1}{x}\right)$ when $x\neq 0$ and $f(0)\equiv 0$. f is continuous[38], twice continuously differentiable[39] and convex[40]. I want to show that $f(b-a)+f(a)<f(b)=f(b)+f(0)$. But $b>a$ so by convexity of f, this holds by Jensen's Inequality:

$$f(b-a)+f(a)=f\left(\frac{b-a}{b}\cdot 0+\frac{a}{b}\cdot b\right)+f\left(\frac{b-a}{b}\cdot b+\frac{a}{b}\cdot 0\right)<\left(\frac{b-a}{b}+\frac{a}{b}\right)[f(b)+f(0)]=f(b)+f(0)$$

In the other case, when $\rho_1<\rho_2$, I need to show that $f(b-a)+f(a)>f(b)$. But f is always weakly positive, so $f(b-a)+f(a)>f(a)>f(b)$ as $a,b\geq 0$ and $a>b$ in this case.

---

[38]$\lim_{x\to 0}f(x)=0$ as it is well known that $\lim_{x\to 0}\frac{e^x-1}{x}=1$.

[39]This requires multiple tedious applications of L'Hôpital's rule and so is omitted here, but is available upon request.

[40]To see this, note that

$$f''(x)=\frac{1}{x^2}+\frac{1}{2-2\cosh(x)}$$

being positive suffices to show convexity given twice continuous differentiability. Because $\cosh(x)\geq 1$ the second term is always negative, so it suffices to show that $2\cosh(x)-2-x^2\geq 0$. When $x=0$ both this expression and its derivative $2\sinh(x)-2x$ are 0. The second derivative of the expression is $2\cosh-2$ which is positive, thus $2\cosh(x)-2-x^2\geq 0$ is convex and minimized at 0 and therefore positive.

Therefore the RHS of expression 14 is increasing in ρ_1. Because the expression is symmetric in $\rho_1 = \rho_2$ it also suffices to show the expression is increasing in ρ_2 (given that $\rho_1 > \rho_2 \geq 1$). Thus if we can bound ρ_1 below by $\rho^{\star\star}$ and we can bound ρ_2 below by ρ', $\forall \Theta \in \Gamma$ then

$$\bar{\sigma}^\star < \frac{\log\left[1 + \bar{\mu}\rho^{\star\star}\right] - \log\left[1 + \bar{\mu}\rho'\right]}{\left(1 - \frac{1}{\rho^{\star\star}}\right)\log\left[1 + \bar{\mu}\rho^{\star\star}\right] - \left(1 - \frac{1}{\rho'}\right)\log\left[1 + \bar{\mu}\rho'\right]} \tag{15}$$

implies that for all experiments where $\lambda'(\Theta) \neq \lambda^{\star\star}(\Theta)$ we have $E_Q\left(\log\left[\beta(\lambda^{\star\star}, \Theta)|\Theta\right]\right) > E_Q\left(\log\left[\beta(\lambda', \Theta)|\Theta\right]\right)$. Because for any $\lambda' \in \Lambda'$ we have by assumption that $\lambda' \supseteq \lambda^{\star\star}$ and $\lambda' \nsim \lambda^{\star\star}$ the argument that established Lemma 1 implies that (given that inequality 15 is violated only at a finite, and therefore zero-probability given that ν is non-atomic, set in Γ), unconditionally, $E_Q\left(\log\left[\beta(\lambda^{\star\star}, \Theta)\right]\right) > E_Q\left(\log\left[\beta(\lambda', \Theta)\right]\right)$. Again by the argument that established Lemma 1 this suffices to establish consistency and prove the forward direction of Lemma 2.

To show that this formula simplifies to expression 7.1 when no ex-ante lower bounds other than 1 can be placed on the precision of the true and competitor theory, note that in this case it is appropriate first to take a limit as ρ' and then $\rho^{\star\star}$ goes to 1, as $\rho' < \rho^{\star\star}$. When $\rho' = 1$, expression 13 simplifies to

$$\frac{\log\left(1 + \bar{\mu}\rho^{\star\star}\right) - \log(1 + \bar{\mu})}{\left(1 - \frac{1}{\rho^{\star\star}}\right)\log\left(1 + \bar{\mu}\rho^{\star\star}\right)}$$

This expression is clearly an indeterminate form ($\frac{0}{0}$) as $\rho^{\star\star} \to 1$. By L'Hôpital's rule, the value of the limit is

$$\lim_{\rho^{\star\star} \to 1} \frac{\frac{\bar{\mu}}{1 + \bar{\mu}\rho^{\star\star}}}{\frac{\bar{\mu}}{1 + \bar{\mu}\rho^{\star\star}} - \frac{\frac{\bar{\mu}}{1 + \bar{\mu}\rho^{\star\star}} - \log\left(1 + \bar{\mu}\rho^{\star\star}\right)}{\rho^{\star\star^2}}} = \frac{\frac{\bar{\mu}}{1 + \bar{\mu}}}{\frac{\bar{\mu}}{1 + \bar{\mu}} - \frac{\bar{\mu}}{1 + \bar{\mu}} + \log(1 + \bar{\mu})} = \frac{\frac{\frac{1 - \bar{\sigma}}{\bar{\sigma}}}{\frac{1}{\bar{\sigma}}}}{\log\left(1 + \frac{1 - \bar{\sigma}}{\bar{\sigma}}\right)} = -\frac{1 - \bar{\sigma}}{\log(\bar{\sigma})}$$

Now consider the converse and assume its hypotheses. Then by the reasoning above, $\exists \lambda' \in \Lambda'$, and infinitely many $\Theta' \in \Gamma$:

$$E_Q\left(\log\left[\beta(\lambda^{\star\star}, \Theta)|\Theta\right]\right) < E_Q\left(\log\left[\beta(\lambda', \Theta)|\Theta\right]\right)$$

So clearly if ν places a probability mass of 1 on this set (which is possible even while remaining non-atomic as it is an infinite set), the reasoning in the proof of Lemma 1 establishes that for any $r \in \mathbb{R}$:

$$\lim_{n \to \infty} Q\left[\frac{P[\lambda^\star = \lambda'|\Xi_n]}{P[\lambda^\star = \lambda^{\star\star}|\Xi_n]} > r\right] = 1$$

So clearly consistency cannot hold.

C   Proof of Theorem 1

The converse is an immediate consequence of Lemma 2.

To establish the first claim, I prove, in the proof Lemmas 1 and 2, in two steps. First I show that for any theory $\lambda' \in \Lambda'$, the expected logarithm of that theory's Bayes Factor for any experiment for which

37

that theory is not identical to the truth is strictly less than the expected logarithm of the true theory's Bayes Factor. Then I invoke the argument used to establish Lemma 1 to prove consistency.

Consider a theory $\lambda' \in \Lambda'$ and an experiment Θ for which $\lambda'(\Theta) \neq \lambda^{\star\star}(\Theta)$. Such an experiment exists as λ' is distinguishable from $\lambda^{\star\star}$. If $\lambda'(\Theta) \supset \lambda^{\star\star}(\Theta)$ or if $\lambda'(\Theta) \subset \lambda^{\star\star}(\Theta)$ then, given the hypotheses of the theorem,

$$E_Q\Big(\log\big[\beta(\lambda^{\star\star}, \bar{\sigma}, \Theta)\big]\Big) > E_Q\Big(\log\big[\beta(\lambda', \bar{\sigma}, \Theta)\big]\Big) \tag{16}$$

by the arguments used to establish Lemmas 1 and 2. I want to show that inequality 16 holds even if $\lambda'(\Theta)$ is neither a superset nor a subset of $\lambda^{\star\star}$. There are three cases to consider:

1. $|\lambda'(\Theta)| = |\lambda^{\star\star}(\Theta)|$: Recall that the expected logarithm of the Bayes Factor of the true theory is

$$\left(1 - \bar{\sigma}^\star + \frac{\bar{\sigma}^\star}{\rho(\lambda^\star, \Theta)}\right)\log\left(1 + \bar{\mu}\rho(\lambda^{\star\star}, \Theta)\right)$$

   On the other hand, the expected logarithm of the Bayes Factor of λ' is

$$\left(\frac{|\lambda'(\Theta) \cap \lambda^{\star\star}(\Theta)|}{|\lambda^{\star\star}(\Theta)|}[1 - \bar{\sigma}^\star] + \frac{\bar{\sigma}^\star}{\rho(\lambda^\star, \Theta)}\right)\log\left(1 + \bar{\mu}\rho(\lambda^{\star\star}, \Theta)\right)$$

   as $\rho(\lambda', \Theta) = \rho(\lambda^{\star\star}, \Theta)$ because $|\lambda'(\Theta)| = |\lambda^{\star\star}(\Theta)|$. But clearly

$$\left(1 - \bar{\sigma}^\star + \frac{\bar{\sigma}^\star}{\rho(\lambda^\star, \Theta)}\right)\log\left(1 + \bar{\mu}\rho(\lambda^{\star\star}, \Theta)\right) > \left(\frac{|\lambda'(\Theta) \cap \lambda^{\star\star}(\Theta)|}{|\lambda^{\star\star}(\Theta)|}[1 - \bar{\sigma}^\star] + \frac{\bar{\sigma}^\star}{\rho(\lambda^\star, \Theta)}\right)\log\left(1 + \bar{\mu}\rho(\lambda^{\star\star}, \Theta)\right)$$

   as $\bar{\sigma}^\star < 1$ $\frac{|\lambda'(\Theta) \cap \lambda^{\star\star}(\Theta)|}{|\lambda^{\star\star}(\Theta)|} < 1$ by construction. Thus inequality 16 holds in this case.

2. $|\lambda'(\Theta)| > |\lambda^{\star\star}(\Theta)|$: Consider some theory $\tilde{\lambda} : \tilde{\lambda}(\Theta) \supset \lambda^{\star\star}(\Theta)$ and $|\tilde{\lambda}(\Theta)| = |\lambda'(\Theta)|$. Recall from the proof of Lemma 2 that the expected logarithm of the Bayes Factor for $\tilde{\lambda}$ is

$$\left(1 - \bar{\sigma}^\star + \frac{\bar{\sigma}^\star}{\rho(\tilde{\lambda}, \Theta)}\right)\log\left(1 + \bar{\mu}\rho(\tilde{\lambda}, \Theta)\right)$$

   On the other hand, the expected logarithm of the Bayes Factor for λ' is

$$\left(\frac{|\lambda'(\Theta) \cap \lambda^{\star\star}(\Theta)|}{|\lambda^{\star\star}(\Theta)|}[1 - \bar{\sigma}^\star] + \frac{\bar{\sigma}^\star}{\rho(\tilde{\lambda}, \Theta)}\right)\log\left(1 + \bar{\mu}\rho(\tilde{\lambda}, \Theta)\right)$$

   as $\rho(\tilde{\lambda}, \Theta) = \rho(\lambda', \Theta)$ given that, by construction, $|\tilde{\lambda}(\Theta)| = |\lambda'(\Theta)|$. By the same reasoning as in case 1 above, this second expression is less than the first. But because, by construction, $\tilde{\lambda}(\Theta) \supset \lambda^{\star\star}(\Theta)$, the proof of Lemma 2 combined with our reasoning here implies that:

$$E_Q\Big(\log\big[\beta(\lambda^{\star\star}, \bar{\sigma}, \Theta)\big]\Big) > E_Q\Big(\log\big[\beta(\tilde{\lambda}, \bar{\sigma}, \Theta)\big]\Big) > E_Q\Big(\log\big[\beta(\lambda', \bar{\sigma}, \Theta)\big]\Big)$$

   establishing inequality 16.

3. $|\lambda'(\Theta)| < |\lambda^{\star\star}(\Theta)|$: Consider some theory $\tilde{\lambda} : \tilde{\lambda}(\Theta) \subset \lambda^{\star\star}(\Theta)$ and $|\tilde{\lambda}(\Theta)| = |\lambda'(\Theta)|$. Recall from the proof of Lemma 1 that the expected logarithm of the Bayes Factor for $\tilde{\lambda}$ is

$$\left(\frac{\rho(\lambda^{\star\star}, \Theta)}{\rho(\tilde{\lambda}, \Theta)} [1 - \bar{\sigma}^{\star}] + \frac{\bar{\sigma}^{\star}}{\rho(\tilde{\lambda}, \Theta)} \right) \log \left(1 + \bar{\mu}\rho(\tilde{\lambda}, \Theta) \right)$$

On the other hand, the expected logarithm of the Bayes Factor for λ' is

$$\left(\frac{\rho(\lambda^{\star\star}, \Theta)}{\rho(\tilde{\lambda}, \Theta)} \frac{|\lambda'(\Theta) \cap \lambda^{\star\star}(\Theta)|}{|\tilde{\lambda}(\Theta) \cap \lambda^{\star\star}(\Theta)|} [1 - \bar{\sigma}^{\star}] + \frac{\bar{\sigma}^{\star}}{\rho(\tilde{\lambda}, \Theta)} \right) \log \left(1 + \bar{\mu}\rho(\tilde{\lambda}, \Theta) \right)$$

as $\rho(\tilde{\lambda}, \Theta) = \rho(\lambda', \Theta)$ given that, by construction, $|\tilde{\lambda}(\Theta)| = |\lambda'(\Theta)|$. Now note that $\frac{|\lambda'(\Theta) \cap \lambda^{\star\star}(\Theta)|}{|\tilde{\lambda}(\Theta) \cap \lambda^{\star\star}(\Theta)|} < 1$ by the assumption that $\lambda'(\Theta)$ is not a subset of $\lambda^{\star\star}(\Theta)$, while $\tilde{\lambda}(\Theta)$ is. Also, clearly $\frac{\rho(\lambda^{\star\star}, \Theta)}{\rho(\tilde{\lambda}, \Theta)} > 0$ so the first expression is great than the second. Again by Lemma 1 and this reasoning

$$E_Q \Big(\log \big[\beta(\lambda^{\star\star}, \bar{\sigma}, \Theta) \big] \Big) > E_Q \Big(\log \big[\beta(\tilde{\lambda}, \bar{\sigma}, \Theta) \big] \Big) > E_Q \Big(\log \big[\beta(\lambda', \bar{\sigma}, \Theta) \big] \Big)$$

establishing inequality 16 in this case as well.

Thus inequality 16 holds so long as $\lambda'(\Theta) \neq \lambda^{\star\star}(\Theta)$. But by exactly the same reasoning as in the proof of Lemma 1 establishes the forward direction of the Theorem.

D   Towards a theory with metric experiment spaces

The basic problem in a continuous metric experiment space is that there is essentially zero probability that the realized outcome will be *exactly* that predicted by a theory. Even if the space is discretized it is not reasonable to see an inflation rate of 3% as a complete falsification of a theory which predicts an inflation of 2%. Thus we require a theory of partial falsification which allows the confirmation or rejection of a model based on an experimental outcome to be essentially continuous in the outcome.

One "natural" way to extend the model discussed above to a metric space would be to imagine that emanating from each value in the space predict by the theory there is a Gaussian distribution centered at that point and that the probability distribution "predicted" by a theory is thus a mixture of Gaussians, where the common variance of the Gaussians is the analog of the error parameter. This model would be nice, because we have simple computational ways of dealing with mixture of Gaussian models. To see the problem with this model, imagine that there is a theory whose predictions are indexed by a continuous parameter $\nu \in [0, 1]$. Suppose that the theory's prediction for $\nu = 0$ is 0, for $\nu = 1$ is 1 but for $\nu \in (0, 1)$ the prediction is always in the interval $[.5, .50001]$. How would one translate this into a mixture of Gaussian model? Would we place a probability weight of .8 essentially on a Gaussian with mean of "the interval" $[.5, .50001]$? How much weight would be placed on the Gaussian centered at 0? How much on the Gaussian centered at 1? There is no reasonable economic sense in which, simply because all parameter values except 0 and 1 generate predictions in $[.5, .50001]$ that 0 and 1 are not equal predictions of the model. Furthermore, despite the fact that each value of $\nu \in (0, 1)$ may give a distinct prediction at the sixth decimal place,

observationally this is nearly equivalent to all of the predictions being .5. Thus the mixture of Gaussians model, despite its initial appeal, is not the right formalization for the idea underlying it.

Instead I want a theory where each prediction of a model has equal probability weight on it and probability weights die off at the same rate around these predictions. I want a maximum of Gaussians, not a mixture of Gaussians. That is the probability weight put on any particular point by a theory should be related to the distance between that point at the nearest point in the space predicted by the theory. This theory can be developed by mimicking the definitions and assumptions used in the development of the simple theory in Section 3, making the relevant changes where necessary.

Definition′ 1. *An* experiment Θ *is a (weak) subset of* $\mathbb{R}^n$.

For simplicity, I assume that there is an upper bound on the dimensionality of an experiment and that experiments have outcomes that can be viewed as subsets of real space. Note that the simple model can obvious be seen as a special case.

Definition′ 2. *The* studied phenomenon *is a (generally quite infinite) set of experiments* Γ.

Definition′ 3. *The* observation Ξ *is a finite subset of* Γ.

Definition′ 4. *A* theory *is a mapping* $\lambda : \Gamma \to 2^{\mathbb{R}^n}$ *such that* $\lambda(\Theta) \subseteq \Theta, \forall \Theta \in \Gamma$. $\lambda(\Theta)$ *is called the set of outcomes predicted by* λ. *Let the* theory space Λ *be the set of all possible theories satisfying this definition.*

Assumption′ 1. *The scientist believes that one theory* $\lambda^\star \in \Lambda$ *is the* true theory *but is uncertain as to its identity. The* plausibility distribution $\pi : \Lambda \to [0,1]$ *be a probability distribution over* Λ *that represents the scientist's priors over the different theories being the true theory.*

Definition′ 5. *A* world *is a mapping* $\omega : \Gamma \to \mathbb{R}^n$ *such that* $\omega(\Theta) \in \Theta, \forall \Theta \in \Gamma$. *Let* Ω *be the set of all possible worlds.*

Assumption′ 2. $P\Big[\omega^\star(\Theta_1) = \theta_1, \ldots, \omega^\star(\Theta_N) = \theta_N \big| \lambda^\star = \lambda\Big] = \prod_{i=1}^N P\Big[\omega^\star(\Theta_i) = \theta_i \big| \lambda^\star = \lambda\Big]$ *for all* $\{\Theta_i\}_{i=1}^N \subseteq \Gamma$ *and all* $N \in \mathbb{N}$ *such that* $\Theta_i \neq \Theta_j$ *for* $i \neq j$.

These are exactly as in the simple theory of Section 3, except for some slight simplifications of domains made possible by the assumption that all $\Theta \subseteq R^n$. Note, however, that Λ is all the more infinite now that each Θ may be an uncountable set. The major differences in this setting emerge in the probabilistic structure.

Assumption′ 3. *For any* $\theta \in \Theta$:

$$P\big[\omega^\star(\Theta) = \theta | \lambda^\star = \lambda\big] = \frac{\phi\Big(D\big[\lambda(\Theta), \theta\big]; \sigma[\lambda, \Theta]\Big)d\chi(\theta; \Theta)}{\int_{z \in \Theta} \phi\Big(D\big[\lambda(\Theta), z\big]; \sigma[\lambda, \Theta]\Big)d\chi(z; \Theta)}$$

where $D : 2^{\mathbb{R}^n} \times \mathbb{R}^n \to \mathbb{R}_+$ *is a metric of distance between subsets of* $\mathbb{R}^n$ *and points in* $\mathbb{R}^n$ *and* $\chi(z; \Theta)$ *a probability distribution over* Θ *called the* prior shape *of* Θ. ϕ *satisfies the following properties:*

1. $\phi(\cdot, \sigma) > 0$ *everywhere and is strictly decreasing* $\forall \sigma \in (0, \infty)$

2. $\int_{y=0}^\infty \phi(y; \sigma)dy$ *exists and is finite*

3. $\phi(\cdot;\sigma)$ is absolutely continuous $\forall \sigma \in (0,\infty)$

4. $\frac{\phi(x;\sigma)}{\phi(x';\sigma)}$ is strictly decreasing in $\sigma, \forall 0 \leq x < x'$

5. $\lim_{\sigma \to 0} \frac{\phi(0,\sigma)}{\int_{y \in \Theta} \phi\big(d(x,y);\sigma\big)dy} = \infty$

6. $\lim_{\sigma \to \infty} \frac{\phi(x;\sigma)}{\phi(x';\sigma)} = 1, \forall x, x' \in \mathbb{R}_+$

Assumption$'$ 3 embodies the basic logic above about extending the model to a metric space. It states that the probability weight put on any particular outcome should be a decreasing function of the "distance" between the predictions of the true theory and the that outcome. The assumptions about the shape of ϕ ensure that it has full support (no outcome entirely disqualifies a theory), is integrable (which ensures that the definition makes sense) and is non-atomic. $\phi(\cdot;\sigma)$ is parameterized by σ (which in turn may depend on the experiment and the true theory). This is analogous to the error parameter in the simple theory; here it is a rough measure of the variance of the distribution. For example, a normal PDF with mean 0 and variance σ would satisfy all the definitions above; one can think of this as a focal example. The properties assumed about this parametrization are meant to capture the notion that σ roughly represents the spread of the distribution: as it gets larger, the relative weight put on points farther away becomes larger; as it goes to infinity, the implied distribution is flat over the space; as it goes to 0 the implied distribution is atomic at points 0 distance from the true theory's predictions.

While this establishes a few properties about ϕ that suggest a reasonable class of parameterizations, D, χ and σ still have little structure. These are provided by the following assumptions.

Assumption$'$ 4. *Two possible assumptions might be used here:*

1. *$\forall \Theta \in \Gamma$, Θ is compact and $\chi(\cdot;\Theta)$ is the uniform distribution over Θ.*

2. *$\forall \Theta \in \Gamma$, $\Theta = \mathbb{R}^m$ for some $m \leq n$ (not necessarily the same for every experiment) and $\chi(\cdot;\Theta)$ is a multivariate normal distribution over Θ.*

The two versions of the assumption correspond to two different perspectives on the distribution over outcomes implied by a fully vague theory (one that makes no predictions). The first corresponds to the assumption that there is some compact set of possible values that the experimental outcome might take on and that any of these is equally likely. The second corresponds to the assumption that the (multidimensional) experimental outcome may take on any value, but that its ex-ante mean (vector) and covariance matrix (roughly) are known. Both versions of the assumption then correspond to the view that the scientist, in the absence of a theory, has no further knowledge (maximum entropy) about the outcome than what is assumed. I think it is likely that in most applications, the second version of assumption 4 will be more useful, as moments are often easier to consider than outer bounds and the resulting space is probably more plausible.

Assumption$'$ 5. $\sigma(\lambda, \Theta) \equiv \sigma(\Theta)$

This is the same theory-independent error assumption in Section 3, though σ now parameterizes the distribution differently.

Assumption′ 6. $D[X, y] \equiv \inf_{x \in X} d(x, y)$, *where d is some metric on* $\mathbb{R}^n$. *A convenient metric is the Euclidean norm.*

Thus D is the Hausdorff extension of the Euclidean metric in the special case when one of the sets (y here) is singleton. Note that the assumption of using the Euclidean metric is not terribly restrictive, as in an application the dimensions may be re-scaled in a reasonable way before applying the theory. Assuming it, however, does provide sufficient structure to allow computations and theoretical investigation.

Given these assumptions, the Bayes Factor for a theory on an experiment Θ will be

$$\beta\big[\omega^\star(\Theta); \lambda, \Theta, \sigma\big] = \frac{\phi\Big(\inf_{y \in \lambda(\Theta)} d\big[\omega^\star(\Theta), y\big]; \sigma\Big) d\chi\big(\omega^\star[\Theta]; \Theta\big)}{\int_{z \in \Theta} \phi\big[\inf_{y \in \lambda(\Theta)} d(z, y); \sigma\big] d\chi(z; \Theta)}$$

This essentially has the same precision-accuracy form as earlier $\phi\Big(\inf_{y \in \lambda(\Theta)} d\big[\omega^\star(\Theta), y\big]\Big); \sigma\Big)$ (up to affine transformations) measures the accuracy of theory λ in experiment Θ when the outcome is $\omega^\star(\Theta)$ and $\frac{1}{\int_{z \in \Theta} \phi\big[\inf_{y \in \lambda(\Theta)} d(z,y); \sigma\big] dz}$ measures the precision of the theory on Θ. For small values of σ, when the distribution collapses towards a point mass, accuracy becomes extremely important, the probability of anything that is not exactly a prediction of the theory goes to 0. For large values of σ precision becomes important because (at least over a large range) the implied distribution is essentially flat, so (conditional on being inside this flat region, that is not totally falsified) what matters is that the integral over the whole space is not too large. Note, though, that when one makes the Gaussian assumption about the prior shape of the space, both precision and accuracy will depend on where the theory's predictions like, not just their shape. Theories that predict things with very low ex-ante probability (far from the center of the normal distribution of the prior space) will be very precise even if they are consistent with a (spatially) wide range of outcomes. They will particularly do well when competing against theories which make no predictions and thus have the same probability distribution as the prior space. Thus one might consider referring now to "falsifiability" rather than "precision", as this captures the fact that precision here is really about the ex-ante probability that the model will be falsified if not (approximately) correct.

Let $\rho(\lambda, \Theta; \sigma) \equiv \frac{1}{\int_{z \in \Theta} \phi\Big(\inf_{y \in \lambda(\Theta)} d(z,y); \sigma\Big) d\chi(z; \Theta)}$. I would like to prove some results analogous to those in the simple model. For example, I want to say that a theory λ is *more precise* than a theory λ' given a parameter value σ if $\rho(\lambda, \Theta; \sigma) \geq \rho(\lambda', \Theta; \sigma)$. Then I can use this to define the idea of a super and subset from before. To establish robustness against subsets for any value of $\bar{\sigma}$ regardless of the value of $\bar{\sigma}^\star$ I would need to show:

$$E_Q\Big[\log\big(\beta[x; \lambda^{\star\star}, \Theta, \bar{\sigma}]\big)\Big] - E_Q\Big[\log\big(\beta[x; \lambda', \Theta, \bar{\sigma}]\big)\Big]$$

for all $\lambda' < \lambda^{\star\star}$ and for all $\bar{\sigma}^\star, \bar{\sigma} \in R_{++}$. I have tried a number of examples and this appears to be true, but I am not sure how to approach proving it...I am sure that there are some standard methods used in relative entropy and likelihood problems like this. It would also be useful to have some measure of the range of parameter values for which consistency against supersets is ensured.

If these results can be proved, and I can develop some reasonable computational methods for computing the Bayes Factor, I think this will make a pretty reasonable model. However, even after solving these problems a few important challenges will remain

1. As discussed earlier, it will be crucial to find some way to relax independence without removing too much structure from the theory.

2. I will need to find a reasonable way to weight re-scale the dimensions from applications to use them with the Euclidean metric.

3. An important problem will be choosing χ, the prior shape of the space, in a plausible manner. In some ways this choice of prior shape faces some of the difficulties that choosing priors in Bayesian models does. However, I am hopeful that some rough guidelines could be developed.

4. It seems natural (both computationally and intuitively) to parameterize ϕ as a normal distribution, at least when $\Theta = \mathbb{R}^m$. However, in some applications when Θ is compact, it may be more reasonable to use a parabolic or (α, β) distribution for ϕ. I hope to develop some guidelines for thinking about this, and hopefully prove some robustness results about choices of ϕ.

5. Finally, it seems particularly important to allow $\sigma(\Theta)$ to depend, in a disciplined way, on the particular experiment Θ. While this is essentially covered by item 2 above, I view that as being primarily about scaling different dimensions relative to one another in a particular experiment, rather than allowing more or less error across different experiments. For this purpose it is important to get a sense of how much "weight" to put on any particular experiment. Finding some principled way of doing this, possibly building off the statistically-based hacks I used in the Charness-Rabin application or possibly departing significantly from them, will be crucial

While all of these challenges taken together are quite substantial, I am hopeful that I will be able to draw heavily on past work to overcome them. In proving the formal results, I believe there are standard relative entropy methods and results in statistics that will help solve some of these problems. On the computational side, I am optimistic that, given the simple structure of the problem as laid out here, techniques can be developed to make the analysis tractable. Finally, in the more fundamental challenges enumerated in the above list, a cursory review of the meta-statistical literature seems to suggest that there are many techniques used there that could be productively adapted to solve many if not most of these problems. While much work remains to be done, I do not think, at least at this stage, that the problem is intractable.

References

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Transactions on Automated Control* **19**, 716–23.

Carnap, R. (1950), *Logical Foundations of Probability*, Chicago: University of Chicago Press.

Charness, G. and M. Rabin (2002), 'Understanding social preferences with simple tests', *Quarterly Journal of Economics* **117**(3), 817–69.

Eddy, D. M., Hasselblad V. and R. Shachter (1990), 'An introduction to a bayesian method for meta-analysis: the confidence profile method', *Medical Decision Making* **10**, 15–23.

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans*, SIAM CBMS-NSF Monograph 38.

Gabaix, X. and D. Laibson (2007), The seven properties of good models, Technical report, NYU Methodology Conference.

Gelman, A., Carlin J. B. Stern H. S. and D. B. Rubin (1995), *Bayesian Data Analysis*, London: Chapman & Hall.

Jaynes, E. T. (1957), 'Information theory and statistical physics', *Physical Review* **106**, 620–630.

Jaynes, E. T. (1982), 'On the rationale of maximum entropy methods'.

Katz, R. and B. Singer (2007), 'Can an attribution assesement be made for yellow rain? systematic reanalysis in a chemical-and-biological-weapons use investigation', *Politics and the Life Sciences* **26**(1), 24–42.

Kelly, J. L. (1956), 'A new interpretation of the information rate', *IEEE Transactions on Information Theory* **2**, 185–9.

Laplace, P. S. (1814), *A Philosophical Essay on Probabilities*, New York: Dover (1951).

Laplace, P.S. (1812), *Theorie Analytiqe des Probabilités*, Paris: Courcier.

Lehmann, E. L. and G. Casella (1998), *Theory of Point Estimation, 2nd Edition*, New York: Springer.

Olszewski, W. and A. Sandroni (2007), Counterfactual predictions, Technical report, manuscript, Northwestern University.

Popper, K. R. (1959), *The Logic of Scientific Discovery*, New York: Hutchinson.

Rissanen, J. (1978), 'Modeling by shortest data description', *Automatica* **14**, 465–71.

Rosenthal, R. (1984), *Meta-analytic Procedures for Social Research*, Newbury Park, CA: Sage.

Sandroni, A. (2003), 'The Reproducible Properties of Correct Forecasts', *International Journal of Game Theory* **32**(1), 151–159.

Vapnik, V. (1995), *The Nature of Statistical Learning Theory*, New York: Springer Verlag.