FACT FREE LEARNING

By

Enriqueta Aragones, Itzhak Gilboa, Andrew Postlewaite and David Schmeidler

November 2004

COWLES FOUNDATION DISCUSSION PAPER NO. 1491

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

YALE UNIVERSITY

Box 208281

New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Fact-Free Learning*

Enriqueta Aragones†, Itzhak Gilboa‡,
Andrew Postlewaite§, and David Schmeidler¶

October 2003

Abstract

People may be surprised by noticing certain regularities that hold in existing knowledge they have had for some time. That is, they may learn without getting new factual information. We argue that this can be partly explained by computational complexity. We show that, given a database, finding a small set of variables that obtain a certain value of R^2 is computationally hard, in the sense that this term is used in computer science. We discuss some of the implications of this result and of fact-free learning in general.

*Earlier versions of this paper circulated under the titles "From Cases to Rules: Induction and Regression" and "Accuarcy versus Simplicity: A Complex Trade-Off". We have benefited greatly from comments and references by Yoav Benjamini, Joe Halpern, Bart Lipman, Yishay Mansour, Nimrod Megiddo, Dov Samet, Petra Todd, Ken Wolpin, as well as the participants of the SITE conference on Behavioral Economics at Stanford, August, 2003 and the Cowles Foundation workshop on Complexity in Economic Theory at Yale, September, 2003.

†Institut d'Anàlisi Econòmica, C.S.I.C. enriqueta.aragones@uab.es. Aragones acknowledges financial support from the Spanish Ministry of Science and Technology, grant number SEC2000-1186.

‡Tel-Aviv University and Cowles Foundation, Yale University. Gilboa gratefully acknowledges support from the Israel Science Foundation (Grant No. 790/00). igilboa@post.tau.ac.il

§University of Pennsylvania; Postlewaite gratefully acknowledges support from the National Science Foundation. apostlew@econ.sas.upenn.edu

¶Tel-Aviv University and the Ohio State University.Schmeidler gratefully acknowledges support from the Israel Science Foundation (Grant No. 790/00). schmeid@post.tau.ac.il

"*The process of induction is the process of assuming the simplest law that can be made to harmonize with our experience.*" – Wittgenstein (1922)

1 Introduction

Understanding one's social environment calls for the collection of information (or data) and for finding regularities in these data. Many theoretical models of learning focus on learning new facts, on their integration in an existing database, and on the way they modify beliefs. Within the Bayesian framework the integration of new facts and the modification of beliefs is done mechanically according to Bayes's rule. However, much of human learning has to do with finding regularities that, in principle, could have been determined using existing knowledge, rather than with the acquisition of new facts.[1] In this paper we model regularities in database and explain the difficulty in finding them.

The immediate consequence of this difficulty is that individuals typically will not discover all the regularities in a database, and may overlook the most useful regularities. Two people with the same database may notice different regularities, and may consequently hold different views about a particular issue. One person may change the beliefs and actions of another without communicating new facts, but simply by pointing to a regularity overlooked by the other person. On the other hand, people may agree to disagree even if they have the same database and are communicating. We elaborate on these consequences in Section 4.

For illustration, consider the following example.

Ann: "Russia is a dangerous country."

[1] To consider an extreme case, assume that an agent follows a mathematical proof of a theorem. The knowledge she thus acquires has always been, in principle, available to her. Yet, mathematics has to be studied.

Bob: "Nonsense."

Ann: "Don't you think that Russia might initiate a war against a Western country?"

Bob: "Not a chance."

Ann: "Well, I believe it very well might."

Bob: "Can you come up with examples of wars that erupted between two democratic countries?"

Ann: "I guess so. Let me see... How about England and the US in 1812?"

Bob: "OK, save colonial wars."

Ann: "Well, then, let's see. OK, maybe you have a point. Perhaps Russia is not so dangerous."

Bob seems to have managed to change Ann's views. Observe that Bob has not provided Ann with any new factual information. Rather, he pointed out a regularity in Ann's database of which she had been unaware: democratic countries have seldom waged war on each other.[2]

Why has Ann failed to notice that the democratic peace phenomenon holds in her own database? It appears most likely that it has simply not occurred to her to categorize wars by the type of regime of the countries involved. For most people, wars are categorized, or "indexed", by chronology and geography, but not by regime. Once the variable "type of regime" is introduced, Ann will be able to reorganize her database and observe the regularity she has failed to notice earlier.

Yet, fact-free learning is not always due to the introduction of a new variable, or categorization, that the individual has not been aware of. Often, one may be aware of all variables involved, and yet fail to see a regularity

[2]In the field of international relations this is referred to as the democratic peace phenomenon. (See Russett (1993), Maoz and Russett (1992, 1993), and Maoz (1998).)

that involves a *combination* of such variables. Consider an econometrician who wants to understand the determinants of the rate of economic growth. She has access to a large database of realized growth rates for particular economies that includes a plethora of variables describing these economies in detail.[3] Assume that the econometrician prefers fewer explanatory variables to more. Her main difficulty is to determine what set of variables to use in her regression. We can formulate her problem as determining whether there exists a set of k regressors that give a particular level of R^2. This is a well-defined problem that can be relegated to a computer software. However, testing all subsets of k regressors out of, say, m variables involves running $\binom{m}{k} = O(m^k)$ regressions. When m and k are of realistic magnitude, it is impractical to perform this exhaustive search. For instance, choosing the best set of $k = 13$ regressors out of $m = 100$ potentially relevant variables involves $\binom{100}{13} \approx 7 * 10^{15}$ regressions. On a computer that can perform 10 million regression analyses per second, this task would take more than 22 years.

But linear regression is a structured and relatively well-understood problem. One may hope that, using clever algorithms that employ statistical analysis, the best set of k regressors can be found without actually testing all $\binom{m}{k}$ subsets. Our main result is that this is not the case. Formally, we prove that finding whether k regressors can obtain a pre-specified value of R^2, r, is, in the parlance of computer science, NP-Complete.[4] Moreover, we show that this problem is hard (NP-Complete) for *every* positive value of r. Thus our regression problem belongs to a large family of combinatorial problems for which no efficient (polynomial) algorithm is known. An impli-

[3] As an example of the variety of variables that may potentially be relevant, consider the following quote from a recent paper by La Porta, Lopez-de-Silanes, Shleifer, and Vishny (1998) on the quality of government: "We find that countries that are poor, close to the equator, ethnolinguistically heterogeneous, use French or socialist laws, or have high proportions of Catholics or Muslims exhibit inferior government performance."

[4] In Section 3 we explain the concept of NP-completeness and provide references to formal definitions.

cation of this result is that, even for moderate size data sets, it will generally be impossible for the econometrician to know the trade-off between increasing the number of regressors and increasing the explanatory power of those regressors.[5]

Our interest lies not in the difficulties facing social scientists, but in the problems encountered by nonspecialists attempting to understand their environment. That is, we wish to model the reasoning of actual economic agents, rather than of economists analyzing data. We contend, however, that a problem that is difficult to solve for a working economist will also be difficult for an economic agent. If an econometrician cannot be guaranteed to find the "best" set of regressors, many economic agents may also fail to find it.[6]

Economic agents, as well as social scientists, do not generally look for the best set of regressors without any guiding principle. That is, they do not engage in data mining. Rather, they espouse and develop various theories that guide their search for regularities. Our econometrician will often have some idea about which variables may be conducive to growth. She therefore need not exhaust all subsets of k regressors in her quest for the "best" regression. By contrast, our model does not capture the development of and selection among causal theories. Yet, even the set of variables that the econometrician deems relevant according to her theory is typically large enough to raise computational difficulties. More importantly, if the econometrician wants to test her scientific paradigm, and if she wants to guarantee that she is not missing some important regularities that might cause a paradigm shift and unveil new causal theories, she cannot restrict her attention to the regressors she has already focused on.

In conclusion, while computational complexity is not the only reason for which individuals may be surprised to discover regularities in their own

[5] In particular, principle components analysis, which finds a set of orthogonal components, is not guaranteed to find the best combination of predictors (with unconstrained correlations).

[6] We support this claim in Section 4 below.

databases, it is one of the reasons that knowledge of facts does not imply knowledge of all their implications. Hence computational complexity, alongside unawareness, is among the reasons that make fact-free learning a rather common phenomenon.

In the next section we lay out our model of individuals' databases. In this context we discuss several notions of regularities and the criteria to choose among them. The difficulty of discovering satisfactory sets of regressors is proven in Section 3. In the last section we discuss the result, its implications and related literature.

2 Regularities in a Data Base

An individual's database consists of her observations, past experiences, as well as observations that were related to her by others. An assumption that greatly simplifies the discussion is that observations are represented as vectors of numbers. An entry in the vector might be the value of a certain numerical variable, or a measure of the degree to which the observation has a particular attribute. Thus, we model the information available to an individual as a database consisting of a matrix of numbers, where rows correspond to observations (distinct pieces of information) and columns to attributes.

We show below a fraction of a conceivable database pertinent to the democratic peace example. Rows correspond to observations in the database, and columns the attributes. The value in a given entry represents the degree to which the attribute (column) holds for the observation (row). (The numbers are illustrative only.)

Observation	$M1$	$M2$	$D1$	$D2$	T	W
WWII[7]	.7	1	1	0	0	1
Cuban missile crisis	1	1	1	0	1	0
1991 Gulf war	1	.3	1	0	1	1

[7]We refer here to England's declaration of war on Germany on September 3, 1939.

M_i – how strong was country i?

D_i – was country i a democracy?

T – was it after 1945?

W – did war result?

The democratic peace regularity states that if, for any given item the attribute W assumes the value 1, then at least one of the attributes $\{D_1, D_2\}$ does not assume that value. (More precisely, this is the contrapositive of the democratic peace regularity.)

It is important to observe that this model is highly simplified in several respects. For instance, it assumes that the individual has access to a complete matrix of data, whereas in reality certain entries in the matrix may not be known or remembered. The model implicitly assumes also that all variables are observed with accuracy. More importantly, in our model observations are already encoded in a particular way. For instance, in this matrix above country "1" is always the democratic one. But, when representing a real-life case by a row in the matrix, one may not know which country should be dubbed "1" and which – "2". This choice of encoding is immaterial in the democratic peace phenomenon, because this rule is symmetric with respect to the countries. If, however, we were to consider the rule "a democratic country would never attack another country", encoding would matter. If the encoding system keeps country "1" as a designator of a democratic country (as long as one of the countries involved is indeed a democracy), this rule would take the form "if $D1 = 1$ then $A1 = 0$", where Ai stands for "country i attacked". If, however, the encoding system does not retain this regularity, the same rule will not be as simple to formulate. In fact, it would require a formal relation between variables, allowing to state "For every i, if $Di = 1$ then $Ai = 0$". Since such relations are not part of our formal model, the model would give rise to different regularities depending on the encoding system. Indeed, finding the "appropriate" encoding is part of the problem of finding regularities in the database. We abstract from this problem here,

and assume that the database is already encoded in a way that suggests the relevant regularities.

We will prove that despite all these simplifying assumptions, it is hard to find regularities in the database. It follows that finding regularities in real databases, which are not so tidy, is an even harder problem.

The democratic peace phenomenon is an example of an *association rule*. Such a rule states that *if,* for any given observation, the values of certain attributes are within stipulated ranges, *then* the values of other attributes are within prespecified ranges. Association rules are used in data mining (see, e.g., Hastie *et al.* (2001)). An association rule does not apply to the entire database: its scope is the set of observations that satisfy its antecedent. It follows that association rules differ from each other in their generality, or scope of applicability. Adding variables to the antecedent (weakly) decreases the scope of such a rule, but may increase its accuracy. For example, we may refine the democratic peace rule by excluding observations prior to the first world war. This will eliminate some exceptions to the rule (e.g., the War of 1812 and the Boer War) but will result in a less general rule.

A second type of regularity is a *functional rule*: a rule that points to a functional relationship between several "explanatory" variables (attributes) and another one (the "predicted" variable). A well-known example of such a rule is linear regression, with which we deal in the formal analysis. All functional rules on a given database have the same scope of applicability, or the same generality. Yet, when different rules are obtained from different databases, they may differ in generality (we return to this point in Section 3 below).

Both association rules and functional rules may be ranked according to three criteria of interest: *accuracy*, *simplicity*, and *generality*. Each criterion admits a variety of measures, depending on the specific model. In the case of linear regression, it is customary to measure accuracy by R^2. Simplicity is often associated with a low number of variables. That is, the number of

variables measures the *complexity* of the rule. Finally, generality might be measured by the number of observations.

Irrespective of the particular measures used, people generally prefer high accuracy, low complexity, and high generality. The preference for accuracy is perhaps the most obvious: rules are supposed to describe the database, and accuracy is simply the degree to which they succeed in doing so. The preference for generality, other things being equal, has obvious pragmatic sources: a more general rule is more likely to come to bear on future cases. The preference for simplicity, is, however, somewhat more intriguing. William of Occam offered simplicity as a guiding normative principle. Wittgenstein (1922) suggested simplicity as a descriptive criterion, modeling the process of induction. The preference for simplicity may be viewed as axiomatic, or as deriving from other principles. For example, simple rules are sometimes more general than complex rules (though this need not always be the case; see Gilboa (1994)). Simple rules may be viewed as identifying causal relationships.[8] For example, in the democratic peace example above, one reason Ann may have been convinced by the democratic peace rule is that she could construct a causal story of why democratic countries might not go to war with each other: politicians who are answerable to the public via elections may be unwilling to go to war unless forced to.[9] A causal story that supports a rule is easier to uncover when there are fewer variables than when there are more, hence simpler rules will generally be preferred.

Another reason to prefer simplicity is the confidence it provides for predictions beyond the given database. Consider the example of linear regression.

[8] See Pearl (2000) who bases his theory of causality on simplicity.

[9] Kant (1795) gave essentially this explanation: "The republican constitution, besides the purity of its origin (having sprung from the pure source of the concept of law), also gives a favorable prospect for the desired consequence, i.e., perpetual peace. The reason is this: if the consent of the citizens is required in order to decide that war should be declared (and in this constitution it cannot but be the case), nothing is more natural than that they would be very cautious in commencing such a poor game, decreeing for themselves all the calamities of war." Observe that Kant wrote in favor of the republican constitution, rather than democracy per se.

For a given accuracy level R^2, one generally prefers a small set of variables to a larger one.[10] It is well known that adding variables to a regression can only increase R^2, and can generically obtain perfect accuracy, that is, $R^2 = 1$ when the number of regressors equals $n - 1$ (where n is the number of observations). But in this case one tends to feel that the theory (rule) is as complex as the data, and that, correspondingly, there is no reason for the theory to have any predictive power outside the given database. Finally, rules or regularities that employ large numbers of variables may be hard to remember or to convey to other people.

In this paper we assume that people generally prefer rules that are as accurate, as simple, and as general as possible. Of course, these three properties present one with non-trivial trade-offs. In the next section we discuss functional rules for a given database, ignoring the criterion of generality, and focus on the accuracy-simplicity trade-off. We will show that the feasible set in the accuracy-simplicity space cannot be easily computed. A similar result can be shown for association rules. We choose to focus on linear regression for two reasons. First, in economics it is a more common technique for uncovering rules. Second, our main result is less straightforward in the case of linear regression.

3 The Complexity of Linear Regression

We devote this section to the study of the trade-off between simplicity and accuracy of functional rules in the case of linear regression. While regression analysis is a basic tool of scientific research, we here view it as an admittedly idealized model of non-professional human reasoning.[11] Given a set of predicting variables, one attempts to provide a good fit to a predicted variable.

[10] This prefence is uncontroversial if "smaller" means "is a subset of". Yet, we will assume that this preference also holds when "smaller" means "has fewer variables than".

[11] See Bray and Savin (1986), who used regression analysis to model the learning of economic agents.

A common measure of accuracy is the coefficient of determination, R^2. A reasonable measure of complexity is the number of explanatory variables one uses. The "adjusted R^2" is frequently used as a measure of the quality of a regression, trading off accuracy, simplicity, and generality. Adjusted R^2 essentially levies a multiplicative penalty for additional variables to offset the spurious increase in R^2 that results from an increase in the number of predicting variables. In recent years statisticians and econometricians mostly use additive penalty functions in model specification (choosing the predicting variables) for a regression problem.[12] The different penalties are associated with different criteria determining the trade-off between parsimony and precision. Each penalty function can be viewed as defining preferences over the number of included variables and R^2, reflecting the trade-off between simplicity and accuracy. Rather than choose a specific penalty function, we assume that an individual can be ascribed a function $v : \mathbb{R}_+ \times [0,1] \to \mathbb{R}$ that represents her preferences for simplicity and accuracy, where $v(k,r)$ is her utility for a regression that attains $R^2 = r$ with k explanatory variables. Thus, if $v(\cdot,\cdot)$ is decreasing in its first argument and increasing in the second, a person who chooses a rule so as to maximize v may be viewed as though she prefers both simplicity and accuracy, and trades them off as described by v.

Our aim is to demonstrate that finding "good" rules is a difficult computational task. We use the concept of NP-Completeness from computer science to formalize the notion of difficulty of solving problems. A yes/no problem is NP if it is easy (can be performed in polynomial worst-case time complexity) to verify that a suggested solution is indeed a solution to it. When an NP problem is also NP-Complete, there is no known algorithm, whose (worst-case time) complexity is polynomial, that can solve it. However, NP-Completeness means somewhat more than the fact that there is no

[12] See, e.g., Hastie *et al.* (2001) for a discussion of model specification and penalty functions.

such known algorithm. The non-existence of such an algorithm is not due to the fact that the problem is new or that little attention has been devoted to it. For NP-Complete problems it is known that, if a polynomial algorithm were found for one of them, such an algorithm could be translated into algorithms for all other problems in NP. Thus, a problem that is NP-Complete is at least as hard as many problems that have been thoroughly studied for years by academics, and for which no polynomial algorithm has yet been found.

We emphasize again that the rules we discuss have no pretense to offer complete theories, identify causal relationships, provide predictions, or suggest courses of action. Rules are merely regularities that happen to hold in a given database, and they may be purely coincidental. Rules may be backed by theories, but we do not purport to model the entire process of developing and choosing among theories.

Assume that we are trying to predict a variable Y given the explanatory variables $X = (X_1, ..., X_m)$. For a subset K of $\{X_1, ..., X_m\}$, let R^2_K be the value of the coefficient of determination R^2 when we regress $(y_i)_{i \leq n}$ on $(x_{ij})_{i \leq n, j \in K}$. We assume that the data are given in their entirety, that is, that there are no missing values.

How does one select a set of explanatory variables? First consider the feasible set of rules, projected onto the accuracy-complexity space. For a set of explanatory variables K, let the degree of complexity be $k = |K|$ and a degree of accuracy – $r = R^2$. Consider the k-r space and, for a given database $X = (X_1, ..., X_m)$ and a variable Y, denote by $F(X, Y)$ the set of pairs (k, r) for which there exists a rule with these parameters. Because the set $F(X)$ is only defined for integer values of k, and for certain values of r, it is more convenient to visualize its comprehensive closure defined by:

$$F'(X, Y) \equiv \{ (k, r) \in \mathbb{R}_+ \times [0, 1] \,|\, \exists (k', r') \in F(X, Y),\, k \geq k',\, r \leq r' \}$$

The set $F'(X,Y)$ is schematically illustrated in Figure 1. Note that it need not be convex.

Insert Figure 1 about here

The optimization problem that such a person with utility function $v(\cdot,\cdot)$ faces is depicted in Figure 2.

Insert Figure 2 about here

This optimization problem is hard to solve, because one generally cannot know its feasible set. In fact, for every $r > 0$, given X, Y, k, determining whether $(k,r) \in F'(X,Y)$ is computationally hard:

Theorem 1 *For every $r \in (0,1]$, the following problem is NP-Complete: Given explanatory variables $X = (X_1, ..., X_m)$, a variable Y, and an integer $k \geq 1$, is there a subset K of $\{X_1, ..., X_m\}$ such that $|K| \leq k$ and $R^2_K \geq r$?*

Theorem 1 explains why people may be surprised to learn of simple regularities that exist in a database they have access to. A person who has access to the data should, in principle, be able to assess the veracity of all linear theories pertaining to these data. Yet, due to computational complexity, this capability remains theoretical. In practice one may often find that one has overlooked a simple linear regularity that, once pointed out, seems evident.

Our discussion here presupposes a fixed database X. In reality, however, one may have to choose among prediction rules that were obtained given different databases. For example, assume that two researchers collected data

in an attempt to predict a variable Y. Researcher A collected 1,000 observations of the variables W, Z, and Y, and obtained $R^2 = .9$ (for Y regressed on W and Z). Researcher B collected two observations of the variables T and Y and, quite expectedly, obtained $R^2 = 1$ (for Y regressed on T). Observe that the two databases cannot be combined into a single database, since they contain information regarding different variables.[13] Which prediction rule should we use?

While database A suggests a rule that is both less accurate and more complex than the rule suggested by database B, one would be expected to prefer the former to the latter. Indeed, obtaining $R^2 = .9$ with two variables and 1,000 observations is a much more impressive feat than obtaining a perfect fit with one variable and two observations. Rules should be accurate and simple, but also general. Other things being equal, a rule that has a higher degree of generality, or a larger scope of applicability, is preferred to a rule that was found to hold in a smaller database. With a given database, all prediction rules have the same scope of applicability, and thus this criterion may be suppressed from the rule selection problem. Yet, in a more general set-up, we should expect accuracy and simplicity to be traded off with generality as well.

We show that, for any positive value of r, it is hard to determine whether a given k is in the r-cut of $F'(X, Y)$ when the input is (X, Y, k). By contrast, for a given k, computing the k-cut of $F'(X, Y)$ is a polynomial problem (when the input is (X, Y, r)), bounded by a polynomial of degree k. Recall, however, that k is bounded only by the number of columns in X. Moreover, even if k is small, a polynomial of degree k may assume large values if m is large.[14]

[13] To be precise, a combination of the databases would result in a database with many missing values. Indeed, a theory of induction that is general enough to encompass databases with missing values will be able to deal with induction given different databases as well.

[14] The number of observations, n, directly affects the computational complexity of the regression analysis for every subset of k variables. If n is smaller than $k + 1$, we would expect, generically, that any k variables will provide a perfect fit, so that $r = 1$ will be

We conclude that, in general, finding the frontier of the set $F'(X,Y)$, as a function of X and Y, is a hard problem. The optimization problem depicted in Figure 2 has a fuzzy feasible set, as described in Figure 3.

Insert Figure 3 about here

A decision maker may choose a functional rule that maximizes $v(k,r)$ out of all the rules she is aware of, but the latter are likely to constitute only a subset of the set of rules defining the actual set $F'(X,Y)$. Hence, many of the rules that people formulate are not necessarily the simplest (for a given degree of accuracy) or the most accurate (for a given degree of complexity).

We conclude this section with the observation that one may prove theorems similar to Theorem 1, which would make explicit reference to a certain function $v(k,r)$. The following is an example of such a theorem.

Theorem 2 *For every $r \in (0,1]$, the following problem is NP-Complete: Given explanatory variables $X = (X_1, ..., X_m)$ and a variable Y, is there a subset K of $\{X_1, ..., X_m\}$ that obtains an adjusted R^2 of at least r?*

As will be clear from the proof of Theorem 2, this result does not depend on the specific measure of the accuracy-simplicity trade-off, and similar results can be proven for a variety of functions $v(k,r)$.[15]

obtained in time $O(k)$.

[15] There are, however, functions v for which the result does *not* hold. For example, consider $v(k,r) = \min(r, 2-k)$. This function obtains its maximum at $k = 1$ and it is therefore easy to maximize it.

4 Discussion

4.1 The relevance of NP-Completeness

We maintain that a problem that is NP-Complete will be hard for economic agents to solve. Agents may obtain or learn the optimal solutions to particular instances of the general problem, especially if they are only interested in instances described by small inputs. But should economic agents encounter new instances of reasonable sizes on a regular basis, high computational complexity implies that it is unlikely to assume that all or most agents in the economy would determine the optimal solutions to these instances.

In the case of fact-free learning, economic agents are called upon to find regularities in large databases. These regularities cannot be uncovered once and for all. The economic and political environment changes constantly and the lore of yesterday does not provide a blueprint for the decisions of tomorrow. It is therefore reasonable to model economic agents as problem solvers who constantly need to cope with new and large problems.

One can argue that NP-Completeness is a concept that relates to the way computers perform computations, and has little or no bearing on human reasoning. Indeed, there are problems such as natural language understanding or face recognition that toddlers perform better than do computers. But these are problems for which finding an appropriate mathematical model is a major part of the solution. By contrast, for well defined combinatorial problems such as those in the class NP it is rarely the case that humans perform better than do computers. It therefore seems safe to assume that neither people nor computers can solve NP-Complete problems optimally.

We do not claim that the human brain is a machine that can be mimicked by a Turing machine, in theory or in practice. Our proposition is much more modest: if a well-defined, combinatorial problem is NP-Complete, then it is probably hard to solve for human beings (for moderate size inputs). This proposition does not imply that the human brain is a machine, let alone a

machine that one may simulate on a digital computer.

Since we do not know what the brain actually does, it is still possible that the brain can efficiently solve problems that are NP-Complete. That is, one cannot rule out the possibility that a statistician or an economist would have an uncanny ability to find an optimal set of k regressors (for every problem and every k). But even if this were the case, they would not be able to share this ability: any description of a procedure by which one may optimally solve the problem would give rise (by Church's thesis[16]) to a Turing machine that can mimic this procedure. Hence, should a human being have such magical ability, it would not be transferable.

One may question the use of complexity concepts that are defined by worst-case analysis. Indeed, why would we worry about an algorithm whose worst-case performance is exponential, if it is polynomial on average? Experience of computer scientists, however, indicates that NP-Complete problems do not tend to be efficiently solvable even in expectation, under any reasonable assumptions on the distribution of inputs.[17]

As all problems in NP, the problems we study have only two possible answers: "yes" or "no".[18] But as is often the case, they are binary manifestations of optimization problems, say "Find the minimal number k of regressors that obtain an R^2 of r", or "Find a set of regressors that maximize the adjusted R^2". When our problems are formulated thus, it is natural to ask whether one can find approximations to the optimal solution. For example, if one can find, in polynomial time, a set of regressors that is guaranteed not to be more than 2% away from the highest possible adjusted R^2, one may be

[16] See Odifreddi (1989) on the Turing-Church theses and the different variants thereof.

[17] See Papadimitriou (1994) who makes this point, and emphasizes that the example of linear programming confirms this experience. Indeed, the simplex algorithm has exponential worst-case time complexity but very good expected complexity. Yet, linear programming is not an NP-Complete problem and it now has algorithms with polynomial worst-case performance.

[18] These are called "decision problems" in the computer science literature. For economists, this term is quite misleading.

content with this result. We do not know whether there exists polynomial algorithms that guarantee such approximations.[19].

We do not claim that the inability to solve NP-Complete problems is necessarily the most important cognitive limitation on people's ability to perform induction. As mentioned above, even polynomial problems can be difficult to solve when the database consists of many cases and many attributes. Moreover, it is often the case that looking for a general rule does not even cross someone's mind. Yet, the difficulty of performing induction shares an important property with NP-Complete problems: while it is hard to come up with a solution to such a problem, it is easy to verify whether a suggested solution is valid. Similarly, it is hard to come up with an appropriate generalization, but it is relatively easy to assess the applicability of such a generalization once it is offered.

We need not assume that people are lazy or irrational to explain why they do not find all relevant rules. Rather, looking for simple regularities is a genuinely hard problem. There is nothing irrational about not being able to solve NP-Complete problems. Faced with the problem of selecting a set of explanatory variables, which is NP-Complete, people may use various heuristics to find prediction rules, but they cannot be sure, in general, that the rules they find are the simplest ones.

[19] A related problem is the satisfaction of a system of linear equalities and inequalities by a minimal number of variables (obtaining non-zero values). Amaldi and Kann (1998) showed that no polynomial algorithm can compute approximations to this problems (unless all problems in NP are polynomial). However, in our case a reasonable definition of approximation will also use the r axis. Generally, all NP-Complete problems are equivalent to each other in the sense that the existence of a polynomial algorithm that perfectly solves one of them implies the existence of such an algorithm for all others. But the existence of a polynomial algorithm that approximates (the optimization version of) one such problem does not imply a similar result for other problems. See Papadimitriou (1994, Ch. 13).

4.2 Implications

Agreeing to disagree. Our model suggests two reasons for which people, who have access to the same database, may have different beliefs, even if these beliefs are defined by rules that are derived from the shared database. First, two people may notice different regularities. Since finding the "best" regularities is a hard problem, we should not be surprised if one person failed to see a regularity that another came up with. Second, even if the individuals share the rules that they found, they may entertain different beliefs if they make different trade-offs between the accuracy and the simplicity of rules. Different people may well have different v functions, with some people more willing to sacrifice accuracy for simpler rules. If two individuals choose different levels of simplicity, they may also disagree on the relevance of a characteristic. In particular, a variable that is important when there are relatively few other variables in a regression may not be important if the number of variables considered increases. Thus, a particular attribute may play a large role in the rule one person uses but no role in the rule another employs.

Locally optimal rules. Our central point is that people use rules that are not fully optimal because of the complexity of the problem of finding such rules. When an individual uses a rule that is less than fully optimal, she may improve upon the rule by considering alternatives to it. A person faced with the regression problem may think of alternatives to her current "best" regression by adding or deleting variables from her current included set, or by replacing variables in the included set with others. While we do not formally model this search and revision process, one can imagine two distinct ways people may update the rules they use. One can search "locally", that is, consider relatively minor changes in the current rule such as adding, deleting, or replacing one or two variables, or one can search globally by considering sets of variables that have no relation whatsoever to the current set of variables. Local search may find local optima that are not global optima. Differently put, people may get "stuck" with suboptimal rules that can be improved

upon only with a "paradigm shift" that considers a completely different way of looking at a problem.

Path dependence. When individuals search locally for improved rules, their reasoning is likely to exhibit path dependence. Two individuals who begin with different initial sets of variables can settle on very different rules, even after very long search times.

Regret. Our model suggests different notions of regret. In a standard model, individuals make optimal choices given the information available to them at the time they decide. In a stochastic environment, an individual may wish *ex post* that she had decided differently. However, a rational person has no reason to regret a decision she had taken since she could have done no better at the time of her decision, given the information available to her at that time. In our model there are two notions in which information can be "given", and correspondingly, two possible sources of regret. As usual, one may learn the realization of a random variable, and wish that she had decided differently. But one can also learn of a rule that one has not been aware of, even though the rule could be derived, in principle, from one's database. Should one feel regret as a result? As argued above, one could not be expected to solve NP-Complete problems, and therefore it may be argued that one could not have chosen optimally. Yet, one might expect individuals to experience a stronger sense of "I could have known" as a result of finding rules that hold in a given database, than as a result of getting new observations.

4.3 Modeling choices

There is an alternative approach to modelling induction that potentially provides a more explicit account of the components of cases. The components should include entities and relations among them. For example, our motivating examples give rise to entities such as countries and governments, and to the relations "fought against" and "exhibits inferior performance", among others. In a formal model, entities would be elements of an abstract set, and

relations, or predicates, would be modeled as functions from sequences of entities into $[0, 1]$. Such a predicate model would provide more structure, would be closer to the way people think of complex problems, and would allow a more intuitive modelling of analogies than one can obtain from our present model. Moreover, while the mathematical notation required to describe a predicate model is more cumbersome than that used for the present model, the description of actual problems within the predicate model may be more concise. In particular, this implies that problems that are computationally easy in the attribute model may still be computationally hard with respect to the predicate model.[20]

Observe that neither the model presented here nor the alternative predicate model attempts to explain how people choose the predicates or attributes they use to describe cases. The importance of this choice has been clearly illustrated by Goodman's (1965) "grue-bleen" paradox.[21] This problem is, however, beyond the scope of the present paper.

4.4 Related literature

Most of the formal literature in economic theory and in related fields adheres to the Bayesian model of information processing. In this model a decision maker starts out with a prior probability, and she updates it in the face of new information by Bayes's rule. Hence, this model can easily capture changes in

[20] In Aragones, Gilboa, Postlewaite and Schmeidler (2001), we present both the attribute and the predicate models for the study of analogies, prove their equivalence in terms of the scope of phenomena they can describe, and show that finding a good analogy in the predicate model is a hard problem.

[21] The paradox is, in a nutshell, the following. If one wishes to test whether emeralds are green or blue, one can sample emeralds and conclude that they seem to be green. Based on this, one may predict that emeralds will be green in the year 2010. Next assume that one starts with two other primitive predicates, "grue" and "bleen". When translated to the more common predicates "green" and "blue", "grue" means "green until 2010 and blue thereafter" and "bleen" – vice versa. With these predicates, emeralds appear to be grue, and one may conclude that they will appear blue after the year 2010. This paradox may be interpreted as showing that inductive inference, as well as the concept of simplicity, depend on the predicates one starts out with.

opinion that result from new information. But it does not deal very graciously with changes of opinion that are not driven by new information. In fact, in a Bayesian model with perfect rationality people cannot change their opinions unless new information has been received. It follows that the example we started out with cannot be explained by such models.

Relaxing the perfect rationality assumption, one may attempt to provide a pseudo-Bayesian account of the phenomena discussed here. For instance, one can use a space of states of the world to describe the subjective uncertainty that a decision maker has regarding the result of a computation, before this computation is carried out. (See Anderlini and Felli (1994) and Al-Najjar, Casadesus-Masanell, and Ozdenoren (1999).) In such a model, one would be described as if one entertained a prior probability of, say p, that "democratic peace" holds. Upon hearing the rhetorical question as in our dialogue, the decision maker performs the computation of the accuracy of this rule, and is described as if the result of this computation were new information.

A related approach employs a subjective state space to provide a Bayesian account of unforeseen contingencies. (See Kreps (1979, 1992), and Dekel, Lipman, and Rustichini (1997, 1998).) Should this approach be applied to the problem of induction, each regularity that might hold in the database would be viewed as an unforeseen contingency that might arise. A decision maker's behavior will then be viewed as arising from Bayesian optimization with respect to a subjective state space that reflects her subjective uncertainty.

Our approach models the process of induction more explicitly. In comparison with pseudo-Bayesian approaches, it allows a better understanding of why and when induction is likely to be a hard problem.

Gilboa and Schmeidler (2001) offer a theory of case-based decision making. They argue that cases are the primitive objects of knowledge, and that rules and probabilities are derived from cases. Moreover, rules and probabilities cannot be known in the same sense, and to the same degree of certitude, that cases can. Yet, rules and probabilities may be efficient and insight-

ful ways of succinctly summarizing many cases. The present paper suggests that summarizing databases by rules may involve loss of information, because one cannot be guaranteed to find the "optimal" rules that a given database induces.

5 Appendix: Proofs

Proof of Theorem 1:

Let there be given $r > 0$. It is easy to see that the problem is in NP: given a suggested set $K \subset \{1, ..., m\}$, one may calculate R^2_K in polynomial time in $|K|n$ (which is bounded by the size of the input, $(m+1)n$).[22] To show that the problem is NP-Complete, we use a reduction of the following problem, which is known to be NP-Complete (see Gary and Johnson (1979), or Papadimitriou (1994)):

Problem EXACT COVER: Given a set S, a set of subsets of S, $\mathfrak{S}$, are there pairwise disjoint subsets in $\mathfrak{S}$ whose union equals S?

(That is, does a subset of $\mathfrak{S}$ constitutes a partition of S?)

Given a set S, a set of subsets of S, $\mathfrak{S}$, we will generate n observations of $(m+1)$ variables, $(x_{ij})_{i\leq n, j\leq m}$ and $(y_i)_{i\leq n}$, and a natural number k, such that S has an exact cover in $\mathfrak{S}$ iff there is a subset K of $\{1, ..., m\}$ with $|K| \leq k$ and $R^2_K \geq r$.

Let there be given, then, S and $\mathfrak{S}$. Assume without loss of generality that $S = \{1, ..., s\}$, and that $\mathfrak{S} = \{S_1, ..., S_l\}$ (where $s, l \geq 1$ are natural numbers). We construct $n = 2(s + l + 1)$ observations of $m = 2l$ predicting variables. It will be convenient to denote the $2l$ predicting variables by $X_1, ..., X_l$ and $Z_1, ..., Z_l$ and the predicted variable – by Y. Their corresponding values will be denoted $(x_{ij})_{i\leq n, j\leq l}$, $(z_{ij})_{i\leq n, j\leq l}$, and $(y_i)_{i\leq n}$. We will use X_j, Z_j, and Y also to denote the column vectors $(x_{ij})_{i\leq n}$, $(z_{ij})_{i\leq n}$, and $(y_i)_{i\leq n}$, respectively. Let $M \geq 0$ be a constant to be specified later. We now specify the vectors $X_1, ..., X_l$, $Z_1, ..., Z_l$, and Y as a function of M.

For $i \leq s$ and $j \leq l$, $x_{ij} = 1$ if $i \in S_j$ and $x_{ij} = 0$ if $i \notin S_j$;

For $i \leq s$ and $j \leq l$, $z_{ij} = 0$;

[22] Here and in the sequel we assume that reading an entry in the matrix X or in the vector Y, as well any algebraic computation require a single time unit. Our results hold also if one assumes that x_{ij} and y_i are all rational and takes into account the time it takes to read and manipulate these numbers.

For $s < i \leq s + l$ and $j \leq l$, $x_{ij} = z_{ij} = 1$ if $i = s + j$ and $x_{ij} = z_{ij} = 0$ if $i \neq s + j$;

For $j \leq l$, $x_{s+l+1,j} = z_{s+l+1,j} = 0$;

For $i \leq s + l$, $y_i = 1$ and $y_{s+l+1} = M$;

For $i > s + l + 1$, $y_i = -y_{i-(s+l+1)}$ and for all $j \leq l$, $x_{ij} = -x_{i-(s+l+1),j}$ and $z_{ij} = -z_{i-(s+l+1),j}$.

Observe that the bottom half of the matrix X as well as the bottom half of the vector Y are the negatives of the respective tops halves. This implies that each of the variables $X_1, ..., X_l$, $Z_1, ..., Z_l$, and Y has a mean of zero. This, in turns, implies that for any set of variables K, when we regress Y on K, we get a regression equation with a zero intercept.

Consider the matrix X and the vector Y obtained by the above construction for different values of M. Observe that the collection of sets K that maximize R^2_K is independent of M. Hence, it is useful to define $\widehat{R}^2_K$ as the R^2 obtained from regressing Y on K, ignoring observations $s + l + 1$ and $2(s + l + 1)$. Obviously, minimizing $\widehat{R}^2_K$ is tantamount to minimizing R^2_K.

We claim that there is a subset K of $\{X_1, ..., X_l\} \cup \{Z_1, ..., Z_l\}$ with $|K| \leq k \equiv l$ for which $\widehat{R}^2_K = 1$ iff S has an exact cover from $\mathfrak{S}$.

First assume that such a cover exists. That is, assume that there is a set $J \subset \{1, ..., l\}$ such that $\{S_j\}_{j \in J}$ constitutes a partition of S. This means that $\sum_{j \in J} \mathbf{1}_{S_j} = \mathbf{1}_S$ where $\mathbf{1}_A$ is the indicator function of a set A. Let α be the intercept, $(\beta_j)_{j \leq l}$ be the coefficients of $(X_j)_{j \leq l}$ and $(\gamma_j)_{j \leq l}$ – of $(Z_j)_{j \leq l}$ in the regression. Set $\alpha = 0$. For $j \in J$, set $\beta_j = 1$ and $\gamma_j = 0$, and for $j \notin J$ set $\beta_j = 0$ and $\gamma_j = 1$. We claim that $\alpha \mathbf{1} + \sum_{j \leq l} \beta_j X_j + \sum_{j \leq l} \gamma_j Z_j = Y$ where $\mathbf{1}$ is a vector of 1's. For $i \leq s$ the equality

$$\alpha + \sum_{j \leq l} \beta_j x_{ij} + \sum_{j \leq l} \gamma_j z_{ij} = \sum_{j \leq l} \beta_j x_{ij} = y_i = 1$$

follows from $\sum_{j \in J} 1_{S_j} = 1_S$. For $s < i \leq s + l$, the equality

$$\alpha + \sum_{j \leq l} \beta_j x_{ij} + \sum_{j \leq l} \gamma_j z_{ij} = \beta_j + \gamma_j = y_i = 1$$

follows from our construction (assigning precisely one of $\{\beta_j, \gamma_j\}$ to 1 and the other – to 0). Obviously, $\alpha + \sum_{j \le l} \beta_j x_{nj} + \sum_{j \le l} \gamma_j z_{nj} = 0 = y_i = 0$. The number of variables used in this regression is l. Specifically, choose $K = \{ X_j \,|\, j \in J \} \cup \{ Z_j \,|\, j \notin J \}$, with $|K| = l$, and observe that $\widehat{R}^2_K = 1$.

We now turn to the converse direction. Assume, then, that there is a subset K of $\{X_1, ..., X_l\} \cup \{Z_1, ..., Z_l\}$ with $|K| \le l$ for which $\widehat{R}^2_K = 1$. Since all variables have zero means, this regression has an intercept of zero ($\alpha = 0$ in the notation above). Let $J \subset \{1, ..., l\}$ be the set of indices of the X variables in K, i.e., $\{X_j\}_{j \in J} = K \cap \{X_1, ..., X_l\}$. We will show that $\{S_j\}_{j \in J}$ constitutes a partition of S. Set $L \subset \{1, ..., l\}$ be the set of indices of the Z variables in K, i.e., $\{Z_j\}_{j \in L} = K \cap \{Z_1, ..., Z_l\}$. Consider the coefficients of the variables in K used in the regression obtaining $\widehat{R}^2_K = 1$. Denote them by $(\beta_j)_{j \in J}$ and $(\gamma_j)_{j \in L}$. Define $\beta_j = 0$ if $j \notin J$ and $\gamma_j = 0$ if $j \notin L$. Thus, we have

$$\sum_{j \le l} \beta_j X_j + \sum_{j \le l} \gamma_j Z_j = Y.$$

We argue that $\beta_j = 1$ for every $j \in J$ and $\gamma_j = 1$ for every $j \in L$. To see this, observe first that for every $j \le l$, the $s + j$ observation implies that $\beta_j + \gamma_j = 1$. This means that for every $j \le l$, $\beta_j \neq 0$ or $\gamma_j \neq 0$ (this also implies that either $j \in J$ or $j \in L$). If for some j both $\beta_j \neq 0$ and $\gamma_j \neq 0$, we will have $|K| > l$, a contradiction. Hence for every $j \le l$ either $\beta_j \neq 0$ or $\gamma_j \neq 0$, but not both. (In other words, $J = L^c$.) This also implies that the non-zero coefficient out of $\{\beta_j, \gamma_j\}$ has to be 1.

Thus the cardinality of K is precisely l, and the coefficients $\{\beta_j, \gamma_j\}$ define a subset of $\{S_1, ...S_l\}$: if $\beta_j = 1$ and $\gamma_j = 0$, i.e., $j \in J$, S_j is included in the subset, and if $\beta_j = 0$ and $\gamma_j = 1$, i.e., $j \notin J$, S_j is not included in the subset. That this subset $\{S_j\}_{j \in J}$ constitutes a partition of S follows from the first s observations as above.

We now turn to define M. We wish to do so in such a way that, for every set of explanatory variables K, $R^2_K \ge r$ iff $\widehat{R}^2_K = 1$. Fix a set K. Denote by

$\widehat{SSR}$ and $\widehat{SST}$ the explained variance and the total variance, respectively, of the regression of Y on K without observations $s+l+1$ and $2(s+l+1)$, where SSR and SST denote the variances of the regression with all observations. Thus, $R^2_K = SSR/SST$ and $\widehat{R}^2_K = \widehat{SSR}/\widehat{SST}$. Observe that $\widehat{SST} = 2(s+l)$ and $SST = 2(s+l) + 2M^2$. Also, $SSR = \widehat{SSR}$ is independent of M.

Note that if K is such that $\widehat{R}^2_K = 1$, then $(SSR =)\widehat{SSR} = \widehat{SST} = 2(s+l)$. In this case, $R^2_K = \frac{2(s+l)}{2(s+l)+2M^2}$. If, however, K is such that $\widehat{R}^2_K < 1$, then we argue that $(SSR =)\widehat{SSR} \leq \widehat{SST} - \frac{1}{9}$. Assume not. That is, assume that K is such that $\widehat{SSR} > \widehat{SST} - \frac{1}{9}$. This implies that on each of the observations $1, ..., s+l, s+l+2, ..., 2(s+l)+1$, the fit produced by K is at most $\frac{1}{3}$ away from y_i. Then for every $j \leq l$, $|\beta_j + \gamma_j - 1| < \frac{1}{3}$. Hence for every $j \leq l$ either $\beta_j \neq 0$ or $\gamma_j \neq 0$, but not both, and the non-zero coefficient out of $\{\beta_j, \gamma_j\}$ has to be in $(\frac{2}{3}, \frac{4}{3})$. But then, considering the first s observations, we find that K is an exact cover. It follows that, if $\widehat{R}^2_K < 1$, then $R^2_K \leq \frac{2(s+l)-\frac{1}{9}}{2(s+l)+2M^2}$.

Choose a rational M in the interval $\left(\sqrt{\frac{(1-r)(s+l)-\frac{1}{18}}{r}}, \sqrt{\frac{(1-r)(s+l)}{r}}\right)$ so that $\frac{2(s+l)-\frac{1}{9}}{2(s+l)+2M^2} < r < \frac{2(s+l)}{2(s+l)+2M^2}$, and observe that for this M, there exists a K such that $R^2_K \geq r$ iff there exists a K for which $\widehat{R}^2_K = 1$, that is, iff K is an exact cover.

To conclude the proof, it remains to observe that the construction of the variables $(X_j)_{j\leq l}$, $(Z_j)_{j\leq l}$, and Y can be done in polynomial time in the size of the input. □

Proof of Theorem 2:

Let there be given $r > 0$. The proof follows that of Theorem 1 with the following modification. For an integer $t \geq 1$, to be specified later, we add t observations for which all the variables ($(X_j)_{j\leq l}$, $(Z_j)_{j\leq l}$, and Y) assume the value 0. These observations do not change the R^2 obtained by any set of regressors, as both SST and SSR remain the same. Assuming that t has been fixed (and that it polynomial in the data), let r' be the R^2 corresponding to an adjusted R^2 of r, with l regressors. That is, $(1-r') = (1-r)\frac{t+2s+2l+1}{t+2s+l+1}$.

Define M as in the proof of Theorem 1 for r'.

We claim that there exists a set of regressors that obtains an adjusted R^2 of r iff there exists a set of l regressors that obtains an R^2 of r' (hence, iff there exists an exact cover in the original problem). The "if" part is obvious from our construction. Consider the "only if" part. Assume, then that a set of regressors obtains an adjusted R^2 of r. If it has l regressors, the same calculation shows that it obtains the desired R^2. We now argue that if no set of l regressors obtains an adjusted R^2 of r, then no set of regressors (of any cardinality) obtains an adjusted R^2 of r.

Consider first a set K_0 with $|K_0| = k_0 > l$ regressors. Observe that, by the choice of M, r' is the upper bound on all R^2_K for all K with $|K| = l$, as r' was computed assuming that an exact cover exists, and that, therefore, there are l variables that perfectly match all the observations but $s + l + 1$ and $2(s + l + 1)$. Due to the structure of the problem, r' is also an upper bound on R^2_K for all K with $|K| \geq l$. This is so because the only observations that are not perfectly matched (in the hypothesized l-regressor set) correspond to zero values of the regressors. It follows that the adjusted R^2 for K_0 is lower than r.

Next consider a set K_0 with $|K_0| = k_0 < l$ regressors. For such a set there exists a $j \leq l$ such that neither X_j nor Z_j are in K_0. Hence, observations $s + j$ and $2s + l + j + 1$ cannot be matched by the regression on K_0. The lowest possible SSE in this problem, corresponding to the hypothesized set of l regressors, is $2M^2$. This means that the SSE of K_0 is at least $2M^2 + 2$. That is, the SSE of the set K_0 is at least $\frac{M^2+1}{M^2}$ larger than the SSE used for the calculation of r. On the other hand, K_0 uses less variables. But if $\frac{t+2s+l+1}{t+2s+k+1} < \frac{M^2+1}{M^2}$, the reduction in the number of variables cannot pay off, and K_0 has an adjusted R^2 lower than r. It remains to choose t large enough so that the above inequality holds, and to observe that this t is bounded by the polynomial of the input size.□

References

[1] Al-Najjar, N., R. Casadesus-Masanell, and E. Ozdenoren (1999), "Probabilistic Models of Complexity," Northwestern University working paper.

Anderlini, L. and L. Felli (1994), "Incomplete Written Contracts: Indescribable States of Nature," *Quarterly Journal of Economics*, **109**: 1085-1124.

Amaldi, E., and V. Kann (1998), "On the Approximability of Minimizing Nonzero Variables or Unsatisfied Relations in Linear Systems", *Theoretical Computer Science*, **209**: 237-260.

Aragones, E., I. Gilboa, A. Postlewaite and D. Schmeidler (2001), "Rhetoric and Analogy," mimeo.

Bray, M. M., and N. E. Savin (1986), "Rational Expectations Equilibria, Learning, and Model Specification", *Econometrica*, **54**: 1129-1160.

Dekel, E., B. L. Lipman, and A. Rustichini (1997), "A Unique Subjective State Space for Unforeseen Contingencies", mimeo.

Dekel, E., B. L. Lipman, and A. Rustichini (1998), "Recent Developments in Modeling Unforeseen Contingencies", *European Economic Review*, **42**: 523–542.

Gary, M. and D. S. Johnson (1979), *Computers and Intractability: A Guide to the Theory of NP-Completeness*. San-Francisco, CA: W. Freeman and Co.

Gilboa, I. (1994), "Philosophical Applications of Kolmogorov's Complexity Measure", in *Logic and Philosophy of Science in Uppsala*, D. Prawitz and D. Westerstahl (eds.), Synthese Library, Vol. 236, Kluwer Academic Press, pp. 205-230.

Gilboa, I. and D. Schmeidler (2001). *A Theory of Case-Based Decisions*. Cambridge: Cambridge University Press.

Goodman, N. (1965). *Fact, Fiction and Forecast*. Indianapolis: Bobbs-Merrill.

Hastie, T., R. Tibshirani and J. Friedman (2001). *The Elements of Statistical Learning*. New York, NY: Springer.

Kant, I. (1795). *Perpetual Peace: A Philosophical Sketch*.

Kreps, D. M. (1979), "A Representation Theorem for 'Preference for Flexibility'," *Econometrica*, **47**: 565– 576.

Kreps, D. M. (1992), "Static Choice and Unforeseen Contingencies" in *Economic Analysis of Markets and Games: Essays in Honor of Frank Hahn*, P. Dasgupta, D. Gale, O. Hart, and E. Maskin (eds.) MIT Press: Cambridge, MA, 259-281.

La Porta, R., F. Lopez-de-Silanes, A. Shleifer, and R. Vishny (1998), "The Quality of Government", mimeo.

Maoz, Z. (1998), "Realist and Cultural Critiques of the Democratic Peace: A Theoretical and Empirical Reassessment", *International Interactions*, **24**: 3-89.

Maoz, Z. and B. Russett (1992), "Alliance, Wealth Contiguity, and Political Stability: Is the Lack of Conflict Between Democracies A Statistical Artifact?" *International Interactions*, **17**: 245-267.

Maoz, Z. and B. Russett (1993), "Normative and Structural Causes of Democratic Peace, 1946-1986", *American Political Science Review*, **87**: 640-654.

Odifreddi, P. (1989). *Classical Recursion Theory*. Vol. I. North-Holland.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference.* Cambridge University Press.

Russett, B. (1993), *Grasping the Democratic Peace: Principles for a Post-Cold War World*. Princeton: Princeton University Press.

Papadimitriou, C. H. (1994), *Computational Complexity*. Addison-Wesley.

Wittgenstein, L. (1922), *Tractatus Logico-Philosophicus*. London: Routledge and Kegan Paul; fifth impression, 1951.

Figure 1

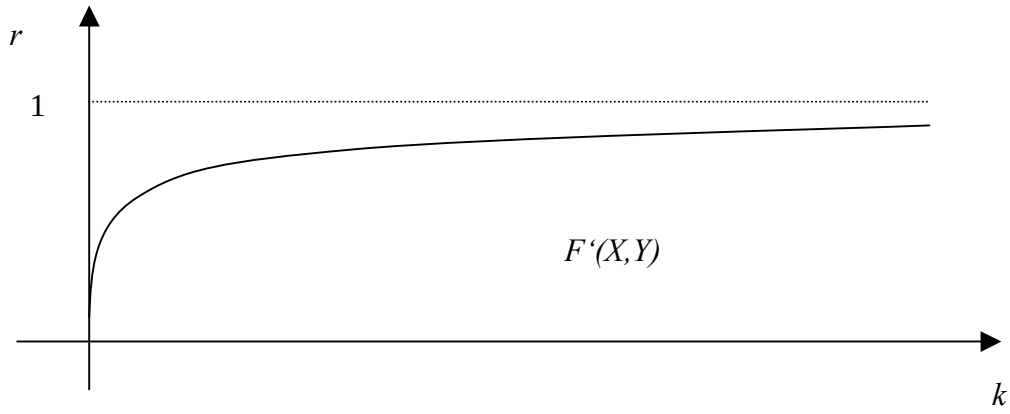

Figure 2

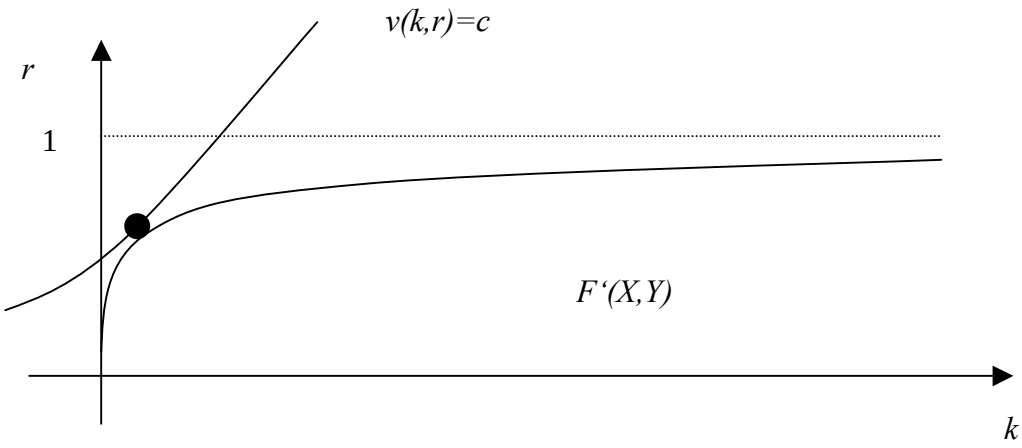

Figure 3